

# UNIVERSITY OF TURIN

---

DOCTORAL SCHOOL OF SCIENCES AND INNOVATIVE TECHNOLOGIES PhD  
PROGRAM IN COMPUTER SCIENCE  
XXXIII CYCLE



PhD Dissertation  
*Endang Wahyu Pamungkas*

*Prominent Challenges in Abusive Language Detection on Social Media:  
Exploring Multidomain and Multilingual Settings*

**Advisor**

**Viviana Patti**

*Università degli Studi di Torino, Italy*

PhD Coordinator  
*Marco Grangetto*

Academic Year 2020-2021



**Abstract** Social media platforms are becoming more and more popular in recent years. The freedom of expression given by social media has a dark side: the growing proliferation of abusive content on these platforms. Several forms of online abusive behaviours are widespread in online communication contexts, a phenomenon that also seems to be influenced by users' anonymity and the lack of regulation provided by social media platforms. Given the current rate of user-generated content produced every minute, manually monitoring abusive behaviour in social media is impractical and not a scalable and long-term solution. Recent studies proposed to automate the detection of abusive language in social media by adopting various approaches. However, the latest approaches developed show that building a robust model to detect abusive language in a social media environment automatically is still challenging. Swear words ambiguity is one of the main challenges which contributes to the difficulty of abusive language detection. On the one hand, swear words could become an important signal to spot abusive content, on the other hand, they could also deceive the abusive detection model when used in not abusive contexts, such as the ones characterized by humor or cathartic use of swearing. Moreover, abusive language detection tasks are multifaceted, and the available datasets featured by various abusive phenomena. Therefore, building a robust model which generalizes across different abusive phenomena is another major challenge in this research field. Finally, providing a robust model to detect abusive language across different languages automatically is also an important challenge, since abusive language is a global phenomenon, while available data in low-resource languages are still very limited.

In this thesis, we conduct a deep exploration of the aforementioned challenges in abusive language detection tasks. First, we investigate the ambiguity issue of swear words in abusive language detection. We build a novel Twitter corpus called SWAD, where a swear word is annotated as either abusive and not abusive, and build a model to automatically predict the abusiveness of a swear word within its context. Furthermore, we also investigate the benefit of resolving swear words abusiveness in several downstream abusive language detection tasks. We experimentally found that classifying swear words as either abusive or not abusive is a challenging task. Meanwhile, resolving the abusiveness of swear words could improve the models' performance to detect abusive content. Second, we explore the further challenge of building a robust model to detect abusive language across domains and languages. To this direction, we develop several models and experiment with various abusive datasets with different topical focuses, targets, and languages. We found that training a model on a topic-generic dataset could provide a more robust model to detect more specific kinds of abusive phenomena in the cross-domain scenario. Furthermore, we also found that the multitask approach could facilitate the knowledge transfer in cross-target classification by allowing the model to learn the abusive detection task and target classification task simultaneously. Meanwhile, on the cross-lingual scenario side, we notice that most of the models obtained a lower performance in the low-resource languages than in other resource-rich languages. We also found that our proposed joint-learning architecture is able to deal with language-shift issues by outperforming other models. Finally, we also explore the aforementioned challenges focusing on a specific category of online abusive language, *misogyny*, which results in several interesting insights.

**Abstract** Le piattaforme di social media stanno diventando sempre più popolari negli ultimi anni. La libertà di espressione che caratterizza questo tipo di contesti comunicativi ha un lato oscuro: la crescente proliferazione di contenuti abusivi su queste piattaforme. I comportamenti abusivi online si ritrovano in quantità copiosa e in forme molteplici nei contesti di comunicazione online, con una possibile influenza di fattori legati all'anonimato degli utenti e alla mancanza di politiche efficaci per contrastare il fenomeno da parte delle piattaforme di social media. Considerata la notevole frequenza di contenuti generati dagli utenti prodotti ogni minuto, il monitoraggio e la moderazione manuale dei comportamenti abusivi nei social media non sono soluzioni praticabili, né scalabili e a lungo termine. In studi recenti possiamo trovare diverse proposte mirate ad automatizzare la rilevazione del linguaggio offensivo nei social media in diverse lingue, adottando vari approcci. Tuttavia, un'analisi della letteratura più recente suggerisce che la creazione di modelli robusti per rilevare automaticamente il linguaggio abusivo nei social media è ancora una sfida aperta. L'ambiguità di insulti, parolacce, parole d'odio è uno dei principali aspetti che contribuiscono a rendere difficile il compito di rilevare il linguaggio abusivo. Da un lato, gli insulti e le cosiddette parolacce potrebbero diventare un segnale importante per individuare contenuti abusivi; dall'altra parte la loro presenza nei messaggi può ingannare i modelli per la rilevazione del linguaggio abusivo, quando vengono utilizzate in contesti non abusivi, come quelli caratterizzati da umorismo o ironia, o in cui le parolacce hanno una funzione catartica. Inoltre, la rilevazione del linguaggio abusivo si presenta come un compito sfaccettato e i dataset disponibili caratterizzano fenomeni abusivi vari. Pertanto, la costruzione di un modello robusto in grado di generalizzare rispetto a diversi fenomeni abusivi è un'altra grande sfida in questo campo di ricerca. Infine, anche sviluppare un modello robusto per rilevare automaticamente il linguaggio abusivo in diverse lingue presenta sfide importanti, soprattutto considerando che il linguaggio abusivo è un fenomeno globale che attraversa paesi e culture, mentre i dati disponibili in lingue meno studiate e con poche risorse linguistiche e computazionali sono ancora molto limitati.

In questa tesi, ci si pone l'obiettivo di esplorare in profondità le suddette sfide nel campo della rilevazione automatica del linguaggio abusivo. Innanzitutto, esploriamo il problema dell'ambiguità delle parolacce in task di *abusive language detection*. Abbiamo creato una nuova risorsa per studiare questo problema, un corpus Twitter di messaggi chiamato SWAD, in cui una parolaccia viene annotata come offensiva e non offensiva, e costruiamo un modello per prevedere automaticamente la carica in termini di abuso di una parolaccia nel suo contesto. Inoltre, investighiamo anche l'utilità di determinare la carica offensiva delle parolacce in diversi task di *abusive language detection*. La sperimentazione condotta ci mostra che classificare automaticamente le parolacce come offensive o non offensive è un compito non banale. Allo stesso tempo, si conferma l'ipotesi che determinare la carica abusiva delle parolacce nel loro contesto d'uso può migliorare le prestazioni dei modelli per rilevare contenuti abusivi, informati di questa conoscenza. In secondo luogo, esploriamo l'ulteriore sfida di costruire un modello robusto per la rilevazione del linguaggio abusivo su diversi domini e lingue. In questa direzione, abbiamo sviluppato diversi modelli neurali e sperimentato con vari dataset di fenomeni d'odio disponibili, caratterizzati da diversi focus, gruppi target, e lingue. I risultati suggeriscono

che l’addestramento di un modello su un dataset generico rispetto alla natura dell’odio espresso fornisce un modello più robusto per rilevare categorie d’odio più specifiche in uno scenario *cross-domain*. Inoltre, abbiamo anche scoperto che l’approccio multitask sembra facilitare il trasferimento di conoscenze nella classificazione *cross-target*, consentendo al modello di apprendere simultaneamente il task di *abusive language detection* e quello di classificazione del target del comportamento abusivo. Inoltre, se consideriamo lo scenario multilingue e cross-lingue, notiamo che la maggior parte delle prestazioni dei modelli su dataset di lingue *low-resource* è bassa rispetto alle prestazioni su lingue *resource-rich*. I risultati mostrano che la nuova architettura neurale di *joint-learning* che viene proposta in questa tesi consente di affrontare problemi relativi al *language-shift*, con prestazioni migliori rispetto a quelle di altri modelli multilingue, anche in scenari *zero-shot*. Infine, proponiamo di declinare l’esplorazione delle sfide affrontate e descritte in termini generali, rispetto a una categoria specifica di linguaggio abusivo online, la misoginia, il che ci porta alla possibilità di discutere diversi spunti interessanti su una forma d’odio online molto diffusa in modo trasversale in diversi paesi, lingue e culture.

# List of Figures

1.1	Relation between topical focuses in abusive language phenomena. Elaborated based on Poletto et al. [2020]	3
1.2	The interaction between targets of abusive phenomena.	6
2.1	Document Collection Methodology	30
2.2	Documents Collection Methodology	42
3.1	Corpus Development Process	55
3.2	Process to Infuse Additional Features	68
5.1	Joint-Learning LSTM Model Architecture	102
5.2	Joint-Learning LSTM-HurtLex Model Architecture	103
5.3	Joint-Learning BERT Model Architecture	104
5.4	Joint-Learning BERT-HurtLex Model Architecture	105
6.1	Top 10 swear words of each dataset	119
6.2	Joint-Learning Model Architecture	136
6.3	Misogynistic Behaviour Classification: Confusion Matrix	144
6.4	Target of Misogyny Classification: Confusion Matrix	145



# List of Tables

1.1	Definition of abusive language and related terms. . . . .	5
1.2	Topical focuses introduced by previous studies. . . . .	7
1.3	Shared tasks in the abusive language detection research field. . . . .	11
2.1	Summarization of Available Abusive Language Dataset Across Different Topical Focuses and Sources (English only). . . . .	28
2.2	Summary of approaches adopted by existing studies for cross-domain abusive language detection tasks. . . . .	33
2.3	Summary of available abusive language datasets across different languages.	39
2.4	Summary of approaches adopted in existing studies on cross-lingual abusive language detection . . . . .	45
2.5	Summary of the AMI shared task systems. . . . .	50
3.1	Corpus statistics after filtering process. . . . .	55
3.2	Label distribution in the SWAD dataset. . . . .	58
3.3	Interaction between the original Holgate’s label with our annotation. . . .	60
3.4	Sequence labeling task: confusion matrix. . . . .	61
3.5	Sequence labeling task: results broken down by label. . . . .	62
3.6	Ablation test on several feature sets. . . . .	63
3.7	Result of Target-based Abusiveness Prediction of Swear Words. . . . .	67
3.8	Result of Investigating Swear Words Role in HatEval Task . . . . .	71
3.9	Result of Investigating Swear Words Role in AMI Evalita Task . . . . .	72
3.10	Result of Investigating Swear Words Role in AMI IberEval Task . . . . .	72
3.11	Result of Investigating Swear Words Role in Davidson Dataset . . . . .	72
4.1	Twitter abusive language datasets in four languages: original labels, language(s) featured, topical focus, distribution of train and test set and positive instance rate (PIR). . . . .	76
4.2	Results on cross-domain abusive language identification (only in English). . . . .	77
4.3	General overview of the datasets along with their topics and targets. . . .	81
4.4	Distribution of instances in topic-generic datasets (used as training). . . .	81
4.5	Distribution of instances in the train/test sets in topic-specific datasets. .	82
4.6	Results for $Top^G \rightarrow Top^S$ configuration when training on Founta. . . .	84
4.7	Results for $Top^G \rightarrow Top^S$ configuration when training on Davidson. .	85

4.8	Results for $Top^S \rightarrow Top^S$ when training on Waseem, HatEval and AMI train sets.	86
4.9	Comparison with related work in terms of accuracy.	86
4.10	Label combination in multitask setting.	88
4.11	Baseline results for $Top^S \rightarrow Top_{seen}^S$ .	90
4.12	Multitask results for $Top^S \rightarrow Top_{seen}^S$ .	91
4.13	Baselines and multitask results for $Tag^S \rightarrow Tag_{seen}^S$ .	91
4.14	Results $Top^S \rightarrow Top_{unseen}^S$ .	92
4.15	Results for $Tag^S \rightarrow Tag_{unseen}^S$ .	92
4.16	Results for $Tag^S \rightarrow Top_{unseen}^S$ .	93
5.1	Size and class distribution of the datasets used in the experiments. HSR is a hate speech instance ratio over all data.	98
5.2	HurtLex Categories.	100
5.3	Results of cross-lingual hate speech detection on the original distribution of training sets.	106
5.4	Results of cross-lingual hate speech detection on the balanced training set.	107
5.5	Results of cross-lingual hate speech detection on individual datasets with the original training set distribution.	108
5.6	Results of cross-lingual hate speech detection of each dataset on the balanced training set.	109
5.7	Results of cross-lingual hate speech detection with additional external resource on the balanced training set.	110
5.8	Dataset topical focuses and its collection time.	112
6.1	AMI IberEval Dataset label distribution.	120
6.2	AMI EVALITA Dataset label distribution (training/test).	120
6.3	HatEval Dataset label distribution. Hate speech target: Women.	122
6.4	HatEval Dataset label distribution. Hate speech target: Immigrant.	122
6.5	HurtLex Categories.	124
6.6	List of features of best-performing systems on each dataset.	126
6.7	Results of Automatic Misogyny Identification Experiment on AMI Dataset Task A.	127
6.8	Result of Experiment on SubTask B.	129
6.9	Dataset label distribution of OLID. OFF : Offensive; NOT : Not Offensive; TIN : Targeted Insult; UNT : Untargeted; IND : Individual; OTH : Other; GRP : Group.	130
6.10	Result of Cross-domain Automatic Misogyny Identification Experiment.	133
6.11	Result of Experiment by Combining Two Datasets in Cross-Domain Classification of Misogyny.	134
6.12	Result of Cross-lingual Automatic Misogyny Identification Experiment.	138
6.13	Ablation test result of the best system on English AMI IberEval.	139
6.14	Ablation test result of the best system on Spanish AMI IberEval.	140
6.15	Top ten features based on SVM weight on Spanish AMI IberEval task.	140

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Abusive Language . . . . .	2
1.2	Abusive Language in Social Media . . . . .	7
1.3	Computational Linguistics Approaches . . . . .	9
1.4	Open Challenges . . . . .	12
1.4.1	Swear Words in Abusive Language Detection . . . . .	12
1.4.2	Abusive Language Detection in Multidomain Settings . . . . .	14
1.4.3	Abusive Language Detection in Multilingual Settings . . . . .	14
1.5	Automatic Misogyny Identification . . . . .	15
1.6	Research Questions . . . . .	16
1.7	Contributions . . . . .	18
1.8	Structure of the Thesis . . . . .	19
<b>2</b>	<b>State of the Art</b>	<b>23</b>
2.1	Swear Words in Abusive Language Detection Studies . . . . .	23
2.1.1	Swearing in Online Content . . . . .	23
2.1.2	Contextual Swearing . . . . .	24
2.1.3	Swear Words Corpora . . . . .	25
2.1.4	Swearing and Abusive Content . . . . .	25
2.2	Abusive Language Detection in Multidomain Settings . . . . .	26
2.2.1	Abusive Language Domain . . . . .	26
2.2.2	Available Datasets for Multidomain Abusive Language Detection .	27
2.2.3	Proposed Approaches in Multidomain Abusive Language Studies .	30
2.3	Abusive Language Detection in Multilingual Settings . . . . .	36
2.3.1	Abusive Language Detection and Cross-lingual Settings . . . . .	37
2.3.2	Available Datasets for Multilingual Abusive Language Detection .	37
2.3.3	Proposed Approaches in Multilingual Abusive Language Studies .	42
2.4	Automatic Misogyny Identification . . . . .	48
2.4.1	The Phenomenon of Misogyny . . . . .	48
2.4.2	Misogyny Detection in Social Media . . . . .	49
2.5	Summary . . . . .	49

<b>3 Swear Words and Abusive Language Detection</b>	<b>53</b>
3.1 Motivation . . . . .	54
3.2 Corpus Creation and Analysis . . . . .	54
3.2.1 Corpus Collection . . . . .	54
3.2.2 Annotation Task and Process . . . . .	56
3.2.3 Annotation Results and Disagreement Analysis . . . . .	58
3.2.4 Corpus Extension . . . . .	59
3.3 Swear Words Abusiveness Prediction . . . . .	61
3.3.1 Sequence Labeling Task . . . . .	61
3.3.2 Simple Text Classification Task . . . . .	62
3.3.3 Target-based Abusiveness Prediction of Swear Words . . . . .	64
3.4 Swear Words in Abusive Language Detection . . . . .	67
3.4.1 Task Description and Experimental Settings . . . . .	67
3.4.2 Results . . . . .	70
3.5 Summary . . . . .	71
<b>4 Multidomain/Multitarget Hate Speech Detection in Social Media</b>	<b>73</b>
4.1 Motivation . . . . .	74
4.2 Cross-dataset Abusive Language Detection . . . . .	75
4.2.1 Datasets . . . . .	75
4.2.2 Experimental Settings . . . . .	75
4.2.3 Results and Analysis . . . . .	76
4.3 Cross-topic and Cross-target Hate Speech Detection . . . . .	78
4.3.1 Datasets . . . . .	79
4.3.2 Generalizing Hate Speech Phenomena Across Multiple Datasets . . . . .	82
4.3.3 Results of Generalizing Hate Speech Across Datasets . . . . .	84
4.3.4 Multitarget Hate Speech Detection . . . . .	86
4.3.5 Results on Multitarget Hate Speech Detection . . . . .	89
4.4 Summary . . . . .	93
<b>5 Multilingual Hate Speech Detection in Social Media</b>	<b>95</b>
5.1 Motivation . . . . .	96
5.2 Objectives . . . . .	96
5.3 Data and Resources . . . . .	97
5.3.1 Datasets . . . . .	97
5.3.2 Language Representation and External Resources . . . . .	99
5.4 Experiments . . . . .	100
5.5 Result and Analysis . . . . .	104
5.6 Discussion . . . . .	107
5.6.1 External Knowledge on Hate Words . . . . .	108
5.6.2 Dataset Topical Focuses . . . . .	112
5.7 Summary . . . . .	112

<b>6 Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study</b>	<b>115</b>
6.1 Motivation . . . . .	116
6.2 Automatic Misogyny Detection: Task and Datasets . . . . .	116
6.2.1 Task Definition . . . . .	117
6.2.2 Datasets . . . . .	117
6.2.3 Related Datasets . . . . .	120
6.3 Automatic Misogyny Identification Experiment . . . . .	121
6.3.1 Traditional Models . . . . .	123
6.3.2 Neural Based Models . . . . .	125
6.3.3 Results . . . . .	126
6.4 Relationship between Misogyny and Other Abusive Phenomena . . . . .	130
6.4.1 Experimental Setup . . . . .	130
6.4.2 Results . . . . .	131
6.5 Cross-Lingual Automatic Misogyny Identification Experiments . . . . .	132
6.5.1 Experimental Setup . . . . .	132
6.5.2 Results . . . . .	136
6.6 Discussion . . . . .	137
6.6.1 Automatic Misogyny Identification Task . . . . .	139
6.6.2 Relationship between Misogyny and other Abusive Phenomena . .	146
6.6.3 Cross-lingual Automatic Misogyny Identification . . . . .	147
6.7 Summary . . . . .	147
<b>7 Conclusion and Future Works</b>	<b>149</b>
7.1 Conclusion . . . . .	149
7.2 Research Contribution . . . . .	154
7.3 Future Works . . . . .	155

# Chapter 1

## Introduction

Abusive language is becoming a relevant issue in social media platforms such as Facebook and Twitter. The rise of the phenomenon also seems to be influenced by anonymity given to users and the lack of effective regulation provided by these platforms. On the one hand, social media provide a facility for improving social connectedness between people by amplifying their relationships. On the other hand, this facility can also be exploited to propagate toxic content such as hate speech or other forms of abusive language. In extreme cases, the hatred promoted in social media could escalate into dangerous criminal acts. Given the current rate of user-generated content produced every minute, manually monitoring abusive behaviour in social media is impractical. Facebook and Twitter also made efforts to moderate and remove abusive contents from their platforms<sup>1</sup> by providing clear policies on hateful conducts<sup>2</sup>, implementing user report mechanisms, and employing human content moderators to filter the abusive posting. However, these efforts are not a scalable and long-term solution to this problem. The latest approaches developed show that building a robust system to detect abusive language automatically is still a challenge. Most current studies in abusive language detection tasks only focus on a single language, mostly English, and tackle a single abusive language phenomenon, e.g., hate speech or sexism or racism, and so on, rather than accounting for multiple phenomena and how they are interconnected.

However, abusive language in social media is not limited to specific languages, and it features multiple abusive phenomena. As a matter of fact, most popular social media are multilingual, as users are encouraged to express themselves spontaneously in their mother tongue, and online social conversations are characterized by multiple different topics. Therefore, in a variety of languages and contexts, there is a considerable urgency to prevent online hate speech from spreading virally, becoming a significant factor in online and offline serious crimes against minorities or vulnerable categories. Specifically, robust approaches are needed for abusive language detection in a multidomain and multilingual environment, which will also enable the implementation of effective tools that could be employed to support both monitoring and content moderation activities such as automatic

---

<sup>1</sup><https://time.com/5739688/facebook-hate-speech-languages/>

<sup>2</sup><https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

moderation and flagging of potentially hateful users and posts, also for guaranteeing better compliance to governments demands to counteract the phenomenon [EU Commission, 2016].

This thesis proposes a deep investigation in building a robust model to automatically detect abusive language in social media, which is articulated in several focuses. First, we explore the use of swear words, which becomes one of the primary challenges in abusive language detection tasks, and provide further analysis of its role in this task. Second, we investigate around the challenge of building domain-agnostic models to detect abusive language. Third, we also explore abusive language detection in a multilingual setting, by focusing on the challenge to develop language-agnostic models. Finally, we focus on a specific task of abusive language related to hate speech against women, called automatic misogyny identification (AMI), to investigate how it is possible to address the aforementioned challenge: swear word use, multidomain, and multilingual aspect.

This chapter is organized as follow. Section 1.1 introduces the theoretical concepts about abusive language and its related phenomena. Section 1.2 describes several problems in identifying abusive language phenomena in social media and the urgency of building models to automatically detect such phenomena. Next, in Section 1.3 we briefly introduce current approaches to detect the abusive language in social media based on previous studies. In Section 1.4, we introduce three different open challenges that guided our research. First, in Subsection 1.4.1 we present a short introduction regarding to swear word role in abusive language detection. Second, in Subsection 1.4.2 we continue providing motivating arguments about the urgency of developing a robust model to detect abusive content across domains. Third, Subsection 1.4.3 presents the open challenges related to the development of a robust model to detect online abusive language across languages. A brief introduction about misogynistic online behaviours and the automatic detection of misogyny in social media is presented in Section 1.5. Finally, Section 1.6 describes research questions, objectives, contributions, and structure of this manuscript.

## 1.1 Abusive Language

There are several definition of abusive language based on the dictionaries including: “using harsh, insulting language”<sup>3</sup>, “using rude and offensive words”<sup>4</sup>, and “using offensive and insulting language”<sup>5</sup>. Therefore we can define abusive language as *verbal messages which use harsh, rude, offensive, and/or insulting words in an inappropriate way and which may also include profanity and slurs to demean the dignity of an individual or group of people*. In the early stage of online abusive language study, the term “harassment” was used to define the abusive phenomena in the online communication [Yin et al., 2009]. Several years later, newer studies proposed to include some terms to describe online abuse, including hate speech, self-harm, sexual violence, and reputation damaging rumours [Pater et al., 2016, Guberman and Hemphill, 2017]. In the further development of the studies, hate

---

<sup>3</sup><https://www.merriam-webster.com/dictionary/abusive>

<sup>4</sup><https://dictionary.cambridge.org/dictionary/english/abusive>

<sup>5</sup><https://www.macmillandictionary.com/dictionary/british/abusive>

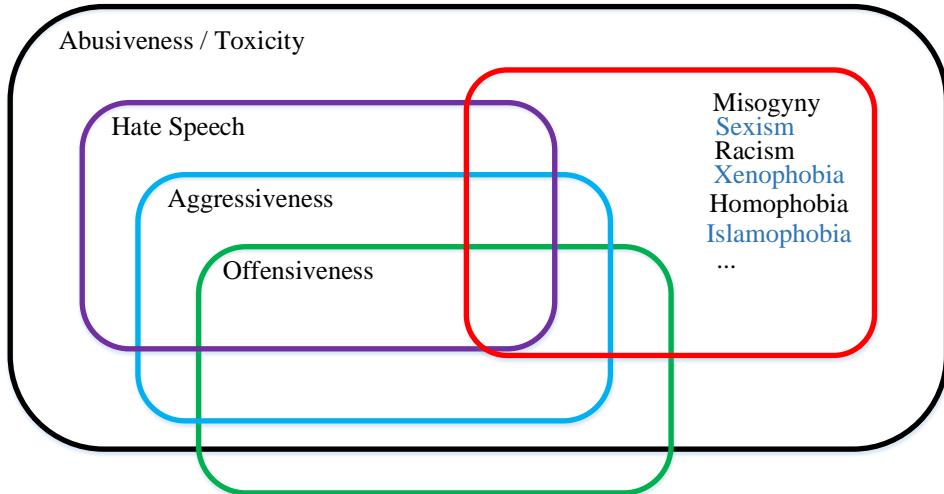


Figure 1.1: Relation between topical focuses in abusive language phenomena. Elaborated based on [Poletto et al. \[2020\]](#)

speech was becoming the main terms used by several authors to focus on investigating the online abuse [[Nobata et al., 2016](#), [Ross et al., 2017](#), [Waseem and Hovy, 2016](#), [Fortuna and Nunes, 2018](#)].

Abusive language is usually used as an umbrella term, which covers several related phenomena from a simple obscene and profanities to threats and severe insults [[Kiritchenko and Nejadgholi, 2020](#)]. Considering the different definitions of terms commonly used to refer to online abusive phenomena reported in Table 1.1, we can see that *abusive language* has a wider definition that covers other phenomena, including hate speech, insulting language, derogatory language, and also profanity. This is also the view depicted in recent studies by [Poletto et al. \[2020\]](#) (see Fig 1.1<sup>6</sup>), which returns a possible way to map the relation between hate speech and other related phenomena., including abusiveness/toxicity, aggressiveness, offensiveness, and other manifestations of hatred towards certain targets such as misogyny, racism, homophobia, and so on. The attempt to design a framework

---

<sup>6</sup>Notice that we elaborated on the original picture in [Poletto et al. \[2020\]](#), by highlighting and positioning three additional abusive phenomena, which will be covered in the exploration conducted in this thesis: sexism, xenophobia, and islamophobia.

to highlight the relationships among different abusive online phenomena was originally developed in [Poletto et al. \[2020\]](#) by analyzing and surveying available datasets in computational linguistics literature. Based on this framework, abusive language has a broader coverage than other forms of abusive behaviour. Their further investigation highlights that there are more than 20 different topical focuses of abusive phenomena introduced by previous works in the field, despite some of them did not provide a clear definition of the phenomenon. Let us also observe that, based on Figure 1.1, offensiveness intersects with abusiveness, but includes also phenomena which are not abusive. This is in tune with the study in [Caselli et al. \[2020b\]](#). Based on the annotation guideline proposed by [Zampieri et al. \[2019b\]](#), [Caselli et al. \[2020b\]](#) concluded that offensive language might not be necessarily abusive. On the other hand, as argued also by [Ibrohim and Budi \[2018\]](#) abusive instances are not necessarily offensive.

Summarizing, considering that most recent studies show that *abusive language* is getting commonly used as the broader context to cover and integrate all other concepts covered in literature [[Pamungkas and Patti, 2019](#), [Waseem et al., 2017](#), [Karan and Šnajder, 2018](#), [Founta et al., 2018](#)], allowing to draw the proper intersections with them, in this thesis, we define our focus and scope as related to the automatic detection on abusive language phenomena, because of the wider coverage and negative impact on society in general.

Example 1 :

“Go kill yourself”, “You’re a sad little f\*ck” [\[Hee et al., 2015\]](#)

Example 2 :

Women who strive to be ‘equal’ to men lack ambition #YesAllMen. [\[Fersini et al., 2018c\]](#)

Example 3 :

All you f\*cking f\*ggots were laughing at her too with that other f\*cking dumb c\*nt liberal Ellen DeGeneres. [\[Basile et al., 2019\]](#)

Example 4 :

“most of them come north and are good at just mowing lawn”. [\[Dinakar et al., 2011\]](#)

Abusive language phenomena are featured by different kinds of *topical focuses*. In their survey [Poletto et al. \[2020\]](#) observed 21 different topical focuses (see Table 1.2) mentioned in the previous studies, focusing on the abusive language field. In abusive language study *topical focus* can be defined as the specific topic or abusive phenomena addressed, which may also be related to the target of abuse [\[Poletto et al., 2020\]](#). For example, several abusive phenomena are closely related to their targets, such as misogyny, which specifically targets women, islamophobia that targets Muslims, and homophobia, which a form of hatred related to sexual-orientation discrimination. However, some topics overlap with each other, i.e., misogyny and sexism or xenophobia and racism, both because they have a common target, but also due to a certain degree of subjectivity inherent in defining these phenomena. Figure 1.2 shows an attempt to map the interaction between targets in several abusive language phenomena. We can see that *target groups*

No.	Terms	Definition
1.	Abusive language	The use of harsh, insulting language. It can include hate speech, derogatory language, and also profanity.
2.	Cyberbullying	The electronic posting of mean-spirited messages about a person (such as a student) often done anonymously.
3.	Toxic Language	Toxic use of language is synonym of <b>aggressive language</b> , used to hurt. It is rude and disrespectful and leads the interlocutors to leave the conversation.
4.	Hate Speech	Speech expressing hatred of a particular group of people
5.	Harassment	The act of systematic and continued unwanted and annoying actions of one party or a group, including threats and demands. The purposes may vary, including racial prejudice, personal malice.
6.	Insulting language	Giving or intended to give offense : being or containing an insult.
7.	Vulgar language	The act of depicting or referring to sexual matters in a way that is unacceptable in polite society
8.	Offensive language	The use of language which causes intense displeasure, disgust, or resentment.

Table 1.1: Definition of abusive language and related terms.

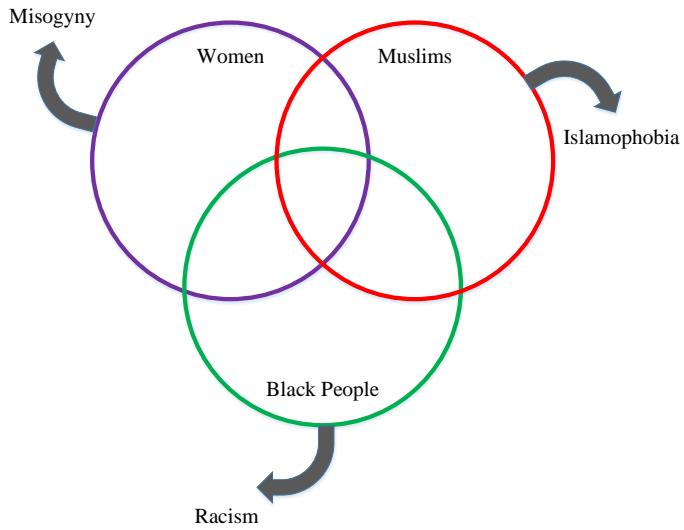


Figure 1.2: The interaction between targets of abusive phenomena.

could be overlapped, and a target entity could be part of more than one target group. Individuals with more than one target identity could experience higher and stronger abusiveness levels, as they are more vulnerable and targeted by more hate attackers, as also suggested by recent studies on intersectionality Crenshaw [2015], a field still little explored in online abusive language detection works in the computational linguistics field. For example, Alberta's Muslim community, which mostly includes Black Muslim Women, experienced a series of racially motivated assaults in Calgary and Edmonton, Canada<sup>7</sup>. Further investigations report that most attackers were men, but then in particular a white supremacist group also became a suspect, based on evidence of several symbols used in the Covid-19 restriction protest during the same period in these regions.

Another definitional attempt to be mentioned can be found in Waseem et al. [2017], where a two-fold categorization of abusive language is proposed according to two dimensions: (i) the target of abuse, and (ii) the degree to which it is explicit. Considering the target, abuse can either be directed towards specific individuals (see Example 1), or can be also generalized towards a group of people based on ethnicity, gender, sexual orientation, or other identities (see Example 2). Cyberbullying and trolling can be categorized as directed abuse, while hate speech could be both directed and generalized. Considering the explicit/implicit dimension, we have *explicit* abusive language when the abusive context of a message is unambiguous, for example in an utterance which contains homophobic and racial slurs (see Example 3). Instead, we have *implicit* abusive language when the abusive context of a message could not be interpreted instantly, for example because

---

<sup>7</sup><https://globalnews.ca/news/7721850/hate-crime-alberta-attacks-black-muslim-women/>

Abusiveness; Aggressiveness; Anti-Roma; Child sexual abuse; Cyberbullying; Flames; Harassment; Homophobia; Hate speech; Islamophobia; Obscenity; Profanity; Offensiveness; Personal attacks; Racism; Sexism; Misogyny; Threats; Violence; Toxicity; White supremacy.
--

Table 1.2: Topical focuses introduced by previous studies.

of the use of figurative devices such as sarcasm (see Example 4). Notice that not only sarcasm, but other subtle form of abusive language could also be found in several texts [Breitfeller et al., 2019, Wiegand et al., 2021], including microaggression and negative stereotype [Wiegand et al., 2021, Kiritchenko and Nejadgholi, 2020, Bodapati et al., 2019, Sanguinetti et al., 2020].

To summarize, based on the literature, abusive language is usually used as an umbrella term several abusive phenomena such as aggressiveness, offensiveness, hate speech, racism, sexism, cyberbullying, homophobia, and etc, which has a broader context covering the full range of inappropriate content, from a simple obscene and profanities to threats and severe insults [Kiritchenko and Nejadgholi, 2020]. The following section introduce the phenomena of abusive language in social media, including various impacts of this phenomena and possible solutions to deal with it.

## 1.2 Abusive Language in Social Media

In the digital era, social media have an integral role in online communication, facilitating its users to publish and share contents providing accessible ways to express their feelings and opinions about anything anytime. Social media is convenient, as sites allow users to reach people worldwide, which could potentially facilitate a positive and constructive conversation between users. Based on Emarsys report<sup>8</sup>, the number of social media users reach 3.2 billion worldwide in 2019, which is equivalent to 42% of the population number. Another statistic provided by GlobalWebIndex report<sup>9</sup> uncovers that an average of 3 hours is spent per day per person on social network and messaging in 2019.

People also increasingly tend to use social media like Facebook and Twitter as their primary source of information and news consumption. There are several reasons behind this tendency, such as the simplicity to gather and share the news and the possibility of staying abreast of the latest news and updated faster than with traditional media. An important factor is also that people can be engaged in conversations on the latest breaking news with their contacts by using these platforms. Pew Research Center's 2017 report<sup>10</sup> shows that two-thirds of U.S. adults gather their news from social media, where Twitter is the most used platform. Meanwhile, Youtube and Facebook have become the most widely used social media among Americans, where more than 50% of them have

<sup>8</sup><https://emarsys.com/learn/blog/top-5-social-media-predictions-2019/>

<sup>9</sup><https://blog.globalwebindex.com/trends/2019-in-review-social-media/>

<sup>10</sup><https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/>

accessed their social media platforms at least once a day<sup>11</sup>.

Within the fields of Artificial Intelligence and Natural Language Processing, the vast amount of data produced by social media allowed the research community to tackle more in-depth long-standing questions such as understanding, measuring and monitoring the sentiment of the users towards certain topics or events [Cambria et al., 2017], expressed in mere texts or also by relying on other visual and vocal modalities [Poria et al., 2018]. Robust and effective approaches are made possible by the rapid progress in supervised learning technologies and by the huge amount of user-generated content available online, especially on social media. Such techniques are typically motivated by purposes such as extracting user opinions on a given product or polling political stance. There is an ever-increasing awareness of the need to take a holistic approach to sentiment analysis by handling the many finer-grained tasks involved in extracting meaning, polarity and specific emotions from the text, like the detection of sarcasm [Majumder et al., 2019, Sulis et al., 2016].

However, there is a downside to the freedom of expression given by social media, as more and more episodes of hate speech and online harassment happen in social media. This is due especially to the freedom and anonymity given to users and to the lack of effective regulations provided by the social network platforms [Rainie et al., 2017, Themeli et al., 2021]. With the huge amount of users, social media becomes a beneficial medium for hate groups to reinforce their views. Social media platforms also offer the opportunity for violent actors to propagate their acts, even with a possibility to gain higher reachability when the post becomes *viral* [Mathew et al., 2018]. Abusive language and hate speech behaviour are becoming significant problems in online communication on social media. It has been proven to be detrimental not only for the mental health of the victims, but also for the society at a larger scale [Langham and Gosha, 2018]. Some countries have developed regulation to prohibit the abusive languages online such as Germany, the United Kingdom, and India. However, inconsistency of policies to regulate online harassment across different social media platforms and countries is also a big issue for the communities to combat abusive language in online environments [Pater et al., 2016]. Abusive language online can also potentially be escalated in real-world problem resulting in dangerous criminal acts. Several acts of violence have been observed as a result of incitement initialized from social media posts and other online speech, such as:

1. In Germany, anti-refugee Facebook posts by the far-right Alternative for Germany party were found to be correlated with the refugees attack.
2. In Rohingya Myanmar, Facebook posts were also exploited by the military leader and Buddhist nationalist to slur and demonize the Rohingya Muslim minority. These incitements have contributed to a growing climate of hatred, resulting in a murder of thousands of civilians.
3. The shooting in Pittsburgh synagogue was also originated from a conspiracy in a social media network called Gab. The shooter falls in with conspiracy that Jews

---

<sup>11</sup><https://www.pewresearch.org/internet/fact-sheet/social-media/>

sought to bring immigrants into the United States, and render white minority.

4. The attack on Tamil Muslim minority in Sri Lanka was inspired by a rumour spread in various social media and messaging platforms including Facebook, WhatsApp, and Viber.

These mentioned criminal acts proved that there is an urgency to fight abusive and hate speech online. Especially, in certain circumstances the abusive utterances should be timely identified and considered for removal before escalated into a serious criminal act in the real-world.

The online abuse problem affects not only the abuse victims but also social medial platforms and governments [Corazza et al., 2020a]. This has determined a growing interest in artificial intelligence and natural language processing tasks related to social and ethical issues, also encouraged by the global commitment to fighting extremism, violence, fake news and other plagues affecting the online environment. In this perspective, let us mention the latest trends of “AI for social good”, with emphasis on developing applications for maximizing the “good” social impacts, while minimizing the likelihood of harm, e.g., suicidal ideation detection for early intervention [Gaur et al., 2019] and recent works on the prevention of sexual harassment [Khatua et al., 2018], sexual discrimination [Khatua et al., 2019], and cyberbullying and trolling [Cambria et al., 2010, Menini et al., 2019], or on hate speech counter-narratives [Chung et al., 2019], with focus on generating positive responses, after tackling with detection of abusive content published online, encouraging the community to adopt a proactive approach to transform the toxic environments into positive ones [Jurgens et al., 2019a].

Several studies have been proposed to combat online abuse. Most studies more focused on the automatic detection of abusive language online, using recent computational linguistics approaches. However, there is also growing interest in studies to prevent online abuse before it is happened by proposing proactive techniques to eliminate abusive language online. In this thesis, we focus mostly on social media data, specifically data gathered from Twitter platforms. This focus is also motivated to the wide availability of corpora, mostly featured by Twitter data, which may be due to the convenience of scraping tweet samples using the available Twitter API and to the Twitter policies on making the data publicly available. The next section presents an overview of the current methods for detecting abusive language in social media texts. A more complete review of state of the art in abusive language detection is included in Chapter 2.

### 1.3 Computational Linguistics Approaches

Given the vast amount of social media contents produced every minute, manually monitoring social media content is impractical. To this end, many studies have been proposed with the spirit to fight online abuse in social media. Some studies were more focused to proactively prevent abusive language. For example, Munger [2017] proposed a simple approach to intervene the use of toxic language (the n-words) by developing a human-looking bot for replying an abusive content with a fixed comment about the harm caused by such

languages to appeal an empathy. Some other works also proved that counter narrative for abusive speech is also effective for limiting the effect of hate speech [Chung et al., 2019, Mathew et al., 2019, Wright et al., 2017]. However, obtaining sufficient and reliable real-world example data to this direction is challenging. Most other works were focusing on the automatic detection of abusive language. In the early stage of development, only a few studies used an unsupervised approach such as proposed by Gitari et al. [2015], where a manually-built lexicon was used to identify hate speech contents. Most works tackle this task by adopting a supervised approach, employing several machine learning models either traditional-based or neural-based models.

**Challenges in Social Media Data.** Working with social media data is a very challenging task. The social media data usually contain a valuable knowledge for such information extraction task, but they are usually very noisy and full of informal language [Baldwin et al., 2013]. There are several properties of social media data based on the study of Baldwin et al. [2013] including: i) code-mixed language is often present; ii) out-of-vocabulary words are a constant as well as iii) grammatical errors. Social media data also usually have very limited context, which is an important issue for abusive language detection task, since it is difficult to classify a text as either abusive or not without context. Other important clues for abusive detection task such as facial expression, gestures, and voice tones (which are recognized in face-to-face communication) are also absent in the social media data. However, social media contents have some signals that can be exploited to partially resolve the context of such texts including emoji, emoticon, hashtag, URL, mention, and etc.

**Supervised Approaches.** As previously mentioned, most works adopted supervised approach to automatically detect abusive content. Among the earliest proposed solutions several works relied on machine learning models with manually engineered features, including decision trees [Burnap and Williams, 2015, Agarwal and Sureka, 2017], naive bayes classifiers [Agarwal and Sureka, 2017, Kwok and Wang, 2013], support vector machines [Badjatiya et al., 2017, Burnap and Williams, 2015, Warner and Hirschberg, 2012], logistic regression [Davidson et al., 2017, Waseem and Hovy, 2016, Badjatiya et al., 2017, Fehn Unsvåg and Gambäck, 2018], and random forest [Badjatiya et al., 2017, Burnap and Williams, 2015, Agarwal and Sureka, 2017]. Different kind of features have been tested, such as lexical features (e.g., bag of words, n-grams, TF-IDF), syntactic features (e.g., part of speech and dependency relation), stylistic features (e.g., number of characters, text length, punctuation), as well as Twitter-specific features (e.g., the number of user mentions, hashtags, URLs, social network information [Mishra et al., 2018], and user-related features [Fehn Unsvåg and Gambäck, 2018, Waseem and Hovy, 2016]). Recent works relied on to use of neural-based approaches such as Long Short-Term Memory (LSTM) [Vigna et al., 2017, Mishra et al., 2018], Bidirectional Long Short-Term Memory (Bi-LSTM) [Qian et al., 2018a], Gated Recurrent Unit (GRU) [Mossie and Wang, 2019], and Convolutional Neural Network (CNN) [Badjatiya et al., 2017]. These models are usually coupled with language representations such as FastText<sup>12</sup>, word2vec<sup>13</sup>,

---

<sup>12</sup><https://fasttext.cc/>

<sup>13</sup><https://code.google.com/archive/p/word2vec/>

Shared Task	Event	Topical Focus	Languages
AMI@IberEval 2018	IberEval 2018	Misogyny	EN, ES
AMI@Evalita 2018	Evalita 2018	Misogyny	EN, IT
HaSpeeDe 2018	Evalita 2018	Hate Speech	IT
Offensive Language Identification	GermEval 2018	Offensiveness	DE
TRAC-1	COLING 2018	Aggresiveness	EN, HI
HASOC 2019	FIRE 2019	Hate Speech, Offensiveness	EN, DE, HI
HatEval	SemEval 2019	Racism, Misogyny	EN, ES
Hate Speech Detection (HSD)	VLSP 2019	Hate Speech	VI
Automatic Cyberbullying Detection	PolEval 2019	Cyberbullying	PL
OffensEval 2019	SemEval 2019	Offensiveness	EN
TRAC-2	LREC 2020	Aggresiveness, Misogyny	EN, BN
AMI@Evalita 2020	Evalita 2020	Misogyny	IT
HaSpeeDe 2020	Evalita 2020	Hate Speech	IT
DankMemes	Evalita 2020	Hate Speech	IT
OffensEval 2020	SemEval 2020	Offensiveness	EN, AR, DA, GR, TR
OSACT4	LREC 2020	Offensiveness	AR
Toxic Spans Detection	SemEval 2021	Toxicity	EN
HaHackathon	SemEval 2021	Offensiveness	EN
EXIST	IberLEF 2021	Sexism	EN, ES
HASOC-ArMI	FIRE 2021	Misogyny	AR
HASOC-Hate Speech	FIRE 2021	Hate Speech	EN, HI
HASOC-Offensive Language Identification	FIRE 2021	Offensiveness	TA, ML
HASOC-Abusive and Threatening Language Detection	FIRE 2021	Abusiveness, Threat	UR
Profiling Hate Speech Spreader	CLEF 2021	Hate Speech	EN, ES

Table 1.3: Shared tasks in the abusive language detection research field.

and ELMo [Peters et al., 2018]. Meanwhile, most of state of the art models in several recent abusive languages detection tasks exploited transformer-based architecture namely BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019]. Additionally, some studies also proposed to exploit external knowledge, by infusing features extracted from several available affective resources into neural-based models, to provide extra information in general abusive language detection tasks [Koufakou et al., 2020] and also domain independent knowledge in the multidomain abusive language detection task [Pamungkas and Patti, 2019, Corazza et al., 2019].

Recent studies show that the use of swear word is very relevant to several abusive language detection subtasks including offensive language [Nobata et al., 2016], cyberbullying [Van Hee et al., 2015, Michal et al., 2010], and hate speech [Malmasi and Zampieri, 2018]. The next section present an introduction related to the role of swear word use in several abusive language detection tasks.

## 1.4 Open Challenges

In this section, we introduce several open challenges in abusive language detection, that guided the thesis path, which covered three main themes.

### 1.4.1 Swear Words in Abusive Language Detection

Swearing is the use of taboo language (also referred to as bad language, swear words, offensive language, curse words, or vulgar words) to express the speaker's emotional state to their listeners [Jay, 1992, 1999]. Not limited to face to face conversation, swearing also occurs in online conversations, across different languages, including social media and online forums, such as Twitter, typically featured by informal language and spontaneous writing. Twitter is considered a particularly interesting data source for investigations related to swearing. According to the study in Wang et al. [2014a] the rate of swear word use in English Twitter is 1.15%, almost double compared to its use in daily conversation (0.5 – 0.7%) as observed in previous work by Jay [1992], Mehl and Pennebaker [2003]. The work by Wang et al. [2014a] also reports that a portion of 7.73% tweets in their random sampling collection is containing swear words, which means that one tweet out of thirteen includes at least one swear word. Interestingly, they also observed that a list of only seven words covers about 90% of all the swear words occurrences in their Twitter sample: *f\*ck, sh\*t, \*ss, b\*tch, n\*gga, h\*ll, and wh\*re*.

Swearing in social media can be linked to an abusive context, when it is intended to offend, intimidate or cause emotional or psychological harm, contributing to the expression of hatred, in its various forms. In such contexts, indeed, swear words are often used to insult, such as in case of sexual harassment, hate speech, obscene telephone calls (OTCs), and verbal abuse [Jay et al., 2006, Jay and Janschewitz, 2008]. However, swearing is a multifaceted phenomenon. The use of swear words does not always result in harm, and the harm depends on the context where the swear word occurs [Jay, 2009a]. Consider for instance the two following tweets containing swearing from the *StackOverflow Offensive Comments* dataset [Fišer et al., 2018]:

If you don't have the answer, move on to the next **f\*cking** question and mind your own **f\*cking** business

Sh\_Khan: **f\*cking** genius. Thank you

In the first example, it is obvious that the swear word is used to insult, thus this is an instance of abusive language. However, the second example shows the use of the same swear word in a casual setting, to emphasize an emotion of gratitude without intention to be offensive [Pinker, 2007, *emphatic swearing*].

Some studies even found that the use of swear words has also several upsides. Using swear words in communication with friends could promote some advantageous social effects, including strengthen the social bonds and improve conversation harmony, when swear word is used in ironic or sarcastic contexts [Jay, 2009a]. Another study by Stephens and Umland [2011] found that swearing in cathartic ways is able to increase pain tolerance.

Furthermore, Johnson [2012] has shown that the use of swear words can improve the effectiveness and persuasiveness of a message, especially when used to express an emotion of positive surprise. Also accounts of appropriated uses of slurs should not be neglected [Bianchi, 2014], that is those uses by targeted groups of their own slurs for non-derogatory purposes (e.g., the appropriation of ‘nigger’ by the African-American community, or the appropriation of ‘queer’ by the homosexual community).

In recent years, more and more studies focused on abusive language detection which covers hate speech, cyberbullying, trolling, and offensive language [Waseem et al., 2017, Schmidt and Wiegand, 2017, Michal et al., 2010]. Swear words play an important role in these tasks, providing a signal to spot an offensive utterance [Malmasi and Zampieri, 2018]. However, the presence of swear words could also lead to false positives when they occur in a casual context [Chen et al., 2012, Nobata et al., 2016, Van Hee et al., 2018, Malmasi and Zampieri, 2018]. Distinguishing between abusive and not-abusive swearing contexts seems to be crucial to support and implement better content moderation practices. Indeed, on the one hand, there is a considerable urgency for most popular social media, such as Twitter and Facebook, to develop robust approaches for abusive language detection, also for guaranteeing a better compliance to governments demands for counteracting the phenomenon (see, e.g., the recently issued EU commission *Code of Conduct on countering illegal hate speech online* [EU Commission, 2016]). On the other hand, as reflected in statements from the Twitter Safety and Security<sup>14</sup> users should be allowed to post potentially inflammatory content, as long as they are not-abusive<sup>15</sup>. The idea is that, as long as swear words are used but do not contain abuse/harassment, hateful conduct, sensitive content, and so on, they should not be censored.

Swear word use varies ranging from one to another abusive phenomenon. Some abusive phenomena are featured by specific swear words, which could become the strong signal to spot the abusive utterance. For example, racism is very closely related to the use of racial slurs such as “n\*gga”, “ch\*nk”, and “t\*welhead”, then homophobia which usually contains some homophobic slurs including “f\*ggot” and “dyk\*”, and finally misogynistic phenomenon is also closely related to the use of specific swear words to attack women such as “b\*tch”, “sl\*t”, and “c\*nt”. These examples prove that swear word use also plays an important role in the abusive language detection task across different domains and topical focuses. Furthermore, swear words are universally used across different languages and nations in the different geographical areas of social media users. However, swear words are very cultural-dependant, and the directly translated swear words from one language to another language could have a completely different sense. This provides a further challenge to transfer knowledge between languages for detecting abusive language in a cross-lingual setting.

---

<sup>14</sup><https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content>

<sup>15</sup>See for instance the Twitter Rules trying to determine what an abusive and hateful conduct is: <https://help.twitter.com/en/rules-and-policies/twitter-rules>

### 1.4.2 Abusive Language Detection in Multidomain Settings

The abusive language behaviour is multifaceted and available datasets are featured by different topical focuses. This makes abusive language detection a domain-dependent task, and building a robust system to detect general abusive content a challenge. Some studies attempted to bridge some of these subtasks by proposing cross-domain classification of abusive content. Some work has been done in the cross-domain classification of abusive language. In Waseem et al. [2018] the first attempt to deal with cross-domain classification in an abusive language detection task is reported, by proposing a multitasks learning (MTL) approach. They argue that MTL has the ability to share knowledge between two or more objective functions, so that it can leverage information encoded in one abusive language dataset to better fit others. They found that the difference of approaches in collecting and annotating datasets is the main factor which influences the performance of such model. In another study, Wiegand et al. [2018a] proposed to use high-level features by combining several linguistic features and lexicons of abusive words in the cross-domain classification of abusive microposts from different sources. Waseem et al. [2018] use multitask learning for domain transfer in a cross-domain hate speech detection task. Recently, Karan and Šnajder [2018] also addressed cross-domain classification in several abusive language datasets, testing the framework of Frustratingly Simple Domain Adaptation (FEDA) [Daumé III, 2007] to transfer knowledge between domains. Similarly, Pamungkas and Patti [2019] proposed a cross-domain classification of abusive language, employing a Long Short Term Memory (LSTM) network and a list of abusive keywords from the lexicon HurtLex [Bassignana et al., 2018], as a proxy to transfer knowledge across different datasets. Their main findings are that i) the model trained on more general abusive language dataset will produce more robust predictions, and ii) HurtLex is able to boost the system performance in cross-domain setting. Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] was also applied to the cross-domain setting in abusive language detection, as proposed by Swamy et al. [2019], Mozafari et al. [2019]. Both studies found that BERT is capable to share knowledge between one domain dataset to other domains, in the context of transfer learning. They argue that the main difficulty in cross-domain classification of abusive language is caused by dataset issues and their biases, with the consequent incapability of the datasets to capture the complete phenomenon of abusive language.

### 1.4.3 Abusive Language Detection in Multilingual Settings

Another prominent challenge in abusive language detection is the multilinguality issue. Even if in the last year abusive language datasets were developed for other languages, including Italian [Bosco et al., 2018, Fersini et al., 2018c], Spanish [Fersini et al., 2018c], and German [Wiegand et al., 2018b], most studies so far focused on English. Since most popular social media such as Twitter and Facebook goes multilingual, fostering their users to interact in their primary language, there is a considerable urgency to develop a robust approach for abusive language detection in a multilingual environment. Cross-lingual classification is an approach to transfer knowledge from resource-rich languages

to resource-poor ones. The main challenge in the cross-lingual setting is to deal with the language shift between one language to another. However, specifically in the abusive language research area, the challenge is not only to deal with language-shift but also domain-shift since different topical focuses and targets feature the available datasets. Meanwhile, cross-lingual abusive language detection has not been much explored yet by NLP scholars. We only found a few works describing participating systems developed for recent shared tasks on the identification of misogynous [Basile and Rubagotti, 2018] and offensive language [van der Goot et al., 2018], where some experiments in a cross-lingual setting is proposed. Basile and Rubagotti [2018] used the *bleaching* approach [van der Goot et al., 2018] to conduct cross-lingual experiments between Italian and English when participating to the automatic misogyny identification task at EVALITA 2018 [Fersini et al., 2018a]. Schneider et al. [2018] used multilingual embeddings in a cross-lingual experiment related to GermEval 2018 [Wiegand et al., 2018b]. Recent work by Pamungkas and Patti [2019] employs Multilingual Unsupervised or Supervised Word Embeddings (MUSE)<sup>16</sup> to build a joint-learning model for cross-lingual classification on the AMI task in three languages, namely Italian, English, and Spanish. In addition, there are studies on cross-lingual classification of abusive language, with a general topical focus. Finally, Ousidhoum et al. [2019] conducted a multilingual experiment on hate speech detection in three languages (i.e., English, France, and Arabic) by using Sluice Network [Ruder et al., 2017] and Babylon multilingual word embeddings [Smith et al., 2017].

This section described the open challenges in abusive language detection task regarding to several focuses including the use of swear words and the detection of abusive language across domains and languages. Specifically, the use of swear word could become an issue for discriminating whether a content is included into abusive utterance or part of freedom of speech. Furthermore, we also explicitly mention the urgency of building robust models which could detect abusive content in different domains/topical focuses and also different languages. A more complete survey of this study can be found in Chapter 2. The next section will introduce a specific online abusive phenomenon, called *misogyny*, which will become our case study to investigate the aforementioned challenges including investigating swear word role and conducting an experiment in both multidomain and multilingual scenario.

## 1.5 Automatic Misogyny Identification

There is a downside to the freedom of expression given by social media, as more and more episodes of hate speech and online harassment happen in social media. In recent years, hateful language and in particular the phenomenon of hate against women are exponentially increasing in social media platforms such as Twitter and Facebook [Poland, 2016, Hewitt et al., 2016], becoming a relevant social problem that needs to be monitored. Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in different and various ways, including social exclusion, sex discrimination, hostility,

---

<sup>16</sup><https://github.com/facebookresearch/MUSE>

androcentrism, patriarchy, male privilege, belittling of women, disenfranchisement of women, violence against women, and sexual objectification [Kramerae and Spender, 2000, Code, 2002]. Based on the recent Online Harassment report from Pew Research Center<sup>17</sup>, women are more likely to be targeted as subject of online harassment because of gender than men (11% vs. 5%). This is a concerning issue, since the study from Fulper et al. [2014] found that there is a strong association between the number of misogynistic tweets and the rape crime statistics in the United States.

The work of Hewitt et al. [2016] is a first study that attempts to detect misogyny in Twitter manually. The authors used several terms related to slurs against women to gather data from Twitter. The Automatic Misogyny Identification (AMI) campaign started by Anzovino et al. [2018] proposed a first benchmark dataset, capturing misogyny phenomena in Twitter. This dataset is a starting point for automatic misogyny identification, leading to two shared tasks focused on the detection of misogyny online, namely AMI IberEval 2018 [Fersini et al., 2018c] and AMI EVALITA 2018 [Fersini et al., 2018a]. AMI IberEval 2018 proposed an automatic misogyny identification task in two languages, Spanish (ES) and English (EN), while AMI EVALITA 2018 included Italian (IT) and English (EN). The task comprises two sub-tasks: i) classification of tweets as either misogynistic or not-misogynistic; ii) classification of misogynistic behaviour into 5 categories (derailing, dominance, discredit, sexual harassment and stereotype), and classification of the target of misogyny as active (individual) or passive (generic or group or women). These shared tasks succeeded in highlighting the barriers and difficulties of automatically detecting misogyny in social media.

In this section, we introduced a specific category of online abusive language, specifically targeting women. We also presented the AMI task, an evaluation benchmark for automatically detecting misogyny content in social media. This section completes the introduction part of this thesis. The next section will summarize the research questions and objectives of this thesis.

## 1.6 Research Questions

Building a robust model to detect abusive language is a challenge. The ambiguity of the swear word context is one of the challenges in the general abusive language detection task. On the one hand, swear words could help the abusive detector to spot abusive contents, but on the other hand, swear words could also deceive the abusive detection model when it is used in not abusive context. The remaining challenges are related to the current social media communication trend where abusive contents exist in both multidomain and multilingual environments. Therefore, developing an abusive language detection model needs to consider also the domain-topic shift and language shift issues. Starting from these challenges, the main objective of this thesis is to investigate the possibility and challenges of building a robust model to detect abusive language in social media. Following this objective, we propose four main research questions, which are articulated into three related questions.

---

<sup>17</sup><https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/>

1. What is the role of swear words in abusive language detection task?
  - How to model the swear word context in social media text as either abusive or not abusive?
  - Is it possible to automatically predict the abusiveness of a swear word within the tweet context?
  - Is the additional information about swear words abusiveness helpful for detecting abusive language?
  
2. How to build a robust model which facilitates domain transfer for detecting abusive language across different topical focuses and targets?
  - How to build a robust architecture to detect abusive languages with different coverage of abusive phenomena?
  - What is the role of domain-independent resource in improving the models' performance in cross-domain abusive language detection?
  - How to build a model which able to predict not only the abusiveness of a tweet but also its target?
  
3. How to build a robust model which facilitates language transfer for detecting abusive language across different languages?
  - What neural architectures are effective for transferring knowledge between language in hate speech detection task?
  - How effective are multilingual pre-trained models for language representation in cross-language hate speech detection?
  - What is the role of external multilingual knowledge in improving the models' performance in cross-lingual hate speech detection?
  
4. How the challenges on swear words use, and on experimenting in cross-domain and cross-lingual settings can be addressed considering a specific abusive phenomenon in social media, namely misogyny, a form of online hatred that is widespread across different countries, languages and cultures?
  - How the use of abusive words could become the important feature in distinguishing between misogynistic and non-misogynistic content in social media?
  - How is misogyny related to other abusive phenomena, and how do they inform each other towards detection of abusive language at large?
  - Is the knowledge about misogyny learned from one language informative to predict misogyny in other languages?

## 1.7 Contributions

Abusive language in social media has been recognized as an important issue which could detriment our society at a larger scale. However, eliminating abusive language from such social media platforms is not an easy task to be done. Detecting abusive language automatically is one of primary step in discarding abusive language in social media. In this thesis, we focus on exploring the possibility of building a robust model to detect abusive language across different domains and languages. In this line, the following contributions are made within the development of the present research:

1. We presented a brief overview of current approaches to deal with abusive language detection in multidomain and multilingual settings. We found that there is still not much model proposed in this direction, despite several datasets are already introduced in the recent studies. Overall, we conclude that developing a robust model to detect abusive language across domains and language is challenging. We discovered several issues which contributes to the difficulties of this task, which could become the main objective for the future works.
2. Swear words could become a problem in abusive language detection task across domain and language. It can be a valuable signal of abusive content, but it could also lead to a false positive when used in not abusive context. Resolving swear words context within an utterance is important for building an accountable model to detect abusive language detection. To this end, we conduct several investigations:
  - We develop a new benchmark Twitter corpus, called SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level, as either abusive or not abusive based on its context within the sentence.
  - We develop and experiment with supervised models to automatically predicting swear words as either abusive or not abusive. Such models are trained on the novel SWAD corpus, to predict the abusiveness of a swear word within a tweet.
  - We quantify the role of swear word use in abusive language detection task, by incorporating the knowledge about the abusiveness of swear words, to improve the performance of abusive language detection task.
3. Abusive language tasks are multifaceted and featured by several abusive phenomena. Having a robust model to detect different kinds of abusive content in social media is crucial. We conduct the following tasks to achieve this objective:
  - We characterize the available datasets as capturing various phenomena related to abusive language, and investigate this characterization in cross-domain classification with several machine learning models.
  - We experiment with the additional feature obtained from external resources as a domain-independent feature to transfer knowledge between domain.

- We adopt a multitask architecture which allows the model to predict not only the hatefulness of a given content but also its target of the hateful language.
4. Abusive language is a global phenomenon, while current studies only focused on English. Having a language-agnostic model to detect hate speech in social media is also a challenge. In this sense, we also explore the possibility of building a robust model to detect abusive language across different languages. The following tasks need to be done to achieve this objective:

- We propose a zero-shot cross-lingual hate speech detection experiment to investigate the feasibility of detecting hate speech in a multilingual setting.
- We propose a joint-learning architecture by adapting several ideas from sentiment analysis studies.
- We test and evaluate three publicly available multilingual pre-trained model in our cross-lingual hate speech detection experiment.
- We incorporate features from multilingual hate lexicon called HurtLex and measure their impact on the model performance.

ov

5. We explore aforementioned contributions into a specific kind of abusive language namely *misogyny*. We show that the cross-domain and cross-lingual classification setting could benefit the detection of abusive language on a smaller scale. We propose to have a deeper investigation on automatic misogyny identification task. Specifically, the following steps are taken:

- We develop state-of-the-art model to detect misogyny in social media, which include features specific to the use of abusive words.
- We explore more deeply regarding the most predictive features for the classifier to detect misogyny content in social media, including to the use of abusive words.
- We explore the relation between misogyny and other abusive phenomena such as sexism and hate speech by running a cross-domain experiment.
- We also experiment with cross-lingual automatic misogyny identification, since the provided dataset featured with three different languages including English, Italian, and Spanish.

## 1.8 Structure of the Thesis

This document comprises a collection of our most relevant research articles submitted and published during my Ph.D period. Four international journals (plus one under revision), two main conference papers, workshop papers describing our participation to relevant shared tasks in evaluation campaigns, as well as not published contents constitute the

substantial material for this thesis. In some parts, we restructured more than one articles into a more complete chapter. In the following, a brief review of each chapter included in this document is provided.

## **Chapter 2 State of The Art**

This chapter contains a brief review of abusive and hate speech detection in social media, specifically focused on the model robustness in multidomain and multilingual environment. We present an extensive review study to obtain the current open problems in these tasks. Most parts of this chapter will be published in Personal and Ubiquitous Computing Journal [[Pamungkas et al., 2021b](#)]. In addition, we also provide related studies in the use of swear words in abusive language detection and automatic misogyny identification tasks. This chapter introduces some state-of-the-art approaches that have been proposed to deal with these aforementioned research topics.

## **Chapter 3 The Role of Swear Word Use in Abusive Language Detection**

This chapter presents the research work published in the LREC 2020 [[Pamungkas et al., 2020a](#)]. However, most contents of this chapter are still not yet published/will be published. In this part, we deeply investigate the role of swear word in abusive language detection task. We build a novel corpora where swear word is annotated as either abusive or not abusive. We also develop supervised models to automatically classify the swear word context within a tweet as abusive or not. Furthermore, we explore the impact of swear word context in abusive language detection task, by incorporating this knowledge into abusive language detection model.

## **Chapter 4 Abusive Language Detection in Multidomain Settings.**

This chapter comprises of two published research articles. One article published in Cognitive Computation Journal [[Chiril et al., 2021](#)] and another one published in Proceeding of ACL SRW 2019 [[Pamungkas and Patti, 2019](#)]. This chapter is mainly focused on the investigation of building robust model to detect abusive language across different domains. We try to characterise the available datasets into different abusive phenomena and topical focuses. We also presents an architecture which allows the supervised model to detect both the abusiveness and the target of abusive of a tweet. Finally, we also explore the use of external resource to transfer knowledge between different abusive language domains.

## **Chapter 5 Abusive Language Detection in Multilingual Settings.**

This chapter presents the research work published in the Proceeding of ACL Student Research Workshopo (SRW) 2019 [[Pamungkas and Patti, 2019](#)] and the Information Processing and Management Journal [[Pamungkas et al., 2021a](#)]. In this chapter we focus on investigating the objective of building a robust model to detect hate speech in multilingual settings. We propose a novel joint-learning architecture which allow the model to learn the task in source and target languages sequentially. We explore the use of several available multilingual language representation to transfer knowledge between

languages. Finally, we also provide a deeper analysis on the difficulties and remain challenges of this task, to provide a better insight for the future works.

## **Chapter 6 Automatic Misogyny Identification: Cross-Domain and Cross-Lingual Study**

This chapter consist of an published articles in Information Processing and Management Journal [[Pamungkas et al., 2020b](#)]. In this chapter we focus on a specific kind of abusive language namely Misogyny. We briefly describe the Automatic Misogyny Identification task, as well as the dataset which is available in three different languages, English, Italian, and Spanish. Then, we propose a state-of-the-art model to detect misogyny in social media, and provide an insightful analysis on the most predictive features to detect it. We also conduct a cross-domain experiment to investigate the relation between misogyny with other related phenomena such as sexism and hate speech. Finally, we also run a cross-lingual experiment to explore the robustness of our model to detect misogyny in three various languages.

**Chapter 7. Conclusion and Future Work** This chapter outlines the conclusions of this thesis. It also includes a list of publication derived from the thesis, and describes some possible directions for future work.



# Chapter 2

## State of the Art

In this chapter we discuss extensively current state of the art studies in abusive language detection task, specifically focusing on four different areas. In Section 2.1, we present a brief review of studies related to the use of swear words in online communication, with a special attention on abusive language detection. Section 2.2 provides an extensive review of abusive language studies in multidomain setting, consisting of review on the available datasets and proposed approaches in this research area. Meanwhile, Section 2.3 also provides an extensive review of available datasets and proposed approaches to deal with abusive language detection in multilingual settings. Finally, in Section 2.4 we discuss about automatic misogyny identification studies, a specific phenomenon of abusive language which will be tackled in this thesis.

### 2.1 Swear Words in Abusive Language Detection Studies

Swearing plays an ubiquitous role in everyday conversations among humans, both in oral and textual communication, and occurs frequently in social media texts, typically featured by informal language and spontaneous writing. Such occurrences can be linked to an abusive context, when they contribute to the expression of hatred and to the abusive effect, causing harm and offense. However, swearing is multifaceted and is often used in casual contexts, also with positive social functions. We discuss this phenomenon in four parts, as following.

#### 2.1.1 Swearing in Online Content

The study of swear word use in online communication was started along with the growing interest of people to engage with social media. Wang et al. [2014b] examines the cursing activity on the social media platform Twitter<sup>1</sup>. They explore several research questions including the ubiquity, utility, and also contextual dependency of textual swearing in Twitter. On the same platform, Bak et al. [2012] found that swearing is used frequently between people who have a stronger social relationship, as a part of their study on

---

<sup>1</sup><https://www.twitter.com>

self-disclosure in Twitter conversation. Furthermore, Gauthier et al. [2015] provide an analysis of swearing on Twitter from several sociolinguistic aspects including age and gender. This study presents a deep exploration of the way British men and women use swear words. A gender- and age-based study of swearing was also conducted by Thelwall [2008], using the social network MySpace<sup>2</sup> to build their corpus. Recently, Cachola et al. [2018] studied vulgar words use in Twitter, by analyzing socio-cultural and pragmatic aspects of vulgarity based on users demographic data. Furthermore, they explored the impact of vulgar words use to the sentiment analysis task, which found that explicitly modeling vulgar words can boost sentiment analysis performance.

Besides social media, the study of swearing is also carried out on online communities. The study by Sood et al. [2012] focused on the use of profanity in an online community called Yahoo! Buzz<sup>3</sup>. They explored several research questions including what are the pitfalls of current profanity detection systems, how profanity differs between different communities, and how different communities receive the swearing in various contexts. Recently, Rojas-Galeano [2017] aimed at tackling the difficulties in detecting obfuscated obscenities on Spanish and Portuguese online news sites. Kwon and Gruzd [2017] studied the contagious diffusion of offensive comments in the Donald Trump's campaign video on YouTube<sup>4</sup>. They examined two kinds of swearing including: public swearing (when swearing has no specific target) and interpersonal swearing (the use of taboo words with a specific target).

### 2.1.2 Contextual Swearing

Swearing is not always abusive — its abusiveness is context-dependant. Swearing context is explored by several prior studies. Fägersten [2012] classifies swearing context into two types, following the dichotomy introduced by Ross [1969]: *annoyance* swearing, “occurring in situations of increased stress”, where the use of swear words appears to be “a manifestation of a release of tension”, and *social* swearing, “occurring in situations of low stress and intended as a solidarity builder”, which is related to a use of swear words in settings that are socially relaxed. Likewise, Allan and Burridge [2006] distinguishes the swearing contexts into *casual* context (when swear words do not cause insult, but are rather cathartic and humorous) and *abusive* context (when swear words are used with an intention to attack or insult).

The work by Jay [2009b] found that the offensiveness of taboo words is very dependant on their context, and postulates that the use of taboo words in conversational context (less offensive) and hostile context (very offensive). These findings support prior work by Rieber et al. [1979] who found that obscenities and swear words used in a *denotative* way are far more offensive than those used in a *connotative* way. Furthermore, Pinker [2007] classified the use of swear words into five categories based on why people swear: *dysphemistic*, exact opposite of euphemistic; *abusive*, using taboo words to abuse or insult someone; *idiomatic*, using taboo words to arouse the interest of listeners without really

---

<sup>2</sup><https://www.myspace.com>

<sup>3</sup>A social news commenting site that is no longer active.

<sup>4</sup><https://www.youtube.com>

referring to the matter; *emphatic*, to emphasize another word; *cathartic*, the use of swear words as a response to stress or pain.

### 2.1.3 Swear Words Corpora

The development of the swear word usage corpus was started by Holgate et al. [2018]. They proposed a novel corpus, consist of tweets containing swear words, where every swear word is annotated by six different labels based on its function. These vulgar function are including “express aggression”, “express emotion”, “emphasize”, “auxiliary”, “signal group identity”, and “non-vulgar”. The annotation process was done by using the crowd-sourced scenario. Furthermore, they build a model based on logistic regression coupled with several handcrafted features to classify the vulgar words function automatically. Pamungkas et al. [2020a] also introduced SWAD (Swear Words Abusiveness Dataset) corpus by filtering tweets from the OLID dataset [Zampieri et al., 2019a] based on swear word presence and annotating them with a binary label including “abusive” and “not-abusive”. They conducted the intrinsic evaluation of SWAD by predicting swear words’ abusiveness within a tweet as a context in two different models of prediction task, including sequence labeling task and text classification. Recently, Kurrek et al. [2020] also proposed a novel corpus that captures the online slur usage. The corpus consists of 39.8k human-annotated comments gathered from Reddit <sup>5</sup>. The annotation guideline outlines four main categories of online slur usage, divided into 12 sub-categories

### 2.1.4 Swearing and Abusive Content

In recent years, abusive language detection is gaining interest from the research community. Swear words play a key role in this task, according to several works in the literature. Razavi et al. [2010] developed an automatic system for discriminating between *regular texts* and *flames*. They built a dictionary for this specific purpose called *Insulting and Abusing Language Dictionary* (IALD), which contains words, phrases, and expressions with several degrees of abuse and insult. Several swear words can be found among IALD entries, which are used as features in the automatic classification. Similarly, Chen et al. [2012] built a dictionary containing pejoratives, obscenities and profanities extracted from Urban Dictionary <sup>6</sup>. By combining both lexical features from their dictionary and syntactic features from dependency relations, their models were able to achieve high precision and recall in detecting both offensive content and offensive users. Mubarak et al. [2017] built a list of Arabic obscene words and hashtags by extracting patterns that are frequently used in offensive Twitter posts. This wordlist is used to classify a tweet into three classes: *obscene*, *offensive*, and *clean*. Wiegand et al. [2018a] also built a manually labeled base abusive lexicon, which then expanded into a large lexicon by classifying the unlabeled negative polar expressions from Wiktionary. The more recent work by Wiegand and Ruppenhofer [2021] also produced an abusive lexicon by using some negative/abusive emojis including *middle finger*, *face vomiting*, *pile of poo*, and etc.

---

<sup>5</sup><https://www.reddit.com/>

<sup>6</sup><https://www.urbandictionary.com/>

as the proxy for obtaining the abusive words. Both lexicons were proven to be effective to provide domain-independent features in cross-domain classification of English microposts.

Recent studies also found that swear words are relevant to several related tasks including abusive language detection [Nobata et al., 2016], cyberbullying detection [Van Hee et al., 2018, Michal et al., 2010], and hate speech detection [Malmasi and Zampieri, 2018]. The study by Holgate et al. [2018] introduced six vulgar word use functions, and built a novel dataset based on them. They filtered their dataset based on presence of swear words from a list taken from the *noswearing* website<sup>7</sup>. Their results show that classifying vulgar word use by its function improves the system performance in detecting hate speech content.

This section presents several background studies related to swear word use, focusing on several different aspects. We notice that swear word context is an important facet for developing a robust model to detect abusive language. Since the use of the swear word is in the grey area between regular speech and abusive language. In the next section, we discuss previous studies that focus on detecting abusive language in multiple domains.

## 2.2 Abusive Language Detection in Multidomain Settings

Abusive language behaviour is multifaceted and available datasets are characterized by different topical focuses. *Abusive language* is generally used as an umbrella term [Waseem et al., 2017], covering several sub-categories, such as cyberbullying [Hee et al., 2015, Sprugnoli et al., 2018], hate speech [Waseem and Hovy, 2016, Davidson et al., 2017], toxic comments [Wulczyn et al., 2017], offensive language [Zampieri et al., 2019a] and online aggression [Kumar et al., 2018]. Several datasets have been proposed having different topical focuses, e.g, misogyny, racism, sexism, and so on, and sourced from different platforms, e.g. Facebook and Twitter. Most studies in this area also tend to focus on one topical focus, which makes difficult to quantify whether a model or feature set which perform well in one dataset is transferrable to other datasets [Schmidt and Wiegand, 2017, Waseem et al., 2018].

### 2.2.1 Abusive Language Domain

Abusive language phenomena are not constrained to one particular topical focus and platform. Therefore, having a robust model to detect abusive language across different topical focuses and platforms is important. Some existing studies proposed cross-domain abusive language detection. A model is trained on one specific dataset with a specific domain and tested in another dataset with a different domain. In this study, the domain term is used to describe both topical focuses and platforms. It has been stated that ensuring that a model can detect abusive language across different domains is one of the main challenges and an important frontier [Vidgen et al., 2019]. The cross-domain setting was also explored by Wiegand et al. [2019] to prevent bias contained in the training data, as they experimentally found several biases in currently popular abusive language

---

<sup>7</sup><http://www.noswearing.com>

datasets, including topic bias and author bias. We divide this section into two main parts. First, we review available datasets that could be exploited for this task, focusing on English. Furthermore, we also describe several approaches that have been proposed in this research direction.

### 2.2.2 Available Datasets for Multidomain Abusive Language Detection

Topical Focus	Sources	Entries	Available	Ref
Hate Speech	Twitter	24,802	Yes	Davidson et al. [2017]
	Twitter	27,330	Yes	ElSherief et al. [2018b]
	Twitter	62 millions	No	Gao et al. [2017]
	Stormfront	10,568	Yes	de Gibert et al. [2018]
	Youtube	24,840	No	Hammer [2016]
	Twitter & Reddit	150 millions	No	Olteanu et al. [2018]
	Gab & Reddit	56,100	Yes	Qian et al. [2019a]
	Twitter	3.5 millions	No	Qian et al. [2018b]
	Twitter	18,667	No	Qian et al. [2019b]
	Twitter	4,000	No	Vidgen and Yasseri [2020]
	Twitter	16,907	Yes	Waseem and Hovy [2016]
	Twitter	13,000	Yes	Basile et al. [2019]
	Facebook	1,288	Yes	Chung et al. [2019]
	Twitter	4,972	Yes	Ribeiro et al. [2018]
	Twitter	5,647	Yes	Ousidhoum et al. [2019]
	Twitter	149,823	Yes	Gomez et al. [2020]
Toxicity	Twitter & Facebook	7,005	Yes	Mandl et al. [2019]
	News Site	1,043	No	Kolhatkar et al. [2019]
	Wikipedia	115,737	Yes	Wulczyn et al. [2017]
	Twitter	6,774	Yes	Radfar et al. [2020]
Cyberbullying	Youtube	2,235	No	Sharma et al. [2018]
	Gaming Platforms	34,329	No	Bretschneider and Peters [2016]
	Formspring <sup>4</sup>	13,160	Yes	Rosa et al. [2018]
	Twitter & Formspring <sup>3</sup>	13,000	No	Zhang et al. [2016]
Offensiveness	Reddit	168 millions	No	Nithyanand et al. [2017]
	Reddit	11 millions	No	Schäfer and Burtenshaw [2019]
Abusiveness	Twitter	14,100	Yes	Zampieri et al. [2019b]
	Twitter	9 millions	Yes	Zampieri et al. [2020]
	News Site	3.1 millions	No	Nobata et al. [2016]
Flames	Twitter	80,000	Yes	Founta et al. [2018]
	News Site	5,077	Yes	Steinberger et al. [2017]
Harassment	Twitter	25,000	Yes	Rezvan et al. [2018]

Misogyny	Twitter	35,000	Yes	Golbeck et al. [2017]
	Twitter	3,977	Yes	Fersini et al. [2018c]
	Twitter	5,000	Yes	Fersini et al. [2018a]
	Twitter	6,000	Yes	Fersini et al. [2020]
Sexism	Twitter	712	Yes	Jha and Mamidi [2017]
Aggressiveness	Facebook	15,000	Yes	Kumar et al. [2018]
	Twitter & Facebook	5,000	Yes	Kumar et al. [2020]

Table 2.1: Summarization of Available Abusive Language Dataset Across Different Topical Focuses and Sources (English only).

We collect information about the available datasets from existing studies on abusive language detection across different domains. Several previous works in abusive language detection defined a domain as a topical focus [Swamy et al., 2019, Pamungkas and Patti, 2019], such as hate speech, cyberbullying, offensiveness, etc. In contrast, some others described it as platforms [Karan and Snajder, 2018, Glavaš et al., 2020] such as Twitter, Facebook, Youtube, etc. We select English datasets by focusing on topical focus and platform variety. We mainly extract this information from the two most recent survey studies on abusive language resources. First, Vidgen and Derczynski [2020] provided the analysis of available training data for abusive language detection tasks and proposes best practices in creating training data of abusive language based on existing studies. Meanwhile, Poletto et al. [2020] presented a more comprehensive study on resources and benchmarks available for hate speech detection tasks based on several aspects. We also add datasets from several shared tasks that are not covered by these works and a few datasets from very recent studies that are not available yet when these articles were published. Table 2.1 summarizes our findings on the available datasets for this research purpose. We discuss a more in-depth comparison between datasets and other aspects we need to consider when using these datasets for multidomain abusive language study based on existing works in the following.

**Topical focus.** The motivation for several multidomain abusive language detection studies is to have a robust model that generalizes the problem across different topical focuses. Topical focus usually includes the addressed abusive phenomena, as well as the specific targets of the abusive behavior. However, some topics overlap with each other, i.e., misogyny and sexism or xenophobia and racism, due to a certain degree of subjectivity in defining these phenomena. The topical focus information presented in Table 2.1 is based on the information provided in the publications which accompany the proposed resources. However, some of these papers did not include a clear definition of the addressed phenomena. We observe that hate speech is the most covered topic by previous studies. However, on some hate speech datasets, we also discover other abusive phenomena such as offensiveness [Davidson et al., 2017], racism [Waseem and Hovy, 2016], and sexism [Waseem and Hovy, 2016]. In this manner, a cross-domain abusive language detection experimental setting means training a model on one or more topical focuses and testing it on different topical focuses unseen in the training data.

**Sources.** Another objective of abusive language detection in the multidomain setting is to have a robust model to detect abusive content across different platforms. This task is also challenging because the available datasets were retrieved from various platforms, and every platform has different characteristics and uniqueness. Based on the information presented in Table 2.1, Twitter is the most studied platform for capturing the abusive phenomena. This is possibly due to the convenience of scraping tweet samples using the available Twitter API and the less strict policy on making the data publicly available. Facebook is another popular social media besides Twitter as a data source in several studies. Other studies exploited news sites, online forums, and Youtube comments for gathering their data. Most studies used several defined keywords to query the data from the platforms mentioned above. Some of them used offensive words [Fersini et al., 2018c,a, Ousidhoum et al., 2019, Davidson et al., 2019], which are usually a strong signal of abusive content, while other studies decided to use more neutral keywords to maintain a real-world approach to the problem [Basile et al., 2019], or even both offensive and neutral keywords [Zampieri et al., 2019a]. Some other works also exploited specific keywords related to some events that trigger abusive phenomena [Waseem and Hovy, 2016].

**Availability.** In Table 2.1, we provide information about the availability of the datasets. We manually check the published papers and mark a dataset as available when the authors explicitly mention the link to the dataset repository or state that the dataset is available for research purposes upon request. We can see that 26 out of 39 datasets were made available by their authors. Most available datasets were obtained from Twitter, likely due to their policy or other regulation restricting data sharing from other sources such as Reddit, Youtube, and news sites. However, we also notice that some Twitter datasets are shared by only providing the tweet identifier [Waseem and Hovy, 2016, Founta et al., 2018] and allow users to download them by using the publicly available Twitter API. In this case, the number of entries could decrease due to the data decay (tweets were already deleted or are simply not available anymore).

**Annotation Scheme.** This information is not provided in Table 2.1, but we perform a manual inspection regarding the annotation scheme of every dataset. Most datasets have binary labels, including abusive and not abusive class. Some other datasets have a multiclass annotation, capturing different abusive phenomena. For example, Davidson et al. [2017] labeled not only the hateful tweets but also their offensiveness. Similarly, Waseem and Hovy [2016] proposed to label racism and sexism separately. Some studies also proposed a finer-grained annotation scheme to capture more in-depth abusive phenomena. For example, Fersini et al. [2018c,a] provided three layers of annotation to capture the misogyny phenomenon (misogyny or not), misogyny category and behavior (stereotype, dominance, derailing, sexual\\_harassment, and discredit), and the target of misogyny (active or passive). In a multidomain or cross-domain classification task, one of the most important steps is to unify the label annotation of every dataset. Most existing works model this task as a binary classification task [Karan and Šnajder, 2018, Pamungkas and Patti, 2019]. Therefore, they casted the multiclass annotation to binary annotation by combining different abusive phenomena into one class. In the case of finer-grained annotation, they only took the first layer of annotation, where the data is mainly annotated

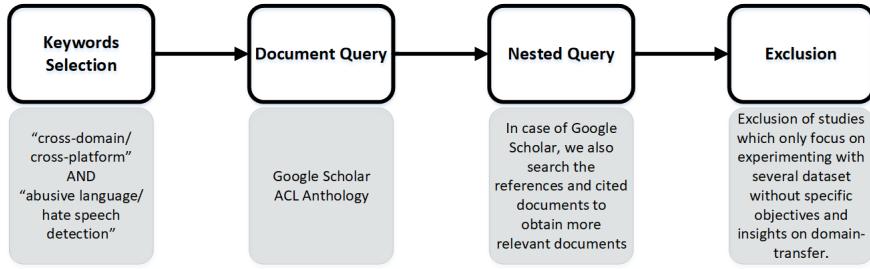


Figure 2.1: Document Collection Methodology

as either abusive or not abusive.

**Data distribution.** Data distribution also needs to be considered in the multidomain and cross-domain classification task, especially the percentage of abusive samples in the dataset. The different label distribution between training and testing sets would make the performance evaluation and comparison between systems is difficult [Pamungkas and Patti, 2019, Pamungkas et al., 2020b]. Specifically, when systems are trained on skewed distributions of labels, with few examples in the abusive class, they will struggle to detect the abusive class on the test set, resulting in a higher rate of false negatives. Pamungkas et al. [2021a] observed that balancing the distribution in the training set improves the f-score of the positive class significantly. Based on our investigation, the class distribution of abusive language datasets varies considerably, mostly depending on how the data is sampled and on the source of the data. However, we found that most abusive language datasets have a lower percentage of abusive content than neutral content, with some datasets only containing less than 20% of abusive instances [Founta et al., 2018, Wulczyn et al., 2017, de Gibert et al., 2018]. Some studies experimentally found that systems often struggle to detect the under-represented class, resulting in low f1-scores on the positive class (abusive label), which is an issue for real-world abusive language detection systems [Pamungkas et al., 2021a, Ibrohim and Budi, 2019a]. Maintaining a uniform label distribution between training and test set is an approach often followed to provide a comparable evaluation in cross-domain classification [Swamy et al., 2019, Ibrohim and Budi, 2019a]. This approach, however, does not necessarily provide an accurate estimate of the robustness of the model in a realistic scenario, where the amount of abusive language could drastically change.

### 2.2.3 Proposed Approaches in Multidomain Abusive Language Studies

This subsection presents studies that have been done in abusive language detection, which focus on building robust models across different domains. We collect any publication found on Google Scholar by using four main keywords, namely "cross-domain abusive language detection", "cross-domain hate speech detection", "cross-platform abusive speech detection", and "cross-platform hate speech detection". These keywords are chosen after several observations using different keyword combinations. We limit our query to the first five pages for each keyword and sort results based on relevance, without a time

filter. Furthermore, we also check each document’s cited documents and references on the first five pages to get more relevant publications. To avoid missing on the very recent works, we also exploit the same keywords on the proceeding of the last three years’ main NLP conferences on ACL Anthology platforms<sup>8</sup>. Finally, we exclude some works which only experiment with different datasets, without any objectives and insights about domain-agnostic models. Figure 2.1 summarizes the methodology for the document collection in this survey study.

Models	Approach	Year	Ref
Traditional Model	Proposed to employ a SVM model with a novel abusive lexicon and exploited it in the cross-domain abusive language detection task, providing domain-independent knowledge.	2018	<a href="#">Wiegand et al. [2018a]</a>
Traditional Model	Employs a linear SVM coupled with a domain adaptation approach called FEDA, which works by duplicating features several times across domains to allow the model to learn domain-dependent weights for each feature.	2018	<a href="#">Karan and Šnajder [2018]</a>
Neural-Based	Experimented with a multitask learning approach, which allows the model to learn the task from two or more tasks sequentially by sharing the learning parameters and combining the loss functions of the respective tasks.	2018	<a href="#">Waseem et al. [2018]</a>
Neural-Based	This work experimented by combining datasets from different platforms to train the GRU-based model and exploit different sets of features.	2019	<a href="#">Corazza et al. [2019]</a>
Neural-Based	Exploited a specific hateful lexicon called HurtLex to provide domain-independent features for two supervised models including a linear SVM and a LSTM in a cross-domain abusive language task.	2019	<a href="#">Pamungkas and Patti [2019]</a>
Neural-Based	Proposed a joint-learning architecture based on ELMo Embeddings, which allows the model to learn the task from two datasets sequentially, obtaining more robust performance.	2019	<a href="#">Rizou et al. [2019]</a>
Transformer-Based	This work aims to study the transferability of the current state of the art BERT model, so no specific approach is proposed to tackle domain transfer.	2019	<a href="#">Swamy et al. [2019]</a>

---

<sup>8</sup><https://www.aclweb.org/anthology/>

Neural Model	This study proposed several LSTM-based models that only focuses on using text information (char n-grams and word embedding) representation for building platform-agnostic hate speech detector, but they did not conduct any cross-domain or multidomain experiment to evaluate their model.	2019	<a href="#">Meyer and Gambäck [2019]</a>
Transformer-Based	Experimented with a BERT-based classifier and topic modeling approach, which show that removing domain-specific instances improve the model's out-domain performance	2020	<a href="#">Nejadgholi and Kiritchenko [2020]</a>
Neural-Based	Proposed several representations including target, content, and linguistic behavior and used cross attention gate flow to refine these representations, providing better domain-transfer knowledge.	2020	<a href="#">Wang et al. [2020]</a>
Transformer-Based	Infused specific hateful lexicon called HurtLex into BERT model to transfer knowledge across domains.	2020	<a href="#">Koufakou et al. [2020]</a>
Multiple Models	Besides experimented with a wide coverage of models including traditional (linear SVM), (LSTM), and (BERT), they also exploited HurtLex as domain-independent features for knowledge transfer between domains.	2020	<a href="#">Pamungkas et al. [2020b]</a>
Neural-Based	Experimented with augmenting all training data from different domains, resulting in the performance improvement of the models based on BERT and RoBERTa representation.	2020	<a href="#">Glavaš et al. [2020]</a>
Transformer-Based	It is proposed to retrain BERT with a big abusive language corpus obtained from Reddit called HateBERT, which shows a promising result in the cross-dataset experiment.	2020	<a href="#">Caselli et al. [2020a]</a>
Traditional Models	They tested the generalisability of wide-coverage traditional models logistic regression, Naïve Bayes, support vector machine, XGBoost, feed-forward neural network) coupled with also a wide range of feature representation in detecting hate speech across different platforms.	2020	<a href="#">Salminen et al. [2020]</a>

Transformer-Based	Experimented by combining several datasets to train the model based on BERT and proven to be effective in detecting uncivil language across multiple domains, it outperformed several fine-tuning strategies.	2020	<a href="#">Ozler et al. [2020]</a>
Transformer-Based	Proposed to use existing regularization method to re-weight input samples which succeeded to decrease the racial bias of the dataset, resulting in the improvement of the BERT-based models' performance in cross-domain classification settings	2020	<a href="#">Mozafari et al. [2020]</a>
Neural-Based	This study reproduced the state of the art models to evaluate the dataset bias issue in abusive language task based on the cross-dataset classification study.	2020	<a href="#">Arango et al. [2020]</a>
Traditional Model	This study proposed a novel multiplatform abusive language dataset. The proposed model for the experiment is the standard SVM without a specific approach to deal with domain-shift issue.	2020	<a href="#">Chowdhury et al. [2020]</a>

Table 2.2: Summary of approaches adopted by existing studies for cross-domain abusive language detection tasks.

We carefully read each work to obtain several key pieces of information to be discussed in this study. Table 2.2 summarizes the full list of works in this direction. Most studies only focused on English, and we only found two studies that work on Italian [Corazza et al., 2019] and Arabic [Chowdhury et al., 2020]. Most of the chosen studies conducted a cross-domain experiment, where the domain can be either abusive phenomena or platforms. We also notice that this research focus is still relatively new, with the earliest works were initiated in 2018 [Karan and Šnajder, 2018, Waseem et al., 2018, Wiegand et al., 2018a]. All studies adopted a supervised approach by training a model on a training set and predicting instances on the test set. Following, we provide a deeper discussion to compare each work based on the models (traditional machine learning-based, neural-based, or transformer-based), features (a very wide variant of features), and approaches adopted to deal with domain-shift specifically.

**Models** A wide variety of models was adopted to deal with this task. Some studies exploited traditional machine learning approaches such as linear support vector machine classifiers (LSVC) [Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Karan and Šnajder, 2018], logistic regression (LR) [Salminen et al., 2020], and support vector machine (SVM) [Wiegand et al., 2018a, Chowdhury et al., 2020]. Their argument for adopting the traditional approach is to provide better explainability of the knowledge transfer between domains. Some other studies adopted several neural-based models, including convolutional

neural networks (CNN) [Wang et al., 2020, Meyer and Gambäck, 2019], long short-term memory (LSTM) [Waseem et al., 2018, Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Meyer and Gambäck, 2019, Arango et al., 2020], bidirectional LSTM (Bi-LSTM) [Rizou et al., 2019], and gated recurrent unit (GRU) [Corazza et al., 2019]. The most recent works more focus on investigating transferability or generalisability of state of the art transformer-based model such as Bidirectional Encoder Representations from Transformers (BERT) [Nejadgholi and Kiritchenko, 2020, Koufakou et al., 2020, Swamy et al., 2019, Glavaš et al., 2020, Pamungkas et al., 2020b, Caselli et al., 2020a, Ozler et al., 2020, Mozafari et al., 2020] and its variant like RoBERTa [Glavaš et al., 2020] in the cross-domain abusive language detection task.

In the early phases of cross-domain abusive language detection, specific models which adopt joint-learning [Rizou et al., 2019] and multitask [Waseem et al., 2018] architectures achieved the best performance. These architectures were proven to be effective for transferring knowledge between domains. However, in the latest studies, transformer-based models succeed in achieving state-of-the-art results. The most recent study by Glavaš et al. [2020] shows that RoBERTa outperformed other models such as BERT in the cross-domain setting of the hate speech detection task. This result confirms a recent finding on other natural language processing tasks [Brown et al., 2020], i.e., that a pre-training language model trained on huge corpora provides a more general representation for knowledge transfer.

**Feature Representation** A wide range of features was also exploited in this particular task, ranging from straightforward n-gram representations to the most recent contextual language representations. Several text representation were used for the traditional machine learning model, including n-grams [Karan and Šnajder, 2018, Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Meyer and Gambäck, 2019, Chowdhury et al., 2020], TF-IDF [Salminen et al., 2020], and word2vec[Salminen et al., 2020]. Some studies also proposed to use linguistic features such as emoji information [Corazza et al., 2019] and lexical [Wiegand et al., 2018a, Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Corazza et al., 2019] features by using a specific lexicon. Most of the neural models in this task used word embedding to represent the text. Several pre-trained models were exploited, such as FastText [Corazza et al., 2019, Pamungkas and Patti, 2019, Pamungkas et al., 2020b] and ELMo [Rizou et al., 2019]. Finally, the transformer-based models use pre-trained models based on a very big corpus such as BERT [Nejadgholi and Kiritchenko, 2020, Koufakou et al., 2020, Swamy et al., 2019, Glavaš et al., 2020, Pamungkas et al., 2020b, Caselli et al., 2020a, Ozler et al., 2020, Mozafari et al., 2020] and RoBERTa [Glavaš et al., 2020]. However, we also found a study that proposes to re-train the BERT representation on a specific corpus related to abusive language [Caselli et al., 2020a]. Finally, the work by Nejadgholi and Kiritchenko Nejadgholi and Kiritchenko [2020] proposed to use unsupervised topic modelling approach to generate the features for obtaining better topic generalization on cross-dataset abusive language detection experiment.

Our study observe that several state-of-the-art pre-trained models provide the best feature representation and better generalization to deal with domain-shift in the cross-domain abusive language detection task. Interestingly, some studies proposed using

external resources to facilitate the knowledge transfer between domains by delivering domain-independent features. These additional features were infused into either traditional models [Wiegand et al., 2018a] or neural-based models [Pamungkas and Patti, 2019] and succeeded in improving the prediction performance. Wiegand et al. [2018a] showed the effectiveness of additional features from their novel abusive words lexicon in a cross-domain abusive language detection setting. The additional features were represented as a score based on the confidence learned by an SVM classifier. Similarly, Pamungkas and Patti [2019], Pamungkas et al. [2020b] exploited the HurtLex lexicon, which contains a list of abusive words in 17 categories. The features were represented as a 17-column binary vector, to indicate the presence of each word category in the document. The vector is then concatenated to the representation of the message computed by LSTM network.

**Domain Transfer** The main challenge of cross-domain classification is the domain shift between training and testing data. Several methods have been proposed by studies in more mature areas, such as sentiment analysis [Pan et al., 2010, Du et al., 2020, Yuan et al., 2018]. These techniques are usually called domain-adaptation or domain-transfer, a specific approach to allow the model to learn domain-independent features, intersecting between two or more different domains. In the abusive language detection task, several features could become an important signal for knowledge transfer between domains, such as the use of swear words, emotional information, and some other linguistic features. In the abusive language detection task, several features could represent an important signal for knowledge transfer between domains, such as the use of abusive words [Wiegand et al., 2018a], emotional information [Rajamanickam et al., 2020, Safi Samghabadi et al., 2020], and some other linguistic features. [Koufakou et al., 2020, Pamungkas et al., 2020b, Pamungkas and Patti, 2019, Corazza et al., 2019]

Table 2.2 shows that studies have different approaches to cope with the domain-shift problem. Some works proposed to **combine the training sets from several different domains dataset** [Corazza et al., 2019, Pamungkas et al., 2020b, Glavaš et al., 2020, Ozler et al., 2020]. This straightforward approach allows the trained model to obtain wider domain coverage for detecting abusive language. Most aforementioned studies found that this simple approach was proven to be effective in this task. However, there is still a possibility that the trained model would struggle when applied to data from the totally unseen domain. Several other studies experimented with the use of lexicon as a domain-independent feature to bridge the domain-transfer. Wiegand et al. [2018a] used their novel lexicon automatically induced from HateBase, a platform that provides several keywords related to hate speech. Meanwhile, Pamungkas et al. [2020b], Pamungkas and Patti [2019], Corazza et al. [2019] exploited HurtLex, a manually built lexicon by De Mauro [2016], which contains offensive words structured in 17 different categories. Additional features from these lexica were also proven helpful to facilitate the transfer of knowledge between domains.

We also notice some works that tried to **modify the input sample for training the model** in order to minimize the domain-shift issue between source and target domains. For example, Nejadgholi and Kiritchenko [2020] used the topic modeling approach and proposed to remove the domain-specific instances from the training set, resulting in

the improvement of the model’s performance. Another effort by Karan and Šnajder [2018] adopted a domain adaptation approach called FEDA, which works by duplicating some key features several times across domains to allow the model to learn domain-dependent weights for each feature. Finally, Mozafari et al. [2020] proposed to deal with the racial bias on the abusive language dataset by re-weighting the input samples using the existing regularization approach. Their approach was shown to be effective in decreasing the dataset bias issue, which was found as one of the main problems in cross-domain classification.

We also notice that some studies focused more on **providing better representation to improve the model’s domain generalization**. Wang et al. [2020] proposed a multiaspect embedding, which combines several representations, including target, content, and linguistic behavior, to provide domain-transfer knowledge. Then, Caselli et al. [2020a] proposed to retrain state-of-the-art BERT with a huge abusive language corpus to obtain a more specific representation for abusive language detection tasks.

Furthermore, we discovered two studies **proposed new architectures to tackle cross-domain abusive language detection task specifically**. Rizoiu et al. [2019] proposed a joint-learning model based on Bi-LSTM, which allows the model to learn from two datasets sequentially, obtaining better generalization. In addition, Waseem et al. [2018] proposed a multitask learning architecture based on LSTM to learn the problem from two or more tasks sequentially, providing a medium for knowledge transfer between domains. The rest of the works more focused on investigating the transferability of some models, including BERT in the cross-domain abusive language detection [Swamy et al., 2019, Salminen et al., 2020]. They found that using BERT [Swamy et al., 2019] only without a specific approach for bridging domain-shift already achieves a solid result.

This section provides an in-depth discussion related to the study of abusive language detection tasks in cross-domain settings. As we mentioned that abusive language phenomena in the online environment are multifaceted and also happened on multiple platforms. Therefore, this research direction is also important to investigate the possibility of developing a robust model in abusive language detection tasks. In the next section, we present related studies in multilingual abusive language detection, another important aspect to have a robust model in this task.

## 2.3 Abusive Language Detection in Multilingual Settings

Another prominent challenge in abusive language detection is the multilinguality issue. Even if in the last years abusive language datasets were developed for other languages, including Italian [Bosco et al., 2018, Fersini et al., 2018c], Spanish [Fersini et al., 2018c], and German [Wiegand et al., 2018b], English remains by far the most represented language. Recently, deep learning approaches have been applied, achieving state-of-the-art results for some languages [Mozafari et al., 2019, Badjatiya et al., 2017]. However, most of the proposed models were tested in monolingual settings, mostly in English. Since the most popular social media such as Twitter and Facebook are highly multilingual, fostering their users to interact in their primary language, there is a considerable urgency to

develop a robust approach for abusive language detection in a multilingual environment, also for guaranteeing a better compliance to governments demands for counteracting the phenomenon — see, e.g., the recently issued EU commission *Code of Conduct on countering illegal hate speech online* [EU Commission, 2016].

### 2.3.1 Abusive Language Detection and Cross-lingual Settings

Similarly to other natural language processing tasks [Joshi et al., 2020], detecting abusive language in less-resourced languages is a prominent and timely challenge. For example, the escalation of hate speech against Muslims in Rohingya Myanmar was also affected by the failure to stop spreading hate comments on Facebook due to the difficulty of processing Burmese text automatically<sup>9</sup>. The current availability of datasets in many languages [Poletto et al., 2020], makes the time ripe for addressing the multilingual challenge. Cross-lingual transfer learning is the common approach to transfer knowledge from one language (usually with more available resources) to another language (usually with less resources) [Lin et al., 2019, Schuster et al., 2019]. In this approach, models are trained and optimized on a dataset from one language (called *source* language), and then tested on another language (called *target* language). Zero-shot learning is an extreme case of transfer learning, where a model trained on one language or one domain is employed to predict samples from a totally unseen language or domain [Goodfellow et al., 2016]. The less extreme form of transfer learning is few-shot learning, where a percentage of samples from unseen data (*target* language) is added to the training set, allowing the model to learn a better generalization between two languages or domains [Schuster et al., 2019]. Similar to the previous section, this section also consists of two main parts. First, we review the available abusive language datasets in languages other than English, which can be exploited for this research purpose. Second, we discuss the previous works on the abusive language detection task in multilingual settings.

### 2.3.2 Available Datasets for Multilingual Abusive Language Detection

Lang.	Topical Focus	Sources	Entries	Available	Ref
AM	Hate Speech	Facebook	4,882	No	Mossie and Wang [2018]
AR	Hate Speech	Twitter	6,000	Yes	Albadi et al. [2018]
	Hate Speech	Multiple Sources	6,039	Yes	Haddad et al. [2019]
	Offensiveness	Twitter	1,100	Yes	Mubarak et al. [2017]
	Offensiveness	Youtube	15,050	Yes	Alakrot et al. [2018]
	Hate Speech	Twitter	5,846	Yes	Mulki et al. [2019]
	Hate Speech	Twitter	3,353	Yes	Ousidhoum et al. [2019]
BN	Offensiveness	Twitter	10,000	Yes	Zampieri et al. [2020]
	Hate Speech	Facebook	5,126	Yes	Ishmam and Sharmin [2019]

<sup>9</sup><https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>

	Aggresiveness	Youtube	5,000	Yes	Kumar et al. [2020]
	Misogyny	Youtube	5,000	Yes	Kumar et al. [2020]
CS	Flamming	News Sites	5,077	Yes	Steinberger et al. [2017]
DA	Offensiveness	Twitter, Facebook, Reddit	3,600	Yes	Sigurbergsson and Derczynski [2020]
DE	Hate Speech	Twitter	541	Yes	Ross et al. [2017]
	Flamming	News Sites	5,077	Yes	Steinberger et al. [2017]
	Hate Speech	Twitter	4,669	Yes	Mandl et al. [2019]
	Offensiveness	Facebook	5,836	Yes	Bretschneider and Peters [2017]
	Offensiveness	Twitter	8,541	Yes	Wiegand et al. [2018b]
EL	Abusiveness	News Sites	1.5 millions	Yes	Pavlopoulos et al. [2017]
	Offensiveness	Twitter	10,287	Yes	Pitenis et al. [2020]
ES	Misogyny	Twitter	4,138	Yes	Fersini et al. [2018c]
	Hate Speech	Twitter	6,600	Yes	Basile et al. [2019]
	Aggresiveness	Twitter	11,000	Yes	Álvarez-Carmona et al. [2018]
	Hate Speech	Twitter	6,000	Yes	Pereira-Kohatsu et al. [2019]
FR	Hate Speech	Other	15,024	Yes	Chung et al. [2019]
	Flamming	News Sites	5,077	Yes	Steinberger et al. [2017]
	Hate Speech	Twitter	4,014	Yes	Ousidhoum et al. [2019]
HI	Offensiveness	Twitter	3,679	No	Mathur et al. [2018]
	Aggresiveness	Facebook	15,000	Yes	Kumar et al. [2018]
	Aggresiveness	Youtube	5,000	Yes	Kumar et al. [2020]
	Misogyny	Youtube	5,000	Yes	Kumar et al. [2020]
HI-EN	Hate Speech	Twitter	4,575	Yes	Bohra et al. [2018]
	Hate Speech	Twitter	5,983	Yes	Mandl et al. [2019]
	Hate Speech	Facebook, Twitter	3,367	Yes	Rani et al. [2020]
HR	Abusiveness	News Site	17 millions	Yes	Ljubešić et al. [2018]
ID	Hate Speech	Twitter	1,100	Yes	Alfina et al. [2017]
	Abusiveness	Twitter	2,016	Yes	Ibrohim and Budi [2018]
	Hate Speech	Twitter	13,169	Yes	Ibrohim and Budi [2019a]
IT	Homophobic	Twitter	1,859	No	Akhtar et al. [2019]
	Hate Speech	Other	15,024	Yes	Chung et al. [2019]
	Hate Speech	Instagram	6,710	No	Corazza et al. [2019]
	Hate Speech	Facebook	6,502	No	Vigna et al. [2017]
	Hate Speech	Twitter	4,000	No	Poletto et al. [2019]

	Flamming	News Sites	5,077	Yes	Steinberger et al. [2017]
	Hate Speech	Twitter	6,009	Yes	Sanguinetti et al. [2018]
	Misogyny	Twitter	5,000	Yes	Fersini et al. [2018a]
	Misogyny	Twitter	6,000	Yes	Fersini et al. [2020]
	Hate Speech	Twitter, Facebook	4,000	Yes	Bosco et al. [2018]
	Hate Speech	Twitter, News Site	8,602	Yes	Sanguinetti et al. [2020]
	Cyberbullying	WhatsApp	14,600	Yes	Sprugnoli et al. [2018]
PL	Cyberbullying	Twitter	11,041	Yes	Ptaszynski et al. [2019]
PT	Offensiveness	Twitter	7,672	Yes	Nascimento et al. [2019]
	Offensiveness	News Site	1,250	Yes	de Pelle and Moreira [2017]
SL	Hate Speech	Twitter	3,059	Yes	Fortuna et al. [2019]
	Abusiveness	News Site	13,000	Yes	Fiser et al. [2017]
	Abusiveness	News Site	7.6 millions	Yes	Ljubešić et al. [2018]
SV	Hate Speech	Web Fora	3,056	No	Fernquist et al. [2019]
SW-EN	Hate Speech	Twitter	25,000	No	Ombui et al. [2019]
TR	Hate Speech	Twitter	36,232	Yes	Çöltekin [2020]
	Offensiveness	Twitter	35,000	Yes	Zampieri et al. [2020]
VI	Hate Speech	Facebook	25,431	Yes	Vu et al. [2020]

Table 2.3: Summary of available abusive language datasets across different languages.

In this section, we present information regarding the available datasets for abusive language detection tasks across different languages. Since we already presented the English datasets in the cross-domain part, in this section we only review the available datasets in languages other than English, which we will call *lower resourced* languages for the rest of this thesis. We obtained this information based on the two most recent reviews [Poletto et al., 2020, Vidgen and Derczynski, 2020] which focused on the available resources in abusive language tasks. In addition, we also add more uncovered resources from the most recent shared tasks in the abusive language field, such as Misogyny@EVALITA2020 [Fersini et al., 2020], HaSpeeDe@EVALITA2020 [Sanguinetti et al., 2020], and OffensEval@SemEval2020 [Zampieri et al., 2020]. We also search for the recently available resources from the last edition of Language Resources and Evaluation Conference (LREC) 2020<sup>10</sup> and Workshop on Online Abuse and Harms (WOAH) 2020<sup>11</sup>, where we collect information of some datasets that are still not covered in these surveys. Table 2.3 summarizes the information of these lower resourced languages datasets for abusive language

<sup>10</sup><https://lrec2020.lrec-conf.org/en/>

<sup>11</sup><https://www.workshoponlineabuse.com/>

detection task. We provide an in-depth discussion focusing on the comparison of these resources in the following.

**Language.** In Table 2.3, we use the ISO 639-1 language code to represent the language names. Based on Table 2.3, the abusive language datasets were already available in 18 different languages. Despite being not as many as in English, we notice that some languages have more resources than others, such as Arabic (AR), Hindi (HI), and Italian (IT). However, some other languages only have one resource available such as Czech (CS), Croatian (HR), Poland (PL), Swedish (SW), Turkish (TR), and Vietnamese (VI). The availability of these lower resourced datasets indicates that this research direction is still growing. However, we observe that these resources are more centered on Indo-European languages. We still could not find datasets in the Niger-Congo language family which are mostly used in some African regions. The datasets in Afro-Asiatic, Austronesian, and other language families are also far less than Indo-European languages. Moreover, we observe Hindi-English (HI-EN) code-mixed datasets, all focusing on detecting hate speech. The first dataset of hate speech in Hindi-English code-mixed was proposed by [Bohra et al. \[2018\]](#). [Mandl et al. \[2019\]](#) presented a new collection created for a shared task, Hate Speech and Offensive Content Identification (HASOC), at FIRE 2019. Recently, [Rani et al. \[2020\]](#) proposed the first Hindi-English hate speech dataset containing tweets written in both Roman and the native Devanagari script. Additionally, a Swahili-English code-mixed hate speech dataset was recently published [[Ombui et al., 2019](#)]. They gathered their dataset from Twitter, mainly related to the 2017 general election in Kenya. It is worth mentioning the work by [Oriola and Kotzé \[2020\]](#) proposing a code-mixed Twitter dataset containing 14,896 tweets written in a mix of four different languages, namely English, Afrikaans, IsiZulu, and Sesotho.

**Topical Focus.** Similarly to the English datasets, these lower resourced languages datasets also feature different topical focuses, where hate speech is the most used phenomenon to describe the resource. Other datasets cover several abusive phenomena such as Offensiveness, Abusiveness, Misogyny, Aggressiveness, and Cyberbullying. The topical focus is also an important aspect to be considered in the cross-lingual abusive language detection task. A study found that topic bias was one of the main issues in cross-lingual abusive language detection [[Arango et al., 2020](#)]. If we do not want to deal with topic-shift between languages, we notice some datasets which only focus on one topic and cover more than one language, such as hate speech and misogyny. We also aware that there are a lot of datasets that have hate speech topics. However, different approaches in collecting the data could potentially introduce another bias issue when exploited in cross-lingual settings. As observed by [Arango et al. \[2020\]](#), several biases such as user bias, racial bias, and sampling bias could be an issue in cross-lingual abusive language detection task. Otherwise, we can freely choose the available datasets if we want to tackle both domain-shift and language-shift.

**Data Source.** Most resources were retrieved from social media platforms such as Twitter, Facebook, and Instagram. Twitter is the most convenient platform which provides API and a more friendly policy to retrieve and distribute the samples gathered from its platforms. We can see from Table 2.3 that almost 60% of abusive language datasets were

obtained from Twitter. Some other datasets were obtained from comments on news sites, online forums such as Reddit and Youtube comments. In a multilingual or cross-lingual setting, we also need to pay attention to the source of the data. Every source has its own specific characteristics, such as stylistic aspects and formality levels. Twitter data have some specific features, such as hashtags and user mentions. Language in social media platforms is usually used more informal language than other sources such as news site comments.

**Availability.** Based on the manual check, most of the abusive language datasets in lower resourced language were made publicly available. We only found 4 out of 60 resources were not shared publicly by their authors. However, some authors decided to provide only the tweet identifier due to some Twitter policies and allowed us to retrieve the tweets by using the Twitter public API. The restricted datasets are mostly obtained from other sources than Twitter, which provides a more strict policy for sharing the data.

**Annotation Scheme.** Similar to the cross-domain setting, in the cross-lingual experiment, we also need to uniform the labels of every dataset. Most previous studies decided to binarize the label into two classes, namely abusive and not abusive. Based on our investigation, some datasets have more than two labels to captured a finer-grained phenomenon instead of merely limiting it to binary labels. Previous studies proposed to combine some labels when some of them can be safely merged into one class [Karan and Šnajder, 2018, Pamungkas and Patti, 2019]. For example, the TRAC-1 datasets [Kumar et al., 2018] have three labels: *overtly aggressive*, *covertly aggressive*, and *not aggressive*. In this case, we can combine *overtly aggressive* and *covertly aggressive* as *aggressive class*. Otherwise, we can remove the data with a specific label when it is too problematic to merge some classes into one class. For example, the dataset proposed by Ousidhoum et al. [2019] introduces some classes, including *hate speech*, *abusive*, *offensive*, *disrespectful*, *fearful*, and *normal*. In this case, we can combine *hate speech*, *abusive*, and *offensive* into one *abusive* class, but it is quite problematic to include the *disrespectful* and *fearful* label in the class, as proposed by Aluru et al. [2020].

**Data Distribution.** In the cross-lingual setting of abusive language detection task, we also need to consider the data distribution of training (in source languages) and testing (in target languages) data. Based on our manual inspection, most of the resources have more positive (abusive) samples than negative (not abusive) ones. As mentioned in the cross-domain part, maintaining the same class distribution of training and testing data is important to have a more reliable evaluation and avoid bias in the models [Pamungkas et al., 2020b, Pamungkas and Patti, 2019]. Therefore, if the test set only contains 20% of abusive instances, a similar distribution can be imposed on the training set in the source language by adding or removing instances.

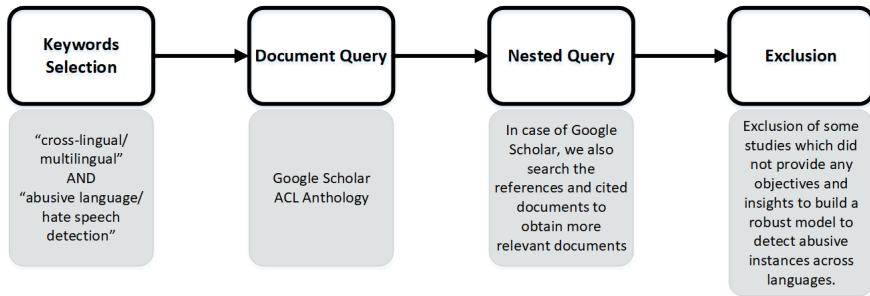


Figure 2.2: Documents Collection Methodology

### 2.3.3 Proposed Approaches in Multilingual Abusive Language Studies

This subsection presents the existing studies focusing on building robust models to detect abusive language across different languages automatically. Overall, we use the same approach, as explained in Section 2.2, to collect related studies from several publication repositories. The only difference is the keywords used to query the relevant publications. For this purpose, we employ four keywords, namely ‘cross-lingual abusive language detection’, ‘cross-lingual hate speech detection’, ‘multilingual abusive language detection’, and ‘multilingual hate speech detection’. We apply these keywords in two scientific publication repositories, namely Google Scholar and ACL Anthology. In the case of Google Scholar, we limit the query only to the first five pages of each keyword, without any limitation on publication time. We also check the cited documents and references for each document shown in the query result. Finally, we remove some studies which did not provide any objective and insight to build a robust model to detect abusive instances across languages. For example, we observe some experiments with different models to cope with datasets in different languages. The summary of the methodology adopted for collecting relevant studies can be seen in Figure 2.2.

Table 2.4 summarizes the existing works found on abusive language detection across different languages. We notice that the study in this direction is still relatively new, with the first study found in 2019. The works are more centered on languages from the Indo-European family, such as English, French, Spanish, Italian, German, and Hindi, in line with the available resources. Most of them were trying to transfer the knowledge from a resource-rich language (English) to other languages with the lower resource available. All studies proposed a supervised approach, where most of them utilized a multilingual language representation as a basis for knowledge transfer between languages. Following, we discuss the gathered studies in this direction, focusing on several aspects, including the model adopted, features used, and approaches proposed to deal with language-shift.

Models	Approach	Year	Ref
Traditional Model	Proposed to use the <i>bleaching</i> approach <a href="#">van der Goot et al. [2018]</a> with a model based on SVM to conduct cross-lingual experiments between Italian and English	2018	<a href="#">Basile and Rubagotti [2018]</a>
Traditional Model	Experimented with a gradient-boosting model and proposed to concatenate two sentence embeddings obtained from LASER Embedding and Multilingual BERT as a language-agnostic representation.	2019	<a href="#">Saha et al. [2019]</a>
Traditional Models	Experimented with the use of machine translation tools to translate the training data to the target language and exploited a wide range of traditional models including SVM, naïve bayes, and random forest.	2019	<a href="#">Ibrohim and Budi [2019b]</a>
Neural Based	Proposed a joint-learning architecture based on LSTM coupled with features from HurtLex to transfer knowledge between domains and languages.	2019	<a href="#">Pamungkas and Patti [2019]</a>
Transformer Based	Proposed multichannel architecture based on BERT model, which learns the task sequentially in three languages: source languages, English, and Chinese.	2019	<a href="#">Sohn and Lee [2019]</a>
Neural Based	Proposed multitask architecture based on Sluice Networks coupled with Babylon cross-lingual embedding.	2019	<a href="#">Ousidhoum et al. [2019]</a>
Transformer Based	Proposed to continue training multilingual BERT and XLM-RoBERTa via masked language modeling (intermediate MLM-ing) as a language and domain adaptation approach.	2020	<a href="#">Glavaš et al. [2020]</a>
Transformer Based	Proposed to use XLM-RoBERTa by inter-language and inter-task language transfer learning for conducting cross-lingual classification of offensive languages.	2020	<a href="#">Ranasinghe and Zampieri [2020]</a>
Transformer Based	Employed multilingual BERT and proposed two data augmentation techniques for the cross-lingual transfer by adding training set with filtered samples from the semi-supervised dataset and samples from languages other than target languages.	2020	<a href="#">Ahn et al. [2020]</a>

Transformer Based	Proposed a hybrid emoji-based masked language model (MLM) on the top of XLM architecture to leverage the common information conveyed by emojis as a language-agnostic feature.	2020	<a href="#">Corazza et al. [2020b]</a>
Multiple Models	Proposed to infuse features from multilingual hate lexicon called HurtLex into traditional (SVM) and neural models (LSTM) for transferring knowledge across different languages.	2020	<a href="#">Pamungkas et al. [2020b]</a>
Transformer Based	Exploited cross-lingual representation based on XLM-RoBERTa for building multilingual models and tested on five different languages.	2020	<a href="#">Dadu and Pant [2020b]</a>
Transformer Based	Proposed a novel architecture consisting of a frozen Transformer Language Model (TLM) and Attention-Maximum-Average Pooling (AXEL) to deal with zero-shot and few-shot cross-lingual learning.	2020	<a href="#">Stappen et al. [2020]</a>
Transformer Based	Proposed a multichannel BERT architecture that learns the task from both source and target languages.	2020	<a href="#">Casula [2020]</a>
Transformer Based	Proposed to convert Hindi-English code-switched data into the high resource languages (English) for exploiting both monolingual and cross-lingual settings by using the state-of-the-art cross-lingual language model XLM-RoBERTa.	2020	<a href="#">Dadu and Pant [2020a]</a>
Multiple Models	Conducted an exploratory work using logistic regression, several deep learning models (CNN-GRU, and BERT-based) and multilingual language representations (LASER, MUSE, and multilingual BERT) to deal with multilingual hate speech detection in nine languages.	2020	<a href="#">Aluru et al. [2020]</a>
Neural Based	Conducted an extensive experiment to build a language-agnostic model based on recurrent neural networks (RNN) by exploiting several language-agnostic features.	2020	<a href="#">Corazza et al. [2020a]</a>
Transformer Based	Proposed a single multilingual hate speech model based on the multilingual BERT model, which is trained on datasets in five different languages.	2020	<a href="#">Pérez et al. [2020]</a>

Multiple Models	Experimented with several models including traditional models (logistic regression), neural models (CNN-LSTM), and transformer models (BERT) to build a multilingual system trained on code-switched datasets in English and Hindi by adopting a transfer learning approach.	2021	Vashistha and Zubiaga [2021]
-----------------	--	------	------------------------------

Table 2.4: Summary of approaches adopted in existing studies on cross-lingual abusive language detection

**Models.** Based on Table 2.4, most studies implemented transformer-based architecture to deal with abusive language detection in a cross-lingual setting. However, we also notice some works that exploited a traditional machine learning approach, such as logistic regression [Basile and Rubagotti, 2018, Vashistha and Zubiaga, 2021, Aluru et al., 2020], linear support vector machines [Pamungkas and Patti, 2019, Pamungkas et al., 2020b], and support vector machines [Ibrohim and Budi, 2019b]. They used multilingual language representation or simple translation tools (to translate the data training to the target languages) for the knowledge sharing between languages. Some studies also exploited several neural-based models such as LSTM [Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Vashistha and Zubiaga, 2021, Corazza et al., 2020a], Bi-LSTMs [Corazza et al., 2020a], and GRU [Aluru et al., 2020, Corazza et al., 2020b]. The more recent works adopted several transformer-based architectures due to the availability of multilingual transformer models such as Multilingual BERT [Glavaš et al., 2020, Ahn et al., 2020, Pamungkas et al., 2020b, Stappen et al., 2020, Vashistha and Zubiaga, 2021, Aluru et al., 2020, Pérez et al., 2020], RoBERTa [Dadu and Pant, 2020b,a], XLM [Corazza et al., 2020b, Stappen et al., 2020], and XLM-RoBERTa [Glavaš et al., 2020, Ranasinghe and Zampieri, 2020, Dadu and Pant, 2020b,a]. Interestingly, we also observe work that proposed a multichannel architecture based on the multilingual BERT model [Casula, 2020, Sohn and Lee, 2019], which allows the model to learn the task in several languages sequentially. Finally, we also discovered a study proposed to adapt a multitask approach to deal with this task [Ousidhoum et al., 2019].

Based on our investigation, transformer-based models with multilingual language representations effectively deal with language-shift in the zero-shot cross-lingual abusive language detection task. Recent studies showed that XLM-RoBERTa provides a more robust performance than other multilingual language models, including multilingual BERT and RoBERTa [Glavaš et al., 2020, Ranasinghe and Zampieri, 2020, Dadu and Pant, 2020b]. However, the most recent study showed that the use of a straightforward English BERT pre-trained model with the help of translation tools already achieved a competitive result. The more complex approaches that adopt joint-learning [Pamungkas and Patti, 2019], multichannel [Sohn and Lee, 2019], or multitask [Ousidhoum et al., 2019] architectures also obtained a competitive results, outperformed the more simpler

models previously mentioned.

**Feature Representation.** For the traditional models, some works used the LASER Embedding model, which provides a language-agnostic representation across 93 languages. A study by Basile and Rubagotti [2018] proposed to use TF-IDF representation of bleached characters n-grams. Other studies simply translated the training data to the target language and used the word n-grams feature representation [Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Ibrohim and Budi, 2019b, Aluru et al., 2020]. Meanwhile, most neural-based models were coupled by multilingual word embedding models, including Facebook MUSE (Multilingual FastText) [Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Aluru et al., 2020, Ousidhoum et al., 2019] and Babylon Embeddings [Ousidhoum et al., 2019]. Finally, the transformer-based architectures exploited the multilingual pre-trained model trained on the very big corpus such as Multilingual BERT [Glavaš et al., 2020, Ahn et al., 2020, Pamungkas et al., 2020b, Stappen et al., 2020, Vashistha and Zubiaga, 2021, Aluru et al., 2020, Pérez et al., 2020], RoBERTa [Dadu and Pant, 2020b,a], ULMFit [Dadu and Pant, 2020a], and the recent XLM-RoBERTa [Glavaš et al., 2020, Ranasinghe and Zampieri, 2020, Dadu and Pant, 2020b,a]. It is worth noting that we also found that some features were introduced to complement the language representation, providing language-agnostic information for knowledge transfer such as a hate-specific lexicon (HurtLex) [Pamungkas and Patti, 2019, Pamungkas et al., 2020b, Corazza et al., 2020a] and emotion features based on emoji presence [Corazza et al., 2020b].

Overall, almost all cross-lingual abusive language detection studies exploited multilingual language models as the main feature representation. In particular, the most recent studies found that a multilingual representation based on XLM-RoBERTa obtained the most robust result and outperformed other multilingual language models [Glavaš et al., 2020, Ranasinghe and Zampieri, 2020, Dadu and Pant, 2020b]. Several studies also present the interesting finding that infusing language-agnostic features extracted from hate-specific lexicons HurtLex, in particular) [Pamungkas and Patti, 2019, Corazza et al., 2020a] and emoji-based features [Corazza et al., 2020b] could improve abusive language detection systems in a multilingual setting. In the case of HurtLex, the feature was represented as a one-hot vector which indicates the word presence in 17 HurtLex categories [Pamungkas and Patti, 2019]. Meanwhile, Corazza et al. [2020b] exploited common information conveyed by emoji for building a pre-trained Masked Language Model (MLM).

**Language Transfer Approaches.** Cross-lingual transfer learning is the common approach to transfer knowledge from one language to another language [Lin et al., 2019, Schuster et al., 2019]. In this approach, models are trained and optimized on a dataset from one language (called *source* language), and then tested on another language (called *target* language). In this task, a specific model or approaches is needed to facilitate the knowledge transfer between language. In this part, we discuss several approaches proposed by existing works to bridge the language-shift in cross-lingual abusive language detection task.

Several works proposed the most straightforward approach by **utilizing machine translation tools to align data training and testing language**. Most of them used

Google Translate, which provides reliable translation results. Pamungkas et al. [2020b], Pamungkas and Patti [2019] exploited Linear Support Vector Classifier with TF-IDF feature representation of translated data by Google Translated. Some other works also tried to align the language of test data to the source language before feeding them to state of the art English BERT pre-trained models [Pamungkas et al., 2020b, Aluru et al., 2020]. The translation tools were also used to obtain parallel corpora in some studies which propose a joint learning or multichannel architecture. These architectures require these corpora to allow the model to learn the task in two or more languages sequentially [Pamungkas et al., 2020b, Pamungkas and Patti, 2019, Casula, 2020, Sohn and Lee, 2019].

Some existing studies proposed to experiment by **infusing language-agnostic features as language-independent information for transferring knowledge between languages**. Pamungkas et al. [2020b], Pamungkas and Patti [2019], Corazza et al. [2020a] used features extracted from HurtLex [Bassignana et al., 2018], a multilingual lexicon that specifically contains abusive words. Another work by Corazza et al. [2020b] exploits a language-agnostic feature provided by emoji in the Twitter data. They argued that emoji could give some signals related to emotion information.

**Novel architectures** was also proposed by several works to obtain a better learning representation across different languages. Glavaš et al. [2020] proposed to continue the training process of Multilingual BERT and XLM-RoBERTa models via masked language modeling. Pamungkas et al. [2020b], Pamungkas and Patti [2019] presented a joint-learning architecture model to learn the task in source and target languages sequentially. Then, Casula [2020], Sohn and Lee [2019] introduced a similar architecture by introducing a multichannel model based on multilingual pre-trained models. Then, Stappen et al. [2020] introduced novel architecture consisting of a frozen Transformer Language Model (TLM) and Attention-Maximum-Average Pooling (AXEL) to deal with the zero-shot cross-lingual classification. Finally, Ousidhoum et al. [2019] proposed a multitask architecture based on Sluice Network [Ruder et al., 2017] coupled with Babylon cross-lingual word embedding Smith et al. [2017], which allows the model to share the same parameters from other related tasks.

The cross-lingual task heavily relied on the machine translation tools for a long time before the emergence of multilingual language representation in recent years. Some prior studies **conduct an exploratory experiment to test the robustness of these multilingual language representation models** in abusive language detection tasks, without any specific knowledge transfer approaches between languages. Pamungkas et al. [2020b], Pamungkas and Patti [2019], Aluru et al. [2020] used a straightforward logistic regression model coupled with Multilingual LASER Embedding. In addition, they also experimented with the Multilingual FastText embedding. Then, Pamungkas et al. [2020b], Pamungkas and Patti [2019], Aluru et al. [2020], Pérez et al. [2020] also test the robustness of the Multilingual BERT model to tackle cross-lingual abusive language detection. Ranasinghe and Zampieri [2020], Dadu and Pant [2020b,a] experimented with the recent state of the art multilingual language representation XLM-RoBERTa to deal with this task. Finally, we found work that proposed two data augmentation techniques for cross-lingual transfer by adding a training set with filtered other data samples and

using an ensemble model based on the Multilingual BERT pre-trained model [Ahn et al., 2020].

This section discusses the development of studies in building robust models to detect abusive language across multiple languages. Specifically, we focus on the abusive language detection task in a cross-lingual setting. In the last section, we present state-of-the-art studies in automatic misogyny identification tasks, and a specific abusive language task focuses on detecting specific abusive phenomena called misogyny. This task will be used as a case study to apply our approaches, including swear word abusiveness analysis, cross-domain experiment, and cross-lingual experiment, in one specific abusive language phenomena.

## 2.4 Automatic Misogyny Identification

In this section, we give a review of the recent literature on automatic misogyny identification. Several studies are connected to descriptive reports of the systems participating in shared tasks, in particular, we identified two closely related shared tasks, reviewed and discussed below. We also provide an overview of the recent literature on abusive language detection, specifically in cross-domain and cross-lingual settings.

### 2.4.1 The Phenomenon of Misogyny

Misogynistic speech is a well-studied phenomenon in social media, and its detection is often cast as a text classification problem. Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in various ways, including social exclusion, discrimination, hostility, threats of violence and sexual objectification. Misogynistic language is a multifaceted phenomenon with its own specificity and it is often imbued with expressions of sexism and offensive language. Moreover, since misogyny is a form of hate, the current studies on the automatic identification of this phenomenon are related to the field of automatic hate speech detection, and studied also in conjunction with other expressions of hate, as in the HatEval shared task proposed in 2019 at SemEval [Basile et al., 2019]. HatEval provided a Twitter data set annotated for hate speech against women and immigrants, where a contrastive comparison of misogynist and xenophobic messages in English and Spanish is possible. While the woman-targeted section of the HatEval dataset could be considered compatible to a misogyny detection benchmark, the distinction was not made explicit and the “target” label was not published. During the competition, the participant systems were evaluated on their capacity to predict hate speech on the whole set of tweets, regardless of the target.

Despite the philosophical debate on whether *sexism* and *misogyny* are distinct concepts [Manne, 2017], or whether misogynistic speech is *hate speech* [Richardson-Self, 2018], there is a strong relation between those phenomena. One of the first studies on sexism detection was proposed by Waseem and Hovy [2016], in conjunction with another abusive phenomenon, namely racism. The dataset has been widely adopted in the broader context of abusive language detection. Jha and Mamidi [2017] proposed another benchmark

dataset, providing a distinction of sexism utterances into two forms: hostile (when sexism is characterized by an explicitly negative attitude) and benevolent (when sexism is more subtle, often expressed as a compliment). A more recent study by Sharifrad and Jacovi [2019] presented a new categorization of sexism including indirect, sexual, and physical sexism, building a CNN model to automatic classify tweets into these three categories. Several studies on abusive language detection used the datasets mentioned above, with different focuses such as hate speech detection [Badjatiya et al., 2017, Kshirsagar et al., 2018, Qian et al., 2018a, Fehn Unsvåg and Gambäck, 2018], author profiling in abuse detection [Mishra et al., 2018], bias in abusive language detection [Park et al., 2018, Wiegand et al., 2019, Davidson et al., 2019], and cross-domain abusive language detection [Karan and Šnajder, 2018, Pamungkas and Patti, 2019, Swamy et al., 2019, Waseem et al., 2018].

#### 2.4.2 Misogyny Detection in Social Media

The earliest work we found specifically on *misogyny* in social media was proposed by Hewitt et al. [2016], where misogynistic tweets are collected by using several terms used to attack women, and coded manually by a single annotator. Research on automatic misogyny identification was boosted by Anzovino et al. [2018], introducing a new benchmark dataset annotated on two levels: i) misogyny identification, and ii) misogynistic behavior and target classification. They also built systems to detect misogynistic tweets automatically, employing several classifiers including random forest, naive bayes, support vector machine, and multilayer perceptron. Two shared tasks investigate the misogyny phenomenon in social media on multiple languages, namely AMI IberEval 2018 [Fersini et al., 2018c] (Spanish and English) and AMI EVALITA 2018 [Fersini et al., 2018a] (Italian and English). Table 2.5 summarizes the participating systems in these shared tasks. Several approaches were proposed, from traditional supervised classifiers such as naive bayes, SVM, and random forest, to deep learning techniques such as Bi-LSTM. Some participants to the shared tasks proposed ensembles of classifiers, by aggregating the output from several classifiers to make the final prediction. However, the best systems in both campaigns are simple classifiers (SVM for AMI IberEval and logistic regression for AMI EVALITA) with manually engineered features.

A philosophical account of misogyny and sexism has been provided by Manne [2017], which argues that they are distinct. On this line, Frenda et al. [2019] presented an approach to detect both misogyny and sexism analyzing collections of English tweets.

### 2.5 Summary

This chapter present literature study of the effort to build a robust model to detect abusive language in online content. Based on previous studies, we observed that swear word has a key role in abusive language detection task. Most works found that the presence of swear word is an important signal to spot abusive content. However, some philosophical studies argue that the use of swear words also has several upsides when they are used not

<b>Authors</b>	<b>Shared Task</b>	<b>Approach</b>
Pamungkas et al. [2018c]	AMI IberEval	SVM with a combination of handcrafted stylistic, structural, and lexical features.
Goenaga et al. [2018]	AMI IberEval	Bi-LSTM with pretrained word embeddings.
Liu et al. [2018a]	AMI IberEval	Average probability of two traditional classifiers trained on doc2vec.
Canós [2018]	AMI IberEval	SVM with tf-idf unigrams.
Nina-Alcocer [2018]	AMI IberEval	SVM, Multilayer Perceptron (MLP) and Multinomial Naive Bayes, with structural, lexical, and syntactical features.
Shushkevich and Cardiff [2018b]	AMI IberEval	Logistic regression, naive bayes, SVM, and ensemble classifier, with tf-idf.
Frenda et al. [2018b]	AMI IberEval	Ensemble of SVM classifiers with character n-grams, sentiment, and lexicons.
Pamungkas et al. [2018b]	AMI EVALITA	Linear and RBF kernel SVM with structural and lexical features, including a multilingual hate lexicon.
Bakarov [2018]	AMI EVALITA	Single Value Decomposition and boosting classifier with tf-idf.
Basile and Rubagotti [2018]	AMI EVALITA	SVM with n-grams and cross-lingual classification with bleaching.
Saha et al. [2018]	AMI EVALITA	Logistic regression trained on concatenated sentence embeddings, tf-idf, and average word embeddings.
Ahluwalia et al. [2018]	AMI EVALITA	Voting ensemble with handcrafted features.
Shushkevich and Cardiff [2018a]	AMI EVALITA	Ensemble of logistic regression, SVM, and naive bayes, with tf-idf.
Buscaldi [2018]	AMI EVALITA	Bi-LSTM with character embedding and random forest with weighted n-grams.
Frenda et al. [2018a]	AMI EVALITA	SVM and random forest with stylistic and lexical features and lexicons.

Table 2.5: Summary of the AMI shared task systems.

in an abusive way. Therefore, we conclude that resolving the swear word context as either abusive or not abusive is important to have a reliable model for detecting abusive language. We also have an extensive review on the abusive language study, both in multidomain and multilingual settings. We observed that this research focus is still relatively new and still has several open problems. One of the main issue is related to the bias both on the datasets or models, which is observed during the cross-domain or cross-lingual experiment. Another issue is mainly related to the insufficient ability of current language resources to deal with both domain-shift and language-shift in cross-domain and cross-lingual settings. Finally, we also decide to have a review study on specific automatic misogyny identification task, since we plan to use this task for conducting investigation related to the building of robust models to detect abusive language. Overall, we found that this task is still new and very relevant for our needs as the task is featured by datasets in multiple languages. The state of the art results are also well-documented, which makes the results comparison is feasible.



## Chapter 3

# Swear Words and Abusive Language Detection

In recent years, more and more studies focused on abusive language detection which covers hate speech, cyberbullying, trolling, and offensive language [Waseem et al., 2017, Schmidt and Wiegand, 2017, Michal et al., 2010]. Swear words play an important role in these tasks, providing a signal to spot an offensive utterance [Malmasi and Zampieri, 2018]. However, the presence of swear words could also lead to false positives when they occur in a casual context [Chen et al., 2012, Nobata et al., 2016, Van Hee et al., 2018, Malmasi and Zampieri, 2018, Nozza, 2021]. Therefore, resolving swearing context would be beneficial to improve the performance of abusive language detection model and also important aspect for building robust model in this task.

In this chapter, we conduct an in-depth investigation on the role of swear words and their context in abusive language detection tasks. We explore the phenomenon of swearing in online conversation, taking the possibility of predicting the abusiveness of a swear word in a tweet context as the main investigation perspective. To achieve this objective, we conduct several contributions. First, we develop a new benchmark Twitter corpus, called SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level. Second, we develop and experiment with supervised models to automatically predicting abusive swearing within the tweet context. Such models are trained on the novel SWAD corpus to predict the *abusiveness* of a swear word within a tweet. Finally, we investigate the impact of resolving swear word abusiveness on downstream abusive language detection tasks.

This chapter is organized as follows. Section 3.1 provide the main motivation of this research focus. Then, Section 3.2 describe our novel swear word abusiveness corpus, including the annotation procedure and the analysis of the overall annotation result. In Section 3.3, we experiment with the automatic prediction of swear word abusiveness, by proposing several experimental settings. Furthermore, we also explore the impact of resolving swear word abusiveness in the downstream abusive language detection tasks in Section 3.4. Finally, Section 3.5 summarize our important findings in this study based on the experimental results.

### 3.1 Motivation

Swearing plays an ubiquitous role in everyday conversations among humans, both in oral and textual communication, and occurs frequently in social media texts, typically featured by informal language and spontaneous writing. Such occurrences can be linked to an abusive context, when they contribute to the expression of hatred and to the abusive effect, causing harm and offense. However, swearing is multifaceted and is often used in casual contexts, also with positive social functions. With the emergence of abusive language study, the usage ambiguity of swear words raises issues with respect to automated abusive language detection models. On the one side, swear words could provide a signal to spot an abusive instances. However, the presence of swear words could also lead to false positive when they are used in a casual context. In this direction, the main goal is to automatically differentiate between abusive swearing, which should be regulated and countered in online communication, and not abusive one, that should be allowed as part of freedom of speech, also recognizing its positive functions, as in the case of reclaimed uses of slurs.

### 3.2 Corpus Creation and Analysis

As the first effort to have a deeper analysis of swear word use in abusive language tasks, we propose building a novel corpus consisting of tweets where swear words are annotated at the word level as either abusive or not abusive on their use within their context. This corpus will become the basis for our further investigation of the swear words' role in abusive language tasks, presented in the next sections.

#### 3.2.1 Corpus Collection

Our starting point was a corpus of tweets selected from the training set of the Offensive Language Identification Dataset (OLID) [Zampieri et al., 2019a], which was proposed in the context of the shared task OffensEval [Zampieri et al., 2019b] at SemEval 2019<sup>1</sup>. This task is aimed at detecting offensive messages as well as their targets. In OLID, Twitter messages were labelled by applying a multilayer hierarchical annotation scheme, which encompasses three dimensions, including tags for marking the presence of offensive language (*offensive* vs *not offensive*), tags for categorizing the offensive language (*targeted* vs *untargeted*), and tags for the offensive target identification (*individual*, *group*, or *other*). The broader coverage of the concept and definition of offensive language are the main reasons we choose this dataset as starting point for our finer grained annotation concerning swearing, rather than other datasets developed around more specific typologies of offensive language, such as hate speech, cyberbullying or misogyny, which we think could introduce a bias in our corpus, undermining the generality of its possible future exploitation.

Some preprocessing has been applied to the OLID data, such as mention and URL normalization. Since our focus is on analyzing swear words in the tweet context, we first filtered out a subset of tweets from OLID based on the presence of swear words,

---

<sup>1</sup><http://alt.qcri.org/semeval2019/index.php?id=tasks>.

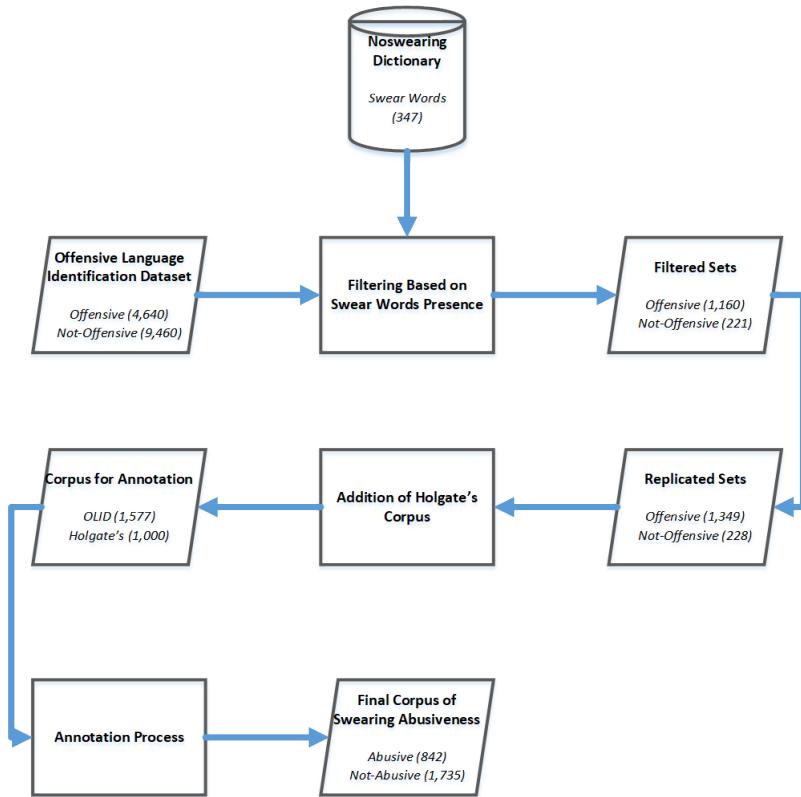


Figure 3.1: Corpus Development Process.

	Original	After Filtering	After Replication
Offensive	4,400	1,111	1,296
Not	8,840	209	215
Total	13,240	1,381	1,511

Table 3.1: Corpus statistics after filtering process.

in order to obtain a collection of tweets that include at least one swear word. At this stage we exploited the list of swear words published on the [noswearing](#) website<sup>2</sup>, an online dictionary site which includes a list of swear words. This dictionary includes 349 swear words covering general vulgarities, slurs, and sex-related terms. We manually checked the list to exclude highly ambiguous words, namely swear words like “ho” and “hard on”<sup>3</sup>. Table 3.1 shows the full statistics of our corpus after the filtering process. We identified 1,320 tweets that contain at least one swear word. Since this annotation task is at the (swear) word level, tweets which have more than one swear word were replicated. We generated as many new instances of the same tweet as the number of swear words occurring in the message, and marked each single swear word with special tags  $< b >$  and  $< /b >$  (e.g.  $< b >$ fuck $< /b >$ ,  $< b >$ shit $< /b >$ , and etc.) so that the abusiveness label on each instance records the context of the marked swear word in the tweet (abusive or not). For instance, given the message *@USER This shit gon keep me in the crib lol fuck it*, two instances will be generated: *@USER This < b >shit< /b > gon keep me in the crib lol fuck it* and *@USER This shit gon keep me in the crib lol < b >fuck< /b > it*.

We found 154 tweets having more than one swear word, with a range of occurrences from 2 to 6 swear words. As a result, we have 1,511 instances to be annotated. Figure 3.1 shows the overall process of our corpus development.

### 3.2.2 Annotation Task and Process

The annotation of 1,511 instances involved three expert annotators, with different gender and ages. All instances were annotated by two independent annotators (A1 and A2). The resulting disagreement was resolved by involving the third annotator (A3), labeling those instances where a disagreement between A1 and A2 was detected. All annotators use English as a second language, with a minimum level of B2.

**Annotation task** Annotators were asked to annotate (with a binary option) whether the highlighted swear word (tagged with the  $< b >$  and  $< /b >$  tags) can be considered *abusive swearing*, contributing to the construction of an abusive context (by using the tag “yes”) or whether the swear word does not contribute to the construction of an abusive context (by using the tag “no”). We first started a trial annotation on a portion of 100 tweets from the collection, to test our annotation guidelines and improve the understanding between annotators. During this trial annotation we also deepened our understanding of the *offensiveness* notion, which underlies the definition of offensive language driving the whole OLID annotation process. There is a crucial difference between the coarse notion of offensive language as defined in OLID and the concept of abusive language we are interested in, given our main goal to reason about abusive swearing. Indeed, according to the OLID definition a tweet can be considered offensive only because of the presence of profanities, even if no occurrence of abusive swearing can be detected.

---

<sup>2</sup><https://www.noswearing.com/>

<sup>3</sup>In the [noswearing](#) site “ho” is a short form of “hoe”, but in the dataset we found that word “ho” is mostly used as a short form of “how”. Similarly, “hard on” is a slang word of “erection” in the [noswearing](#) site, but this word is frequently used to express hard effort, as in “...I’m working *hard on* this task right now,...”

Such considerations have driven our decision to annotate the abusiveness of swear words on tweets belonging to both classes (offensive and not-offensive) of the OLID data. Another issue discovered during the trial annotation consisted in some cases where the swear word is used for indirect insult: the swear word itself is used to insult, but the overall context of the tweet is not abusive. This mostly happened in the reported speech such as in the example below, where we determined this tweet as not abusive:

[Example of indirect insult.]

@USER Everyone saying f\*ck Russ dont know a damn thing about him or  
watched the interview 

Therefore, in the final annotation guidelines, we decided to include the author *intention* to resolve the swear word context, especially to deal with this kind of swear word use. We consider abusive swearing those uses where *swearing contributes to the construction of an abusive context such as name-calling, harassment, hate speech, and bullying, involving several sensitive topics including physical appearance, sexuality, race and culture, and intelligence, with intention from the author of tweet to insult or abuse a target (person or group of persons)*. Let us notice that one tweet can have more than one swear word, but for every tweet, only one swear word will be highlighted as relevant for the annotation in each row (see the replication process explained above). Therefore, the annotator only needs to focus on the marked swear words (e.g., < b >fuck< /b >). We remark again that abusive swearing can be found on both offensive and not-offensive tweets, therefore during the application of our annotation layer, we decided to ignore the original message-level layer of annotation from the original OLID (offensive vs not-offensive), in order to avoid confusing the annotators during the annotation process. Indeed, we observed four possible cases, when we consider the OLID original labels on the offensiveness of a tweet, namely: i) the message is offensive and the swear word is abusive, ii) the message is offensive but the swear word is not abusive, iii) the message is not offensive but the swear word is abusive, and iv) the message is not offensive and the swear word is not abusive. Let us provide an example for each case to get a better understanding on such circumstances:

[Ex. i): offensive tweet & abusive swearing]

@USER You are an absolute **d\*ck** 

[Ex. ii): offensive tweet & not abusive swearing]

@USER I was definitely drunk as **sh\*t**

[Ex. iii): not offensive tweet & abusive swearing]

@USER **bullsh\*t** there's rich liberals too so what are you saying ???

[Ex. iv): not offensive tweet & not abusive swearing]

@USER Haley thanx! you know how to brighten up my **sh\*tty** day 

	Original Abusive OLID	Not- Abusive
Offensive	1,296	568
Not	215	52
Total	1,511	620

Table 3.2: Label distribution in the SWAD dataset.

### 3.2.3 Annotation Results and Disagreement Analysis

Referring to the application of two independent annotations on the whole dataset of tweets (A1 and A2), we can say that annotators achieved a good agreement, selecting the same value in a large portion of the annotated tweets being only 216 out of 1,511 the messages where they disagreed by marking in a different way the presence of abusive swearing. The average pairwise agreement percentage amounts to 85.70%. The inter-annotator agreement is 0.652 (Cohen’s kappa coefficient), which corresponds to a substantial agreement. The final SWAD annotated corpus consists of 1,511 swear words immersed in the context of 1,320 tweets, where 620 swear words are marked as abusive and 891 are rated as not-abusive<sup>4</sup>. Table 3.2 shows the detailed distribution of our annotation result. Interestingly, we found more not-abusive swearing than abusive ones in tweets belonging to the offensive class of OLID (728 versus 568). In addition, we also found 52 cases of abusive swearing in tweets belonging to the OLID not-offensive class.

In the following we list and share some interesting findings and elements of discussion related to the annotation task and outcome.

**Most of the non-abusive contexts of swearing are dominated by emphatic and cathartic swearing function.** Cathartic swearing is a swear word function when it is used as a response to pain or misfortune, while emphatic swearing is another swear word function when a swear word is used to emphasize another word in order to draw more attention. Two examples, one for each swearing function mentioned, follow:

[Cathartic function]  
*@USER d\*mn I felt this shit Why you so loud lol*

[Emphatic function]  
*@USER I AM FUCKING SO F\*CKING HAPPY*

**Emojis could become an important signal to resolve the context of a swear word within the tweet.** In some tweets when the context of swear word use is difficult to be resolved, the presence of emojis could give key information. As shown in the following example, without the presence of the emoji, the swear word  *fucking*  seems to contribute to the construction of an abusive context, but the presence of the  *Face with Tears of Joy*  emoji helped annotators to understand the real context of the whole tweet.

---

<sup>4</sup>The corpus is available for research purpose at the following URL: <https://github.com/dadangewp/SWAD>

*@USER ur a f\*cking dumbass fr. there's no way she is anyone else's 😂*

**Irony and sarcasm could provide an issue for automatic prediction based on machine learning approach.** We found some tweets which contain sarcasm and irony, most of the times in not-abusive context. As in other related tasks such as sentiment analysis, irony and sarcasm could contribute to the difficulties of this task. An example follows:

*@USER Yeah we need some more made up bullsh\*t protestors and antifa lol time for an epic beatdown 😊*

Furthermore, we analyzed cases of disagreement between annotators. We conducted a manual analysis of 216 disagreement cases with the aim to extract the most common patterns, which contribute to the difficulty of the annotation task. As a result, we found several difficult cases:

**Missing context.** We found that some tweets are very short, resulting in the context missing. Other instances are also challenging to understand due to the presence of grammatical errors. These issues are very dominant in the annotator disagreement cases. In the following we show two examples where the context is hard to resolve:

[Very short tweet]

*@USER Lmfaoo! 🎉 b\*tch*

[Noisy text with grammatical errors]

*@USER d\*m that headgear is lit sucks im not on pc ubi plz for console to*

**Need of world knowledge to understand the context.** Some tweets are also very difficult to understand due to the lack of world knowledge. Sometimes annotators need to gather more information by using search engine to understand the context. The presence of hashtags usually becomes the key to understand the nature of the context. Let us see an example for this issue:

*@USER @USER It's probably better to have an ✘ next to my name than a pink p\*ssy hat on my head 🙄🙄🙄🙄 #MAGA #MakeAmericaGreatAgain*

### 3.2.4 Corpus Extension

After the full annotation process, SWAD consists of 1,511 instances. We realize that this collection is still relatively small to obtain reliable performance for machine learning models. Therefore, we extended SWAD by conducting another round of annotation. We also included in the collection the test set of the OLID dataset, which contains 860 tweets, and we re-annotated part of instances from Holgate's dataset [Holgate et al., 2018] according to the SWAD guidelines. Similar to SWAD, tweets in Holgate's dataset were filtered based on the presence of vulgar words. Then, all instances of vulgar words were annotated with one of the six categories of vulgar word use by using crowdsourcing

	<b>AGG</b>	<b>EMO</b>	<b>EMPH</b>	<b>AUX</b>	<b>SGI</b>	<b>NV</b>
Abusive	66	62	21	33	18	5
Not Abusive	61	253	142	230	59	50

Table 3.3: Interaction between the original Holgate’s label with our annotation.

approach. They introduced six mutually exclusive labels, namely **express aggression** (AGG), **express emotion** (EMO), **emphasis** (EMPH), **auxiliary** (AUX), **signal group identity** (SGI), and **non-vulgar** (NV) use. The idea of including the Holgate’s dataset in our collection, by applying to the data the SWAD annotation scheme, was stimulated by the possibility of investigating the interaction of our swear word abusiveness label with the swear word function as introduced in Holgate’s study.

We annotated the new data by following the same annotation guidelines described in previous subsection and involving the same pool of three expert annotators. In the case of the OLID test set, we got 66 instances after the filtering and replicating process. For Holgate’s dataset, we only selected the first 1,000 instances to be re-annotated. We re-annotated all tweets regardless of their original labels. To avoid bias in the annotation process, we hide the original label based on Holgate’s study from our annotators’ view. Our effort was therefore towards adding another layer of annotation on the swear word. After the annotation process, we obtained 204 instances annotated as abusive and 796 as not abusive for Holgate’s data. Meanwhile, for the OLID test, we obtained 18 instances annotated as abusive and 48 instances as not abusive. The inter-annotator agreement on this corpus extension is 0.516 based on the Cohen Kappa coefficient from the annotation of the first and second annotators on 1,066 instances. Therefore, we have 2,577 tweets in total after this extension process. Table 3.3 shows the interaction between the original label from Holgate’s work and our new label addition.

Before the annotation process, we expected that most tweets with AGG label will be classified into abusive class. However, we found that these AGG tweets were distributed in both abusive and not abusive classes in a similar proportion. We were interested in AGG tweets, which are categorized into not abusive class. Some examples of these instances are reported in the following. Based on our annotation guidelines, the first example (Ex. i) is not labeled as abusive because there is not an insulted target. Similarly, the second example (Ex. ii) shows an expression of humor and catharsis, which is not classified as abusive based on our annotation task.

[Ex. i): not abusive based on our guidelines]  
*My **bullsh\*t** radar is on full force today*

[Ex. ii): humor and catharsis]  
*I gained ten pounds this summer. **D\*mn...** LOL*

### 3.3 Swear Words Abusiveness Prediction

In this section, we provide an intrinsic evaluation of the corpus by conducting cross-validation experiments. We build supervised machine learning models to predict the abusiveness of swear words in SWAD. We model this prediction task in three different tasks, namely sequence labeling, simple text classification, and target-based swear word abusiveness prediction. The main objective of the sequence labeling experiment is to test the consistency of the annotation of the corpus. Meanwhile, we devise the classification experiment to shed some light on the most predictive feature to differentiate between abusive and not-abusive swearing. We also propose to adopt a target-based sentiment analysis task, a more well-explored task in the sentiment analysis area, into our experiment.

#### 3.3.1 Sequence Labeling Task

In order to test the robustness of the annotation of swear words in SWAD, we devise a cross-validation test based on a sequence labeling task. Given a sequence of words (i.e., a tweet from our dataset), the task consists in correctly labeling each word with one of three possible labels: abusive swear word (**SWA**), non-abusive swear word (**SWNA**) or not a swear word (**NSW**). The task is carried out in a supervised fashion, by splitting the dataset in a training set (90% of the instances) and a test set (the remaining 10%).

##### 3.3.1.1 Model Description

For this experiment, we adapt the BERT Transformer-based architecture [Devlin et al., 2019] with the pre-trained model for English `bert-base-cased`. We train the model for 5 epochs, with learning rate  $10^{-5}$  and a batch size of 32.

##### 3.3.1.2 Results

predicted ground truth	SWNA	SWA	NSW
SWNA	1,366	217	68
SWA	455	322	35
NSW	139	52	38,950

Table 3.4: Sequence labeling task: confusion matrix.

Table 3.4 shows the confusion matrix resulting from the cross-validation. Unsurprisingly, the majority of classification errors are due to SWA/SWNA confusion, while the distinction between swear words and non-swear words is basically trivial. The classifier is slightly biased towards abusive swear words (217 SWA→SWNA misclassifications) than non-abusive swear words (455 SWNA→SWA misclassifications). These results are confirmed by the performance measured in terms of per-class macro-average precision, recall and *F*-score, shown in Table 3.5, where the SWA class has a higher recall than

	precision	recall	<i>F</i> -score
SWNA	.705	.829	.753
SWA	.532	.389	.421
NSW	.997	.994	.995
macro avg	.745	.737	.723

Table 3.5: Sequence labeling task: results broken down by label.

precision, while the opposite is true for the SWNA class. In absolute terms, the per-class and macro-average *F*-score confirms that our annotation is stable when tested in a supervised learning setting. In our test, only one abusive swear word was misclassified as NSW. Interestingly, the word is *sk\*nk*, which is semantically ambiguous, conveying the offensive sense as well as the animal sense. Even more interestingly, the few NSW instances misclassified as SWA are all borderline cases of abusive language: *sh\*ticago* (an offensive slang for Chicago), *messed*, *c\*mming*, and *c\*mslave*.

### 3.3.2 Simple Text Classification Task

In this setting, we explicitly predict the abusiveness of swear words (as the target word) in given tweets as context. We employ several machine learning models including a linear support classifier (L SVC), logistic regression (LR), and random forest (RF) classifier. We use different features, at the word level (focusing on the target word) and at the tweet level (identifying the context). We only implement these traditional models for better interpretability, allowing us to conduct feature analysis to explore important features in this task.

#### 3.3.2.1 Features

**Lexical Features -** In this feature set, we focus on the word-level features. We include the **Swear Word** feature, that is, the unigram of the marked swear word, as we aim to investigate whether the abusiveness of a swear word could be predicted only from the word choice. We also use the **Bigrams** feature, obtained from bigrams of the target word with its next and previous words.

**Twitter Features -** Since our corpus consists of tweets, we also employ several features which are particular to the Twitter data. This feature set include **Hashtag Presence**, **Emoji Presence**, **Mention Presence**, and **Link Presence**. We use regular expressions to extract hashtags, mentions and URLs, and a specialized library <sup>5</sup> for emoji extraction.

**Sentiment Features -** This feature is proposed in order to resolve the context of the tweet. We use two features: **Text Sentiment**, to model the polarity of the text, and **Emoji Sentiment** to model the overall sentiment of the emojis in the tweet. We use the VADER dictionary [Hutto and Gilbert, 2014] to extract the polarity score of the text and *emoji sentiment ranking* <sup>6</sup> to get the sentiment value for emojis.

<sup>5</sup><https://pypi.org/project/emoji/>

<sup>6</sup>[http://kt.ijs.si/data/Emoji\\_sentiment\\_ranking/](http://kt.ijs.si/data/Emoji_sentiment_ranking/)

Feature Set	LSVC				LR				RF			
	P	R	F	Acc	P	R	F	Acc	P	R	F	Acc
ALL	.660	.618	.638	.694	.724	.641	<b>.680</b>	.727	.657	.582	.617	.684
ALL - Unigram SW	.576	.503	.537	.651	.586	.539	.561	.651	.581	.537	.558	.649
ALL - Bigram	.652	.626	.639	.690	.732	.635	<b>.680</b>	.727	.667	.590	.626	.690
ALL - Twitter	.723	.578	.642	.696	.723	.608	.661	.711	.702	.580	.635	.694
ALL - Sentiment	.576	.501	.536	.651	.723	.628	.672	.721	.670	.588	.626	.690
ALL - Stylistic	.667	.585	.623	.688	.719	.635	.674	.723	.642	.572	.605	.676
ALL - Syntactic	.710	.623	.664	.715	.722	.624	.670	.719	.667	.597	.630	.692

Table 3.6: Ablation test on several feature sets.

**Stylistic Features** - In this feature set, we consider several common stylistic features for text classification task such as **Capital Word Count**<sup>7</sup>, **Exclamation Mark Count**, **Question Mark Count**, **Text Length**. In addition, we also exploit another word-level feature, namely **Swear Word Position**, indicating the index position of the marked swear word in the tweet.

**Syntactic Features** - In this feature set, we focus on the word-level features, including **Part of Speech** and the **Dependency Relation** of the target word with its next and previous words. We extract part-of-speech tags with the NLTK library<sup>8</sup>, while dependency relations are extracted with SpaCy<sup>9</sup>.

### 3.3.2.2 System Description and Evaluation

We build our models by using the Scikit-learn library<sup>10</sup>. We split the dataset into 80% and 20% for the training and testing respectively. We use several evaluation metrics, including accuracy, macro average precision, macro average recall, and macro average *F*-score. An ablation test is performed to investigate the role of each feature set in the classification result. The swear word unigram feature is used as a baseline in this experimental setting.

### 3.3.2.3 Results

Table 3.6 shows the full results of the text classification experiment by using LSVC, LR, and RF models. We start the experiment by using all feature groups altogether. Then, we remove one feature at a time to see the importance of each feature group in the model performance. Overall, RF is under-performing compared to the two other classifiers. The results also show that LR performed the best compared to two other models. Based on the macro average *F*-score, the best performance is achieved using all the features coupled with LR. With the same model and by removing Bigrams feature also obtained similar performance, but a lower macro average recall. Our goal is to investigate the most predictive feature set in the ablation experiment by removing one feature set at a time. We found that the unigram of a swear word is the most informative feature in

<sup>7</sup>This feature consider all capital words on the tweet

<sup>8</sup><https://www.nltk.org/>

<sup>9</sup><https://spacy.io/>

<sup>10</sup><https://scikit-learn.org/stable/>

this classification task. Sentiment, Emotion, Twitter, Stylistic and Syntactic features all contribute to the classification performance of LR, while the Bigrams, Twitter, and Syntactic features have a detrimental effect on the LSVC and RF models. The main issue of this task is the lower recall compared to the precision, which is consistent across all models. It denotes that such models struggle to deal with false-negatives. We argue that this happens due to the dataset imbalance, where the swear words percentage over both classes is dominated by not-abusive class (negative class).

### 3.3.3 Target-based Abusiveness Prediction of Swear Words

This setting is similar to the text classification task presented in Section 3.3.2. However, here we explicitly model the task by adopting a similar setting as the target-dependent sentiment analysis task. The main objective of target-dependent sentiment analysis is to identify the sentiment polarity of a given target in an utterance [Jiang et al., 2011]. This task is also related to aspect-based sentiment analysis. However, in target-dependent sentiment analysis, the target word is known and mentioned explicitly in the given utterance. Meanwhile in aspect based sentiment analysis, the target aspect could be expressed implicitly, where the aspect detection is also part of the task. Adopting the similar idea of target-dependent sentiment analysis, we use the swear word as the target word, with the main objective to predict its abusiveness in a given utterance as a context. For instance, see the example below:

[Example from Davidson’s dataset]  
*@USER damn I hate a bitch that like to argue and shit*

In the example, we can find three swear words in the tweet, i.e., “damn”, “bitch”, and “shit”. Therefore, there are three target words, and the task is to predict the abusiveness of each swear word in the tweet individually. Based on our manual investigation, the first swear word is not abusive, the second one is abusive, while the third one is more difficult to assess. The first swear word is used to express catharsis, which is not abusive in most of the cases. The second swear is abusive because it can insult some targets. The last swear word is a bit problematic since the swear word is used as an idiomatic expression. The abusiveness of a given swear word is highly dependent on its context in the tweet, which is identical to the target-dependent sentiment analysis task.

#### 3.3.3.1 System Description

In this experiment, we adopt several state-of-the-art models from the target-dependent sentiment analysis task as baseline models. In addition, we also implement a BERT model by applying a simple masking approach to mark the target words. Following is a short description of each model we use in our experiment:

- **TD-LSTM (Target-dependent LSTM)** The basic idea of this architecture is to model the preceding and following context surrounding the target word so that the feature representation consists of the left part (preceding the target word) and

the right part (following the target word) [Tang et al., 2016]. Specifically, this architecture consists of two LSTMs (LSTM left and LSTM right), which model the preceding and following target word context, respectively. The output of these LSTMs is then concatenated to the softmax layer to predict the label.

- **TC-LSTM (Target-connection LSTM)** This architecture is a further development of TD-LSTM, which tries to incorporate a target connection component. The additional component explicitly models the connection between the target word and each context of the word when building the sentence representation [Tang et al., 2016]. This component was implemented as a target word vector obtained by averaging the vectors of context words it contains. This vector is then concatenated to the word representation before feeding it to the LSTM network. The rest of the architecture is similar to the TD-LSTM.
- **AE-LSTM (Aspect Embedding LSTM)** This architecture [Wang et al., 2016] tries to learn the embedding vector of each aspect, or in our study, is the target word. This vector is then concatenated to the sentence embedding representation, which is followed by the LSTM network. The additional vector representation of the target word gives vital information to the model to learn the sentiment for each target word.
- **AT-LSTM (Attention-based LSTM)** The standard LSTM is not able to model the important part of aspect-based sentiment classification. This particular model (AT-LSTM) [Wang et al., 2016] has an attention mechanism which captures the important part of a sentence by focusing on the given aspect. This attention mechanism took input from the hidden layer produced by LSTM and aspect embedding vector and produce an attention weight vector and a weighted hidden representation.
- **ATAE-LSTM (Attention-based LSTM with Aspect Embedding)** Basically, this architecture [Wang et al., 2016] is AT-LSTM which is concatenated with aspect embedding vector as implemented in AE-LSTM.
- **CABASC (Content Attention Based Aspect Based Sentiment Classification)** This architecture consist of two enhanced attention mechanisms [Liu et al., 2018b], including sentence-level content attention mechanism which captures the important information about given aspects from a global perspective and the context attention mechanism which simultaneously takes the order of the words and their correlations into account, by embedding them into a series of customized memories.
- **IAN** This architecture uses two LSTM networks to model the sentences and the target words [Ma et al., 2017]. Then, the target word’s hidden state and the hidden state of context sentence are placed in parallel to generate an attention vector interactively. Finally, these attention vectors provide a sentence representation and target representation.

- **RAM (Recurrent Attention on Memory)** This framework implements a multiple-attention mechanism that captures sentiment features separated by a long distance so that it is more robust against irrelevant information [Chen et al., 2017b]. The outputs of these multiple attentions are non-linearly combined with the LSTM network, strengthening the model for handling more complications.
- **TD-BERT (Target-dependent BERT)** We also propose to adopt the idea of TD-LSTM and exploit the state-of-the-art pre-trained model BERT as language representation. Therefore, our model consists of two BERT layers (BERT left and BERT right) to represent the context of preceding and following target words, respectively. The output of these BERT layers is passed into a fully connected dense layer with RELU activation before going into the last sigmoid layer to produce the final prediction. This model is optimized using Adam Optimizer with a learning rate of 1e-5 and trained with three epochs and batch size at 32.
- **TM-BERT (Target-masked BERT)** BERT model has an attention mechanism to model many downstream tasks which involve single text or even text pairs. BERT encodes multiple text segments using two special tokens ([SEP] and [CLS]). [SEP] token is used to separate two or more text segments in case of multiple text segment processing. In the single text, the encoded text is started by [CLS] token and ended by [SEP] token. In this model, we introduce a special marker [SW] and [SW] to mark the swear word in the sentence [Boualili et al., 2020]. The intuition for doing so is to inform the important part (target word) of the text for the model. We expect the BERT model able to construct the representation by focusing on this special token. We use an open-source implementation of BERT by HuggingFace<sup>11</sup>, which provides a special method to add a new special token in the BERT masking process.<sup>12</sup>

### 3.3.3.2 Result

As shown in Table 3.7, the **TM-BERT** obtained the best result with .665 in *F*-score in positive class, .843 in *F*-score in negative class, and .754 in macro *F*-score. Overall, the BERT-based models achieved a better result than other models, where **TD-BERT** also obtained a competitive result in all evaluation metrics. We also notice that **TD-LSTM** and **TC-LSTM** get better results than the rest of non-BERT models, including **CABASC** and **RAM**, which achieved better performance in several benchmarks for the aspect-based sentiment analysis task Liu et al. [2018b]. We also compare our result in this experiment with the results in our previous experiment, as presented in Table 3.6. The overall result shows that our models presented in this experiment, which are based on neural architecture, outperformed the traditional models. We also notice that most of the models exploited in this experiment are able to cope with the dataset imbalance issue,

---

<sup>11</sup><https://huggingface.co>

<sup>12</sup>[https://huggingface.co/transformers/model\\_doc/bert.html](https://huggingface.co/transformers/model_doc/bert.html)

Model	$P_0$	$P_1$	$R_0$	$R_1$	$F_0$	$F_1$	$F_{avg}$	Acc
TD-LSTM	0.610	0.793	0.617	0.789	0.613	0.791	0.702	0.729
TC-LSTM	0.628	0.801	0.628	0.801	0.628	0.801	0.714	0.740
AT-LSTM	0.611	0.661	0.061	0.979	0.111	0.789	0.450	0.659
AE-LSTM	0.721	0.708	0.272	0.943	0.395	0.809	0.602	0.709
ATAE-LSTM	0.603	0.786	0.600	0.789	0.602	0.788	0.695	0.723
IAN	0.661	0.735	0.400	0.890	0.498	0.805	0.652	0.719
CABASC	0.747	0.723	0.328	0.940	0.456	0.818	0.637	0.727
RAM	0.628	0.744	0.450	0.857	0.524	0.797	0.660	0.715
TD-BERT	0.719	0.814	0.580	0.848	0.636	0.827	0.731	0.784
TM-BERT	0.708	0.825	0.625	0.859	<b>0.665</b>	<b>0.843</b>	<b>0.754</b>	0.806

Table 3.7: Result of Target-based Abusiveness Prediction of Swear Words.

as we discovered in previous experiments with traditional models, where we obtained lower recall than precision.

## 3.4 Swear Words in Abusive Language Detection

As part of the extrinsic evaluation of our corpus and to have a deeper investigation on the swear words’ role in abusive language detection tasks, in this section we explore the impact of swear word context prediction in several downstream tasks of abusive language detection. We do so by infusing the swear word context prediction as an additional feature to the baseline models.

### 3.4.1 Task Description and Experimental Settings

We explore the usefulness of the swear word abusiveness information feature on several downstream abusive language detection tasks. We reiterate that our assumption is that knowing to swear word context as either abusive or not could help the system to resolve the abusiveness of the whole utterance. Therefore, our idea is to explicitly infuse the swear word abusiveness prediction into the abusive language detection model to help the model dealing with swear word ambiguity. The overall experimental scenario is illustrated in Figure 3.2.

First, we need to select some abusive language benchmarks which contain a high frequency of swear words. We found four dataset collections, including HatEval [Basile et al., 2019], AMI@IberEval [Fersini et al., 2018d], AMI@Evalita [Fersini et al., 2018b], and Davidson [Davidson et al., 2017] datasets. These datasets contain a fairly high frequency of swear words. Around half of their instances are containing swear words, specifically 42.26%, 56.74%, 62.79%, and 69.2% for HatEval, AMI@Evalita, AMI@IberEval, and Davidson dataset respectively. Following is a short description of each dataset.

**HatEval Dataset.** The dataset focuses on the detection of hate speech in Twitter on two specific targets, namely immigrants and women, in a multilingual perspective [Basile et al., 2019]. The HatEval shared task introduced a dataset in two languages, English and

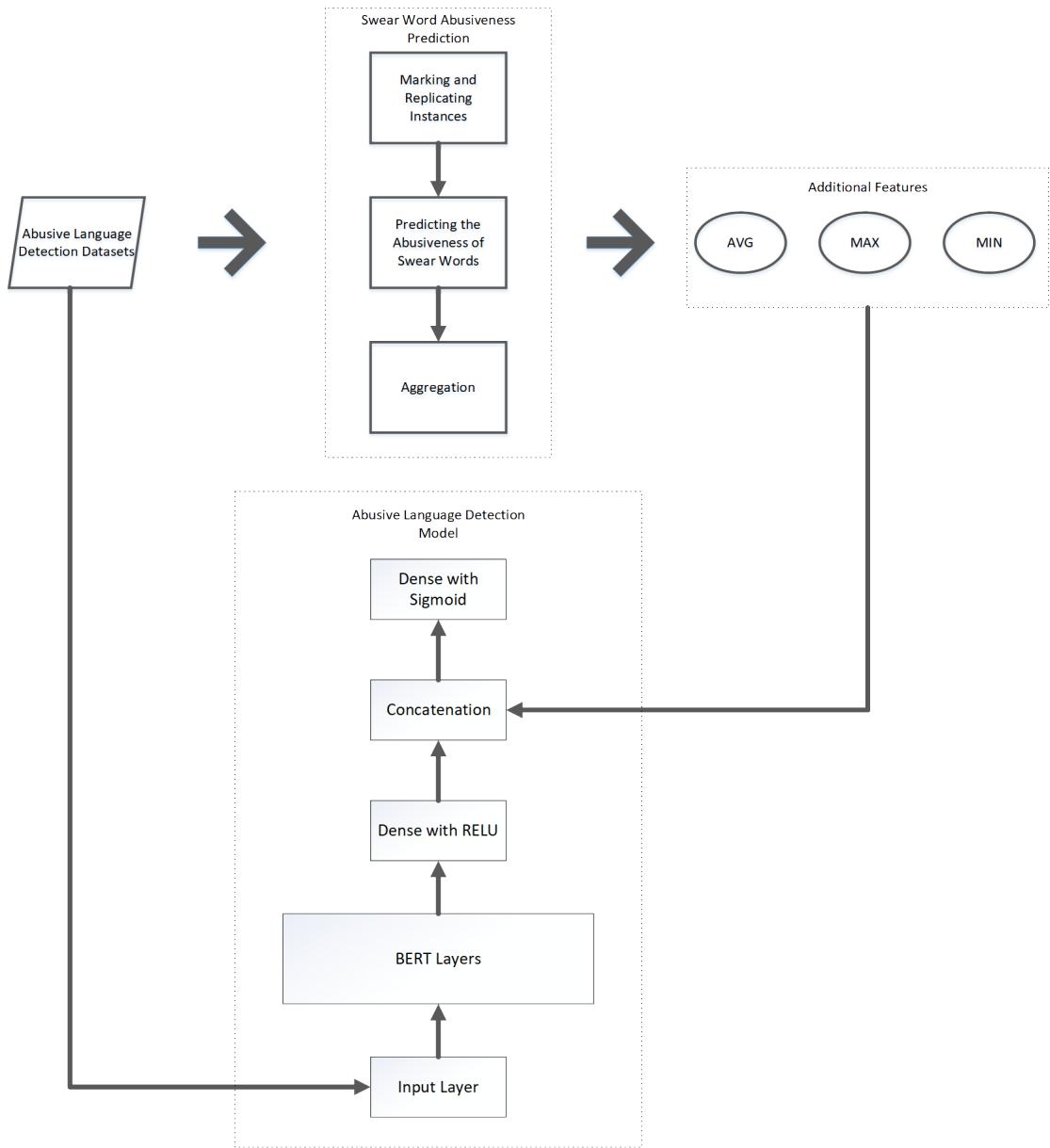


Figure 3.2: Process to Infuse Additional Features.

Spanish. However, we will only focus on the English collection. The HatEval collection was gathered by using several keywords, including neutral keywords, pejorative words towards targets, and highly polarized hashtags. This dataset was annotated by relying on judges from a crowdsourcing platform, which applied an annotation scheme including three binary labels: hate speech (hate speech or not), target range (generic or individual), and aggressiveness (aggressive or not). The final dataset used for the English HatEval shared task contains 13,000 (about 10,000 for training and for 3,000 testing).<sup>13</sup>

**AMI Datasets.** Basically, datasets for AMI@IberEval [Fersini et al., 2018d] and AMI@Evalita [Fersini et al., 2018b] were selected from the same collection of tweets, which were filtered using three approaches including querying from streaming API based on some keywords, monitoring account of online harassment victims, and downloading tweets from misogynist accounts. AMI@IberEval dataset contains 3,977 tweets (3,251 training and 831 testing), while AMI@Evalita collection contains 5,000 tweets (4,000 training and 1,000 testing). Originally, AMI dataset is available in three languages including English, Italian, and Spanish, but here we only focus on English.

**Davidson Dataset.** The dataset has been built by Davidson et al. [2017] and contains 24,783 tweets<sup>14</sup> manually annotated with crowdsourcing scenario. Differently from the other datasets considered, in this dataset a multilabel annotation is applied, with three labels including *hate speech*, *offensive*, and *neither*. These tweets were sampled from a collection of 85.4 million tweets gathered using the Twitter search API, focusing on tweets containing keywords from HateBase<sup>15</sup>. Only 5.8% of the total tweets were labeled as *hate speech* and 77.4% as *offensive*, while the remaining 16.8% were labelled as *not offensive*.

The second process is to predict the abusiveness of swear words in each instance of these datasets. We pre-processed all instances of these datasets similarly as we did to the SWAD (see Fig 3.1), including marking the swear word and replicating instance when more than one swear words are found. After a preprocessing step, we immediately predict all preprocessed instances by employing our best performing system based on results presented in the previous section, which is TM-BERT. We aggregate the prediction score for instances which contain more than one swear words by taking *minimum*, *maximum*, and *average* score. In case of instances which do not containing swear word, we set the prediction score to 0.

The final process is to infuse the swear word abusiveness prediction score into the base model for detecting abusive language in these respective tasks. However, note that this work aims not to produce the best possible system for these shared tasks but rather to test our hypothesis on the usefulness of predicting the pragmatics of swear word use. For this experiment, we employ a straightforward BERT model with a minimum hyperparameter

---

<sup>13</sup>Upon manual investigation, organizers decided to exclude 1,000 tweets from the English training set, 29 tweets from the English test set due to duplicated instances.

<sup>14</sup>Although in the original paper the authors mention that the dataset consists of 24,802 annotated tweets, we only found this number of instances in the shared GitHub repository: <https://github.com/t-davidson/hate-speech-and-offensive-language>

<sup>15</sup>A multilingual repository, which allows for the identification of HS terms by region: <https://hatebase.org>

tuning. We use (`bert-base-cased`) model available on TensorFlow-hub<sup>16</sup>, which allows us to integrate BERT with the Keras functional layer<sup>17</sup>. Our network starts with the BERT layer, which takes three inputs consisting of id, mask, and segment before passing into a dense layer with RELU activation (256 units) on top and an output layer with sigmoid activation. We train the network with the Adam optimizer with a learning rate of  $2^{-5}$ . We tune this model by trying several combinations of batch size (32, 64, 128) and the number of epochs (1-5). We infuse the additional feature by simply concatenating the swear word abusiveness probability into the dense layer after the BERT embedding layer.

### 3.4.2 Results

We apply standard evaluation metrics in this experiment, including a wide coverage of evaluation metrics such as precision, recall,  $F$ -score, and accuracy. We present precision, recall, and  $F$ -score on both positive and negative classes to picture the system performance better. Table 3.8, Table 3.9, Table 3.10, and Table 3.11 present the result of the experiments on HatEval task, AMI@Evalita task, AMI@IberEval, and Davidson dataset, respectively. As mentioned before that, we experiment with three additional features, namely MIN, MAX, and AVG. These additional features depict the approach in aggregating the abusiveness score when more than one swear words exists in the tweet. We marked with superscript (\*) results where the performance improvement is statistically significant compared to baseline models ( $F_{avg}$  and Acc columns).<sup>18</sup>

On the HatEval task, the additional feature was able to improve the model performance. The best result is obtained using the MIN score with .482 in the macro average  $F$ -score with a statistical significance compared to the baseline model. A similar result is observed on both AMI@Evalita and AMI@IberEval task datasets. All models infused by additional features are experiencing performance improvement significantly, where the best result was obtained by using the MIN aggregation score. The performance improvement is consistent in both classes, as observed from the  $F$ -score in the positive ( $F_1$ ) and negative ( $F_0$ ) class. However, a different result was observed in the experiment of the Davidson dataset as presented in Table 3.11. We found that the additional features were not able to augment the model performance.

It was an interesting finding that the MIN aggregation is recognized as the most effective approach on most datasets. Based on our further investigation, we found two possible reasons which lead to this result. First, we found several examples where two or more swear words were used in different abusiveness degrees within one tweet. As shown in the Example below, which is taken from AMI Evalita collection. Our model predicted the first swear word with a high abusiveness degree, while the second one with a low degree of abusiveness. With the MIN aggregation, the additional feature informs the model that there is an not-abusive swear word, which could become an important signal to resolve the context of the whole message. On the contrary, if we use MAX aggregation,

---

<sup>16</sup><https://www.tensorflow.org/hub>

<sup>17</sup><https://keras.io/>

<sup>18</sup>We used bootstrap sampling significance test tools publicly available at <https://github.com/fornaciari/boostsa>.

Model	$P_0$	$P_1$	$R_0$	$R_1$	$F_0$	$F_1$	$F_{avg}$	Acc
BERT	.695	.442	.225	.838	.340	.579	.459	.502
BERT + Features (MAX)	.708	.443	.227	.846	.343	.582	.462	.491
BERT + Features (MIN)	.690	.454	.261	.822	<b>.379</b>	<b>.585</b>	<b>.482*</b>	.513
BERT + Features (AVG)	.720	.436	.184	.885	.294	.584	.439	.485

Table 3.8: Result of Investigating Swear Words Role in HatEval Task

the additional feature could also deceive the model. In this case, MIN aggregation would provide a better knowledge for the model. Second, there are many instances of HatEval, AMI@Evalita, and AMI@IberEval, which contain more than one swear word. Therefore, aggregating the score in a better way would heavily influence the prediction result, where in this case MIN aggregation provides better information for the models.

[Example Not Misogyny tweet from AMI Evalita dataset]  
*everytime i reach the highlights of smut im reading me. ok ho\* calm down  
calm down sit your \*ss relax its just a smut*

Regarding to the peculiar result on Davidson dataset, we conducted a deeper investigation. We notice that our models struggle to detect the hate speech class as observed in Table 3.11, where the micro  $F$ -score in hate speech class was very low. Furthermore, our additional feature also failed to improve the model performance in determining the hate speech instances. Our manual inspection of the dataset highlights that our swear word abusiveness prediction model struggles to differentiate between the swear word in the offensive class and the hate speech class. For example, as shown in the examples below, we can see that our model predicts the swear word use in both classes with a high abusiveness degree. Even with human reasoning, we also could not differentiate the abusiveness degree of the swear words in both messages. We argue that this issue is the main reason for the less impact of our additional features in the Davidson dataset.

[Example Offensive tweet from Davidson’s dataset]  
*@USER @USER so you was in a female DMs talking to another n\*gga...  
You’re a f\*ggot...*

[Example Hate Speech tweet from Davidson’s dataset]  
*Vanessa is such a f\*ckin f\*ggot.*

### 3.5 Summary

In this chapter, we explore the phenomenon of swearing in Twitter conversations, by automatically predicting the *abusiveness* of a swear word in a tweet as the main investigation perspective. We developed the Twitter English corpus SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level. Our collection

Model	$P_0$	$P_1$	$R_0$	$R_1$	$F_0$	$F_1$	$F_{avg}$	Acc
BERT	.604	.635	.761	.458	.674	.532	.603	.606
BERT + Features (MAX)	.617	.664	.756	.478	.680	.555	.618	.637*
BERT + Features (MIN)	.627	.676	.762	.486	<b>.688</b>	<b>.565</b>	<b>.627*</b>	.636*
BERT + Features (AVG)	.587	.647	.764	.469	.664	.544	.604	.616

Table 3.9: Result of Investigating Swear Words Role in AMI Evalita Task

Model	$P_0$	$P_1$	$R_0$	$R_1$	$F_0$	$F_1$	$F_{avg}$	Acc
BERT	.701	.746	.869	.540	.776	.627	.701	.740
BERT + Features (MAX)	.734	.765	.904	.543	<b>.810</b>	.636	.723*	.747
BERT + Features (MIN)	.715	.785	.931	.555	.809	<b>.650</b>	<b>.730*</b>	.766*
BERT + Features (AVG)	.718	.765	.912	.542	.803	.634	.719	.773*

Table 3.10: Result of Investigating Swear Words Role in AMI IberEval Task

consists of 2,577 instances in total from two phases of manual annotation. We developed models to automatically predict abusive swearing, to provide an intrinsic evaluation of SWAD and confirm the robustness of the resource. We model this prediction task as three different tasks, namely sequence labeling, text classification, and target-based swear word abusiveness prediction. We experimentally found that our intention to model the task similarly to aspect-based sentiment analysis leads to promising results. Subsequently, we employ the classifier to improve the prediction of abusive language in several standard benchmarks. The results of our experiments show that additional abusiveness feature of the swear words is able to alleviate the false positive issue in the classification of abusive language, improving the performance across several benchmark datasets. All resources and source code developed in this work are publicly available on GitHub.<sup>19</sup>

<sup>19</sup><https://github.com/dadangewp/SWAD-Repository>

Model	$P_0$	$P_1$	$P_2$	$R_0$	$R_1$	$R_2$	$F_0$	$F_1$	$F_2$	$F_{avg}$	Acc
BERT	.428	.926	.719	.250	.925	.830	<b>.316</b>	<b>.925</b>	.771	<b>.671</b>	.869
BERT + Features (MAX)	.375	.911	.757	.172	.940	.792	.236	<b>.925</b>	<b>.774</b>	.645	.870
BERT + Features (MIN)	.392	.923	.719	.203	.927	.832	.267	<b>.925</b>	.771	.655	.868
BERT + Features (AVG)	.288	.912	.755	.193	.928	.782	.231	.920	.768	.640	.860

Table 3.11: Result of Investigating Swear Words Role in Davidson Dataset

## Chapter 4

# Multidomain/Multitarget Hate Speech Detection in Social Media

Abusive language and harassment are widespread in online communication, due to users' freedom and anonymity and the lack of regulation provided by social media platforms. Abusive language behaviour is multifaceted and available datasets are featured by different topical focuses and targets. Hate speech is one kind of abusive language that is topically-focused (misogyny, sexism, racism, xenophobia, homophobia, etc.) and each specific manifestation of hate speech targets different vulnerable groups based on characteristics such as gender (misogyny, sexism), ethnicity, race, religion (xenophobia, racism, Islamophobia), sexual orientation (homophobia), and so on. These characteristics make abusive language detection a domain-dependent task, and building a robust system to detect general abusive content is a challenge.

This chapter presents an investigation about building robust models to detect abusive language across different domains and targets. Specifically, this chapter includes two main investigations. First part focuses on experimenting with the detection of abusive language across different datasets. Second part focuses on the more specific hate speech phenomenon, by experimenting with the detection of hate speech across different topical focuses and also across different hate speech targets. This chapter is organized as follows. Section 4.1 introduces the motivation of this research study. Section 4.2 presents our exploration on cross-dataset experiments towards a broad abusive language detection task. Then, Section 4.3 further explores the experiment on the detection of more specific kind of abusive language, namely hate speech, across different topical focuses and targets. Meanwhile, Section 4.4 discusses several important findings based on our experimental results.

## 4.1 Motivation

In recent years several datasets have been proposed for abusive language detection, having different topical focuses and specific targets, e.g., misogyny, sexism, xenophobia, racism and so on (see again the illustration in Figure 1.1 and Figure 1.2). This diversity contributes to make the task to detect general abusive language difficult. Some studies attempted to bridge some of these subtasks by proposing cross-domain classification of abusive content [Wiegand et al., 2018a, Karan and Šnajder, 2018, Waseem et al., 2018]. To this end, many studies in the field exploited supervised approaches generally casting abusive language or hate speech detection as a binary classification problem (i.e., abusive/hateful vs. not abusive/not hateful) [Schmidt and Wiegand, 2017, Jurgens et al., 2019b, Fortuna and Nunes, 2018] relying on several manually annotated datasets that can be grouped into one of these categories:

- *Topic-generic* datasets, with a broad range of abusive language without limiting it to specific targets [Founta et al., 2018, Golbeck et al., 2017, Chatzakou et al., 2017]. For example, [Chatzakou et al., 2017] consider aggressive and bullying in their annotation scheme, while Founta et al. [2018] looks, in addition, for other expressions of online abuse such as offensive, abusive and hateful speech.
- *Topic-specific* datasets, where the abusive language category (racism, sexism, etc.) is known in advance (i.e., drives the data gathering process) and is often labelled. The abusive targets, either person-directed or group-directed, can be considered as *oriented*, containing, as they do, hateful content towards groups of targets or specific targets. For example, Waseem et al. [2017] sampled data for multiple targets, that is racism and sexism for, respectively, religious/ethnic minorities hate speech and sexual/gender (male and female) hate speech. Others focus on single targets including, for instance, sampling for the misogyny topic, targeting women [Fersini et al., 2018c,a, Chiril et al., 2020]. Similarly, for the xenophobia and racism topics the target are groups discriminated against on the grounds of ethnicity (e.g., immigrants [Basile et al., 2019], ethnic minorities [Waseem and Hovy, 2016, Tulkens et al., 2016], religious communities [Vidgen and Yasseri, 2020], Jewish communities [Zannettou et al., 2020], etc.).

Independently from the datasets that are used, all existing systems share two common characteristics. First, they are trained to predict the presence of general, target-independent abusive language, without addressing the problem of the variety of aspects related to both the topical focus and target-oriented nature of abusive. Second, systems are built, optimized, and evaluated based on a single dataset, one that is either topic-generic or topic-specific. The idea of cross-domain setting consists in using a one-to-one configuration by training a system on a given dataset and testing the system on another one, using domain adaptation techniques. Most existing works mapped between fine-grained schemes (that are specific for each dataset) and a unified set of tags, usually composed of a positive and negative label to account for the heterogeneity of labels

across datasets. Again, this binarization fails to discriminate among the multiple abusive targets. Thus, it has become difficult to measure the generalization power of such systems and, more specifically, their ability to adapt their predictions in the presence of novel or different topics and targets [Vidgen and Derczynski, 2020].

## 4.2 Cross-dataset Abusive Language Detection

In this section, we conduct an experiment to explore cross-domain abusive language classification in social media data, by proposing several machine learning models. We exploit several available Twitter datasets with different topical focuses, capturing different abusive phenomena. We also try to characterize the available datasets as capturing various phenomena related to abusive language, and investigate this characterization in cross-domain classification. Furthermore, we explore the use of a domain-independent lexicon of abusive words called *HurtLex* [Bassignana et al., 2018] to investigate its impact for transferring knowledge between different datasets.

### 4.2.1 Datasets

We consider four different publicly abusive language datasets and benchmark corpora in abusive language detection tasks. Table 4.1 summarizes the datasets’ characteristics. We binarize the label of these datasets into abusive (**bold**) and not-abusive. We split all datasets into training and testing by keeping the original split when provided, and splitting the distribution randomly (70% for training and 30% for testing) otherwise.

We also provide further information about the captured phenomena of every dataset. Based on this information, we can compare the nature and topical focus of the dataset, which potentially affect the cross-domain experimental results. Some datasets have a broader coverage than the others, focussing on more general phenomena, such as OffensEval [Zampieri et al., 2019b], and GermEval [Wiegand et al., 2018b]. However, there are also some shared phenomena between datasets, such as racism and sexism in Waseem [Waseem and Hovy, 2016] and HatEval [Basile et al., 2019]. AMI datasets contain the most specific phenomenon, only focusing on misogyny. The positive instance rate (PIR) denotes the ratio of abusive instances to all instances of the dataset.

### 4.2.2 Experimental Settings

In this experiment, we investigate the performance of machine learning classifiers which are trained on a particular dataset and tested on different datasets ones. We focus on investigating the influence of captured phenomena coverage between datasets. We hypothesize that a classifier which is trained on a broader coverage dataset and tested on narrower coverage dataset will give better performance than the opposite. Furthermore, we analyse the impact of using the HurtLex lexicon [Bassignana et al., 2018] to transfer knowledge between domains. HurtLex is a multilingual lexicon of hate words, originally built from 1,082 Italian hate words compiled in a manual fashion by the linguist Tullio De Mauro [De Mauro, 2016]. This lexicon is semi-automatically extended and translated

<b>Dataset</b>	<b>Label</b>	<b>Topical Focus</b>	<b>Train</b>	<b>Test</b>	<b>PIR</b>
Harassment <a href="#">Golbeck et al. [2017]</a>	<b>harassing</b> non-harassing	Harassing content, including racist and misogynistic contents, offensive profanities and threats	14,252	6,108	0.26
Waseem <a href="#">Waseem and Hovy [2016]</a>	<b>racism</b> <b>sexism</b> none	Racism and Sexism	11,542	4,947	0.31
OffensEval <a href="#">Zampieri et al. [2019b]</a>	<b>offensive</b> not offensive	Offensive content, including insults, threats, and posts containing profane language or swear words	13,240	860	0.33
HatEval <a href="#">Basile et al. [2019]</a>	<b>hateful</b> not hateful	Hate speech against women and immigrants	9,000	2,971	0.42

Table 4.1: Twitter abusive language datasets in four languages: original labels, language(s) featured, topical focus, distribution of train and test set and positive instance rate (PIR).

into 53 languages by using BabelNet [[Navigli and Ponzetto, 2012](#)], and the lexical items are divided into 17 categories such as homophobic slurs, ethnic slurs, genitalia, cognitive and physical disabilities, animals and more<sup>1</sup>.

**Model.** In this experiment, we employ two models. First, we exploit a simple traditional machine learning approach by using linear support vector classifier (LSVC) with unigram representation as a feature. Second, we utilize a long short-term memory (LSTM) neural model consisting of several layers, starting with a word embedding layer (32-dimensions) without any pre-trained model initialization<sup>2</sup>. This embedding layer is followed by LSTM networks (16-units), whose output is passed to a dense layer with ReLU activation function and dropout (0.4). The last section is a dense layer with sigmoid activation to produce the final prediction. We experiment with HurtLex by concatenating its 17 categories as one hot encoding representation to both LSVC-based and LSTM-based systems.

**Data and Evaluation** We use four English datasets, namely Harassment, Waseem, HatEval, and OffensEval<sup>3</sup>. We evaluate the system performance based on precision, recall, and *F*-score on the positive class (abusive class)<sup>4</sup>.

#### 4.2.3 Results and Analysis

Table 4.2 shows the results of the cross-domain experiment. We test every dataset with three systems which are trained on three other datasets. We also run in-domain scenario to compare the delta between in-domain and out-domain performance and measure the

<sup>1</sup><http://hatespeech.di.unito.it/resources.html>

<sup>2</sup>We experimented the use of pre-trained models (i.e. GloVe, word2vec, and FastText), but the result is lower compared to a self-trained model based on training set.

<sup>3</sup>AMI datasets are excluded due to the low number of instances.

<sup>4</sup>We used this metric in order to get a better insight on the interpretation of the results, since our objective is more focused on detecting the positive class

Dataset		LSVC + BoW				LSVC + BoW + HL			
Test	Train	P	R	$F_1$	$\Delta$	P	R	$F_1$	$\Delta$
Harassment	Waseem	.325	.233	.271	.103	.337	.264	.296	.079
	HatEval	.389	.119	.183	.191	.374	.116	.177	.198
	OffensEval	.320	.508	.393	-.019	.322	.516	.396	-.021
	Harassment	.547	.284	.374		.540	.288	.375	
Waseem	Harassment	.729	.022	.043	.688	.720	.034	.065	.669
	HatEval	.620	.109	.186	.545	.672	.113	.194	.540
	OffensEval	.461	.390	<b>.422</b>	.309	.453	.391	.420	.314
	Waseem	.817	.662	.731		.819	.665	.734	
HatEval	Harassment	.485	.181	.264	.339	.513	.229	.317	.290
	Waseem	.505	.490	.497	.106	.477	.558	.514	.093
	OffensEval	.450	.646	.531	.072	.451	.656	.534	.073
	HatEval	.449	.919	.603		.453	.919	.607	
OffensEval	Harassment	.301	.104	.155	.422	.321	.113	.167	.406
	Waseem	.440	.246	.316	.261	.462	.254	<b>.328</b>	.245
	HatEval	.372	.225	.281	.296	.381	.233	.289	.284
	OffensEval	.616	.542	.577		.626	.529	.573	
Dataset		LSTM + WE				LSTM + WE + HL			
Test	Train	P	R	$F_1$	$\Delta$	P	R	$F_1$	$\Delta$
Harassment	Waseem	.291	.467	.359	.033	.290	.524	<b>.373</b>	.045
	HatEval	.341	.308	.324	.068	.332	.379	<b>.354</b>	.064
	OffensEval	.333	.443	.380	.012	.314	.567	<b>.404</b>	.014
	Harassment	.510	.319	.392		.464	.380	.418	
Waseem	Harassment	.464	.111	.179	.587	.491	.149	<b>.229</b>	.520
	HatEval	.496	.213	.299	.467	.453	.318	<b>.374</b>	.375
	OffensEval	.444	.282	.345	.421	.419	.411	.415	.334
	Waseem	.760	.771	.766		.711	.790	.749	
HatEval	Harassment	.523	.308	.387	.216	.514	.394	<b>.446</b>	.158
	Waseem	.481	.636	<b>.548</b>	.055	.494	.609	.546	.058
	OffensEval	.452	.603	.516	.087	.457	.704	<b>.554</b>	.050
	HatEval	.444	.939	.603		.441	.955	.604	
OffensEval	Harassment	.525	.133	.213	.395	.406	.179	<b>.249</b>	.349
	Waseem	.403	.225	.289	.319	.400	.175	.244	.354
	HatEval	.392	.371	.381	.227	.371	.529	<b>.436</b>	.162
	OffensEval	.667	.558	.608		.551	.654	.598	

Table 4.2: Results on cross-domain abusive language identification (only in English).

drop in performance. Not surprisingly, the performance on out-domain datasets is always lower (except in two cases when the Harassment dataset is used as test set). Overall, LSTM-based systems performed better than LSVC-based systems. The use of HurtLex also succeeded in improving the performance on both LSVC-based and LSTM-based systems. We can see that HurtLex is able to improve the recall in most of the cases. Our further investigation shows that systems with HurtLex are able to detect more abusive contents, noted by the increases of true positives. The OffensEval training set always achieves the best performance when tested on three other datasets. On the other hand, the Harassment dataset always presents the larger drop in performance when used as training data. Training the models on the Harassment dataset lead to a very low result even in the in-domain setting. The highest result on the Harassment dataset is only .418  $F$ -score, achieved by LSTM with HurtLex <sup>5</sup>, while when trained on the other datasets our models are able to reach above .600  $F$ -score. Upon further investigation, we found, that [Golbeck et al. \[2017\]](#) only used a limited set of keywords, which contributes to limit their dataset coverage. Overall, we argue that there are good arguments in favor of our hypothesis that a system trained on datasets with a broader coverage of phenomena will be more robust to detect other kinds of abusive language (see the OffensEval results).

As described in Subsection 4.2.1, the datasets we considered have different focuses w.r.t. the abusive phenomena captured, and this impacts on the lexical distribution in each dataset. Based on a further analysis we observed that in datasets with a general topical focus such as OffensEval, the abusive tweets are marked by some common swear words such as “fuck”, “shit”, and “ass”. While in datasets featured by a specific hate target, such as the AMI dataset (misogyny), the lexical keywords in abusive tweets are dominated by specific sexist slurs such as “bitch”, “cunt”, and “whore”. This finding is consistent with the study of [ElSherief et al. \[2018a\]](#), which conducted an analysis on hate speech in social media based on its target. Furthermore, the pragmatics of swearing could also change from one dataset to another, depending on the topical features.

### 4.3 Cross-topic and Cross-target Hate Speech Detection

In this section, we explore two different objectives. First, we investigate the ability of hate speech detection models to capture common properties from generic hate speech datasets and to transfer this knowledge to recognize specific manifestations of hate. We propose several deep learning models and experiment with binary classification using two generic corpora. We evaluate their ability to detect hate speech in four topically focused datasets: sexism, misogyny, racism, and xenophobia (see again Figure 1.1).

Furthermore, we also experiment with the development of models for detecting both the topics (racism, xenophobia, sexism, misogyny) and the targets (gender, ethnicity) of hate speech going beyond standard binary classification. We aim to investigate (a) *how to detect hate speech at a finer level of granularity* and (b) *how to transfer knowledge across different types of hate speech*. We rely on multiple topic-specific datasets and develop, in

---

<sup>5</sup>[Marwa et al. \[2018\]](#) claimed to get a higher result, but that paper did not give a complete information about system configuration they used.

addition to the deep learning models designed to address the first challenge, a multitask architecture that has been shown to be quite effective in cross-domain sentiment analysis [Zhang et al., 2019, Cai and Wan, 2019]. We consider several experimental scenarios. First, ones where the topics/targets that will be classified in a multilabel fashion are present in the training data; and second, in cross-topic/target scenarios, where we try to predict a specific target/topic, training on data where that particular topic/target is unseen.

### 4.3.1 Datasets

We experiment with seven available hate speech corpora from previous studies among which two are topic-generic (Davidson [Davidson et al., 2017] and Founta [Founta et al., 2018]), and four are topic-specific about four different topics: *misogyny* (the AMI dataset collection from both IberEval [Fersini et al., 2018c] and Evalita [Fersini et al., 2018a]), *misogyny and xenophobia* (the HatEval dataset [Basile et al., 2019]), and *racism* and *sexism* (the Waseem dataset [Waseem and Hovy, 2016]). Each of these topics target either gender (sexism and misogyny) and/or ethnicity, religion or race (xenophobia and racism) (see again Figure 1.2).

In this section, we first detail the characteristics of each of the seven datasets, then provide general statistics.

- **Davidson.** The dataset has been built by Davidson et al. [2017] and contains 24,783 tweets<sup>6</sup> manually annotated with three labels including *hate speech*, *offensive*, and *neither*. These tweets were sampled from a collection of 85.4 million tweets gathered using the Twitter search API, focusing on tweets containing keywords from HateBase<sup>7</sup>. The dataset was manually labeled by using the CrowdFlower platforms<sup>8</sup>, where at least three annotators annotated each tweet. With an inter-annotator agreement of 92%, the final label for each instance was assigned according to a majority vote. Only 5.8% of the total tweets were labeled as *hate speech* and 77.4% as *offensive*, while the remaining 16.8% were labelled as *not offensive*.
- **Founta.** The dataset consists of 80,000 tweets<sup>9</sup> annotated with four mutually exclusive labels including *abusive*, *hateful*, *spam* and *normal* [Founta et al., 2018]. The original corpus of 30 millions tweets was collected from 30 March 2017 to 9 April 2017 by using the Twitter Stream API. For each tweet, the authors also extracted the meta-information and linguistic features in order to facilitate the

---

<sup>6</sup>Although in the original paper the authors mention that the dataset consists of 24,802 annotated tweets, we only found this number of instances in the shared GitHub repository: <https://github.com/t-davidson/hate-speech-and-offensive-language>

<sup>7</sup>A multilingual repository, which allows for the identification of hate speech terms by region: <https://hatebase.org>

<sup>8</sup>Now Figure Eight <https://www.figure-eight.com/>

<sup>9</sup>At the moment of collecting the data, from the original dataset <http://ow.ly/BqCf30jqffN> we were able to retrieve only 44,898 tweets, though in a recent shared task (<https://sites.google.com/view/icwsm2020datachallenge/home>) the full dataset was made available.

filtering and sampling process. Annotation was done by five crowdworkers and the final dataset was composed of 11% tweets labeled as *abusive* 7.5% as *hateful*, 59% as *normal*, and 22.5% as *spam*.

- **Waseem.** It consists of tweets collected over a period of two months by using representative keywords (common slurs) that target religious, sexual, gender and ethnic minorities [Waseem and Hovy, 2016]. The authors manually annotated the dataset with a third expert annotator reviewing their annotations. The final dataset consists of 16,914 tweets, with 3,383 instances from **Sexism<sub>Waseem</sub>** targeting gender minorities, 1,972 from **Racism<sub>Waseem</sub>** with racist instances , and 11,559 tweets that were judged to be neither sexist nor racist<sup>10</sup>.
- **AMI corpora.** The main goal of the AMI task consists in identifying tweets that convey hate or prejudice against women while categorizing forms of misogynous behavior (stereotype & objectification, dominance, derailing, sexual harassment & threats of violence, discredit), as well as classifying the target of a given instance (specific individual or a generic group). We use in this study the two AMI datasets: **IberEval** [Fersini et al., 2018c] containing 3,977 tweets collected over a period of four months (from 20th of July until 30th of November 2017) and **Evalita** [Fersini et al., 2018a] that comprises 5,000 tweets. Below are two examples of tweets annotated as misogyny taken respectively from **IberEval** and **Evalita**. Their associated misogynistic behavior are "sexual harassment" in the first example and "derailing" in the second.
- **HatEval.** The dataset consists of 13,000 tweets distributed across two different targets: immigrants and women [Basile et al., 2019]. Most of the tweets that target women were derived from the **AMI corpora**, while the remainder of the dataset was collected over a period of three months (from July to September 2018) by employing the same approaches as AMI. The dataset was annotated by using the Figure Eight crowdsourcing platform. In each instance, the annotators were asked to specify whether a tweet conveys hate speech or not towards any given targets. The annotators were also asked to indicate whether the author of the tweet was aggressive and to identify the target of the tweet (i.e., a specific individual or a group of people). Although the inter-annotator agreement obtained for each category (respectively 0.83, 0.73, and 0.70 respectively) was quite high, the final label was assigned based on a majority vote by adding two expert annotations to the crowd-annotated data. The final distribution of the dataset includes 13,000 tweets (6,500 for each target).

Table 4.3 provides a general overview of the datasets, along with the labels used in their annotation schemes. We can observe that the classes are imbalanced in most datasets, where the majority class is the negative class (non hate speech), except for the AMI collection (**AMI-IberEval** and **AMI-Evalita**) and **Davidson**.

---

<sup>10</sup>When collecting the data, we were able to retrieve only 16,488 instances (3,216 targeting gender minorities, 1,957 racist and 11,315 that were neither racist nor sexist).

Table 4.3: General overview of the datasets along with their topics and targets.

<b>Dataset</b>	<b>Labels</b>	<b># of instances</b>	<b>Topic</b>	<b>Target</b>
Davidson	hate speech	1,430	generic	none
	offensive	19,190		
	neither	4,163		
Founta	abusive	27,037	generic	none
	hateful	4,948		
	spam	14,024		
	normal	53,790		
Waseem	racism	1,957	specific	race gender
	sexism	3,216		
	none	11,315		
Evalita	misogyny	2,245	specific	women
	not misogyny	2,755		
IberEval	misogyny	1,851	specific	women
	not misogyny	2,126		
HatEval	immigrant	2,427	specific	women ethnicity
	women	2,608		
	not hate speech	6,936		

For our experiments, the corpora have been divided into train and test sets keeping the same tweet distribution as the original papers. This was done in order to make better comparisions with the state-of-the-art results<sup>11</sup>. Table 4.4 and Table 4.5 provide the distribution of instances in these two sets. As one of the research questions that we want to address involves the possibility of transferring knowledge from several topic-specific datasets into another topic-specific dataset where the topic is unseen, we decided to merge under the same topic (i.e., misogyny) both the `AMI corpora` and `HatEval dataset`<sup>12</sup>.

Table 4.4: Distribution of instances in topic-generic datasets (used as training).

<b>Dataset</b>	<b>Labels</b>	<b>N. of instances</b>	
Founta	hateful	1,930	39,700
	not-hateful	37,770	
Davidson	hateful	1,430	5,593
	not-hateful	4,163	

<sup>11</sup>The only difference with the original paper appears in the training set of the `HatEval` dataset as we found duplicate instances (already there in the `AMI corpora`).

<sup>12</sup>We recall that these two datasets used the same approach for collecting the data and for annotation guidelines.

Table 4.5: Distribution of instances in the train/test sets in topic-specific datasets.

Topic	Racism (Waseem)			Sexism (Waseem)		
	Racism	Non-racism	Total	Sexism	Non-sexism	Total
Train	1,346	7,943	9,289	2,253	7,943	10,196
Test	611	3,373	3,984	963	3,373	4,336
Topic	Misogyny (AMI corpora + HatEval)				Xenophobia (HatEval)	
	Misogyny	Non-misogyny	Total	Hateful	Non-hateful	Total
Train	Evalita	1,785	2,215	4,000		
	HatEval	1,305	1,396	2,701		
	IberEval	1,568	1,683	3,251	1,988	3,012
	Total	4,658	5,294	9,952		5,000
Test	Evalita	460	540	1,000		
	HatEval	623	849	1,472		
	IberEval	283	443	726	629	870
	Total	1,366	1,832	3,198		1,499

### 4.3.2 Generalizing Hate Speech Phenomena Across Multiple Datasets

In this subsection, we focus to investigate two objectives which articulate into two different questions including: i) *Are models able to capture common properties of hate speech and transfer this knowledge from topic-generic datasets to topic-specific datasets?* ii) *How do these models compare with ones that are trained on topic-specific datasets?* To this end, we propose the following two experiment configurations:

- $Top^G \rightarrow Top^S$ : Train on topic-general hate speech datasets (i.e., Davidson and Founta)<sup>13</sup> and test on *all* topic-specific datasets (i.e.,  $\text{Racism}_{\text{Waseem}}$ ,  $\text{Sexism}_{\text{Waseem}}$ ,  $\text{Misogyny}_{\text{Evalita}}$ ,  $\text{Misogyny}_{\text{IberEval}}$ ,  $\text{Misogyny}_{\text{HatEval}}$ , and  $\text{Xenophobia}_{\text{HatEval}}$ ) without splitting them into train/test.
- $Top^S \rightarrow Top^S$ : Train on the combined training sets of all topic-specific datasets (i.e., Waseem, HatEval, Evalita, and IberEval) and test on the test set of each topic-specific dataset.

These two configurations are cast as a binary classification task, where the system needs to predict whether a given tweet is hateful (1) or not (0). To this end, we experiment with several performing state of the art models for hate speech detection. This is a necessary first step in measuring to what extent existing models are capable of transferring knowledge across different hate speech datasets, be they topic-generic or topic-specific.

<sup>13</sup>We only use the *hateful* and *not-hateful* instances, although the data is annotated as *hate-speech*, *offensive* and *none* (for the Davidson dataset) and annotated as *hate-speech*, *abusive*, *normal* and *spam* (for the Founta dataset).

To deal with this experiment, we propose adopt several models as follows<sup>14</sup>:

- **Baseline.** This model is straight-forward based on a linear support vector classifier (LSVC). The use of linear kernel is based on [Joachims \[1998\]](#), who argue that the linear kernel has an advantage for text classification. They observe that text representation features are frequently linearly separable. Hereby, the baseline is an LSVC with unigrams, bigrams, and trigrams TF-IDF.
- **LSTM.** This model uses a LSTM network [[Hochreiter and Schmidhuber, 1997](#)] with an architecture consisting of several layers, starting with an embedding layer representing the input to the LSTM network (128 units), followed by a dense layer (64 units) with ReLU activation function. The final layer consists of a dense layer with sigmoid activation producing the final prediction. In order to get the best possible results, we optimized the batch size (16, 32, 64, 128) and the number of epochs (1-5). We used as input either randomly initialized embeddings (**LSTM**) or FastText<sup>15</sup> English word vectors with an embedding dimension of 300 [[Grave et al., 2018](#)] pre-trained on Wikipedia and Common Crawl (**LSTM<sub>FastText</sub>**). LSTM, a type of Recurrent Neural Network, has already been proven as a robust architecture in hate speech detection [[Badjatiya et al., 2017](#)].
- **CNN<sub>FastText</sub>.** This model was inspired by [Badjatiya et al. \[2017\]](#), [Gambäck and Sikdar \[2017\]](#). It uses FastText English word vectors (with the dimension of 300) and three 1D convolutional layers, each one using 100 filters and a stride of 1, but with different window sizes (respectively 2, 3, and 4) in order to capture different scales of correlation between words, with a ReLU activation function. We further downsample the output of these layers by a 1D max-pooling layer and we feed its output into the final dense layer. All the experiments run for a maximum of 100 epochs, with a patience of 10 and a batch size of 32<sup>16</sup>.
- **ELMo.** This model employs ELMo [[Peters et al., 2018](#)], a deep contextualized word representation, which shows a significant improvement in the study of hate speech [[Rizoiu et al., 2019](#)]. Since we implement ELMo as a Keras layer<sup>17</sup>, we were able to add more layers after the word embedding layer. The latter is followed by a dense layer (256 units) and a dropout rate of 0.1, before being passed to another dense layer (2 units) with a sigmoid activation function, which produces the final prediction. This architecture is fine-tuned based on the number of epochs (1-15) and batch-size (16, 32, 64, and 128), and optimized by using Adam optimizer.<sup>18</sup>

---

<sup>14</sup>In an exploratory attempt at finding the best way of representing the data, we included a standard pre-processing step (i.e., URLs and user mentions replacement with replacement tokens, RT removal) as well as emoji replacement with their detailed description [[Singh et al., 2019](#)]. However, the results were inconclusive.

<sup>15</sup><https://fasttext.cc/>

<sup>16</sup>All the hyperparameters were tuned on the validation set (20% of the training dataset), such that the best validation error was produced.

<sup>17</sup><https://keras.io/>

<sup>18</sup>We use the default parameter of Adam optimizer as described in <https://www.tensorflow.org/>

– **BERT**. This model uses the pre-trained BERT model (BERT-Base, Cased), [Devlin et al., 2019] on top of which we added an untrained layer of neurons. We then used the HuggingFace’s PyTorch implementation of BERT [Wolf et al., 2019] that we trained for three epochs with a learning rate of 2e-5 and AdamW optimizer. It is based on Swamy et al. [2019] where it achieved the best results for the task of abusive language detection.

### 4.3.3 Results of Generalizing Hate Speech Across Datasets

In this subsection, we provide the results of our effort to generalize hate speech phenomena by experimenting on the hate speech classification across available datasets. Table 4.6 and Table 4.7 present our results when training respectively on **Founta** and **Davidson**. We provide our results in terms of accuracy ( $A$ ), macro-averaged F-score ( $F_1$ ), precision ( $P$ ) and recall ( $R$ ) with the best results in terms of  $F_1$  presented in bold.

Table 4.6: Results for  $Top^G \rightarrow Top^S$  configuration when training on **Founta**.

Dataset	Baseline				LSTM				LSTM <sub>FastText</sub>			
	$P$	$R$	$F_1$	$A$	$P$	$R$	$F_1$	$A$	$P$	$R$	$F_1$	$A$
<b>Racism<sub>Waseem</sub></b>	0.680	0.601	0.638	0.850	0.613	0.533	0.570	0.842	0.666	0.585	0.623	0.846
<b>Sexism<sub>Waseem</sub></b>	0.555	0.516	0.534	0.760	0.585	0.517	0.549	0.771	0.624	0.543	0.581	0.773
<b>Xenophobia<sub>HatEval</sub></b>	0.632	0.542	0.583	0.622	0.602	0.507	0.550	0.601	0.589	0.509	0.546	0.601
<b>Misogyny<sub>Evalita</sub></b>	0.627	0.582	0.603	0.612	0.692	0.634	0.662	0.661	0.679	0.649	<b>0.664</b>	0.669
<b>Misogyny<sub>IberEval</sub></b>	0.622	0.569	0.594	0.592	0.669	0.610	0.638	0.630	0.662	0.625	0.643	0.641
<b>Misogyny<sub>HatEval</sub></b>	0.615	0.584	0.599	0.615	0.632	0.616	0.624	0.636	0.636	0.631	0.633	0.642
<b>Misogyny<sub>all</sub></b>	0.645	0.584	0.613	0.616	0.655	0.619	0.636	0.643	0.651	0.632	<b>0.641</b>	0.649
Dataset	CNN <sub>FastText</sub>				BERT				ELMo			
	$P$	$R$	$F_1$	$A$	$P$	$R$	$F_1$	$A$	$P$	$R$	$F_1$	$A$
<b>Racism<sub>Waseem</sub></b>	0.700	0.627	0.661	0.855	0.705	0.742	<b>0.723</b>	0.84	0.584	0.568	0.575	0.806
<b>Sexism<sub>Waseem</sub></b>	0.622	0.563	<b>0.591</b>	0.767	0.528	0.501	0.514	0.712	0.543	0.524	0.533	0.736
<b>Xenophobia<sub>HatEval</sub></b>	0.624	0.517	0.565	0.607	0.651	0.652	<b>0.651</b>	0.611	0.581	0.520	0.548	0.604
<b>Misogyny<sub>Evalita</sub></b>	0.649	0.612	0.629	0.637	0.651	0.659	0.654	0.663	0.635	0.608	0.621	0.630
<b>Misogyny<sub>IberEval</sub></b>	0.629	0.590	0.609	0.609	0.661	0.639	<b>0.649</b>	0.661	0.602	0.571	0.586	0.590
<b>Misogyny<sub>HatEval</sub></b>	0.609	0.595	0.601	0.616	0.632	0.637	<b>0.634</b>	0.639	0.620	0.602	0.610	0.625
<b>Misogyny<sub>all</sub></b>	0.628	0.615	0.621	0.630	0.643	0.637	0.639	0.647	0.627	0.597	0.612	0.621

We recall here that we focus on learning topic-generic hate speech properties and test how neural models are able to extrapolate this information in order to detect topic-specific hate speech. The results show that **ELMo** outperformed other models in the **Waseem** dataset (**Racism<sub>Waseem</sub>**, **Sexism<sub>Waseem</sub>**) when trained on **Davidson**. When trained on **Founta**, **CNN<sub>FastText</sub>** obtained the best results for **Sexism<sub>Waseem</sub>** and **BERT** for **Racism<sub>Waseem</sub>**. For most of the topic-specific testing datasets (**AMI corpora** in particular), the results are comparable across the two general hate speech training datasets (**Davidson** and **Founta**), with higher disparities being observed in the **Waseem** results.

Table 4.8 presents the results obtained when focusing on learning topic-specific hate speech properties by combining all training sets of all datasets. The overall picture of the results shows that our baseline (i.e., **LSVC**) performed quite well when compared

Table 4.7: Results for  $Top^G \rightarrow Top^S$  configuration when training on Davidson.

Dataset	Baseline				ELMo				LSTM			
	P	R	$F_1$	A	P	R	$F_1$	A	P	R	$F_1$	A
<b>Racism</b> <sub>Waseem</sub>	0.585	0.560	0.572	0.814	0.665	0.661	<b>0.663</b>	0.833	0.573	0.535	0.553	0.852
<b>Sexism</b> <sub>Waseem</sub>	0.558	0.528	0.542	0.747	0.628	0.586	<b>0.606</b>	0.761	0.574	0.526	0.549	0.761
<b>Xenophobia</b> <sub>HatEval</sub>	0.601	0.541	0.569	0.615	0.616	0.544	0.577	0.620	0.604	0.517	0.557	0.605
<b>Misogyny</b> <sub>Evalita</sub>	0.668	0.666	0.667	0.672	0.623	0.624	0.624	0.626	0.680	0.681	<b>0.680</b>	0.682
<b>Misogyny</b> <sub>IberEval</sub>	0.638	0.633	0.635	0.639	0.632	0.631	0.631	0.635	0.678	0.676	<b>0.677</b>	0.680
<b>Misogyny</b> <sub>HatEval</sub>	0.635	0.636	0.635	0.630	0.621	0.622	0.621	0.619	0.638	0.636	0.637	0.623
<b>Misogyny</b> <sub>all</sub>	0.653	0.654	0.654	0.657	0.623	0.617	0.620	0.628	0.657	0.658	0.657	0.656

Dataset	LSTM <sub>FastText</sub>				CNN <sub>FastText</sub>				BERT			
	P	R	$F_1$	A	P	R	$F_1$	A	P	R	$F_1$	A
<b>Racism</b> <sub>Waseem</sub>	0.613	0.656	0.634	0.775	0.622	0.617	0.619	0.812	0.605	0.561	0.582	0.819
<b>Sexism</b> <sub>Waseem</sub>	0.544	0.540	0.542	0.699	0.586	0.557	0.571	0.744	0.544	0.531	0.537	0.741
<b>Xenophobia</b> <sub>HatEval</sub>	0.635	0.547	0.588	0.624	0.641	0.551	<b>0.592</b>	0.628	0.635	0.527	0.575	0.607
<b>Misogyny</b> <sub>Evalita</sub>	0.635	0.620	0.627	0.602	0.652	0.653	0.652	0.652	0.676	0.678	0.677	0.673
<b>Misogyny</b> <sub>IberEval</sub>	0.649	0.635	0.643	0.623	0.653	0.653	0.653	0.654	0.663	0.661	0.662	0.661
<b>Misogyny</b> <sub>HatEval</sub>	0.619	0.593	0.606	0.562	0.659	0.647	<b>0.652</b>	0.626	0.639	0.644	0.641	0.624
<b>Misogyny</b> <sub>all</sub>	0.633	0.614	0.623	0.594	0.658	0.657	<b>0.658</b>	0.648	0.654	0.654	0.654	0.649

to other models: it presents a decrease of anywhere in between 1% and 11% in terms of  $F1$  score, when compared to the best-performing models for a specific topic. For most topics, the best results were obtained by **BERT**, with the only exception being for the **Misogyny**<sub>HatEval</sub> dataset, where **ELMo** obtained the best results (with a difference of almost 2% in terms of  $F1$  score). We note that **Misogyny**<sub>HatEval</sub> is the only dataset for which **ELMo** achieved good results. For all the other datasets, the results are low, even lower than the baseline<sup>19</sup>. We also note that state of the art models achieved good results for both topics in the **Waseem** dataset, whereas they attain lower results when tested on the xenophobia topic from the **HatEval** dataset. However, our results are similar to the ones obtained by state-of-the-art baselines for **Waseem** ( $F1=0.739$  [Waseem and Hovy, 2016]) and **HatEval** ( $F1=0.451$  [Basile et al., 2019])<sup>20</sup>.

In order to assess whether training on topic-specific data improves the results beyond those achieved by training on topic-generic data, we compare our results with both the baselines and the best-submitted systems in the shared task competition where these data has been used (only available for AMI corpora). The comparison was made by training either on a topic-general dataset (i.e.,  $Top^G \rightarrow Top^S$ ) or on all topic-specific datasets (i.e.,  $Top^S \rightarrow Top^S$ ), and testing the test data provided by the organizers of AMI-IberEval and AMI-Evalita. Table 4.9 shows our results.

When compared to the AMI **Misogyny**<sub>Evalita</sub> and **Misogyny**<sub>IberEval</sub> baselines<sup>21</sup> provided in terms of accuracy (respectively 0.605 and 0.783), we observe that using a topic-specific training approach, **BERT** achieved more than a 10% increase for both datasets, while for the topic-generic training approach the only improvement of (0.5%) is

<sup>19</sup>The baseline achieved better results in all datasets, except the topics in the **HatEval** dataset.

<sup>20</sup>The baseline for the **Waseem** dataset is a LR coupled with character n-grams and the gender information of the tweet author, while the baseline for the **HatEval** shared task is a straightforward SVM with TF-IDF features.

<sup>21</sup>SVM with linear kernel trained on the unigram representation of the tweets.

Table 4.8: Results for  $Top^S \rightarrow Top^S$  when training on Waseem, HatEval and AMI train sets.

Dataset	Baseline				LSTM				LSTM <sub>FastText</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
Racism <sub>Waseem</sub>	0.786	0.798	0.792	0.889	0.796	0.765	0.779	0.878	0.783	0.783	0.783	0.887
Sexism <sub>Waseem</sub>	0.815	0.790	0.801	0.868	0.787	0.795	0.791	0.857	0.758	0.807	0.775	0.855
Xenophobia <sub>HatEval</sub>	0.572	0.546	0.470	0.497	0.530	0.560	0.427	0.471	0.546	0.589	0.447	0.488
Misogyny <sub>Evalita</sub>	0.645	0.646	0.645	0.646	0.652	0.652	0.648	0.648	0.661	0.660	0.657	0.658
Misogyny <sub>IberEval</sub>	0.803	0.732	0.742	0.778	0.709	0.754	0.717	0.750	0.739	0.793	0.749	0.779
Misogyny <sub>HatEval</sub>	0.659	0.551	0.421	0.487	0.613	0.688	0.534	0.561	0.564	0.665	0.447	0.502
Misogyny <sub>all</sub>	0.630	0.624	0.601	0.602	0.650	0.654	0.631	0.631	0.636	0.644	0.612	0.614
Dataset	CNN <sub>FastText</sub>				BERT				ELMo			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
Racism <sub>Waseem</sub>	0.764	0.800	0.782	0.827	0.775	0.844	<b>0.802</b>	0.884	0.616	0.833	0.651	0.874
Sexism <sub>Waseem</sub>	0.793	0.798	0.795	0.816	0.807	0.829	<b>0.817</b>	0.869	0.589	0.815	0.599	0.810
Xenophobia <sub>HatEval</sub>	0.492	0.471	0.481	0.462	0.619	0.543	<b>0.578</b>	0.577	0.562	0.596	0.543	0.609
Misogyny <sub>Evalita</sub>	0.673	0.684	0.678	0.684	0.704	0.705	<b>0.704</b>	0.706	0.562	0.672	0.496	0.594
Misogyny <sub>IberEval</sub>	0.713	0.742	0.727	0.735	0.841	0.840	<b>0.840</b>	0.848	0.538	0.774	0.460	0.639
Misogyny <sub>HatEval</sub>	0.603	0.532	0.565	0.553	0.694	0.523	0.596	0.573	0.618	0.643	<b>0.615</b>	0.649
Misogyny <sub>all</sub>	0.671	0.640	0.655	0.651	0.703	0.697	<b>0.676</b>	0.677	0.583	0.646	0.557	0.630

Table 4.9: Comparison with related work in terms of accuracy.

System	Misogyny <sub>Evalita</sub>		Misogyny <sub>IberEval</sub>	
	A	A	A	A
Competition Baseline	0.605		0.783	
Competition Best System	0.704		<b>0.913</b>	
Best $Top^G$ (Founta) $\rightarrow Top^S$ (ELMo/BERT)	0.597		0.697	
Best $Top^G$ (Davidson) $\rightarrow Top^S$ (BERT/ELMo)	0.610		0.658	
Best $Top^S$ (all) $\rightarrow Top^S$ (BERT)	<b>0.706</b>		0.848	

brought by **BERT** trained on the Davidson dataset (for Misogyny<sub>Evalita</sub>). When comparing the results with the best-submitted systems (0.704 and 0.913<sup>22</sup>) we still observe a small improvement achieved by **BERT** trained on topic-specific data for the Misogyny<sub>Evalita</sub> task, though all the other system results were lower. These results confirm that a model trained with a combination of several datasets with different topical focuses is more robust than a model trained on a topic-generic dataset.

#### 4.3.4 Multitarget Hate Speech Detection

Based on the experimental result presented in the previous section, we observed that the topic-generic datasets are not adequate for capturing specific instances of hate speech using state of the art hate speech detection models. Therefore, this subsection will focus

<sup>22</sup>The best-submitted system for the AMI Evalita competition is an LR with a vector representation that concatenates sentence embedding, TF-IDF and average word embeddings, while for the AMI IberEval competition it was an SVM with a combination of structural, stylistic and lexical features.

to evaluate how topically focused datasets can be used to detect hate speech across different hate speech targets. Our objective is to answer two main questions including: i). *Is combining topic-specific datasets better for predicting hate speech towards a given seen topic/target?;* ii). *What happens when the models are tested on a topic-specific dataset where the topic and/or the target are unseen?* Let  $T$  be either a topic ( $Top$ ) or a target ( $Tag$ ). We propose the following experiment configurations:

- $T^S \rightarrow T_{seen}^S$ : We model the task as a multilabel classification problem with two sub-configurations:
  - (a)  $Top^S \rightarrow Top_{seen}^S$ : Detect the harmfulness of a given tweet and the topic to which the hate speech belongs. Each tweet is thus classified into eight different classes, representing the combination of the four topics (racism, sexism, misogyny, xenophobia) and two hate speech classes (hate speech vs. non hate speech). As in the previous experiments (cf. Section 4.3.2), we combine all the training sets of the topic-specific datasets for training. Then, all the models are tested on the test set of each topic-specific datasets.
  - (b)  $Tag^S \rightarrow Tag_{seen}^S$ : It is similar to (a), except that it concerns the multilabel classification of targets. Therefore, we merge topic-specific train and test sets that share the same target (i.e. *women*:  $\text{Sexism}_{\text{Waseem}}$  and  $\text{Misogyny}_{\text{all}}$  and *ethnicity*:  $\text{Racism}_{\text{Waseem}}$  and  $\text{Xenophobia}_{\text{HatEval}}$ ).
- $T^S \rightarrow T_{unseen}^S$ : We model the task as a binary classification task to predict the topic/target not previously seen during training time. We also design two experiments here:
  - (c)  $Top^S \rightarrow Top_{unseen}^S$ : It uses three out of the four topic datasets for training and the remaining topic dataset for testing (i.e., the dataset left out at training time). For example, to detect the harmfulness of misogynistic messages, we train on the following topics: racism ( $\text{Racism}_{\text{Waseem}}$ ), sexism ( $\text{Sexism}_{\text{Waseem}}$ ) and xenophobia ( $\text{Xenophobia}_{\text{HatEval}}$ ), then we test on the misogyny topic (i.e., comprising AMI corpora and  $\text{Misogyny}_{\text{HatEval}}$ ).
  - (d)  $Tag^S \rightarrow Tag_{unseen}^S$ : It is similar to (c), except that it concerns targets. For example, to detect the hateful messages that target women, we train by using the datasets related to the target race (i.e.,  $\text{Racism}_{\text{Waseem}}$  and  $\text{Xenophobia}_{\text{HatEval}}$ ) and test on the four datasets related to the target *women* (i.e.,  $\text{Sexism}_{\text{Waseem}}$ , the two AMI corpora and  $\text{Misogyny}_{\text{HatEval}}$ ).

Both  $T^S \rightarrow T_{seen}^S$  (multilabel classification) and  $T^S \rightarrow T_{unseen}^S$  (binary classification) rely on the six models presented in Section 4.3.2 (i.e., **LSVC**, **LSTM**, **LSTM<sub>FastText</sub>**, **CNN<sub>FastText</sub>**, **ELMo**, and **BERT**). In addition, for  $T^S \rightarrow T_{seen}^S$  we propose a multitask setting that consists of two classifiers that are trained jointly by multitask objectives. The first classifier predicts whether the tweet is hateful or not (0 and 1), while the second one

the topic of hate speech (racism (0), sexism (1), misogyny (2), and xenophobia (3)). The final label prediction is broken down into eight classes (cf. Table 4.10). The multitask systems are compared to the previous six models used here as strong baselines.

Table 4.10: Label combination in multitask setting.

Target Label	Hate Speech Label	Final Label
Racism (0)	Not Hate Speech (0)	Not Racism (0)
	Hate Speech (1)	Racism (1)
Sexism (1)	Not Hate Speech (0)	Not Sexism (2)
	Hate Speech (1)	Sexism (3)
Misogyny (2)	Not Hate Speech (0)	Not Misogyny (4)
	Hate Speech (1)	Misogyny (5)
Xenophobia (3)	Not Hate Speech (0)	Not Hate Speech towards immigrants (6)
	Hate Speech (1)	Hate Speech towards immigrants (7)

MTL has already been successfully applied in cross-domain aspect-based sentiment analysis and is used here for the first time in an hate speech detection task, making a parallel between the sentiment domain (e.g., restaurant, book, hotel, etc.) and the topic/target of hate speech. Indeed, the main problem in sentiment analysis is the big performance decline in the out-domain setting (when a system is trained and tested with different dataset domains) compared to the in-domain setting (when a system is trained and tested on dataset within the same domain). Similar challenges also arise in the abusive language detection task, where a system is struggling to obtain a robust performance when trained and tested with different datasets. These usually have different focuses on the phenomena they want to capture.

To investigate this objective, we experiment with state of the art models (i.e., **L SVC**, **LSTM**, **LSTM<sub>FastText</sub>**, **CNN<sub>FastText</sub>**, **ELMo**, and **BERT**, as described in Section 4.3.2) and extend them with a multitask architecture, as described below:

–**LSTM<sub>multitask</sub>**. First, we investigate successful approaches in multidomain sentiment analysis, a research area that is more mature in dealing with multidomain classification. For example, Liu et al. [2018c] used Bi-LSTM networks with adversarial training [Ganin and Lempitsky, 2015, Goodfellow et al., 2014] for learning general representation from all domains data. Peng et al. [2018] proposed a co-training approach for jointly learning the representation from both domain-invariant and domain-specific representations, while Zhang et al. [2019], Cai and Wan [2019] adopted a MTL approach. Among existing models, we decided to re-implement the system proposed in Cai and Wan [2019], as it has been shown to outperform existing models in one of the most used multidomain sentiment classification benchmark dataset [Liu et al., 2017]. This system consists of two Bi-LSTM classifiers, each of them classifying the domain (domain classifier) and the

sentiment (sentiment classifier) of the tweets at the same time, with the loss of both tasks being added up. The output of the Bi-LSTM domain classifier is concatenated to the word embedding layer of the sentiment classifier to acquire a domain-aware representation. Then, the output of average pooling (after Bi-LSTMs) of the domain classifier is also concatenated to the sentiment classifier to obtain domain-aware attention.

We extend the architecture proposed in [Cai and Wan, 2019]. The first Bi-LSTM predicts whether a given tweet is hateful or not, while the second one predicts the topic/target of hate speech. In this way, we obtain both topic/target-aware representation and topic/target-aware attention when predicting whether the tweet is hateful or not. For experiments, we fine-tune this model by varying the number of epochs (1-15) and batch-sizes (16, 32, 64, and 128) while keeping the same configurations as in [Cai and Wan, 2019]. The model input is either embeddings randomly initialized (**LSTM<sub>multitask</sub>**) or FastText pre-trained embeddings, (**LSTM<sub>multitask</sub>** (**FastText**))<sup>23</sup>.

**-ELMo<sub>multitask</sub>.** We also modify our **ELMo** system (cf. Section 4.3.2) in order to be able to use it in multitask setting. Therefore, we built two ELMo-based architectures to predict the hatefulness and topic/target of tweets. Each architecture starts with the ELMo embedding layer, followed by a dense layer with a ReLU activation function, before being passed into another dense layer with a sigmoid activation function to produce the final prediction. Since ELMo embeddings are not trainable, we could not get the topic/target-aware representation as in the previous Bi-LSTMs model. We can only transfer knowledge by concatenating the output of the first dense layer of the topic/target classifier to the dense layer of the hateful classifier. In this way, we expect to get meaningful information about the topic/target to classify the hatefulness of tweets. Again, we only tune the systems by optimizing the number of epochs and batch-sizes.

**-BERT<sub>multitask</sub>.** This model is similar to Liu et al. [2019], where all tasks share and update the same low layers (i.e., **BERT** layers), except for the task-specific classification layer. In this architecture, after transferring the text to contextual embeddings in the shared layers and retrieving the first token hidden state of the shared **BERT** model, we apply a dropout of 0.1 and connect it to two different layers (corresponding to the two classification tasks: topic/target and hatefulness). To preserve individual task-specific loss functions and to perform training at the same time, we defined the losses for the two tasks separately and optimized them jointly (by backpropagating their sum through the model). This model was trained for three epochs with a learning rate of 2e-5 and AdamW optimizer.

#### 4.3.5 Results on Multitarget Hate Speech Detection

Table 4.11 and Table 4.12 present the results obtained in the  $Top^S \rightarrow Top_{seen}^S$  configuration in which the testing topic was previously seen during training. Table 4.11 presents the baseline results while Table 4.12 the multitask results. We can observe

---

<sup>23</sup>GloVe used in the original paper gives lower results.

that multitask models are the best, outperforming all the baselines, the best systems being **LSTM<sub>multitask</sub>** (**FastText**) and **BERT<sub>multitask</sub>**. The results obtained on the **Waseem** dataset surpass all the others, which could be a consequence of the higher number of instances in this particular dataset when compared to the others. Overall, the best performance for the multi-topic hate speech detection task is achieved by **BERT<sub>multitask</sub>**, which attains the best result in eight out of nine test datasets.

Table 4.11: Baseline results for  $\text{Top}^S \rightarrow \text{Top}_{\text{seen}}^S$ .

Dataset	LSVC				LSTM				<b>LSTM<sub>FastText</sub></b>			
	P	R	$F_1$	A	P	R	$F_1$	A	P	R	$F_1$	A
<b>Racism<sub>Waseem</sub></b>	0.701	0.844	0.766	0.610	0.841	0.827	0.834	0.856	0.816	0.856	<b>0.835</b>	0.855
<b>Sexism<sub>Waseem</sub></b>	0.694	0.852	0.765	0.545	0.781	0.859	0.818	0.827	0.782	0.869	<b>0.826</b>	0.832
<b>Xenophobia<sub>HatEval</sub></b>	0.474	0.544	0.507	0.404	0.459	0.601	0.521	0.387	0.496	0.651	0.563	0.421
<b>Misogyny<sub>Evalita</sub></b>	0.614	0.653	0.633	0.612	0.598	0.657	0.626	0.599	0.609	0.661	0.634	0.604
<b>Misogyny<sub>IberEval</sub></b>	0.642	0.841	0.728	0.643	0.504	0.716	0.592	0.502	0.607	0.782	0.684	0.582
<b>Misogyny<sub>HatEval</sub></b>	0.518	0.578	0.546	0.452	0.595	0.644	0.618	0.551	0.536	0.662	0.592	0.468
<b>Misogyny<sub>all</sub></b>	0.576	0.638	0.605	0.545	0.574	0.638	0.604	0.555	0.573	0.645	0.607	0.536
Dataset	<b>CNN<sub>FastText</sub></b>				<b>BERT</b>				<b>ELMo</b>			
	P	R	$F_1$	A	P	R	$F_1$	A	P	R	$F_1$	A
<b>Racism<sub>Waseem</sub></b>	0.703	0.754	0.727	0.855	0.847	0.597	0.701	0.791	0.819	0.840	0.829	0.859
<b>Sexism<sub>Waseem</sub></b>	0.841	0.810	0.825	0.826	0.876	0.666	0.757	0.812	0.675	0.854	0.754	0.788
<b>Xenophobia<sub>HatEval</sub></b>	0.532	0.491	0.510	0.422	0.667	0.527	<b>0.588</b>	0.516	0.356	0.567	0.437	0.312
<b>Misogyny<sub>Evalita</sub></b>	0.653	0.586	0.618	0.595	0.723	0.672	<b>0.697</b>	0.670	0.427	0.650	0.516	0.431
<b>Misogyny<sub>IberEval</sub></b>	0.865	0.725	0.788	0.724	0.857	0.783	<b>0.818</b>	0.780	0.484	0.738	0.585	0.531
<b>Misogyny<sub>HatEval</sub></b>	0.602	0.563	0.582	0.505	0.681	0.581	<b>0.627</b>	0.632	0.529	0.624	0.573	0.488
<b>Misogyny<sub>all</sub></b>	0.656	0.612	0.633	0.643	0.702	0.654	<b>0.677</b>	0.657	0.488	0.634	0.551	0.479

Table 4.13 presents the results obtained for the  $\text{Tag}^S \rightarrow \text{Tag}_{\text{seen}}^S$  experiments in which the testing target was previously seen during training. The best result for the target women was obtained by **CNN<sub>FastText</sub>**, while for the target race **LSTM<sub>multitask</sub>** (**FastText**) outperformed all the other models. Our results confirm our assumption that the multitask approach is capable of a robust performance in a multi-topic experiment, proving its ability in transferring knowledge between different topics, as reported in previous cross-domain sentiment analysis studies.

We begin by presenting the results in the  $\text{Top}^S \rightarrow \text{Top}_{\text{unseen}}^S$  experiments in which the testing topic was unseen during training. As shown in Table 4.14, we observe that in the absence of data annotated for a specific type of hate speech, one can use (already existing) annotated data for different kinds of hate speech.

As this experiment is cast as a binary classification task, we compare the results with the ones presented in Table 4.8 that concern  $\text{Top}^S \rightarrow \text{Top}^S$  when training on **Waseem**, **HatEval** and **AMI** train sets and where topics are seen in the test sets. We noticed that **CNN<sub>FastText</sub>** was able to achieve a similar performance for the topic misogyny (0.655 in both  $\text{Top}^S \rightarrow \text{Top}_{\text{unseen}}^S$  and  $\text{Top}^S \rightarrow \text{Top}^S$ ), improving almost 2% for the target xenophobia (moving from 0.578 in  $\text{Top}^S \rightarrow \text{Top}^S$  with **BERT** to 0.595 in terms of  $F_1$ ). However, lower results were obtained for the **Waseem** dataset, where the drop in terms of  $F_1$  is between 15% and 20%. The overall results also show that **CNN<sub>FastText</sub>** was

Table 4.12: Multitask results for  $Top^S \rightarrow Top_{seen}^S$ .

Dataset	LSTM <sub>multitask</sub>				LSTM <sub>multitask (FastText)</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>Racism</b> <sub>Waseem</sub>	0.787	0.851	0.818	0.877	0.839	0.811	0.825	0.828
<b>Sexism</b> <sub>Waseem</sub>	0.774	0.867	<b>0.818</b>	0.848	0.763	0.842	0.801	0.797
<b>Xenophobia</b> <sub>HatEval</sub>	0.475	0.534	0.503	0.407	0.495	0.621	0.551	0.422
<b>Misogyny</b> <sub>Evalita</sub>	0.573	0.639	0.604	0.560	0.621	0.687	<b>0.653</b>	0.605
<b>Misogyny</b> <sub>IberEval</sub>	0.556	0.774	<b>0.647</b>	0.542	0.644	0.792	<b>0.710</b>	0.621
<b>Misogyny</b> <sub>HatEval</sub>	0.551	0.650	0.597	0.489	0.554	0.682	<b>0.612</b>	0.489
<b>Misogyny</b> <sub>all</sub>	0.560	0.651	0.602	0.523	0.597	0.684	<b>0.637</b>	0.555

Dataset	ELMo <sub>multitask</sub>				BERT <sub>multitask</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>Racism</b> <sub>Waseem</sub>	0.677	0.862	0.758	0.827	0.835	0.667	<b>0.742</b>	0.865
<b>Sexism</b> <sub>Waseem</sub>	0.599	0.862	0.707	0.764	0.870	0.703	<b>0.777</b>	0.874
<b>Xenophobia</b> <sub>HatEval</sub>	0.356	0.617	<b>0.451</b>	0.340	0.650	0.585	<b>0.616</b>	0.513
<b>Misogyny</b> <sub>Evalita</sub>	0.457	0.594	<b>0.517</b>	0.472	0.725	0.65	<b>0.704</b>	0.684
<b>Misogyny</b> <sub>IberEval</sub>	0.479	0.714	0.573	0.541	0.865	0.774	0.817	0.774
<b>Misogyny</b> <sub>HatEval</sub>	0.580	0.615	<b>0.597</b>	0.580	0.701	0.598	<b>0.646</b>	0.642
<b>Misogyny</b> <sub>all</sub>	0.520	0.613	<b>0.563</b>	0.538	0.721	0.648	<b>0.682</b>	0.683

 Table 4.13: Baselines and multitask results for  $Tag^S \rightarrow Tag_{seen}^S$ .

System	WOMEN				ETHNICITY			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>LSVC</b>	0.530	0.704	0.605	0.431	0.548	0.632	0.587	0.457
<b>LSTM</b>	0.678	0.713	0.695	0.711	0.650	0.608	0.628	0.728
<b>LSTM</b> <sub>FastText</sub>	0.677	0.721	0.698	0.707	0.656	0.621	0.638	0.737
<b>CNN</b> <sub>FastText</sub>	0.732	0.716	<b>0.724</b>	0.731	0.580	0.435	0.497	0.613
<b>BERT</b>	0.772	0.660	0.712	0.681	0.652	0.638	0.645	0.651
<b>ELMo</b>	0.582	0.654	0.616	0.657	0.588	0.656	0.620	0.710
<b>LSTM</b> <sub>multitask</sub>	0.667	0.719	0.692	0.710	0.631	0.649	0.640	0.774
<b>LSTM</b> <sub>multitask (FastText)</sub>	0.680	0.725	0.701	0.694	0.667	0.673	<b>0.670</b>	0.717
<b>ELMo</b> <sub>multitask</sub>	0.559	0.678	0.613	0.668	0.516	0.694	0.592	0.694
<b>BERT</b> <sub>multitask</sub>	0.772	0.671	0.718	0.692	0.649	0.642	0.645	0.657

the best in predicting unseen topics for the four topics we experiment on. By capturing different scales of correlation between words (i.e., bigrams, trigrams, and unigrams), the CNN model can detect different patterns in the sentence, regardless of their position [Shirbandi and Moradi, 2019].

Finally, Table 4.15 presents the results obtained when the models are trained on all the available data belonging to a target and tested on all the available data belonging to a different target (i.e.,  $Tag^S \rightarrow Tag_{unseen}^S$ ). In line with the previous experiment, the

Table 4.14: Results  $\text{Top}^S \rightarrow \text{Top}_{\text{unseen}}^S$ .

System	Racism <sub>Waseem</sub>				Sexism <sub>Waseem</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>LSVC</b>	0.458	0.490	0.474	0.820	0.491	0.498	0.494	0.761
<b>LSTM</b>	0.481	0.462	0.471	0.790	0.525	0.543	0.534	0.731
<b>LSTM<sub>FastText</sub></b>	0.489	0.460	0.473	0.787	0.507	0.518	0.513	0.740
<b>ELMo</b>	0.492	0.489	0.491	0.769	0.502	0.506	0.504	0.745
<b>CNN<sub>FastText</sub></b>	0.742	0.506	<b>0.602</b>	0.853	0.882	0.545	<b>0.674</b>	0.798
<b>BERT</b>	0.507	0.500	0.504	0.842	0.693	0.537	0.605	0.785

System	Misogyny <sub>all</sub>				Xenophobia <sub>HatEval</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>LSVC</b>	0.580	0.581	0.581	0.577	0.629	0.536	0.579	0.603
<b>LSTM</b>	0.562	0.563	0.562	0.545	0.541	0.557	0.549	0.583
<b>LSTM<sub>FastText</sub></b>	0.564	0.572	0.568	0.535	0.508	0.560	0.535	0.583
<b>ELMo</b>	0.510	0.556	0.532	0.583	0.511	0.542	0.526	0.573
<b>CNN<sub>FastText</sub></b>	0.659	0.652	<b>0.655</b>	0.638	0.598	0.593	<b>0.595</b>	0.617
<b>BERT</b>	0.634	0.628	0.631	0.639	0.617	0.531	0.571	0.614

 Table 4.15: Results for  $\text{Tag}^S \rightarrow \text{Tag}_{\text{unseen}}^S$ .

System	WOMEN				ETHNICITY			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>LSVC</b>	0.399	0.491	0.440	0.676	0.438	0.491	0.463	0.753
<b>LSTM</b>	0.423	0.489	0.453	0.670	0.500	0.500	0.500	0.744
<b>LSTM<sub>FastText</sub></b>	0.445	0.487	0.465	0.659	0.476	0.489	0.482	0.722
<b>ELMo</b>	0.420	0.486	0.451	0.665	0.437	0.486	0.460	0.743
<b>CNN<sub>FastText</sub></b>	0.579	0.513	<b>0.544</b>	0.660	0.665	0.543	<b>0.598</b>	0.773
<b>BERT</b>	0.514	0.501	0.507	0.656	0.596	0.506	0.548	0.766

best results were achieved by **CNN<sub>FastText</sub>**. In order to better interpret these results, we conducted another experiment in which a model is trained only on data belonging to a target and tested on data belonging to a topical focus on a different target (e.g., training on the target women and testing on the topic xenophobia belonging to the target race). When comparing these results (cf. Table 4.16) with the ones presented in Table 4.14, one can observe the importance for the system of having learned some information regarding the target, even if the data belongs to a different topical focus. In the absence of such information, a drop of anywhere in between 1% and 12% can be observed for the best-performing models.

To conclude, the results confirm that the multitask approach is able to achieve a robust performance, especially for the multi-topic hate speech detection task. These

Table 4.16: Results for  $Tag^S \rightarrow Top_{unseen}^S$ .

System	Train on target: <b>women</b> and test on:							
	Racism <sub>Waseem</sub>				Xenophobia <sub>HatEval</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>LSVC</b>	0.446	0.488	0.466	0.819	0.494	0.499	0.497	0.577
<b>LSTM</b>	0.432	0.478	0.451	0.805	0.469	0.486	0.478	0.548
<b>LSTM<sub>FastText</sub></b>	0.434	0.475	0.451	0.798	0.480	0.492	0.486	0.557
<b>ELMo</b>	0.445	0.481	0.462	0.805	0.510	0.501	0.505	0.577
<b>CNN<sub>FastText</sub></b>	0.716	0.504	<b>0.592</b>	0.852	0.563	0.534	<b>0.548</b>	0.600
<b>BERT</b>	0.553	0.502	0.526	0.849	0.547	0.505	0.525	0.597

System	Train on target: <b>ethnicity</b> and test on:							
	Sexism <sub>Waseem</sub>				Misogyny <sub>a11</sub>			
	P	R	F <sub>1</sub>	A	P	R	F <sub>1</sub>	A
<b>LSVC</b>	0.391	0.486	0.431	0.756	0.498	0.470	0.484	0.569
<b>LSTM</b>	0.395	0.484	0.431	0.753	0.500	0.500	0.500	0.571
<b>LSTM<sub>FastText</sub></b>	0.403	0.479	0.431	0.741	0.474	0.495	0.484	0.560
<b>ELMo</b>	0.419	0.479	0.436	0.737	0.452	0.495	0.472	0.565
<b>CNN<sub>FastText</sub></b>	0.843	0.504	<b>0.631</b>	0.780	0.576	0.532	<b>0.553</b>	0.570
<b>BERT</b>	0.446	0.498	0.470	0.774	0.483	0.498	0.490	0.546

results are encouraging as they can constitute the first step towards targeted hate speech detection. This would be especially true for languages that lack annotated data for a particular target or in the aftermath of a triggering event.

## 4.4 Summary

In this chapter, we focus on three main investigation: (1) we conduct a cross-dataset experiment by exploiting several abusive language datasets with different topical focuses; (2) we explore the ability of abusive language detection models to capture common properties from topic-generic datasets and transfer this knowledge to recognize specific manifestations of abusive language; and (2) we experiment with the development of models to detect both topics (racism, xenophobia, sexism, misogyny) and target of abusive, going beyond standard binary classification, to investigate *how to detect abusive language at a finer level of granularity* and *how to transfer knowledge across different topics and targets*. We experimented with different neural models including multitask approaches. Our study shows that: (1) training a model on datasets featured by more general abusive phenomena able to produce a more robust model to detect other more specific kinds of abusive languages; (2) training a model on a combination of several (training sets from several) topic-specific datasets is more effective than training a model on a topic-generic dataset;

and (3) the multitask approach outperforms a single-task model when detecting both the abusiveness of a tweet and its topical focus in the context of a multilabel classification approach; Our results demonstrate that multitarget abusive language detection from existing datasets is feasible, which is a first step towards abusive language detection for a specific topic/target when dedicated annotated data are missing. All resources and source code developed in this work are publicly available on GitHub.<sup>24</sup>

---

<sup>24</sup><https://github.com/dadangewp/Multidomain-Abusive-Language-Detection>

## Chapter 5

# Multilingual Hate Speech Detection in Social Media

While hate speech is a global phenomenon, current studies on automatic hate speech detection are typically framed in a monolingual setting. There is an urgent need to develop robust systems to identify online hate speech across multiple languages, considering how it is a global issue. As a matter of fact, most popular social media, such as Twitter and Facebook, are multilingual, fostering their users to interact in their primary language. There is a considerable urgency to prevent online hate speech from spreading virally, becoming a significant factor in grave crimes committed against minorities or vulnerable categories. Specifically, robust approaches are needed for abusive language detection in a multilingual environment, which will enable the implementation of effective tools for guaranteeing better compliance to governments demands to counteract the phenomenon.

In the previous Chapter, we present some investigation related to the building of robust model to detect hate speech across different domains. Meanwhile, this Chapter focus on the investigation of the possibility to build a robust model to detect hate speech across different languages. This Chapter is organized as follows. Section 5.1 introduces some background motivation of this research direction. Then, Section 5.2 describes the objective of this study including the mention of approaches adopted in this work. Section 5.3 describes the datasets and resources used in this work. In Section 5.4 we present the models employed in the experiments of this work, including the novel joint-learning models and the baseline models. The experimental results and analysis are described in Section 5.5. Section 5.6 discusses and highlights the main findings of the experiments presented in the previous sections. Finally, Section 5.7 summarize the whole chapter, focusing on the important finding of this research study.

## 5.1 Motivation

The increasing number of social media users has several upsides and downsides. Hate speech online is one of the prominent issues, especially due to the freedom and anonymity given to users and the lack of effective regulations provided by the social network platforms. This problem affects not only the abuse victims but also social medial platforms and governments [Corazza et al., 2020a]. Hate Speech (HS) can be defined as any type of communication that is abusive, insulting, intimidating, harassing, and/or inciting violence or discrimination, disparaging a person or a vulnerable group based on some characteristics such as ethnicity, gender, sexual orientation, religion, or other characteristics [Erjavec and Kovačić, 2012]. Hate speech is becoming a significant problem in online communication on social media, potentially resulting in dangerous criminal acts [Williams et al., 2020, Müller and Schwarz, 2019], as in Rohingya, Myanmar, in 2017, resulting in the murder of thousands of civilians.<sup>1</sup>

Similar to other natural language processing (NLP) tasks [Joshi et al., 2020], detecting hate speech in less-resourced languages is a prominent and timely challenge. For example, the escalation of hate speech against Muslims in Rohingya Myanmar was also affected by the failure to stop spreading hate comments on Facebook due to the difficulty of processing Burmese text automatically<sup>2</sup>. Furthermore, the current availability of datasets in many languages [Poletto et al., 2020], makes the time ripe for addressing the multilingual challenge. Therefore in this part of the thesis, we focus on investigating the hate speech detection task in multilingual environment.

## 5.2 Objectives

This Chapter focuses on investigating the cross-lingual transfer of hate speech detection from a resource-rich language to a lower-resource language. In this direction, we implement zero-shot learning of cross-lingual hate speech detection, by training a model on one language and using it to predict the harmfulness in an unseen language. We focus on English (EN) as a resource-rich source language and six different lower-resource languages as targets, namely French (FR), German (DE), Indonesian (ID), Italian (IT), Portuguese (PT), and Spanish (ES). We propose a novel joint-learning approach to detect hate speech in a cross-lingual setting by exploiting multilingual language representations, including Facebook MUSE [Conneau et al., 2017] and Multilingual BERT<sup>3</sup>. This architecture allows the model to learn simultaneously from both source and target languages, thus transferring knowledge from the resource-rich language (EN) to the less-resourced languages. In addition, we also explore the use of a domain-independent, multilingual lexicon of abusive words called HurtLex [Bassignana et al., 2018], as a proxy to transfer knowledge across different languages. Abusive words have been proven to be powerful signals to detect hate speech, and almost all languages have an arsenal of hateful words that vary in quantity,

---

<sup>1</sup><https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html>

<sup>2</sup><https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>

<sup>3</sup><https://github.com/google-research/bert/blob/master/multilingual.md>

content, and degree of vulgarity. Although such words represent, in some sense, the dark side of language, they are also often prime examples of creative use of language. Hateful expressions often make use of rhetorical figures (e.g., metaphors, synecdoche, metonymy) and idiomatic expressions, and they are highly sensitive to geographical, temporal, and cultural variations, especially when the derogatory meaning is linked to stereotype and prejudice. Our working hypothesis is that in this scenario the injection of additional linguistic knowledge on hateful words from the multilingual lexicon HurtLex can be helpful to improve the multilingual representation provided by BERT-based models. Indeed, a wide range of hateful words are included in HurtLex, organized in general categories sometimes related to cultural stereotypes, ranging from ethnic slurs to insulting words that target physical disabilities, and derogatory senses in different languages have been linked.

## 5.3 Data and Resources

### 5.3.1 Datasets

We collected 11 publicly available datasets in 7 different languages from previous studies that explicitly mention “hate speech” from the ones listed on the Hate Speech Data website<sup>4</sup>. Some of the chosen datasets contain more than the two labels *hate speech* and *not hate speech*, including [Davidson et al., 2017] (offensive), [Founta et al., 2018] (offensive, abusive, aggressive, cyberbullying, and spam), and [Ousidhoum et al., 2019] (abusive, offensive, disrespectful, and fearful). We exclude these labels from the respective datasets and only focus on the binary HS classification. Table 5.1 shows that most datasets have more negative samples (not hate speech) than positive samples (hate speech), reaching extreme imbalance in Founta et al. [2018] and Mandl et al. [2019] with a hate speech ratio (HSR) below 10%. We combine all datasets in the same language, resulting in seven language-specific datasets. In the following we describe each dataset.

**Davidson et. al. Dataset.** This dataset [Davidson et al., 2017] contains 24,783 tweets in English and annotated with three labels: *hate speech*, *offensive*, and *neither*. Further description of this dataset has been presented in subsection 3.4.1.

**Basile et. al. Dataset.** This corpus [Basile et al., 2019] contains 13,000 tweets in English and Spanish, distributed across two different hate speech targets including *immigrant* and *women*. Further description of this dataset has been presented in subsection 4.3.1.

**Founta et. al. Dataset.** This dataset [Founta et al., 2018] contains 80,000 English tweets, tagged with seven mutually exclusive labels, namely *offensive*, *abusive*, *hateful*, *aggressive*, *cyberbullying*, *spam*, and *normal*. Further description of this dataset has been provided in subsection 4.3.1.

**Ousidhoum et. al. Dataset.** This dataset [Ousidhoum et al., 2019] contains 13,014 tweets and consists of three different languages: English (5,647), French (4,014), and Arabic (3,353). The dataset was annotated by using a crowdsourcing with the

---

<sup>4</sup><http://hatespeechdata.com/>

Lang.	Dataset	Label	Total	HSR
EN	Davidson et al. [2017]	hate speech, offensive, neither	5,593	0.26
	Basile et al. [2019]	hate speech, not	12,971	0.42
	Founta et al. [2018]	offensive, abusive, aggressive, cyberbul-	58,722	0.08
	Ousidhoum et al. [2019]	lying, spam, and none hate speech, abusive, offensive, disre- spectful, fearful, and normal	1,939	0.66
FR	Ousidhoum et al. [2019]	hate speech, abusive, offensive, disre- spectful, fearful, and normal	1,220	0.33
DE	Mandl et al. [2019]	hate speech, no	4,743	0.03
	Ross et al. [2017]	hate speech, no	369	0.15
ID	Ibrohim and Budi [2019a]	hate speech, abusive, and no	13,169	0.42
	Alfina et al. [2017]	hateful and normal	713	0.36
IT	Bosco et al. [2018]	hate speech, no	4,000	0.32
PT	Fortuna et al. [2019]	hate speech, no	5,670	0.31
ES	Basile et al. [2019]	hate speech, no	6,599	0.42
	Pereira-Kohatsu et al. [2019]	hate speech, no	6,000	0.26

Table 5.1: Size and class distribution of the datasets used in the experiments. HSR is a hate speech instance ratio over all data.

Amazon Mechanical Turk platform.<sup>5</sup> The average Krippendorff scores for inter-annotator agreement are 0.153, 0.244, and 0.202 for English, French, and Arabic respectively. The original dataset has six labels, while in this study we only use *hateful* and *normal*.

**Mandl et. al. Dataset.** This dataset [Mandl et al. \[2019\]](#) sampled from Twitter and partially from Facebook contains 17,657 instances in three different languages covering English (7,005), Hindi (5,983), and German (4,669).<sup>6</sup> The original dataset was annotated with three different annotation layers as part of the Hate Speech and Offensive Content Identification in Indo-European Languages shared task in FIRE 2019. In this work, we only use the first layer of annotation, which consists of two labels, hate speech or not hate speech.

**Ross et. al. Dataset.** The original collection of this dataset [\[Ross et al., 2017\]](#) contains 469 tweets, where two raters annotated each tweet. In this work, we only use tweets where there is agreement between annotator 1 and annotator 2, resulting in 369 tweets. This corpus contains tweets mostly related to the refugee crisis in Germany, collected by using ten specific hashtags roughly dating from February to March 2016.

**Ibrohim et. al. Dataset.** This dataset [\[Ibrohim and Budi, 2018\]](#) contains 13,169 tweets in Indonesian, crawled from Twitter with the Search API by using several keywords related to hate speech towards categories including religion, race, physical disability, and gender, in the span of 7 months (March – September 2018). Several annotation layers were introduced, mainly focusing on hate speech and abusive language. In this work, we only use the hate speech layer annotation, where each tweet is labeled as hate speech or not hate speech.

**Alfina et. al. Dataset.** This dataset [\[Alfina et al., 2017\]](#) consists of 713 tweets in

<sup>5</sup><https://www.mturk.com/>

<sup>6</sup>We combine training and testing data and obtain the number as presented in Table 5.1

Indonesian, 260 tweets labeled as hate speech, and 453 as not hate speech. The tweets were gathered from Twitter with the Twitter Streaming API using hashtags related to political events in Indonesia from the beginning of February until April 2017. The annotation process involved 30 college students, 43.3% men and 56.7% women.

**Bosco et. al. Dataset.** This dataset [Bosco et al., 2018] contains 4,000 tweets in Italian sampled from 6,928 tweets crawled from Twitter with a keyword-based approach. The keywords were chosen based on three social groups, considered potential targets of hate speech in Italy, namely *Immigrant*, *Muslim*, and *Roma*. This collection was annotated with the Figure Eight platform. The dataset was used in the hate speech detection (HaSpeDe) shared task in EVALITA 2018.

**Fortuna et. al. Dataset.** This dataset [Fortuna et al., 2019] comprises 5,670 tweets in Portuguese and was collected based on keywords and profiles using the Twitter Search API. Most tweets were posted from January until March 2017. The dataset was rated using a finer-grained hierarchical annotation scheme with 81 hate speech categories. We only use the first layer of annotation in this work, which consists of a binary label (hate speech vs. not hate speech).

**Pereira et. al. Dataset.** This corpus [Pereira-Kohatsu et al., 2019] contains 6,000 tweets in Spanish and was filtered from 2 million tweets gathered from Twitter from February to December 2017. The filtering process involved several keywords, which were categorized as *absolute hate* or *relative hate*. The dataset was annotated with a binary label (hate speech vs. not hate speech). The annotation process includes four annotators, where the final label was decided based on a majority vote. In the case of disagreement, a fifth annotator cast the deciding vote.

### 5.3.2 Language Representation and External Resources

In this subsection, we will describe the language representation models used in this work. We use three different multilingual pre-trained models, namely LASER, Facebook MUSE, and Multilingual BERT. In addition, we also use one linguistic resources of abusive words, namely HurtLex. Below are the description of each model:

**LASER Embeddings.** Language-Agnostic SEntence Representations (LASER) [Artetxe and Schwenk, 2019] is a multilingual language representation covering 93 languages, belonging to 30 different language families and written in 28 different scripts. This language representation is obtained from max-pooling over a Bi-LSTM encoder output trained on publicly available parallel corpora. This model has been applied to several cross-lingual benchmark tasks such as cross-lingual natural language inference (XNLI dataset), cross-lingual classification (MLDoc dataset), and bitext mining (BUCC dataset). In this work, we use the pre-trained model, which is publicly available without re-training the model.

**Facebook MUSE.** Multilingual Unsupervised and Supervised Embeddings (MUSE) is a multilingual word embedding model obtained by aligning monolingual word embeddings in an unsupervised way. Unlike several state-of-the-art cross-lingual embeddings that rely on the use of parallel corpora, MUSE was built using a bilingual dictionary between

Category	Description
PS	Ethnic Slurs
RCI	Location and Demonyms
PA	Profession and Occupation
DDP	Physical Disabilities and Diversity
DDF	Cognitive Disabilities and Diversity
DMC	Moral Behavior and Defect
IS	Words Related to Social and Economic antage
OR	Words Related to Plants
AN	Words Related to Animals
ASM	Words Related to Male Genitalia
ASF	Words Related to Female Genitalia
PR	Words Related Prostitution
OM	Words Related Homosexuality
QAS	Descriptive Words with Potential Negative Connotations
CDS	Derogatory Words
RE	Felonies and Words Related to Crime and Immoral Behavior
SVP	Words Related to the Seven Deadly Sins of the Christian Tradition

Table 5.2: HurtLex Categories.

pairs of languages to align the embedding representation. As shown on the Github page<sup>7</sup>, the recent development of this multilingual model covers 30 different languages.

**Multilingual BERT.** Multilingual BERT is a multilingual version of original English BERT [Devlin et al., 2019], which is trained on a Wikipedia dump (excluding user and talk pages) in 104 languages. The languages were chosen based on the top 100 languages with the largest Wikipedias. This pre-trained model obtained a competitive result on cross-lingual natural language inference (XNLI dataset). Two multilingual models are publicly available at the current stage, including `bert-multi-uncased` and `bert-multi-cased`. In this work, we use the `bert-multi-cased` as the newer and recommended model at the current stage.<sup>8</sup>

**HurtLex.** HurtLex is a multilingual lexicon of hate words, originally built from 1,082 Italian hate words compiled in a manual fashion by the linguist Tullio De Mauro [De Mauro, 2016]. This lexicon is semi-automatically extended and translated into 53 languages, and the lexical items are divided into 17 categories such as homophobic slurs, ethnic slurs, genitalia, cognitive and physical disabilities, animals and more.<sup>9</sup> The full description and abbreviation of each category is presented in Table 5.2. In this work, we only rely on seven languages of HurtLex.

## 5.4 Experiments

We model hate speech detection as a binary classification task, where we use English data as the training set and the other languages as test sets. We evaluate the performance in terms of Precision ( $P_0$ ), Recall ( $R_0$ ),  $F$ -score on the negative class ( $F_0$ ), and Precision

<sup>7</sup><https://github.com/facebookresearch/MUSE>

<sup>8</sup><https://github.com/google-research/bert/blob/master/multilingual.md>

<sup>9</sup><http://hatespeech.di.unito.it/resources.html>

$(P_1)$ , Recall ( $R_1$ ),  $F$ -score on the positive class ( $F_1$ ), and also macro-averaged  $F$ -score ( $M$ ) and accuracy ( $Acc$ ). In this section, we describe the models that we use in the experiment.

We experiment with five different models, including two novel models based on a joint-learning approach. The rest of the models are adapted from several previous works as baselines to compare our results. All the models are built by using previously presented multilingual language representations.

**Logistic Regression with LASER Embedding.** This model is based on Logistic Regression (LR) coupled with LASER embeddings [Artetxe and Schwenk, 2019]. Based on a previous study [Aluru et al., 2020], this model performed well in cross-lingual hate speech detection, specifically on low-resource languages, where the size of the training dataset is limited. We use the default hyperparameters as initialized by the Scikit-Learn library.<sup>10</sup>

**Neural Model based on English BERT with Translation.** We employ a state-of-the-art model for several natural language processing tasks in English, that is, the Transformer-based architecture BERT (`bert-base-cased`) available on TensorFlow-hub<sup>11</sup>, which allows us to integrate BERT with the Keras functional layer<sup>12</sup>. Our network starts with the BERT layer, which takes three inputs consisting of id, mask, and segment before passing into a dense layer with RELU activation (256 units) on top and an output layer with sigmoid activation. We train the network with the Adam optimizer with a learning rate of  $2^{-5}$ . Since we use the English pre-trained BERT model, we translate the language-specific datasets into English using the Google Translate API.<sup>13</sup> We tune this model by trying several combinations of batch size (32, 64, 128) and number of epochs (1-5).

**Neural Model based on Multilingual BERT.** This model also uses a pre-trained Multilingual BERT model available in TensorFlow-hub (`bert-multi-cased`). The rest network architecture is similar to the previous model, which used the English BERT model, where we also stack dense with RELU activation and dense with sigmoid activation. The use of the multilingual BERT model allows us to feed the text in any language to the architecture, without the translation process. This model is also optimized with Adam optimizer with a learning rate of  $2^{-5}$ . We vary the number of batch sizes (32, 64, 128) and epochs (1-5) to tune this model.

**Joint-Learning Model Based on LSTM with MUSE.** We propose a joint-learning model employing the Multilingual Unsupervised and Supervised Embeddings (MUSE).<sup>14</sup> Figure 5.1 shows the architecture of this model. We translate the data in both directions from the source language to the target language and vice versa to create bilingual training and test data. The architecture consists of two LSTM networks followed by a dense layer with RELU activation and dropout (0.3), one to learn the task in the

---

<sup>10</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

<sup>11</sup><https://www.tensorflow.org/hub>

<sup>12</sup><https://keras.io/>

<sup>13</sup><https://translate.google.com/>

<sup>14</sup><https://github.com/facebookresearch/MUSE>

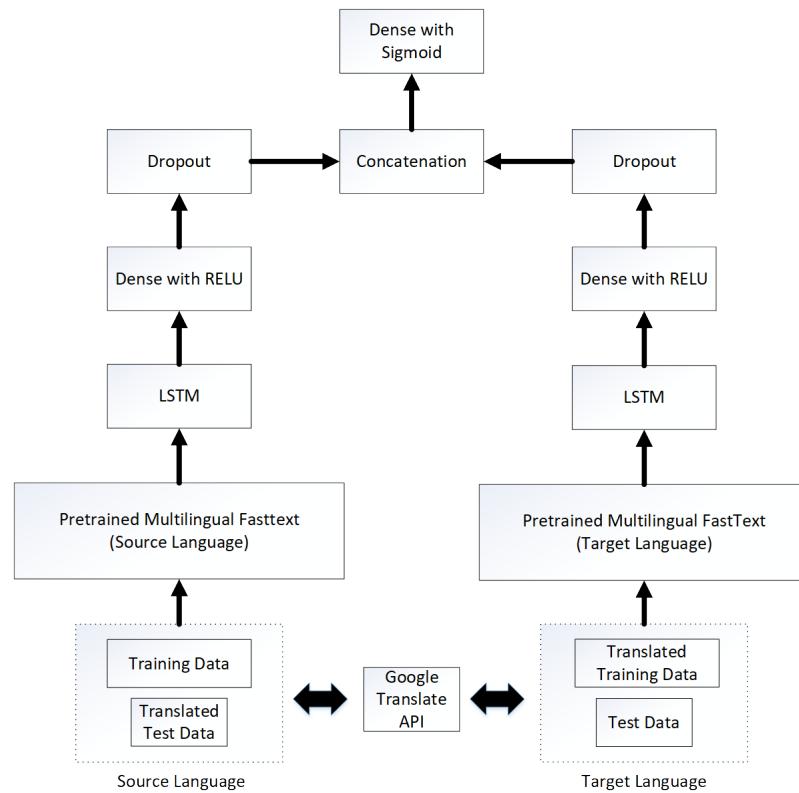


Figure 5.1: Joint-Learning LSTM Model Architecture.

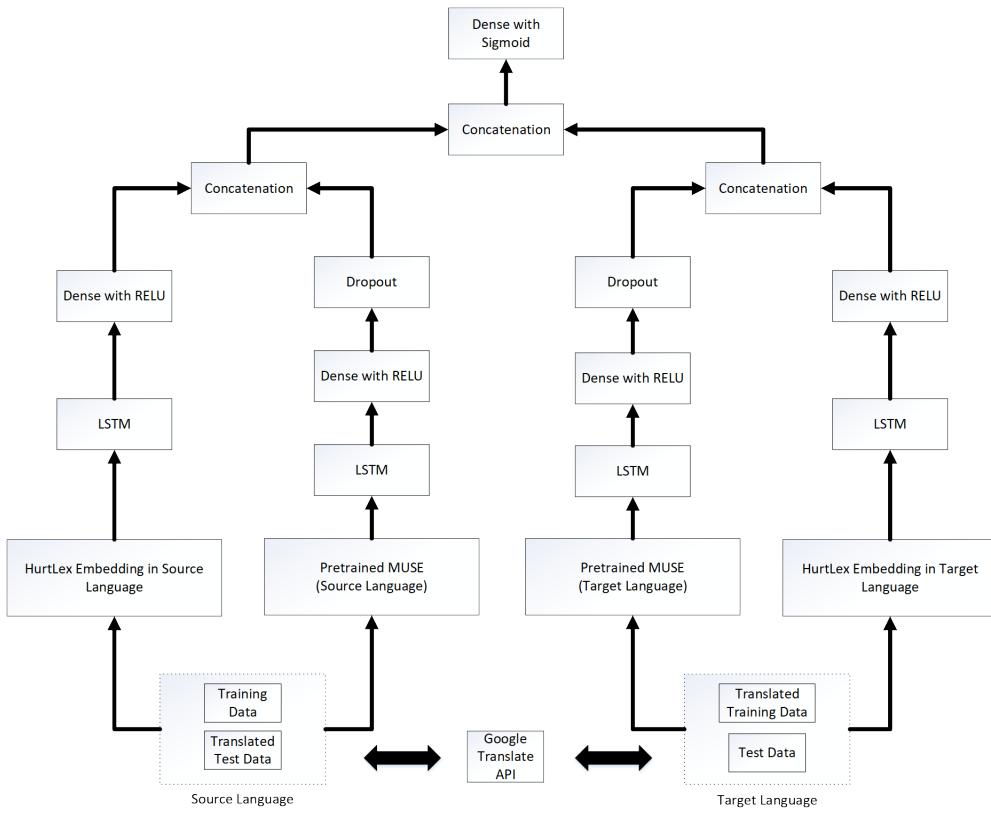


Figure 5.2: Joint-Learning LSTM-HurtLex Model Architecture.

source language, and the other to learn the task in the target language. The output of these networks is concatenated and fed to a dense layer with sigmoid activation as the output layer. This architecture is optimized by an RMS optimizer with default parameters and fine-tuned by varying the number of epochs (1-5) and batch sizes(16, 32, and 64).

In addition, we experiment with the addition of external information provided by a publicly available hate speech-specific lexicon called HurtLex. We build an extra layer consisting of 17-dimension one-hot encoding of the word presence in each of the lexicon categories. Therefore, every word in the comment has one 17-dimensional vector representation. This embedding takes sequential input, passes through an LSTM and a dense layer before being concatenated to the output of BERT, as shown in Figure 5.4. We use two HurtLex embeddings in each architecture to accommodate the input from the source and target languages.

**Joint-Learning Model Based on Multilingual BERT.** We incorporated Multilingual BERT (`bert-multi-cased`) into our joint-learning architecture — see Figure 5.3. Similarly to the Joint-LSTM model, this architecture consists of two main classifiers that learn the task in the source and target language, which are concatenated to produce the final prediction. This model is optimized using the Adam optimizers with a learning rate at  $2^{-5}$ , and fine-tuned by varying the number of epochs (1-5) and batch size (16, 32, and

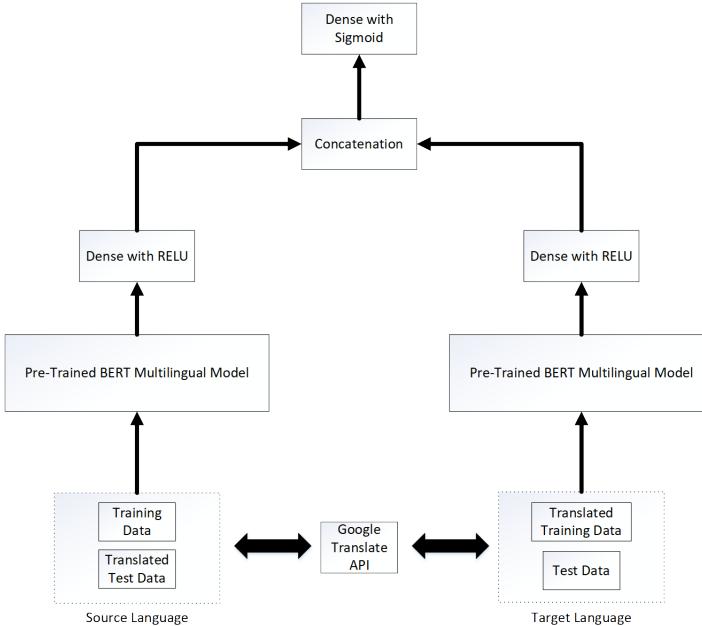


Figure 5.3: Joint-Learning BERT Model Architecture.

64).

Similarly to the previous joint-learning models based on LSTM and MUSE, we employ HurtLex embeddings in this model. The rest of the architecture, with respect to how HurtLex embeddings are integrated with the joint-learning with Multilingual BERT, is the same as the joint-learning LSTM model. The full illustration of this model can be seen in Figure 5.4.

## 5.5 Result and Analysis

Table 5.3 shows the results of our experiment. First, we focus on the comparison between **LR + LASER** and **BERT Multilingual**. We can observe that **LR + LASER** outperforms **BERT Multilingual** in all languages settings, in terms of Macro  $F$ -score. Despite using a traditional machine learning model (logistic regression), this result proves that LASER embeddings provide a better representation for the cross-lingual case. This result is in line with [Reimers and Gurevych, 2020], where Multilingual BERT obtained a poor performance in cross-lingual transfer learning for semantic textual similarity (STS). The study suggests that Multilingual BERT only predicts a single token vector value rather than a sentence, which causes errors in aligning the vectors due to lexical differences between languages.

Another interesting result was obtained by **BERT + Translation**, which outperformed the other systems in two languages settings, namely French (FR) and Spanish (ES). These results raise two arguments. First, the pre-trained BERT model for English is

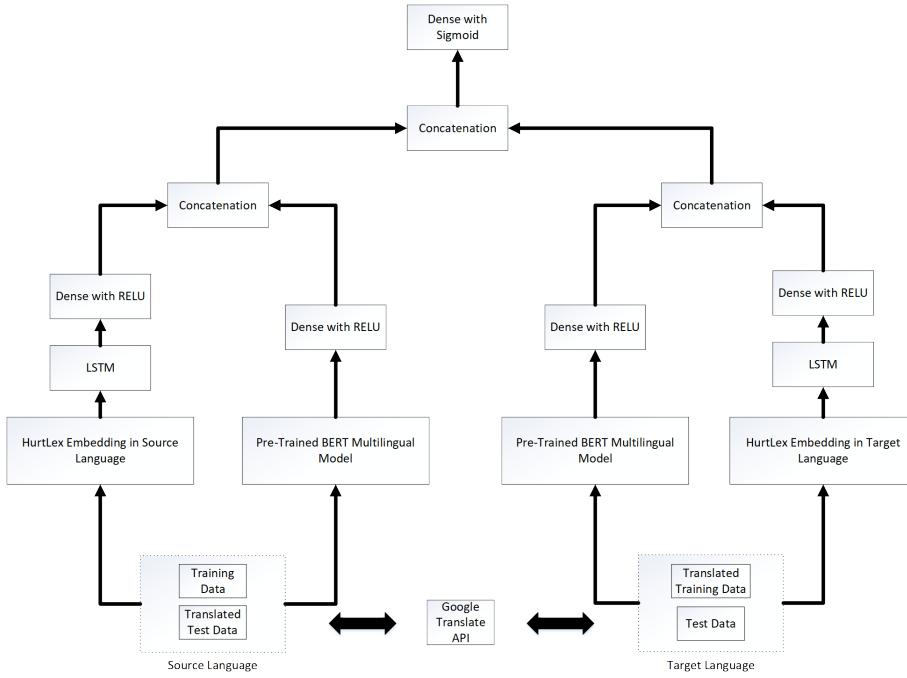


Figure 5.4: Joint-Learning BERT-HurtLex Model Architecture.

a robust language representation model in cross-lingual hate speech detection task, when a good translation is provided. Second, translation tools are quite reliable in providing good translations to English. For comparison, the issue of automatic translation was raised by a recent study where the translation is applied from English to other languages [Pamungkas et al., 2020b], resulting in poor performance.

Our joint-learning models achieved a better performance than the other models in most experimental settings (4 out of 6) in terms of Macro  $F$ -score. The **Joint-learning MUSE** got the best result when tested on German (DE), Indonesian (ID), and Portugal (PT), while **Joint-learning BERT** only outperformed other systems when tested on Spanish (ES). The results in Table 5.3 indicate that all models struggle with the positive class, which is an issue for real-world hate speech detection systems. We believe that this is mainly due to the unbalanced distribution of the training sets. Therefore, we ran another experiment where we balanced the training sets by randomly under-sampling the negative class, keeping the other settings fixed. It is worth noting that our **Joint-learning BERT** model still obtained a competitive performance despite the low performance of the pre-trained model of Multilingual BERT, as observed with the **BERT Multilingual** model. This result indicates that our joint-learning architecture can help improve the Multilingual BERT language model for the cross-lingual task.

Table 5.4 shows the result of this experiment when each system is trained on a balanced training set. **BERT Multilingual** obtained a better performance than when it is trained on the original distribution training set, where it succeeded to outperform

	LR + Laser							BERT Multi.								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
FR	.752	.721	.736	.471	.511	.490	.613	.652	.700	.776	.736	.406	.316	.355	.546	.625
DE	.964	.954	.959	.180	.223	.199	.579	.922	.962	.968	.965	.191	.165	.177	.571	.933
ID	.607	.949	.740	.677	.147	.242	.491	.613	.585	.979	.733	.575	.040	.074	.403	.585
IT	.768	.909	.833	.693	.428	.529	.681	.753	.721	.959	.823	.728	.227	.346	.585	.722
PT	.723	.931	.814	.601	.224	.326	.570	.708	.694	.958	.805	.474	.083	.141	.473	.682
ES	.704	.817	.756	.490	.337	.399	.578	.653	.664	.976	.790	.508	.047	.086	.438	.659
	BERT + Translation							Joint-learning MUSE								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
FR	.765	.741	.752	.499	.531	.515	<b>.634</b>	.672	.740	.657	.696	.427	.526	.471	.584	.614
DE	.973	.903	.937	.178	.456	.254	.595	.883	.971	.928	.949	.201	.398	.268	<b>.608</b>	.905
ID	.625	.858	.723	.593	.285	.385	.554	.618	.661	.645	.653	.524	.543	.533	<b>.593</b>	.602
IT	.821	.829	.825	.635	.623	.629	.727	.762	.759	.928	.835	.718	.386	.502	.668	.752
PT	.730	.930	.818	.625	.253	.361	.589	.717	.733	.936	.822	.651	.261	.373	<b>.598</b>	.723
ES	.748	.845	.794	.602	.453	.517	<b>.655</b>	.711	.716	.843	.774	.541	.3567	.430	.602	.677
	Joint-learning BERT															
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$								
FR	.750	.702	.725	.458	.519	.486	.606	.642								
DE	.967	.955	.961	.226	.286	.253	.607	.926								
ID	.607	.923	.733	.621	.173	.271	.502	.609								
IT	.812	.864	.837	.673	.583	.624	<b>.731</b>	.773								
PT	.732	.917	.814	.600	.272	.375	.594	.713								
ES	.751	.771	.761	.535	.508	.521	.641	.681								

Table 5.3: Results of cross-lingual hate speech detection on the original distribution of training sets.

**LR + LASER** two out of six settings, including when tested on German (DE) and Spanish (ES). However, the performance of both systems is still lower than the three other systems. Again, our joint-learning based model outperformed the other systems in most settings. Only in one setting, testing on Indonesian (ID), **BERT + translation** got better results. We observe a significant improvement in the  $F$ -score of the positive class for most models compared to a system trained with the original distribution — only in German  $F_1$  does not improve, possibly due to an extreme imbalance distribution of the test set. In most cases, a significant improvement can be observed on the recall score of the positive class ( $R_1$ ), which is an important metric for a monitoring system for abusive language [Chen et al., 2017a].

The overall results indicate that this task is difficult and still far from being resolved. The dataset bias is one of the main problems observed when the dataset distribution heavily influences the model performance. Different approaches in collecting and annotating the datasets are also a potential source of bias that impacts the model performance. To show this issue more clearly, we tested the systems on the datasets when more than one language is available (DE, IN, and ES). The results when the system is trained on the original distribution of the training set is presented in Table 5.5, while Table 5.6 presents the results when the systems are trained on a balanced training set. We can observe that our systems do not have uniform performance across two different datasets in the same language, in all three languages. Upon further investigation, we found that several datasets have more specific focuses than others, such as [Ross et al., 2017] (related to

	LR + Laser							BERT Multi.								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
<i>FR</i>	.796	.305	.441	.370	.840	.513	.477	.480	.782	.196	.314	.349	.887	.501	.407	.422
<i>DE</i>	.981	.748	.849	.109	.680	.188	.519	.745	.973	.799	.877	.103	.510	.172	.525	.786
<i>ID</i>	.663	.699	.680	.549	.507	.527	.604	.619	.617	.785	.691	.521	.324	.400	.545	.592
<i>IT</i>	.872	.642	.740	.519	.804	.631	.685	.695	.859	.581	.693	.478	.801	.599	.646	.653
<i>PT</i>	.790	.714	.750	.486	.588	.532	.641	.674	.784	.668	.721	.454	.600	.517	.619	.646
<i>ES</i>	.771	.493	.601	.424	.719	.533	.567	.570	.723	.710	.716	.460	.475	.467	.592	.630
	BERT + Translation							Joint-learning MUSE								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
<i>FR</i>	.794	.305	.440	.369	.837	.512	.476	.479	.772	.268	.398	.357	.837	.501	.449	.454
<i>DE</i>	.987	.691	.813	.105	.796	.185	.499	.695	.975	.866	.917	.146	.505	.227	<b>.572</b>	.851
<i>ID</i>	.718	.636	.675	.565	.653	.606	<b>.640</b>	.643	.716	.517	.600	.517	.716	.600	.600	
<i>IT</i>	.889	.566	.691	.485	.852	.618	.655	.659	.847	.660	.742	.514	.751	.610	.676	.689
<i>PT</i>	.795	.732	.762	.504	.589	.543	.653	.687	.789	.791	.790	.544	.541	.543	<b>.666</b>	.712
<i>ES</i>	.814	.507	.624	.450	.776	.569	.597	.599	.765	.640	.697	.473	.622	.537	<b>.617</b>	.634
	Joint-learning BERT															
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$								
<i>FR</i>	.774	.334	.466	.368	.799	.504	<b>.485</b>	.486								
<i>DE</i>	.973	.813	.886	.107	.495	.176	.531	.799								
<i>ID</i>	.673	.639	.656	.533	.571	.551	.604	.610								
<i>IT</i>	.868	.681	.763	.541	.784	.640	<b>.702</b>	.715								
<i>PT</i>	.779	.775	.777	.516	.522	.519	.648	.695								
<i>ES</i>	.793	.562	.658	.460	.718	.560	.609	.615								

Table 5.4: Results of cross-lingual hate speech detection on the balanced training set.

anti-refugee), [Alfina et al., 2017] (related to political hate speech), and [Basile et al., 2019] (related to hate speech towards women and immigrants). Based on the results in Table 5.6, our models perform consistently better on these datasets compared to datasets with more general topics in the respective language. Indeed, the dataset bias issue is already raised by several studies in hate speech detection [Wiegand et al., 2019, Arango et al., 2020].

Table 5.7 presents the results of our two joint-learning based systems with the additional features from HurtLex. We only run this experiment on the balanced training set. To see the impact of additional features from HurtLex, we also provide **Joint-learning MUSE** and **Joint-learning BERT** model results without HurtLex features in the same table. We see how HurtLex features only improve the performance in three out of six settings with **Joint-learning MUSE**. However, the bigger impact of the HurtLex feature can be seen on the **Joint-learning BERT** model. HurtLex features improve the model performance in four out of six settings in terms of macro  $F$ -score, while in terms of  $F_1$  they succeed in improving the performance in all settings.

## 5.6 Discussion

The results of the experiments presented in the previous section clearly show the advantage of employing the proposed methods for cross-lingual hate speech classification. However, they also show how several issues remain open, especially if looked in terms of absolute figures. In this section, we present the results of a series of additional, qualitative

	LR + Laser							BERT Multi.								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
$DE_{Ross}$	.883	.794	.836	.244	.389	.300	.568	.734	.878	.752	.810	.212	.389	.275	.542	.699
$DE_{HASOC}$	.970	.966	.968	.149	.164	.156	.562	.938	.972	.938	.955	.128	.250	.169	.562	.914
$ID_{Ibrohim}$	.602	.947	.736	.664	.142	.234	.485	.607	.583	.971	.728	.549	.049	.089	.409	.581
$ID_{Alfina}$	.703	.978	.818	.880	.281	.426	.622	.724	.646	.989	.782	.750	.058	.107	.444	.649
$ES_{Basile}$	.632	.743	.683	.519	.391	.446	.565	.597	.600	.970	.741	.677	.087	.155	.448	.604
$ES_{Pereira}$	.767	.882	.821	.422	.244	.309	.565	.715	.744	.961	.839	.371	.066	.112	.475	.727
	BERT + Translation							Joint-learning MUSE								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
$DE_{Ross}$	.907	.680	.778	.241	.593	.342	.560	.668	.932	.740	.825	.311	.685	.428	.626	.732
$DE_{HASOC}$	.975	.954	.964	.198	.316	.244	.604	.932	.972	.961	.967	.185	.243	.210	.589	.936
$ID_{Ibrohim}$	.596	.935	.728	.596	.131	.215	.471	.596	.635	.800	.708	.575	.370	.450	.579	.619
$ID_{Alfina}$	.654	.989	.787	.821	.088	.160	.474	.661	.667	.980	.794	.809	.146	.248	.521	.676
$ES_{Basile}$	.680	.829	.747	.651	.450	.532	.640	.672	.687	.738	.711	.587	.525	.554	.633	.650
$ES_{Pereira}$	.770	.923	.840	.504	.222	.308	.574	.740	.757	.904	.824	.400	.181	.249	.537	.715
	Joint-learning BERT															
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$								
$DE_{Ross}$	.894	.749	.815	.248	.481	.327	.571	.710								
$DE_{HASOC}$	.975	.907	.940	.123	.362	.183	.562	.888								
$ID_{Ibrohim}$	.605	.885	.719	.572	.210	.307	.513	.600								
$ID_{Alfina}$	.687	.951	.798	.744	.246	.370	.584	.694								
$ES_{Basile}$	.677	.826	.744	.645	.445	.526	.635	.668								
$ES_{Alfina}$	.778	.871	.822	.448	.296	.357	.589	.721								

Table 5.5: Results of cross-lingual hate speech detection on individual datasets with the original training set distribution.

analysis that attempt to shed light on the reasons why some of our models obtain better performance than previous works, but also where to look for venues to improve cross-lingual hate speech classification.

### 5.6.1 External Knowledge on Hate Words

Based on the results in Table 5.7, we observe that the use of HurtLex can improve the  $F_1$ , especially with the **Joint-learning BERT** model. The improvement of  $F_1$  is due to an increased number of true positives. More true positives, in turn, means that HurtLex is able to successfully catch hate speech instances which are misclassified by the model without HurtLex.

Derogatory words are often powerful signals to detect hate speech. Hurtful words vary in an imaginative way from one language to another, giving rise to expressions that often sound bizarre or incomprehensible when observed under the lens of one’s mother tongue. This is especially true in the case of words which are literally descriptive of some entity, but also have some markedly derogatory meaning, often linked to a negative stereotype associated to the entity. Such links are very culture-dependant and vary from a language to another. Moreover, hateful expressions often make use of figurative language, rhetorical figures and idiomatic expressions, which are also language-specific. Recently, a study by Nozza [2021] also observed a similar finding related to the use of taboo words, which are not directly translatable between languages. In fact, a swear word could be linked to an

	LR + Laser							BERT Multi.								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
$DE_{Ross}$	.951	.311	.469	.184	.907	.306	.388	.398	.894	.619	.732	.205	.574	.302	.517	.612
$DE_{HASOC}$	.982	.781	.870	.090	.599	.156	.513	.775	.975	.852	.910	.089	.401	.146	.528	.836
$ID_{Ibrohim}$	.654	.691	.672	.542	.500	.520	.596	.610	.607	.792	.688	.513	.300	.378	.533	.584
$ID_{Alfina}$	.818	.834	.826	.701	.677	.689	.758	.777	.702	.863	.774	.603	.362	.452	.613	.680
$ES_{Basile}$	.718	.425	.534	.486	.765	.594	.564	.566	.612	.891	.726	.572	.205	.302	.514	.606
$ES_{Pereira}$	.811	.552	.657	.334	.638	.439	.548	.0574	.758	.843	.798	.350	.238	.283	.541	.685
	BERT + Translation							Joint-learning MUSE								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
$DE_{Ross}$	.980	.307	.467	.192	.963	.320	.394	.403	.910	.768	.833	.291	.556	.382	.608	.737
$DE_{HASOC}$	.985	.732	.840	.086	.697	.152	.496	.730	.982	.694	.813	.071	.645	.127	.470	.692
$ID_{Ibrohim}$	.666	.707	.686	.563	.515	.538	.612	.626	.694	.425	.527	.486	.744	.588	.557	.559
$ID_{Alfina}$	.790	.890	.837	.754	.588	.661	.749	.780	.786	.894	.837	.758	.577	.655	.746	.778
$ES_{Basile}$	.789	.466	.586	.523	.825	.640	.613	.615	.725	.614	.665	.552	.671	.606	.635	.638
$ES_{Pereira}$	.816	.619	.704	.360	.606	.452	.578	.616	.801	.657	.723	.357	.538	.429	.575	.626
	Joint-learning BERT															
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$								
$DE_{Ross}$	.943	.317	.475	.183	.889	.303	.389	.401								
$DE_{HASOC}$	.978	.834	.901	.095	.480	.158	.529	.822								
$ID_{Ibrohim}$	.644	.731	.685	.548	.446	.492	.588	.611								
$ID_{Alfina}$	.787	.817	.802	.658	.616	.636	.719	.743								
$ES_{Basile}$	.828	.546	.658	.346	.680	.459	.559	.581								
$ES_{Alfina}$	.829	.566	.672	.353	.669	.462	.567	.593								

Table 5.6: Results of cross-lingual hate speech detection of each dataset on the balanced training set.

abusive context in one language, but its translated counterpart could be not linked to an abusive context in other languages, which will contribute to the difficulty of the task. We hypothesize that, in such contexts, the knowledge infused by the multilingual lexicon HurtLex, which map such links, can be crucial to recognize the presence of hate speech.

To provide a more in-depth insight, we performed an error analysis on the samples that were predicted as not containing hate speech by the model without HurtLex and were instead correctly classified as hate speech by the model augmented with HurtLex. In this analysis, we only focus on the **Joint-learning BERT** model, where the improvement is more consistent. The analysis is done in two languages, namely Indonesian and Italian, where native speakers of the respective languages are available in our research group.

Example 1 :

Twitter Maju lu sini **anjing** URL

UK Come on here **dog** URL

Example 2 :

Twitter Ahok : memilih pemimpin Berdasarkan agama melanggar konstitusi. Sebaiknya **BABI INI DI BUNGKUS** aja, CONGORNYA PECAH BELAH UMAT BERAGAMA.

UK Ahok: choosing a leader based on religion violates the constitution. It

	Joint-learning MUSE							Joint-learning BERT								
	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$	$P_0$	$R_0$	$F_0$	$P_1$	$R_1$	$F_1$	$M$	$Acc$
<i>FR</i>	.772	.268	.398	.357	.837	.501	.449	.454	.774	.334	.466	.368	.799	.504	.485	.486
<i>DE</i>	.975	.866	.917	.146	.505	.227	.572	.851	.973	.813	.886	.107	.495	.176	.531	.799
<i>ID</i>	.716	.517	.600	.517	.716	.600	.600	.600	.673	.639	.656	.533	.571	.551	.604	.610
<i>IT</i>	.847	.660	.742	.514	.751	.610	.676	.689	.868	.681	.763	.541	.784	.640	<b>.702</b>	.715
<i>PT</i>	.789	.791	.790	.544	.541	.543	.666	.712	.779	.775	.777	.516	.522	.519	.648	.695
<i>ES</i>	.765	.640	.697	.473	.622	.537	.617	.634	.793	.562	.658	.460	.718	.560	.609	.615
Joint-learning MUSE + HurtLex							Joint-learning BERT + HurtLex									
<i>FR</i>	.788	.421	.549	.392	.767	.519	.534	.534	.804	.330	.468	.377	.835	.520	.494	.495
<i>DE</i>	.973	.858	.912	.132	.476	.207	.559	.842	.979	.778	.867	.113	.626	.192	.529	.771
<i>ID</i>	.708	.444	.546	.492	.747	.594	.570	.571	.704	.576	.634	.531	.665	.591	.612	.613
<i>IT</i>	.837	.697	.760	.531	.717	.610	.685	.703	.879	.685	.763	.550	.803	.653	.708	.723
<i>PT</i>	.785	.765	.775	.517	.546	.531	.653	.696	.806	.685	.741	.485	.643	.553	.647	.672
<i>ES</i>	.769	.654	.707	.483	.623	.544	.625	.643	.793	.576	.667	.465	.710	.562	.615	.622

Table 5.7: Results of cross-lingual hate speech detection with additional external resource on the balanced training set.

*is better if this **PIG IS IN WRAPPING**, it is better to break up with religious beliefs.*

Example 3 :

🐦USER Pengungsi asing bukan penfungsi aseng **dodol**

🇬🇧 USER Foreign refugees are not **dodol** foreign refugees

The first three examples are originally written in Indonesian. The first two tweets contain offensive words, marked in bold, denoting animals in Indonesian. However, these words usually have a neutral sense in English, rarely used in an abusive context. As in Example 1 and Example 2, the words “anjing” (dog) and “babi” (pig) are the main trigger for the abusiveness. Based on the experimental results, both tweets are classified as not hate speech without HurtLex but corrected to hate speech when the HurtLex features are added. We believe that this is due to HurtLex, since these words (“anjing” and “babi”) are covered in the HurtLex Indonesian set. Example 3 is different, where the triggering word could not be translated into English properly. The word “dodol” is the word that triggers the abusive context of the tweet, which can roughly be translated as “stupid” in English. Notice that originally, “dodol” is the name of a traditional snack in Indonesia, but a figurative and creative use of this term exists in colloquial Indonesian and social media to refer to a person as being ’stupid’ or ’illogical’, as a slang for ’bodoh’ (stupid). Since “dodol” is also contained in HurtLex, our model with HurtLex succeeds to classify the tweet as hate speech, while without HurtLex, it is classified as not hate speech.

Example 4 :

🐦E meno male che dovevano pagare le nostre pensioni... #migranti #parassiti

#invasione <https://t.co/2MIVO59LDw>

🇬🇧 And thank goodness they had to pay our pensions ... #migrants #parasites #invasion <https://t.co/2MIVO59LDw>

Example 5 :

🐦 Napoli, **branco** di rom investe con l'auto tre carabinieri e fugge | Ripuliamo l'Italia <https://t.co/36oZGYRXxd>

🇬🇧 Naples, Napoli, Roma **herd** invests with the car three policemen and flees / Let's clean up Italy <https://t.co/36oZGYRXxd>

Example 6 :

🐦 @USER Ho sempre ragione amica mia. A presto e lascia perdere la citta' dei **Polentoni** immigrati meridionali. Roma impera

🇬🇧 @USER I'm always right my friend. See you soon and forget the city of southern immigrant **Polentoni**. Rome reigns

A similar pattern is confirmed when we consider the Italian samples. The examples from 4 to 6 are tweets originally written in Italian. They include words used in a derogatory sense, marked in bold. In Examples 4 and 5, “parassiti” and “branco” (*parasites* and *herd*, respectively) are words which can be neutral when referred to animals, but here they have a clear derogatory meaning, which triggers the abusive reading of the post. They are included in the Italian HurtLex, and we believe that this knowledge infusion from HurtLex has a decisive impact on the correct classification of such cases. Example 6 includes “polentoni”, which cannot be translated into English properly by Google translate, since it makes use of figurative language with reference to issues specific to Italian culture. The term indeed could be translated as “polenta eater”, but is commonly used in a derogatory way by Italians from southern Italy to offend Italians from the north. The word belongs to the group of HurtLex terms evoking a negative stereotype, where the offence does not target a single individual but rather an entire category (in this case geographically connotated). Again, we notice that our model with HurtLex succeeds to classify the tweet as hate speech, while without HurtLex, the tweet is classified as not hate speech.

Overall, the manual comparative inspection of the predictions of our Joint-learning BERT models with and without HurtLex, in two languages, confirms that the additional knowledge from HurtLex allows the model to refine its multilingual representation of the hate words. This is particularly relevant to account for the cases where the derogatory meaning is conveyed by a creative use of language, such as figurative languages or linking to culturally negative stereotypes.

<b>Lang.</b>	<b>Dataset</b>	<b>Topical Focus</b>	<b>Collection</b>
EN	Davidson et. al.	Topic generic	-
	Basile et. al.	Hate speech towards immigrant and women	Jul - Sept 2018
	Founta et. al.	Topic generic	Mar - Apr 2017
	Ousidhoum et. al.	Some controversial topics including feminism, immigrant and islamic-leftism	-
FR	Ousidhoum et. al.	Some controversial topics including feminism, immigrant and islamic-leftism	-
DE	Mandl et. al.	Topic generic	-
	Ross et. al.	Related to refugee crisis	Feb - Mar 2016
ID	Ibrohim et. al.	HS related to religion, race, physical disability, and gender	Mar - Sept 2018
	Alfina et. al.	HS related to political event	Feb - Apr 2017
IT	Bosco et. al.	HS related to Immigrant, Muslim, and Roma	-
PT	Fortuna et. al.	Generic topic	Jan - Mar 2017
ES	Basile et. al.	Hate speech towards immigrant and women	Jul - Sept 2018
	Pereira et. al.	Generic topic	Feb - Dec 2017

Table 5.8: Dataset topical focuses and its collection time.

### 5.6.2 Dataset Topical Focuses

As observed in the experimental results, we found that topical bias in the dataset also influences the performance of our models across different languages. To better understand this issue, we investigated the description of each dataset as provided by the original papers presenting them. Table 5.8 summarized the datasets, including their topical focus and collection period. As shown Tables 5.5 and 5.6, our model obtained a different results on different datasets in the same language, for German, Indonesian, and Spanish. We discovered that each of these datasets has a different topical focus. Some datasets are general, while others focus on more specific topics such as anti-refugee hate, immigrants, politics, and religion. We believe that this difference heavily affects the model performance on several datasets. Similar findings were presented in [Stappen et al., 2020], showing how out-of-domain (different topical focus) samples could hurt the model performance in cross-lingual classification. This study also argues that the temporal aspect could influence the performance as well. The triggering event [Downs, 1973] in different periods of time could result in a dataset with a different topical focus, as reported, e.g., in [Florio et al., 2020]. The datasets in the same language shown in Table 5.8 were collected at different times, which we believe affects their topical focus and therefore the cross-dataset classification results.

## 5.7 Summary

In this chapter, we explore hate speech detection in low-resource languages by transferring knowledge from a resource-rich language, English, in a zero-shot learning fashion. We ex-

periment with traditional and recent neural architectures, and propose two joint-learning models, using different multilingual language representations to transfer knowledge between pairs of languages. We also evaluate the impact of additional knowledge in our experiment, by incorporating information from a multilingual lexicon of abusive words. The results show that our joint-learning models achieve the best performance on most languages. However, a simple approach that uses machine translation and a pre-trained English language model achieves a robust performance. In contrast, Multilingual BERT fails to obtain a good performance in cross-lingual hate speech detection. We also experimentally found that the external knowledge from a multilingual abusive lexicon is able to improve the models' performance, specifically in detecting the positive class. The results of our experimental evaluation highlight a number of challenges and issues in this particular task. One of the main challenges is related to the issue of current benchmarks for hate speech detection, in particular how bias related to the topical focus in the datasets influences the classification performance. The insufficient ability of current multilingual language models to transfer knowledge between languages in the specific hate speech detection task also remain an open problem. However, our experimental evaluation and our qualitative analysis show how the explicit integration of linguistic knowledge from a structured abusive language lexicon helps to alleviate this issue. All resources and source code developed in this work are publicly available on GitHub.<sup>15</sup>

---

<sup>15</sup><https://github.com/dadangewp/Multilingual-Abusive-Language-Detection>



## Chapter 6

# Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study

The freedom of expression given by social media has a dark side: the growing proliferation of Hate Speech (HS) contents on these platforms. Hate speech can be based on race, skin color, ethnicity, gender, sexual orientation, nationality, or religion, it incites to violence and discrimination, abusive, insulting, intimidating, and harassing. Hateful language is becoming a huge problem in social media platforms such as Twitter and Facebook [Poland, 2016]. In particular, a type of cyberhate that is increasingly worrying nowadays is the use of hateful language that specifically targets women, which is normally referred to as: *misogyny* [Bartlett et al., 2014].

Chapter 3 already discussed the role of offensive words in abusive language detection tasks. Meanwhile, Chapter 4 and Chapter 5 presented the investigation of possibility in building a robust model to detect abusive language across different domains/targets and languages. This Chapter focuses in implementing all aforementioned contributions into a specific kind of abusive language, namely Misogyny. This Chapter is organized as follows. Section 6.1 provides the research motivation in conducting deep exploration of Misogyny phenomenon. Section 6.2 describes the Automatic Misogyny Identification (AMI) task and dataset. Section 6.2 also describes other datasets which will be used in this work. Then, Section 6.3 presents the proposed experimental settings for the AMI task as well as the results. Section 6.4 provides an investigation of the relationship between misogyny and other related phenomena by conducting a cross-domain classification experiment. Meanwhile, Section 6.5 presents the experimental settings and results of our cross-lingual experiments on the AMI task. Section 6.6 discusses, analyses, and highlights the results and the main findings of the experiments presented in previous sections. Finally, Section 6.7 summarizes our works and important findings in this study.

## 6.1 Motivation

Considering that more and more episodes of misogynistic hate speech and online harassment happen in social media, which stems from sexist stereotypes, prejudices and intolerance and which can lead to episodes of violence, discrimination and persecution also offline, our contribution is devoted to advance the understanding of online misogynistic behaviours. We propose for the first time a computational and multilingual study where the emphasis on a better conceptualization of misogyny and its relation to other abusive phenomena such as sexism, which are more subtle but contribute to a negative environment, is combined with the development of models for detecting misogynous contents in different languages and domains. In this way, we address the open challenge to enhance the robustness and accuracy of tools to contrast the harmful effects of misogynistic behaviors, e.g., tools for automatic support to moderation or for monitoring and mapping the dynamics and the diffusion of hate speech dynamics over a territory, which is only possible at a large scale by employing computational methods.

More specifically, we provide a deep analysis of the automatic misogyny identification task. In particular, we investigate the most predictive features for capturing misogynistic content in social media. In this direction, we explore the state of the art approaches on several available benchmark datasets provided by shared tasks. We experiment with three families of supervised classification models: i) Support Vector Machines (SVM) using word ngrams as features; ii) Recurrent Neural Networks initialized with pre-trained word embeddings; iii) Transformer-based Neural Models, with pre-trained multilingual language models and fine-tuned for each classification task. We further include our own novel method, augmenting both the SVM and the deep learning models with knowledge from a multilingual abusive lexicon. We aim at studying the relation between misogyny and other kinds of hateful language online such as sexist and hate speech in the datasets we collected. To this aim, we experiment in a cross-domain classification setting, to explore the interaction between misogyny and other kinds of hateful language phenomena in terms of what information can be retained across tasks (transfer learning). Finally, as corpora on misogyny are only available in a limited number of languages, developing tools which work cross-lingually is particularly important. To this aim, we conduct experiments on automatic misogyny identification in a multilingual setting.

## 6.2 Automatic Misogyny Detection: Task and Datasets

In this section, we present a detailed description of the Automatic Misogyny Identification shared tasks (AMI), including their definition, evaluation procedure, and the datasets provided to the participants. The data in particular form the basis of the experimental work of the present Chapter. We also include two additional datasets to further validate our hypotheses, namely one widely used benchmark for abusive language detection and the corpus from the hate speech detection evaluation campaign HatEval, both comprising subsets of messages with misogynistic content.

### 6.2.1 Task Definition

AMI is organized as a text classification task across different dimensions. The shared task comprises two subtasks, A and B. The main objective of the AMI task is to discriminate between misogynistic and not-misogynistic content in a binary classification fashion (subtask A). As a secondary goal, systems are asked to categorize misogynistic content into five different misogynistic behaviours and to classify the target of the misogynistic instances (subtask B). The five categories of misogynistic behaviours can be defined as follows:

1. **Stereotype and Objectification:** over-generalization of the women' image, including personality, preferences, and abilities to a very narrow standard.
2. **Dominance:** the intention to show that men are superior to women in a context of gender inequality.
3. **Derailing:** confirming abuse towards women by rejecting male responsibility, or an effort to disrupt conversation in order to redirect women's conversations on something more comfortable for men.
4. **Sexual harassment and threat of violence:** an action to harass women that relates to sexual and inappropriate promise of rewards in exchange for sexual favors. Also includes intent to physically assert power over women through violent threats.
5. **Discredit:** lack of respect toward women, which could also contain slurs.

Target classification is a binary classification task where the categories are defined as follows:

1. **Active:** when the misogyny is specifically target an individual.
2. **Passive:** when the misogyny targets more than one individual or a group of woman.

Subtask A is evaluated in terms of accuracy, while subtask B is evaluated by using the macro-average F-score for misogynistic behaviour and target, and their arithmetic mean. The reason for using a different metric for subtask B is the unbalance within both the Misogynistic Category Classification and the Target Classification, whereas accuracy would not measure performance as fairly in subtask B with respect to subtask A.

### 6.2.2 Datasets

The datasets for AMI IberEval and AMI EVALITA were collected from Twitter following the same procedure, consisting of three different approaches:

1. From the Streaming API, downloading tweets containing representative keywords frequently used to harass women, as introduced in [Hewitt et al., 2016], such as "whore", "cunt", and "bitch".

2. Monitoring a selection of Twitter account of potential victim of harassment and known feminist activists, such as personalities involved in the Gamergate scandal <sup>1</sup>.
3. Downloading tweets from the history of misogynist accounts. These users declare that they are misogynists based on the information shown on the account profile or screen name.

The collection of AMI IberEval was gathered in the span of more than 4 months, starting from 20th of July 2017 until 30th of November 2017, resulting in 83 million tweets for English and 72 millions tweets for Spanish. The shared task organizers queried subsets of tweets for English and Spanish based on co-presence of some keywords. These subsets were then partially annotated fully by two annotators, and by a third annotator to solve the disagreement cases, to build a gold standard set. The rest of tweets in these subsets were annotated by crowd-sourcing with CrowdFlower (now called Appen<sup>2</sup>), where the gold standard was used as test question set. The labels of the crowd-sourced data were decided by using a majority vote approach. The final dataset of AMI IberEval consists of 3,977 tweets (3,251 for training and 726 for testing) for English and 4,138 tweets (3,307 for training and 831 for testing) for Spanish. The detailed distribution of the dataset is shown in Table 6.1.

The AMI EVALITA collection was gathered in the same time period as the one collected for AMI IberEval. The organizers queried the initial collection with a set of predefined keywords, obtaining 10,000 tweets for each language, English and Italian. The annotation process involved six experts using CrowdFlower.

The collection strategy adopted to construct the AMI dataset is partially keyword-based. As recently highlighted in [Wiegand et al., 2019], the adoption of keyword-based data collection processes can introduce biases in the data, in terms of the topics they cover, and therefore it impacts the representativeness of the corpora. Concerning the AMI dataset, the problem is partially mitigated by embracing a combined approach where the keyword-based filtering of Twitter streams is combined with the retrieval of tweets obtained by monitoring potential victims of hate accounts and downloading the history of identified haters and filtering Twitter. However, also this combined strategy presents some limitation in terms of coverage of misogynistic behavior, probably leaving out interesting samples of misogynistic behaviours like benevolent misogyny and disfranchisement. Such phenomena, with few exceptions [Jha and Mamidi, 2017], are often neglected in current studies for being either too subtle or quite rare.

The final collection for the AMI EVALITA shared task comprises 5,000 tweets (4,000 for training and 1,000 for testing) for English and 5,000 tweets (4,000 for training and 1,000 for testing) for Italian. The overall inter-annotation agreement on the English set for “misogynistic”, “misogyny behaviour”, and “misogyny target” is 0.81, 0.45 and 0.49 respectively, while for Italian are slightly higher 0.96, 0.68 and 0.76. The detailed distribution of AMI EVALITA dataset is shown in Table 6.2. Interestingly, it can be observed that the label distribution are very imbalanced for the task B, where *discredit*

---

<sup>1</sup><https://www.theguardian.com/technology/2016/dec/01/gamergate-alt-right-hate-trump>

<sup>2</sup><https://appen.com/>

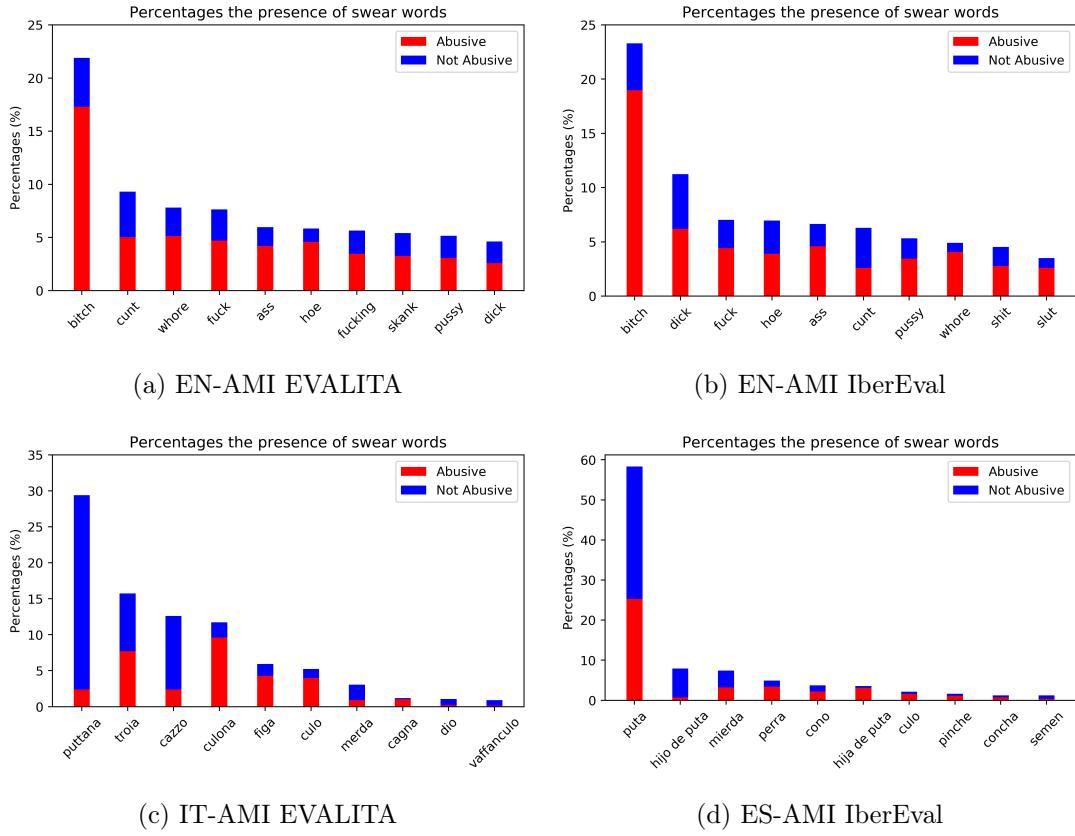


Figure 6.1: Top 10 swear words of each dataset

is the most dominant category of misogynistic behaviour. Some classes were naturally under-represented in these data, such as derailing and dominance. Notice that such resulting class imbalance could have been also affected by the collection strategy applied to construct the data.

Moreover, the *active* class is definitely more represented than the *passive* one when we consider the target of misogyny. Notice that this result fits the most recent theoretical accounts of misogyny from philosophy [Manne, 2017]: “Most misogynistic behavior is about hostility towards women who violate patriarchal norms and expectations, who aren’t serving male interests in the ways they’re expected to. So there’s this sense that women are doing something wrong: that they’re morally objectionable or have a bad attitude or they’re abrasive or shrill or too pushy”. In fact, in the AMI datasets it can be observed that it is the individual woman that, violating the patriarchal norms and expectations, triggers the misogynistic verbal attack online.

We also carried out a lexical analysis on the AMI datasets in all the languages, with the aim of gaining insight about predictive features for the detection task. Figure 6.1 depicts the distribution of offensive words in all four AMI datasets of the EVALITA and IberEval 2018 evaluation campaigns. The red part of the bars shows the frequency of

Task A		Task B		
	English	Spanish	English	Spanish
<b>Misogynistic</b>	1,568/283	1,649/415	Stereotype	137/72
			Dominance	49/28
			Derailing	29/28
			Sexual Harassment	410/32
			Discredit	943/123
			Active	942/104
			Passive	626/179
<b>Not misogynistic</b>	1,683/443	1,658/416	No class	1,683/443
<b>Total</b>				3,251/726
				3,307/831

Table 6.1: AMI IberEval Dataset label distribution.

Task A		Task B		
	English	Italian	English	Italian
<b>Misogynistic</b>	1,785/460	1,828/512	Stereotype	179/140
			Dominance	148/124
			Derailing	92/11
			S. Harassment	352/44
			Discredit	1,014/141
			Active	1,058/401
			Passive	727/59
<b>Not misogynistic</b>	2,215/540	2,172/488	No class	2,215/540
<b>Total</b>				4,000/1,000
				4,000/1,000

Table 6.2: AMI EVALITA Dataset label distribution (training/test).

each swear word when it is used in misogynistic tweet, while the blue one is the frequency when used in messages labeled as not-misogynistic. We took the list of swear words from the online swear word dictionary of noswearing<sup>3</sup>. Based on these figures, we found that the use of specific slurs related to prostitution, female and male genitalia, physical disability and diversity (basically the same words in every languages) is very dominant in all dataset collections across languages<sup>4</sup>. Focusing on the English datasets, misogynistic slurs are mainly used in an abusive/misogynistic context, while in two other languages they are more evenly distributed.

### 6.2.3 Related Datasets

Besides the AMI task datasets, in our study we considered two additional datasets with topical focus on the related notions of sexism and hate speech in social media.

The **Waseem and Hovy Hate Speech Dataset** was collected in the duration of 2 months, for a total of 136,052 tweets. This collection was bootstrapped by conducting a manual search based on several common slurs and terms related to sexual, religious, gender, and ethnic minorities, using public Twitter API search. The authors identified a final set of keywords<sup>5</sup> frequently used in tweets that contain hate speech and references to

<sup>3</sup><https://www.noswearing.com/>

<sup>4</sup>We adopt this categorization from HurtLex [Bassignana et al., 2018].

<sup>5</sup>“MKR”, “asian drive”, “feminazi”, “immigrant”, “nigger”, “sjw”, “WomenAgainstFeminism”, “blame-

specific entities. The portion of collected tweets were manually annotated by the authors, and the annotation were reviewed by a student in gender studies, in order to mitigate annotator bias. The detailed annotation guideline is available in [Waseem and Hovy, 2016]. The final annotated dataset consists of 16,914 tweets, coded in three categories: racism (1,972 tweets), sexism (3,383 tweets), and none (11,559 tweets). The overall inter-annotator agreement based on Cohen’s Kappa coefficient is 0.84, where 85% of the disagreement cases occur in the annotation of sexism. Besides the tweet text, the authors also collected the demographic information of the authors, but this information was not publicly released. The tweet IDs and final annotation labels are available in the Github page<sup>6</sup>. Due to data decay, on our latest effort to retrieve the dataset, we were able to obtain 16,488 out of the 16,914 tweets, of which 1,957 marked as racism, 3,216 as sexism and 11,315 as none.

**HatEval** was introduced at SemEval 2019 [Basile et al., 2019] and focuses on the detection of hate speech in Twitter on two specific targets, namely immigrants and women, in a multilingual perspective. This shared task introduced a dataset in two languages, English and Spanish. The keywords which used to collect the dataset include neutral words [Sanguinetti et al., 2018], pejorative words towards the targets, and highly polarized hashtags. Based on the retrieved collection, the distribution of the keywords over the collection is skewed, with some keywords more frequently occurring than others, including “migrant”, “refugee”, “#buildthatwall”, “bitch”, “hoe”, “women” for English and “immigraarabe”, “sudaca”, “puta”, “callate”, “perra” for Spanish. The collected tweets were annotated by non-trained contributors using Figure Eight with three binary labels: hate speech (HS), target range (TR: generic or individual), and aggressiveness (AG)<sup>7</sup>. The average confidence score as reported by Figure Eight for these three labels is 0.83, 0.70 and 0.73 respectively for English and 0.89, 0.47 and 0.47 for Spanish. The final dataset used for the HatEval shared task contains 13,000 (about 10,000 for training and for 3,000 testing) tweets for English and 6,600 (about 5,000 for training and 1,600 for testing) tweets for Spanish<sup>8</sup>. Table 6.3 and Table 6.4 show the detailed label distribution of the dataset for each target.

### 6.3 Automatic Misogyny Identification Experiment

In this section, we present our experiment at building a system with a comparable or better performance than the state of the art to detect misogyny. We use the AMI IberEval and AMI EVALITA benchmark datasets for all languages, namely English (EN), Spanish (ES), and Italian (IT), to evaluate our model. We explore several approaches, including

---

“onenotall”, “islam terrorism”, “notallmen”, “victimcard”, “victim card”, “arab terror”, “gamergate”, “jsil”, “racecard”, “race card”

<sup>6</sup><https://github.com/zeerakw/hatespeech>

<sup>7</sup>the detailed description of annotation guidelines is available at [https://github.com/msang/hateval/blob/master/annotation\\_guidelines.md](https://github.com/msang/hateval/blob/master/annotation_guidelines.md)

<sup>8</sup>Upon manual investigation, organizers decided to exclude 1,000 tweets from the English training set, 29 tweets from the English test set due to duplicated instances, and 500 tweets from the Spanish training set, due to duplicated instances.

Main class	Fine-grained class	Training	Development	Test
	English			
Hate Speech		1,985	237	623
	Aggressive	558	110	214
	Not-Aggressive	1,427	127	409
	Generic	752	27	122
	Individual	1,233	210	501
Not Hate Speech		2,515	263	849
Total (HS+not HS)		4,500	500	1,472
	Spanish			
Hate Speech		1,185	143	336
	Aggressive	1,036	127	311
	Not-Aggressive	149	16	25
	Generic	149	16	17
	Individual	1,036	127	319
Not Hate Speech		1,697	184	463
Total (HS+not HS)		2,882	327	799

Table 6.3: HatEval Dataset label distribution. Hate speech target: Women.

Main class	Fine-grained class	Training	Development	Test
	English			
Hate Speech		1,798	190	629
	Aggressive	1,001	94	376
	Not-Aggressive	797	96	253
	Generic	1,690	181	608
	Individual	108	9	21
Not Hate Speech		2,702	310	870
Total (HS+not HS)		4500	500	1499
	Spanish			
Hate Speech		672	79	324
	Aggressive	466	49	163
	Not-Aggressive	206	30	161
	Generic	579	69	220
	Individual	93	10	104
Not Hate Speech		946	94	476
Total (HS+not HS)		1,618	173	800

Table 6.4: HatEval Dataset label distribution. Hate speech target: Immigrant.

traditional machine-learning models and more recent deep learning techniques. The model performance is evaluated along several metrics such as precision, recall,  $F$ -score, and accuracy for subtask A, and accuracy and macro-averaged  $F_1$ -score for subtask B, as explained in Section 6.2.

### 6.3.1 Traditional Models

We built two Support Vector Machine (SVM) models using different kernel functions, namely linear and radial basis function (RBF). The use of linear kernel is based on [Joachims, 1998], who argue that linear kernel has an advantage for text classification, based on the observation that text representation features are frequently linearly separable. The RBF kernel is preferable to a linear kernel for some text classification task due to its better performance, despite it having higher complexity [Pamungkas et al., 2018c,a].

We employ several stylistic and lexical features, performing a straightforward pre-processing step including tokenization and stemming by using Natural Language Toolkit (NLTK)<sup>9</sup>. Specifically, we employ the features detailed in the following sections.

#### 6.3.1.1 Lexical Features

This set of features aims at representing the semantic content of the tweets at the lexical level.

**Bag of Words.** This feature includes unigram, bigram, and trigram representation of the tweets, where all characters were changed to lower case.

**Bag of Hashtags.** We observed that hashtags<sup>10</sup> were frequently used in both AMI datasets. This feature is built by using the same technique as bag of words which includes unigram, bigrams, and trigrams (some tweets have more than one hashtags), focusing on the hashtag presence.

**Bag of Emojis.** Similarly to hashtags, emojis were also utilized in many instances in the AMI datasets. We normalize every emoji into its Unicode Common Locale Data Repository (CLDR) short name by using the *emoji* library<sup>11</sup>.

**Swear Words.** This feature includes the presence of swear words which are often indicative of abusive content. The list of keywords is gathered from the *noswearing* website<sup>12</sup>, an online dictionary which contains 349 English swear words. For the other languages, we translate the swear words automatically by using Google Translate<sup>13</sup>, and including other sources such as the list of bad words from Wikipedia page<sup>14</sup> and a list of manually checked swear words by a popular linguist blog<sup>15</sup> for Italian. We encode the

---

<sup>9</sup><https://www.nltk.org/>

<sup>10</sup>We also experimented by splitting the hashtags into their constituent words using Ekphrasis [Baziotis et al., 2017], but this did not improve the system performance.

<sup>11</sup><https://pypi.org/project/emoji/>

<sup>12</sup><https://www.noswearing.com/>

<sup>13</sup><https://translate.google.com/>

<sup>14</sup>[https://it.wikipedia.org/wiki/Turpiloquio\\_nella\\_lingua\\_italiana](https://it.wikipedia.org/wiki/Turpiloquio_nella_lingua_italiana)

<sup>15</sup><https://www.parolacce.org/2016/12/20/dati-frequenza-turpiloquio/>

Category	Description
PS	Ethnic Slurs
RCI	Location and Demonyms
PA	Profession and Occupation
DDP	Physical Disabilities and Diversity
DDF	Cognitive Disabilities and Diversity
DMC	Moral Behavior and Defect
IS	Words Related to Social and Economic antage
OR	Words Related to Plants
AN	Words Related to Animals
ASM	Words Related to Male Genitalia
ASF	Words Related to Female Genitalia
PR	Words Related Prostitution
OM	Words Related Homosexuality
QAS	Descriptive Words with Potential Negative Connotations
CDS	Derogatory Words
RE	Felonies and Words Related to Crime and Immoral Behavior
SVP	Words Related to the Seven Deadly Sins of the Christian Tradition

Table 6.5: HurtLex Categories.

information about swear words into two individual features: swear word presence (binary feature) and swear word count (the number of swear words).

**Sexist Slurs.** We include the list of sexist words proposed by [Fasoli et al. \[2015\]](#), which are often used in hate speech messages against women. We manually translate and expand these words for Italian and Spanish. This feature has a binary value of 0 if there is no sexist slur in the tweet, or 1 if there is at least one sexist slur in the tweet.

**Women-related words.** We also manually built a list of words containing synonyms and related words to “woman” (for English), “donna” (for Italian), and “mujer” (for Spanish). This list of words represents a feature to detect the target of hateful content, in this case towards women. Similarly to the sexist slur feature, this feature is also represented as binary number, 0 (there is no woman-related word in the tweet) and 1 (there is at least one woman-related word in the tweet).

**Hate Words Lexicon.** This feature captures the presence of words contained in multilingual hate lexicon HurtLex [[Bassignana et al., 2018](#)]. This lexicon was built starting from a list of words compiled manually by the Italian linguist Tullio De Mauro [[De Mauro, 2016](#)] in Italian, then semi-automatically translated into 53 languages. The lexical items are divided into 17 categories. For our system configuration, we exploited the presence of the words in each category as a single feature, thus obtaining 17 single features, one for each HurtLex category. The full list of HurtLex categories can be seen in Table 6.5. We included this feature because our preliminary lexical analysis suggests that a specific subset of the HurtLex categories can be relevant to detect the misogynistic speech in social media, such as PR (words related to prostitution), ASF (words related to female genitalia), DDP (physical disability and diversity), and DDF (cognitive disability

and diversity).

### 6.3.1.2 Stylistic Features

This set of features aims at capturing the structure of the tweets in terms of the type of some of its constituent elements.

**Hashtag Count.** The number of hashtags contained in the tweets.

**Upper Case Count.** The number of upper case characters in tweets.

**Link Counts.** The number of URLs in the tweets.

**Tweet Length.** The total number of characters of every tweet.

## 6.3.2 Neural Based Models

We adopt two kind of deep learning architectures including recurrent neural networks (RNN) based and transformer based, where we employ BERT. RNN was recognized as an effective architecture for learning text, also in text classification tasks. In this study we will implement two variants of RNN, namely long short term memory (LSTM) and gated recurrent unit (GRU). BERT is a transformer-based architecture which gained a lot of attention in NLP because of its superiority in most standard benchmarks. Here we describe both architectures.

### 6.3.2.1 RNN-based

We use straightforward Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] and Gated Recurrent Units (GRU) [Cho et al., 2014] networks. Our architecture consists of several layers, starting with an embedding layer (300 dimensions), where we experiment with and without pre-trained word embeddings. We employ the readily available embeddings provided by FastText<sup>16</sup> in three languages, i.e., English, Spanish, and Italian. The embedding layer is input to either LSTM or GRU network (64 units), followed by a dense layer (16 units) with ReLU activation function. The final layer consists of a dense layer with sigmoid activation producing the final prediction. We only optimize the batch size (16, 32, 64, 128) and number of epochs (1-5) to tune our architecture in order to get the best possible result.

### 6.3.2.2 BERT

We also adapt BERT [Devlin et al., 2019] for this experiment. We utilize the pre-trained models available on tensorflow-hub<sup>17</sup>, which allows us to integrate BERT in the Keras library<sup>18</sup>. For English, we use the `bert-cased` model, while for Italian and Spanish we use the `bert-multi-cased` model. Our network starts with the BERT layer, which takes three inputs consisting of *id*, *mask* and *segment*. The output of this layer connects to a dense layer with RELU activation (256 units), before passing into a dense layer with

---

<sup>16</sup><https://fasttext.cc/>

<sup>17</sup><https://www.tensorflow.org/hub>

<sup>18</sup><https://keras.io/>

EN-AMI IberEval	ES-AMI IberEval
SVM with RBF Kernel - Swear words count - Swear words presence - Hashtags presence - Links count - Sexist slurs presence - Women words presence	SVM with Linear Kernel - Bag of words - Bag of hashtags - Bag of emojis - Sexist slurs presence - Women words presence - ASF presence - ASM presence - DDF presence - DDP presence - PR presence
EN-AMI EVALITA	IT-AMI EVALITA
BERT - With <code>bert-cased</code> model - Dense layer units = 256 - Batch size = 32 - Epoch = 2	BERT - With <code>bert-multi-cased</code> - Dense layer units = 256 - Batch size = 32 - Epoch = 2

Table 6.6: List of features of best-performing systems on each dataset.

*sigmoid* activation as the predictor layer. We train our network with the Adam optimizer with learning rate  $2^{-5}$ . We fine-tune the model only on the number of epochs (1-5) and batch size (16, 32, and 64).

### 6.3.3 Results

Table 6.7 shows the results of our experiment on subtask A. Since the task organizers only provide accuracy score as the competition baseline, we also built a baseline system with the same configuration as the competition baseline, that is, a linear SVM with word unigram representations as features. Despite not obtaining the exact score provided by the organizers, the score is still relatively comparable. We optimize this model on the training set by testing several combinations of features. We selected the best-performing model based on 10-fold cross evaluation, to be evaluated on the test set. Therefore, our best system configuration is not always containing all the features mentioned in the previous subsection. The deep learning models were optimized by fine-tuning only on the number epochs and batch sizes. Overall, we got the best results on all benchmark datasets. The features and system configurations of our best-performing model for the respective datasets can be found in Table 6.6.

	English AMI IberEval			
	P	R	F <sub>1</sub>	Acc
Baseline of the shared task	-	-	-	78.37
Best system of the shared task	-	-	-	91.32
New baseline	73.63	71.02	72.30	78.79
Support vector classifier with linear kernel	82.70	69.26	75.38	82.37
Support vector classifier with RBF kernel	<b>87.16</b>	91.16	<b>89.12</b>	<b>91.32</b>
LSTM without pre-trained embedding	74.63	71.73	73.15	79.48
LSTM with FastText embedding	74.12	59.72	66.14	76.17
GRU without pre-trained embedding	65.34	81.27	72.44	75.9
GRU with FastText embedding	68.35	76.33	72.12	77.00
LSTM Attention without pre-trained embedding	65.33	69.26	67.24	73.69
LSTM Attention with FastText embedding	74.86	46.29	57.21	73.00
BERT	77.31	<b>91.52</b>	83.82	86.23
	Spanish AMI IberEval			
	P	R	F <sub>1</sub>	Acc
Baseline of the shared task	-	-	-	76.78
Best system of the shared task	-	-	-	81.47
New baseline	72.92	73.98	73.44	73.29
Support vector classifier with linear kernel	80.71	82.65	<b>81.67</b>	<b>81.47</b>
Support vector classifier with RBF kernel	54.26	46.02	49.80	53.67
LSTM without pre-trained embedding	76.40	75.66	76.03	76.17
LSTM with FastText embedding	76.42	78.07	77.23	77.02
GRU without pre-trained embedding	<b>81.95</b>	68.92	74.87	76.90
GRU with FastText embedding	77.03	82.41	79.62	78.94
LSTM Attention without pre-trained embedding	74.59	77.11	75.83	75.45
LSTM Attention with FastText embedding	75.22	82.65	78.76	77.74
BERT	70.84	<b>87.23</b>	78.18	75.69
	English AMI EVALITA			
	P	R	F <sub>1</sub>	Acc
Baseline of the shared task	-	-	-	60.50
Best system of the shared task	-	-	-	70.40
New baseline	55.70	65.87	60.36	60.20
Support vector classifier with linear kernel	44.44	34.78	39.24	50.00
Support vector classifier with RBF kernel	57.54	67.17	61.99	62.10
LSTM without pre-trained embedding	64.39	49.13	55.73	64.39
LSTM with FastText embedding	63.61	57.39	60.34	65.30
GRU without pre-trained embedding	52.12	<b>69.57</b>	59.59	56.6
GRU with FastText embedding	56.85	66.74	61.40	61.40
LSTM Attention without pre-trained embedding	56.61	63.26	59.75	60.80
LSTM Attention with FastText embedding	57.88	63.04	60.35	61.90
BERT	<b>70.37</b>	66.09	<b>68.16</b>	<b>71.6</b>
	Italian AMI EVALITA			
	P	R	F <sub>1</sub>	Acc
Baseline of the shared task	-	-	-	83.00
Best system of the shared task	-	-	-	84.40
New baseline	77.92	93.75	85.11	83.20
Support vector classifier with linear kernel	77.24	<b>97.46</b>	<b>86.18</b>	83.90
Support vector classifier with RBF kernel	76.52	78.91	77.69	76.80
LSTM without pre-trained embedding	79.70	92.77	85.74	84.20
LSTM with FastText embedding	82.10	88.67	85.26	84.30
GRU without pre-trained embedding	78.05	92.38	84.62	82.80
GRU with FastText embedding	78.35	94.73	85.76	83.90
LSTM Attention without pre-trained embedding	82.28	88.87	85.45	84.50
LSTM Attention with FastText embedding	79.05	94.34	86.02	84.30
BERT	<b>83.93</b>	87.11	85.44	<b>84.80</b>

Table 6.7: Results of Automatic Misogyny Identification Experiment on AMI Dataset Task A.

For the traditional model, we use the same model as our contributed system in AMI IberEval [Pamungkas et al., 2018c], which obtained the top ranking in the competition on both English and Spanish. In English the best result was obtained by a support vector machine (SVM) classifier with RBF kernel and several handcrafted features including *hashtags presence*, *links presence*, *swear words count*, *swear words presence*, *sexist slurs presence*, and *woman words presence*. We used the default hyper-parameters as defined by the scikit-learn library<sup>19</sup>. Our system achieves an accuracy of 91.32, a significant improvement compared to the baseline. Meanwhile, our system for Spanish was also developed based on SVM but with linear kernel, coupled by some classic text representation as shown in Table 6.6. This model obtained 81.47 in accuracy. Our BERT models also achieved the best performance on both the English and Italian sets of AMI EVALITA, outperforming the best performing systems of the respective shared task. Our BERT model obtained 71.6 and 84.8 in accuracy on English and Italian respectively.

We also experimented with the four AMI tasks on the subtask B: Misogynistic Behaviour and Target Classification. We used the same systems as in the subtask A experiment, but different evaluation metrics are applied, namely accuracy and macro-averaged  $F_1$ -score. We need to clarify that in the official AMI shared tasks, subtask A and subtask B are treated as a pipeline process, where the prediction of subtask B will be fully-dependent on the subtask A results. Rather, in this experiment, we handle subtask B as an independent multi-class classification task. Table 6.8 shows the full results of the experiments on the subtask B on all four AMI datasets. We compare our models performance with the AMI competition baseline and the best systems. The results show that our proposed systems were able to outperform the best performing systems on all the AMI tasks, based on the average of the macro-averaged  $F_1$ -scores on the two classification tasks of subtask B (misogynistic behaviour and target classification). Overall, BERT was the most consistent model, which gave the best performance on all dataset collections. Only in Italian AMI EVALITA, SVM with linear kernel perform slightly better than BERT. Most systems based on SVM with RBF kernel were under-performing on all datasets, compared to other systems. The big picture of the results also tells us that classifying the target of misogynistic behaviour is an easier task than determining its category, maybe due to the unbalanced distribution of classes in category of misogyny. The low annotator agreement on the “misogyny behaviour”, and “misogyny target” layers in the AMI dataset could also contribute to the difficulty of subtask B, especially on English AMI EVALITA, where the inter-annotator agreement on the dataset is only 0.45 and 0.49 for target classification and category of misogyny, respectively. On the one hand, the low annotator agreement can be a signal for the difficulty of this finer-grained tasks, especially concerning the detection of misogyny behaviours: drawing a sharp separation between the different categories has been difficult also for humans. On the other hand, it can be an alert for a possible inconsistency in the data annotation, that could cause problems to the model to learn the overall phenomena.

---

<sup>19</sup><https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

	Category		Target		Average	
	Acc	Macro $F_1$	Acc	Macro $F_1$	Acc	Macro $F_1$
English AMI EVALITA						
Baseline of Shared Task	-	.342	-	.399	-	.371
Best System of Shared Task	-	<b>.361</b>	-	.451	-	.406
SVM Linear Kernel	.544	.355	.579	.484	.562	.419
SVM RBF Kernel	.461	.164	.552	.446	.507	.305
LSTM without Pre-trained Emb.	.515	.299	.594	.506	.555	.403
LSTM with FastText Emb.	.489	.258	.608	.501	.549	.380
GRU without Pre-trained Emb.	.488	.295	.587	.498	.538	.396
GRU with FastText Emb.	.474	.275	.578	.475	.526	.375
LSTM Att. without Pre-trained Emb.	.496	.336	.563	.475	.530	.405
LSTM Att. with FastText Emb.	.483	.301	.559	.480	.521	.390
BERT	<b>.568</b>	.278	<b>.680</b>	<b>.580</b>	<b>.624</b>	<b>.429</b>
Italian AMI EVALITA						
Baseline of Shared Task	-	.543	-	.440	-	.492
Best System of Shared Task	-	.501	-	.579	-	.540
SVM Linear Kernel	<b>.751</b>	<b>.596</b>	.793	.558	<b>.772</b>	<b>.577</b>
SVM RBF Kernel	.488	.109	.445	.237	.467	.173
LSTM without Pre-trained Emb.	.743	.584	.753	.564	.748	.574
LSTM with FastText Emb.	.721	.516	.770	.575	.746	.546
GRU without Pre-trained Emb.	.710	.480	.767	<b>.607</b>	.739	.543
GRU with FastText Emb.	.729	.538	<b>.797</b>	.571	.763	.554
LSTM Att. without Pre-trained Emb.	.738	.549	.783	.553	.761	.551
LSTM Att. with FastText Emb.	.721	.470	.795	.553	.758	.512
BERT	.739	.508	.777	.537	.758	.522
English AMI IberEval						
Baseline of Shared Task	-	.157	-	.518	-	.337
Best System of Shared Task	-	<b>.293</b>	-	.593	-	.443
SVM Linear Kernel	.674	.259	.759	.642	.716	.451
SVM RBF Kernel	.674	.228	.709	.545	.692	.387
LSTM without Pre-trained Emb.	.572	.274	.674	.577	.623	.425
LSTM with FastText Emb.	.623	.256	.663	.579	.643	.417
GRU without Pre-trained Emb.	.596	.262	.606	.536	.601	.399
GRU with FastText Emb.	.606	.248	.664	.601	.635	.424
LSTM Att. without Pre-trained Emb.	.619	.229	.643	.553	.631	.391
LSTM Att. with FastText Emb.	.605	.227	.663	.527	.634	.377
BERT	<b>.703</b>	.285	<b>.814</b>	<b>.714</b>	<b>.758</b>	<b>.499</b>
Spanish AMI IberEval						
Baseline of Shared Task	-	.281	-	.537	-	.409
Best System of Shared Task	-	.339	-	.553	-	.446
SVM Linear Kernel	<b>.698</b>	<b>.371</b>	<b>.770</b>	.566	<b>.734</b>	.469
SVM RBF Kernel	.501	.111	.460	.261	.480	.186
LSTM without Pre-trained Emb.	.633	.344	.668	.585	.650	.465
LSTM with FastText Emb.	.658	.328	.728	<b>.614</b>	.693	.471
GRU without Pre-trained Emb.	.608	.332	.706	.582	.657	.457
GRU with FastText Emb.	.661	.351	.732	.577	.696	.464
LSTM Att. without Pre-trained Emb.	.609	.311	.666	.545	.637	.428
LSTM Att. with FastText Emb.	.584	.328	.718	.568	.651	.448
BERT	.666	<b>.371</b>	.744	.577	.705	<b>.474</b>

Table 6.8: Result of Experiment on SubTask B.

Subtask A	Subtask B	Subtask C	Train	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	-	524	27	551
NOT	-	-	8,840	620	9,460
All			13,240	860	14,100

Table 6.9: Dataset label distribution of OLID. OFF : Offensive; NOT : Not Offensive; TIN : Targeted Insult; UNT : Untargeted; IND : Individual; OTH : Other; GRP : Group.

## 6.4 Relationship between Misogyny and Other Abusive Phenomena

In this section, we present the results of an experiment carried out with the goal of studying the relationship between misogyny and other abusive language phenomena including sexism, hate speech, and offensive language. In essence, we train models on additional datasets (different abusive phenomena) and test their prediction capability for misogyny detection on the AMI benchmark. Furthermore, we train models on misogyny datasets and test their classification of other abusive phenomena.

### 6.4.1 Experimental Setup

Basically, we use the same system as in the misogyny detection experiment in Section 6.3. We employ two classifiers: a Linear Support Vector Classifier (L SVC) and a Long Short-Term Memory (LSTM) architecture with additional features extracted from HurtLex. Our motivation is that L SVC has a higher degree of interpretability, while deep learning is capable of better generalization. Furthermore, HurtLex, being a domain-neutral lexicon, is used as an aid for transferring knowledge between datasets with different domains. In addition to these systems, we also build a BERT-based model, which is reported as the best model in generalizing different tasks of abusive language detection [Swamy et al., 2019]. All these systems are trained and optimized with similar approach, as explained in Section 6.3.

This experiment is restricted to English datasets, namely the two collection AMI datasets from AMI IberEval and AMI EVALITA, and three other related datasets, Waseem [Waseem and Hovy, 2016], HatEval [Basile et al., 2019], and OffensEval [Zampieri et al., 2019b]. Based on the description of each dataset, we assume that the Waseem and HatEval datasets are partly related to AMI topic-wise (sexism in Waseem and hate speech toward women in HatEval), while OffensEval has a very different and broader focus on offensive language.

The **OffensEval** corpus, also known as Offensive Language Identification Dataset (OLID [Zampieri et al., 2019a]) is a collection of 14,200 English tweets where abuse is represented and annotated according to a hierarchical framing for the following dimensions:

presence of offensiveness (binary labels OFF vs NOT, Subtask A), offensive type (binary labels TIN and UNT for targeted vs not targeted offenses, Subtask B), target type (labels IND, GRP and OTH for individual, group or other types of target). Table 6.9 shows the label distribution for the three layers. The data were collected by filtering Twitter with keywords for topics on which significant amount of offensive language was observed (e.g., MAGA, antifa) as well as patterns correlated to direct insults (e.g., “she is”, “you are”). The dataset was annotated by two to three annotators per instance, reporting a relatively high agreement (.83 Fleiss’ kappa on a trial set of 21 tweets). Notice that the class distribution for all the layers is very imbalanced, as the authors claim that did not alter the natural distribution resulting from the adopted data collection criteria.

In this work, we only use the “sexism” class of the Waseem dataset (which we will call “WaseemS” in the rest of the Chapter) and the “hate targeting women” subset of the HatEval dataset (which we will call “HatEvalM” in the rest of the Chapter), to observe the shared characteristics and relations between phenomena contained in these datasets with misogyny.

The main procedure for this experiment is to train a system in one dataset, and test it on the other datasets. In addition to the main experiment, we also experiment by combining two datasets as a training set to extend the coverage of the dataset, then test it on the test set of each dataset. Similarly to the previous experiment on the AMI task, this experiment is evaluated in terms of precision, recall, *F*-score, and accuracy. In case of the WaseemS dataset, where the partition of training and testing set is not specified, we split randomly the dataset in a 70%/30% proportion for training and testing, respectively.

#### 6.4.2 Results

Table 6.10 shows the full results of cross-domain classification with five different classifiers on five different datasets. We evaluate the models’ performance by using standard evaluation metric including precision, recall, macro F-score, and accuracy. The systems are based on LSVC and LSTM either with and without HurtLex, and also BERT. Datasets were chosen based on their relation with misogyny phenomena, where HatEvalM contains a similar phenomena (hate speech towards women), WaseemS covers a related phenomena (sexism), and OffensEval has a quite different focus, related to offensive language in general. Based on the description of each dataset, AMI IberEval, AMI EVALITA, and HatEvalM were collected and annotated with the same approach. Based on our manual investigation on these three datasets, we found duplicate instances across the collections. We identified 489 tweets in the EN-HatEval training set identical to tweets in the EN-AMI IberEval test set and 636 tweets in the EN-AMI EVALITA test set. 656 duplicated tweets are also shared between the EN-AMI EVALITA training set and the EN-AMI IberEval test set. For cross-domain classification purposes, we excluded these duplicates from the training sets. The final HatEvalM training set contains 3,355 tweets, while the training set of EN-AMI EVALITA consists of 3,344 tweets.

The underlined numbers in the table indicate the basic classification setting, where a system is trained and tested on the same dataset. Overall, the deep learning models (LSTM and BERT) achieved better performance than traditional classifiers such as

LSVC in the cross-domain classification setting (16 out of 18 runs) in terms of  $F_1$ -score. Specifically, both models almost always obtain a better score in term of recall, resulting often in a better  $F_1$ -score. LSVC obtained better results than LSTM and BERT only when the system is tested on the WaseemS dataset. Meanwhile, the comparison between BERT and LSTM shows that BERT has a better performance when tested on EN-IberEval, EN-EVALITA, and WaseemS, while LSTM outperforms BERT when tested on HatEvalM and OffensEval. The results also indicate that our systems obtain lower results in most of out-domain settings with respect to in-domain. An exception is when LSVC is trained on WaseemS and tested on HatEvalM, where it obtained the highest performance compared to the other runs. The lowest result was obtained when our systems are tested on OffensEval, the dataset which has the most different focus from misogyny phenomena. The final important finding is that the use of HurtLex boosts the systems performance, both LSVC and LSTM. Most of the improvement was measured in the recall score.

Table 6.11 depicts the results of an additional experiment where we combined the training sets of two datasets at a time, to augment the coverage. We compared the classification results of this setting with the basic setting, where only the original training set is used. The experiment results show that a performance improvement is measured on all test sets, except for OffensEval. When systems are tested on AMI EVALITA, almost all the additional training sets succeeded to enhance the classification result, whether on  $F_1$ -score or accuracy. When tested on AMI IberEval, the performance improvement is only achieved when the in-domain training sets are added. On the contrary, the addition of out-domain training sets (WaseemS and OffensEval) was able to boost the system performance when tested on HatEvalM. When tested on WaseemS, the extra training set from OffensEval was the only one which could not improve the system performance. In the last experiment setting, testing on OffensEval, there was no additional training set able to enhance the system performance.

## 6.5 Cross-Lingual Automatic Misogyny Identification Experiments

In this section, We propose an experiment in cross-lingual automatic misogyny identification. We take advantage of the AMI task datasets, which contain tweets in three different languages: English, Spanish, and Italian.

### 6.5.1 Experimental Setup

In this cross-lingual classification experiment, we train models on one language and test it on datasets in a different language. Specifically, we build four systems:

1. **Linear Support Vector Classifier (LSVC).** With this classifier, we only use unigrams as features. Therefore, we need to translate the training set from the source language (the original language of the training set) to the target language

#### Linear Support Vector Classifier

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	.828	.629	.715	.804	-	-	-	-	.695	.305	.424	.815	.438	.962	.602	.461	.563	.113	.188	.728
EN-EVALITA	-	-	-	-	.584	.670	.624	.629	.621	.150	.242	.790	.442	.974	.608	.469	.553	.088	.151	.726
WaseemS	.892	.205	.333	.680	.559	.370	.445	.576	.874	.626	.730	.896	.503	.750	.602	.581	.526	.042	.072	.722
HatEvalM	.869	.537	.664	.788	.597	.665	.629	.639	.627	.150	.242	.790	.449	.973	.614	.483	.561	.096	.164	.727
OffensEval	.591	.484	.532	.668	.534	.639	.582	.577	.395	.261	.314	.746	.431	.990	.602	.442	.710	.479	.572	.800

#### Linear Support Vector Classifier and HurtLex

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	.822	.622	.708	.800	-	-	-	-	.684	.302	.419	.813	.442	.928	.605	.471	.532	.104	.174	.724
EN-EVALITA	-	-	-	-	.584	.652	.616	.626	.606	.155	.247	.789	.452	.970	.617	.490	.636	.117	.197	.735
WaseemS	.848	.276	.416	.698	.565	.417	.480	.584	.869	.632	.732	.897	.464	.955	.625	.514	.615	.067	.120	.728
HatEvalM	.851	.527	.651	.780	.592	.659	.624	.634	.618	.159	.253	.790	.456	.963	.620	.499	.650	.108	.186	.735
OffensEval	.569	.513	.539	.658	.521	.650	.578	.564	.391	.270	.320	.743	.429	.995	.600	.438	.707	.483	.574	.800

#### Long Short Term Memory

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	.746	.717	.732	.795	-	-	-	-	.426	.398	.412	.746	.444	.891	.593	.482	.372	.267	.311	.670
EN-EVALITA	-	-	-	-	.598	.652	.624	.638	.396	.110	.172	.764	.546	.827	.657	.635	.443	.146	.219	.711
WaseemS	.750	.286	.414	.685	.612	.498	.549	.624	.855	.697	.768	.906	.458	.957	.619	.502	.511	.096	.161	.722
HatEvalM	.678	.587	.629	.730	.569	.657	.610	.613	.275	.275	.275	.676	.484	.910	.632	.552	.389	.358	.373	.664
OffensEval	.611	.555	.582	.689	.561	.678	.614	.608	.285	.215	.245	.704	.433	.986	.602	.448	.746	.513	.607	.815

#### Long Short Term Memory and HurtLex

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	.701	.739	.719	.776	-	-	-	-	.403	.430	.416	.730	.441	.920	.596	.472	.338	.288	.311	.644
EN-EVALITA	-	-	-	-	.536	.763	.630	.587	.264	.362	.306	.632	.460	.960	.622	.505	.343	.425	.380	.613
WaseemS	.695	.290	.409	.674	.606	.478	.535	.617	.855	.676	.755	.902	.454	.958	.616	.495	.542	.108	.181	.726
HatEvalM	.745	.505	.602	.740	.617	.585	.600	.642	.328	.172	.225	.736	.517	.867	.649	.601	.414	.250	.312	.692
OffensEval	.612	.804	.695	.660	.593	.786	.676	.664	.292	.271	.281	.690	.428	.995	.599	.435	.712	.567	.631	.815

#### BERT

	EN-IberEval				EN-EVALITA				WaseemS				HatEvalM				OffensEval			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	.773	.915	.838	.862	-	-	-	-	.722	.229	.348	.808	.486	.913	.634	.554	.386	.142	.207	.698
EN-EVALITA	-	-	-	-	.716	.704	.661	.682	.645	.322	.430	.809	.504	.918	.651	.584	.444	.100	.163	.714
WaseemS	.864	.201	.327	.676	.589	.657	.621	.631	.846	.692	.761	.903	.532	.621	.573	.608	.406	.054	.096	.714
HatEvalM	.829	.802	.815	.858	.698	.670	.684	.715	.679	.328	.442	.815	.509	.957	.664	.590	.404	.088	.144	.709
OffensEval	.538	.748	.626	.588	.508	.811	.624	.551	.407	.574	.476	.718	.429	.995	.599	.437	.815	.550	.657	.840

Table 6.10: Result of Cross-domain Automatic Misogyny Identification Experiment.

Test on AMI EVALITA	LSVC				LSTM				BERT			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
AMI EVALITA Only	.584	.670	.624	.629	.658	.489	.561	.648	.704	.661	.682	.716
+ AMI IberEval	.626	.639	.632	.658	.648	.504	.567	.646	.697	.639	.667	.706
+ HatEvalM	.598	.663	.629	.640	.549	.726	.626	.600	.628	.713	.668	.674
+ WaseemS	.603	.587	.595	.632	.595	.519	.555	.616	.678	.696	.687	.708
+ OffensEval	.547	.654	.596	.592	.559	.680	.614	.606	.586	.667	.624	.630
Test on AMI IberEval	LSVC				LSTM				BERT			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
AMI IberEval Only	.828	.629	.715	.804	.746	.717	.732	.795	.773	.915	.838	.862
+ AMI EVALITA	.837	.636	.723	.810	.808	.580	.675	.782	.795	.919	.853	.876
+ HatEvalM	.845	.636	.726	.813	.664	.746	.702	.754	.842	.696	.762	.831
+ WaseemS	.836	.488	.616	.763	.776	.601	.677	.777	.824	.841	.832	.868
+ OffensEval	.728	.576	.643	.751	.712	.785	.746	.792	.788	.774	.781	.831
Test on HatEvalM	LSVC				LSTM				BERT			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
HatEvalM Only	.449	.973	.614	.483	.484	.910	.632	.552	.479	.981	.643	.539
+ AMI EVALITA	.444	.971	.609	.472	.478	.857	.613	.543	.541	.908	.678	.635
+ AMI IberEval	.444	.954	.606	.475	.466	.925	.620	.520	.491	.974	.652	.561
+ WaseemS	.461	.979	.627	.507	.481	.968	.643	.545	.488	.971	.650	.557
+ OffensEval	.461	.958	.623	.508	.463	.934	.620	.514	.450	.990	.619	.483
Test on WaseemS	LSVC				LSTM				BERT			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
WaseemS Only	.874	.626	.730	.896	.855	.697	.768	.906	.826	.749	.785	.909
+ AMI EVALITA	.874	.633	.735	.898	.854	.652	.739	.897	.847	.674	.750	.899
+ AMI IberEval	.874	.649	.745	.901	.830	.706	.763	.902	.750	.775	.762	.892
+ HatEvalM	.875	.632	.734	.897	.819	.703	.757	.899	.806	.725	.763	.899
+ OffensEval	.797	.636	.707	.882	.724	.701	.712	.873	.839	.690	.757	.901
Test on OffensEval	LSVC				LSTM				BERT			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
OffensEval Only	.710	.479	.572	.800	.746	.513	.607	.815	.694	.679	.686	.827
+ AMI EVALITA	.710	.388	.501	.785	.752	.492	.595	.813	.708	.617	.659	.822
+ AMI IberEval	.703	.404	.513	.786	.648	.567	.604	.793	.734	.621	.673	.831
+ HatEvalM	.692	.383	.493	.780	.806	.450	.578	.816	.888	.329	.480	.801
+ WaseemS	.709	.346	.465	.778	.752	.392	.515	.794	.778	.350	.483	.791

Table 6.11: Result of Experiment by Combining Two Datasets in Cross-Domain Classification of Misogyny.

(the language of the test set). We used Google Translate<sup>20</sup> as translation service.

**2. Long Short Term Memory (LSTM) with Monolingual Word Embedding.**

We implement a LSTM architecture with monolingual word embeddings as word representation. The pre-trained word embeddings provided by FastText are used to initialize the embedding layer of the network, followed by LSTM layers consisting of 64 units. The output of the LSTM network is connected to a dense layer (16 units) with ReLu activation function. The last layer is a dense layer with sigmoid activation which provides the final prediction of the label. Similar to LSVC system, in this setting we also translate the training set to the target language by using Google Translate.

**3. Long Short Term Memory (LSTM) with Multilingual Word Embedding.**

We employ the multilingual word embeddings developed by Facebook research group published as MUSE (Multilingual Unsupervised or Supervised word Embeddings), a supervised word embedding model aligned across 30 languages [Lample et al., 2018]. With this representation, we do not need to translate the training set to the target language. The rest of the configuration of this model is the same as the LSTM with monolingual word embedding.

**4. Joint-Learning Model with Multilingual Word Embeddings.** We also propose a joint-learning model with a focus on transfer knowledge between languages in a cross-lingual classification setting. Figure 6.2 shows the full process of how the data is transformed and learned by architecture. We start the process by creating a bilingual dataset automatically by using Google Translate. The training and test set are translated in both directions (source to target language and target to source language), then used to train two LSTM-based models in two languages independently. We concatenate the output of the two models before the final layer (output layer), which provides the final prediction. In this architecture, we expect to reduce some of the noise from the translation while keeping the original structure of the training and test set. The configuration of each single LSTM architecture is the same to the two previous models (monolingual and multilingual LSTM).

**5. Joint-Learning Model with Multilingual Word Embedding and HurtLex.**

Finally, we experiment by adding HurtLex to the joint-learning model. We concatenate a feature representation obtained with the lexicon to the input of each LSTM networks in both languages. In this architecture, HurtLex provides a 17-dimension vector, i.e., a one-hot encoding of the word presence in each of the lexicon categories.

**6. BERT with Multilingual Model.** We also propose a model similar to LSTM with multilingual embedding, by substituting LSTM and MUSE with a pre-trained multilingual BERT model. In particular, we use the `bert-multi-cased` model. The rest of the configuration is the same as the BERT model in the previous experiment,

---

<sup>20</sup><https://translate.google.com/>

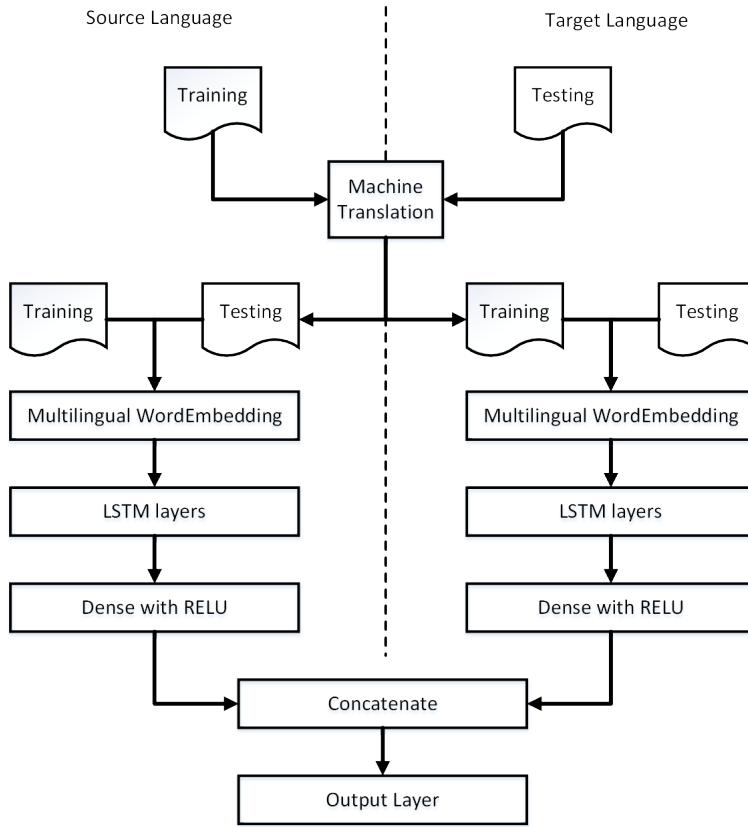


Figure 6.2: Joint-Learning Model Architecture.

with a dense layer with ReLU activation function, followed by a dense layer with sigmoid activation function as the final layer. This model is optimized by using Adam optimizer and trained on different combinations of batch size (16,32,64) and epochs (1-5).

7. **Joint-Learning BERT with Multilingual Model.** Adopting a similar idea to our joint-learning LSTM model, we also propose to build a model by substituting the LSTM with the BERT `bert-multi-cased` model. This model is optimized and trained with the same configuration as BERT with monolingual model.

### 6.5.2 Results

Table 6.12 depicts the results of cross-lingual automatic misogyny identification experiments, where we train a system on one language and test it on another language. We report the result in several standard evaluation metrics including precision, recall, macro F-score, and accuracy. However in this analysis, we only focus on the system performance based on  $F_1$  and accuracy score. We mark the highest  $F_1$  in each run in bold face, and the highest accuracy by underline. We start the analysis of the result by focusing on the comparison

between LSVC and LSTM with monolingual embeddings, where both systems only rely on the use of machine translation to deal with the multilingual environment. We found out that the use of traditional models does not always give a lower performance than deep learning. LSCV achieved better performance in some of the settings, including *ES-IberEval* → *EN-IberEval*, *ES-IberEval* → *EN-EVALITA*, *IT-EVALITA* → *EN-EVALITA*, *EN-IberEval* → *ES-IberEval*, and *EN-EVALITA* → *ES-IberEval*. However, the LSVC performance is much lower compared to LSTM in settings where the translation from English to Italian is needed.

The second analysis focuses on the performance comparison between LSTM with monolingual embedding and machine translation, and LSTM with multilingual embeddings where no translation is needed. Surprisingly, LSTM with monolingual embeddings are able to outperform LSTM with multilingual embeddings, which use pre-trained word embeddings that are specifically developed for cross-lingual learning. In terms of  $F_1$ -score, monolingual LSTM has a better performance in 6 out of 10 run settings, while based on accuracy it outperformed LSTM with multilingual embeddings in 9 out of 10 settings. However, a different outcome emerges when we compare LSTM with monolingual embedding against the multilingual BERT model. BERT tends to have more robust performance on two languages, namely English and Spanish, but not on Italian.

The third analysis focuses on the comparison between LSTM with multilingual embedding, the joint-learning model with multilingual embeddings, and the respective BERT-based variants, combining the machine translation ability and multilingual embeddings. In terms of accuracy, joint-learning always outperforms LSTM with multilingual embedding in all settings. Both systems achieve the best performance in half the runs, in terms of  $F_1$ -score. However, the overall results show that joint-learning has a more robust performance across the settings. We observe that in some settings, including *EN-IberEval* → *ES-IberEval*, *EN-IberEval* → *IT-EVALITA*, and *EN-EVALITA* → *IT-EVALITA*, LSTM with multilingual embeddings experienced a big drop in performance. With BERT, our joint-learning model also performs consistently better than the multilingual BERT model in term of  $F_1$ -score. Also in term of accuracy, the joint-learning models outperform the normal multilingual BERT configuration in 7 out of 10 runs.

The last analysis is a comparison between using and not using HurtLex in the joint learning model with multilingual embeddings. Based on the experimental results, the use of HurtLex succeeded to improve the model performance in term of  $F_1$ -score.

## 6.6 Discussion

In this section, we present the discussion and analysis of the results of all our proposed experiments. The discussion is organized in three subsections reflecting the different experimental settings, namely automatic misogyny identification (subtask A and subtask B of the AMI challenge), relationship between misogyny and other abusive phenomena, and cross-lingual classification.

Linear Support Vector Classifier																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.675	.771	.720	.704	.198	.135	.160	.277
EN-EVALITA	-	-	-	-	-	-	-	-	.640	.545	.589	.620	.205	.121	.152	.309
ES-IberEval	.409	.477	.441	.528	.566	.704	.524	.610	-	-	-	-	.621	.621	.621	.612
IT-EVALITA	.376	.686	<b>.486</b>	.434	.492	.739	.591	.529	.568	.542	.555	.566	-	-	-	-
Long-Short Term Memory with Monolingual Embedding																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.672	.745	.706	.691	.458	.606	<b>.521</b>	.431
EN-EVALITA	-	-	-	-	-	-	-	-	.712	.246	.366	.575	.448	.295	.356	.453
ES-IberEval	.406	.237	.299	.568	.676	.404	.506	.637	-	-	-	-	.658	.846	<b>.740</b>	.696
IT-EVALITA	.339	.519	.410	.417	.536	.557	.546	.574	.589	.598	.593	.591	-	-	-	-
Long-Short Term Memory with Multilingual Embedding																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.644	.157	.252	.536	.324	.065	.102	.454
EN-EVALITA	-	-	-	-	-	-	-	-	.554	.523	.538	.551	.257	.094	.137	.397
ES-IberEval	.376	.933	.536	.371	.481	.885	.623	.508	-	-	-	-	.571	.922	.706	.606
IT-EVALITA	.299	.558	.389	.317	.428	.315	.363	.491	.544	.774	.639	.563	-	-	-	-
Joint Learning Model with Multilingual Embedding																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.749	.648	.695	.716	.472	.502	.487	.458
EN-EVALITA	-	-	-	-	-	-	-	-	.676	.407	.508	.607	.584	.102	.173	.503
ES-IberEval	.396	.516	.448	.504	.572	.576	.574	.607	-	-	-	-	.587	.920	.717	.628
IT-EVALITA	.423	.283	.339	.570	.566	.380	.455	.581	.409	.299	.409	.569	-	-	-	-
Joint Learning Model with Multilingual Embedding and HurtLex																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.624	.868	.726	.673	.480	.506	.492	.466
EN-EVALITA	-	-	-	-	-	-	-	-	.724	.542	<b>.620</b>	.668	.553	.287	.377	.516
ES-IberEval	.395	.643	.490	.478	.530	.702	.604	.577	-	-	-	-	.637	.842	.725	.673
IT-EVALITA	.372	.686	.483	.427	.486	.911	<b>.633</b>	.512	.622	.448	.521	.588	-	-	-	-
Multilingual BERT																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.648	.641	.645	.647	.226	.146	.177	.306
EN-EVALITA	-	-	-	-	-	-	-	-	.652	.451	.533	.605	.393	.324	.356	.398
ES-IberEval	.463	.647	.540	.570	.704	.491	.579	.671	-	-	-	-	.650	.443	.527	.593
IT-EVALITA	.277	.357	.312	.386	.528	.626	.573	.571	.536	.708	.610	.548	-	-	-	-
BERT Joint Learning																
	EN-IberEval				EN-EVALITA				ES-IberEval				IT-EVALITA			
	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc	P	R	F <sub>1</sub>	Acc
EN-Ibereval	-	-	-	-	-	-	-	-	.710	.754	<b>.731</b>	.723	.457	.115	.184	.477
EN-EVALITA	-	-	-	-	-	-	-	-	.754	.525	.619	.678	.574	.281	<b>.378</b>	.525
ES-IberEval	.499	.657	<b>.567</b>	.609	.589	.733	<b>.653</b>	.642	-	-	-	-	.639	.488	.554	.597
IT-EVALITA	.286	.445	.348	.350	.517	.722	.603	.562	.575	.810	<b>.673</b>	.607	-	-	-	-

Table 6.12: Result of Cross-lingual Automatic Misogyny Identification Experiment.

Features	Accuracy	$\Delta$ (delta)
All features	91.32	-
- Swear words count	89.81	1.51
- Swear words presence	90.50	0.82
- Hashtags presence	90.63	0.69
- Links count	85.40	5.92
- Sexist slurs presence	75.90	15.42
- Women words presence	73.83	17.49

Table 6.13: Ablation test result of the best system on English AMI IberEval.

### 6.6.1 Automatic Misogyny Identification Task

In order to get a deeper insight, we performed an ablation test on our best models on the AMI IberEval dataset, removing each feature to measure the impact on the system performance. Table 6.13 presents the ablation test results of our English AMI IberEval, which shows that *sexist slurs* and *women words presence* are the most predictive features on this task. These figures confirm the findings of the lexical distribution analysis in Section 6.2, where sexist slurs were found to be mainly used in misogynistic instances. Moreover, the importance of the women-related words feature indicates that the detection of target gender is highly informative for the detection of misogyny.

Similar to English part, our system was also top ranked in the Spanish AMI IberEval task. While the best system for English is a SVM classifier with RBF kernel, for Spanish the best system is a SVM with linear kernel including several features such as bags of words (1-gram to 3-grams), bags of hashtags, bags of emojis, sexist slurs presence, woman words presence, and the presence of some HurtLex categories, including words related to female genitalia (ASF), words related to prostitution (PR), words related to cognitive disabilities and diversity (DDF), words related to physical disabilities and diversity (DDF), and words related to male genitalia (ASM). In summary, these results show that HurtLex helps informing the model, but only some of its categories are actually related to the misogynistic action. As shown in Table 6.14, bags of words are the most informative feature of this model. Therefore, we decides to conduct a further analysis by extracting the SVM classifier weights when only token n-grams are used as features, to obtain a clearer picture of what is the most predictive features in Spanish AMI IberEval task. Table 6.15 shows the top ten features for the Spanish AMI IberEval task based on the SVM weight. The use of sexist slurs such as *zorra* (bitch), *perra* (bitch/slut), *guarra* (slut), and *coño* (pussy/cunt) is a clear signal of misogynistic content. This finding is consistent with the results on the English dataset, where sexist slurs is also the most important feature to detect misogyny instances. It also confirms that the use of swear words has an impactful role, specifically in this task and dataset.

Our BERT models performed very well on both English and Italian AMI EVALITA, outperforming the best systems on the respective tasks. In English AMI EVALITA, it achieved better performances than the best system participating in the shared task, with

Features	Accuracy	$\Delta$ (delta)
All features	81.47	-
- Bag of words	65.98	15.49
- Bag of hashtags	81.40	0.07
- Bag of emojis	80.44	1.03
- Sexist slurs presence	80.85	0.62
- Women words presence	81.13	0.34
- ASF presence	81.27	0.2
- ASM presence	81.13	0.34
- DDF presence	81.13	0.34
- DDP presence	81.27	0.2
- PR presence	81.13	0.34

Table 6.14: Ablation test result of the best system on Spanish AMI IberEval.

Setting No.	Offensive	
	Features	Coefficient
1.	zorra	2.143
2.	perra	1.499
3.	callate	1.427
4.	hija	1.061
5.	guarra	1.028
6.	callate puta	0.935
7.	mi polla	0.905
8.	callate perra	0.904
9.	tu cono	0.806
10.	mujer	0.706

Table 6.15: Top ten features based on SVM weight on Spanish AMI IberEval task.

an accuracy of 71.60 (the best system obtained an accuracy of 70.40). Concerning the Italian part, our BERT model also able to surpass the competition best system, obtaining 84.80 in accuracy, slightly higher than best system with 84.40 in accuracy.

The results on AMI subtask A show that traditional models obtain a good performance, especially on the IberEval tasks, with the advantage of being far more transparent than deep learning. However, these models fail to have a stable performance across different datasets, as highlighted by their low performance on the EVALITA tasks. Here, BERT achieves the best results, both in English and Italian. The overall result signifies that deep learning approaches have a more stable performance on both shared tasks, where they always obtain a competitive results. We also notice that SVM with RBF kernel always obtains a good performance when applied to English datasets, but a much lower performance on the other languages. Similarly, our BERT model also tend to have better performance when applied to English. Finally, SVM with linear kernel tends to achieve comparably better results when applied to languages other than English.

**Error Analysis.** We conducted a manual error analysis on the misclassified instances, to explore the most common pitfalls in detecting misogyny. Our investigation found that at least five issues contribute to the difficulties of this task:

1. **Presence of swear words.** We encountered a lot of “bad words” in the dataset of this shared task for both English and Italian. In case of abusive context, the presence of swear words can help to spot abusive content such as misogyny. However, they could also lead to false positives when the swear word is used in a casual, not offensive context [Malmasi and Zampieri, 2018, Van Hee et al., 2018, Nobata et al., 2016]. Consider the following two examples containing the swear word “bitch” in different contexts:

1.  Im such a fucking cunt bitch and i dont even mean to be goddammit
2.  Bitch you aint the only one who hate me, join the club, stand in the corner, and stfu.

In Example 1, the swear word “bitch” is used just to arouse interest/show off, thus not directly insulting the other person. This is a case of *idiomatic swearing* [Pinker, 2007]. In Example 2, the swear word “bitch” is used to insult a specific target in an abusive context, an instance of *abusive swearing* [Pinker, 2007]. The ambiguity of swearing context is still a problem which contributes to the difficulties of this task and abusive language detection task in general.

2. **Reported speech.** Tweets may contain misogynistic content as an indirect quote of someone else’s words, such as in the following example:

3.  Quella volta che mia madre mi ha detto quella cosa le ho risposto "Mannaggia! Non sarò mai una brava donna schiava zitta e lava! E adesso?!" Potrei morire per il dispiacere.  
→ *That time when my mom told me that thing and I answered “Holy s\*\*t! I will never be a good slave who shuts up and cleans! What now?”*

According to task guidelines this should not be labeled as a misogynistic tweet, because it is not the user himself who is misogynistic. Therefore, instances of this type tend to confuse a classifier based on lexical features.

3. **Irony and world knowledge.** In Example 3, the sentence “Potrei morire per il dispiacere.”<sup>21</sup> is ironic. Humor is very hard to model for automatic systems — sometimes, the presence of figurative language even baffles human annotators. Moreover, external world knowledge is often required in order to infer whether an utterance is ironic [Wallace et al., 2014].
4. **Preprocessing and tokenization.** In computer-mediated communication, and specifically on Twitter, users often resort to a language type that is closer to speech, rather than written language. This is reflected in less-than-clean orthography, with forms and expressions that imitate the verbal face-to-face conversation.

4.  @ ██████████ @ ██████████ @ ██████████ x me glob prox2aa  
colpiran tutti incluso nemicinterno.. esterno colpopiúduro saràculogrande che bevetropvodka  
e inoltre x questiondisoldi progetta farmezzofallirsudfinitestampe: ciò nnvàben xrcheinde-  
bolis  
→ 4 me glob next2aa will hit everyone included internalenemy.. external harderhit  
willbebigass who drinkstoomuchvodka and also 4 mattersofmoney isplanning tomakethe-  
southfailwithprintings: dis notgood causeweaken

In Example 4, preprocessing steps like tokenization and stemming are particularly hard to perform, because of the lack of spaces between one word and the other and the confused orthography. Consequently all the classification pipeline is compromised and error-prone.

5. **Gender of the target.** As defined in the Introduction, we know that misogyny is a specific type of hateful language, targeting women. However, detecting the gender of the target is a challenging task in itself, especially in Twitter datasets.

5.  @realDonaldTrump shut the FUCK up you infected pussy fungus.

6.  @TomiLahren You're a fucking skank!

Both examples use bad words to abuse their targets. However, the first example is labeled as not misogyny since the target is Donald Trump (man), while the second example is labeled as misogyny with the target Tomi Lahren (woman).

On subtask B, overall results indicates that treating subtask B as an independent multiclass classification is more effective than handling it as a pipeline classification task, as a sequential task to the results of subtask A, which is a mostly used approach by all AMI task participants. We argue that in a pipeline classification scenario, the results on subtask B would be highly dependant on the system performance in subtask A. In

---

<sup>21</sup>Translation: I could die for heartbreak.

addition, we undertook a deeper analysis to get more insight regarding common issues in task B classification. We produced a confusion matrix of the classification results based on the best performing system on each dataset, consisting of two classification tasks, namely misogyny behaviour classification (in Figure 6.3) and target of misogyny classification (in Figure 6.4). Based on Figure 6.3, we can see that detecting derailing class is the most challenging aspect of this task. On English, our systems were able to classify discredit quite well. In the EN-AMI EVALITA dataset, the dominance class was also classified quite well, as well as the sexual harassment class. On the Italian dataset, our system got a promising result, balanced across the classes. On Spanish, stereotype and derailing were the only two classes which are difficult to be detected. For the classification of target of misogyny, most of our systems only performed well in detecting the active class. We argue that this result is influenced by the label distribution on the gold standard where most of misogyny tweets are labeled as active.

In addition to the analysis of the confusion matrix, we also performed a manual error analysis on the dataset to find other difficulties of this particular task. After a manual inspection of the data, it emerged that there is no clear demarcation line between one category and the others in classification of misogyny behaviour task. The single label introduced for the misogyny behaviour classification task forces tweets to only have one label, representing the dominant category. We argue that it is possible for one tweet to express more than one misogyny behaviour phenomenon. For example, *dominance* and *discredit* are both highly correlated to high presence of swearing, with varying focus (e.g., the agent (man) vs. the wounded part, the target (woman)). Similarly, *stereotype & objectification* is not so conceptually distant from *sexual harassment*, due to a strong use of language referring to sexual body parts or vulgar sexual practices. These insights are reflected in the examples we provide in the following.

STEREOTYPE & OBJECTIFICATION (EN): No girl is even capable of developing morals until they get the slut fucked out of them. Welcome to my generation

DOMINANCE (EN): Bad girls get spankings

DERAILING: Women want u to automatically believe women who scream rape they don't understand our position....

SEXUAL HARASSMENT & THREATS OF VIOLENCE (EN): @ [REDACTED] hey bitch  
wassup bitch suck my dick bitch

DISCREDIT (EN): @ [REDACTED] Fuck that money whore @HillaryClinton Too stupid to know consensual touching or grabbing is not assault. Only @ChelseaClinton is dumber

STEREOTYPE & OBJECTIFICATION (ES): Que cuza antes la calle, una mujer inteligente o una tortuga vieja? Una tortuga vieja porque las mujeres inteligentes no existen ...

DOMINANCE (ES): "Voy a enseñarle a esta perra como se trata a un hombre"

LMAO IN LOVE WITH EL TITI 😂

SEXUAL HARASSMENT & THREATS OF VIOLENCE (ES): @ [REDACTED] Me gustaría abrirte las piernas y clavarte toda mi polla en tu culo.

DISCREDIT (ES): Porque ladra tanto mi perra? La puta madre cállate un poco

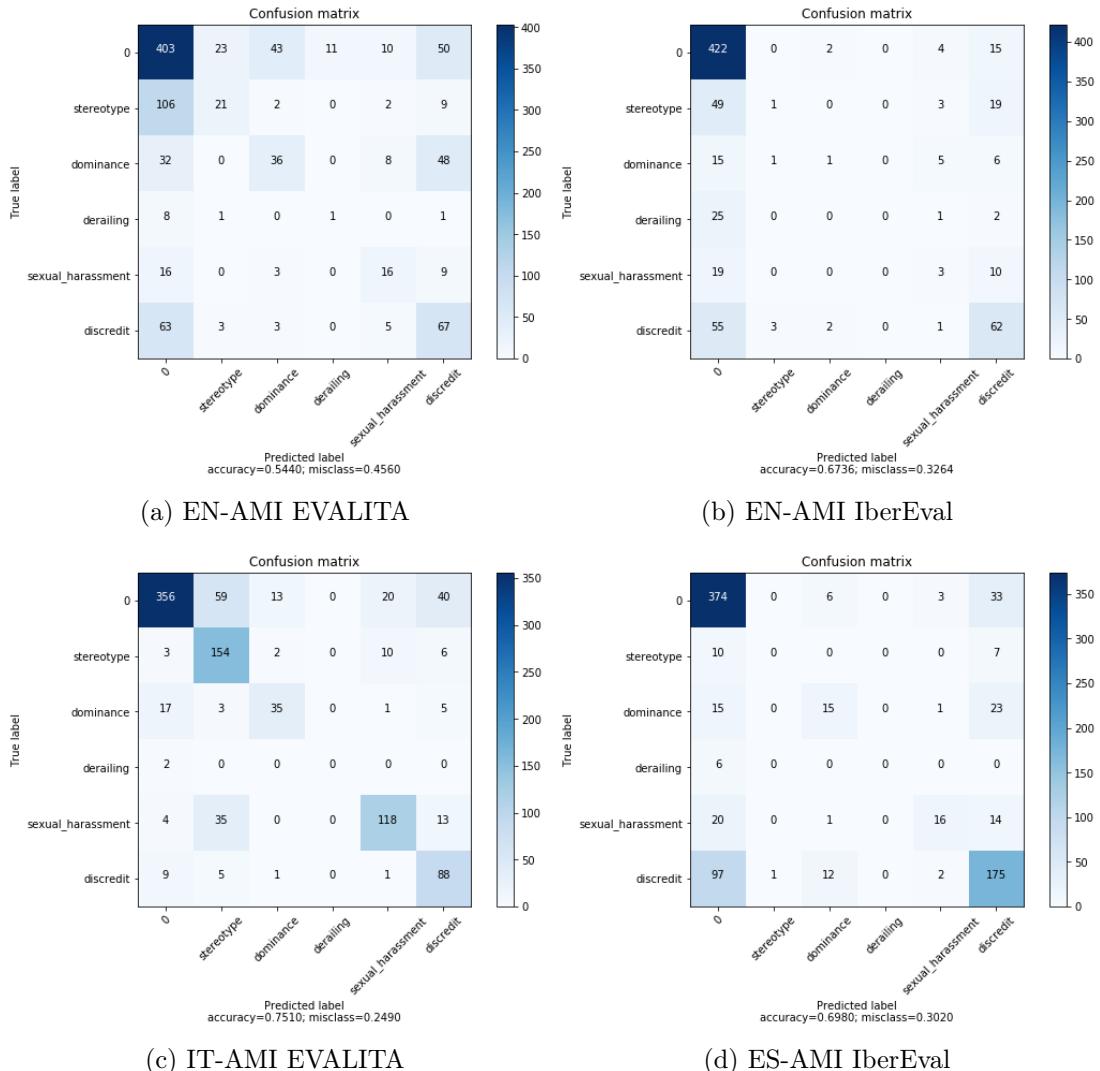


Figure 6.3: Misogynistic Behaviour Classification: Confusion Matrix

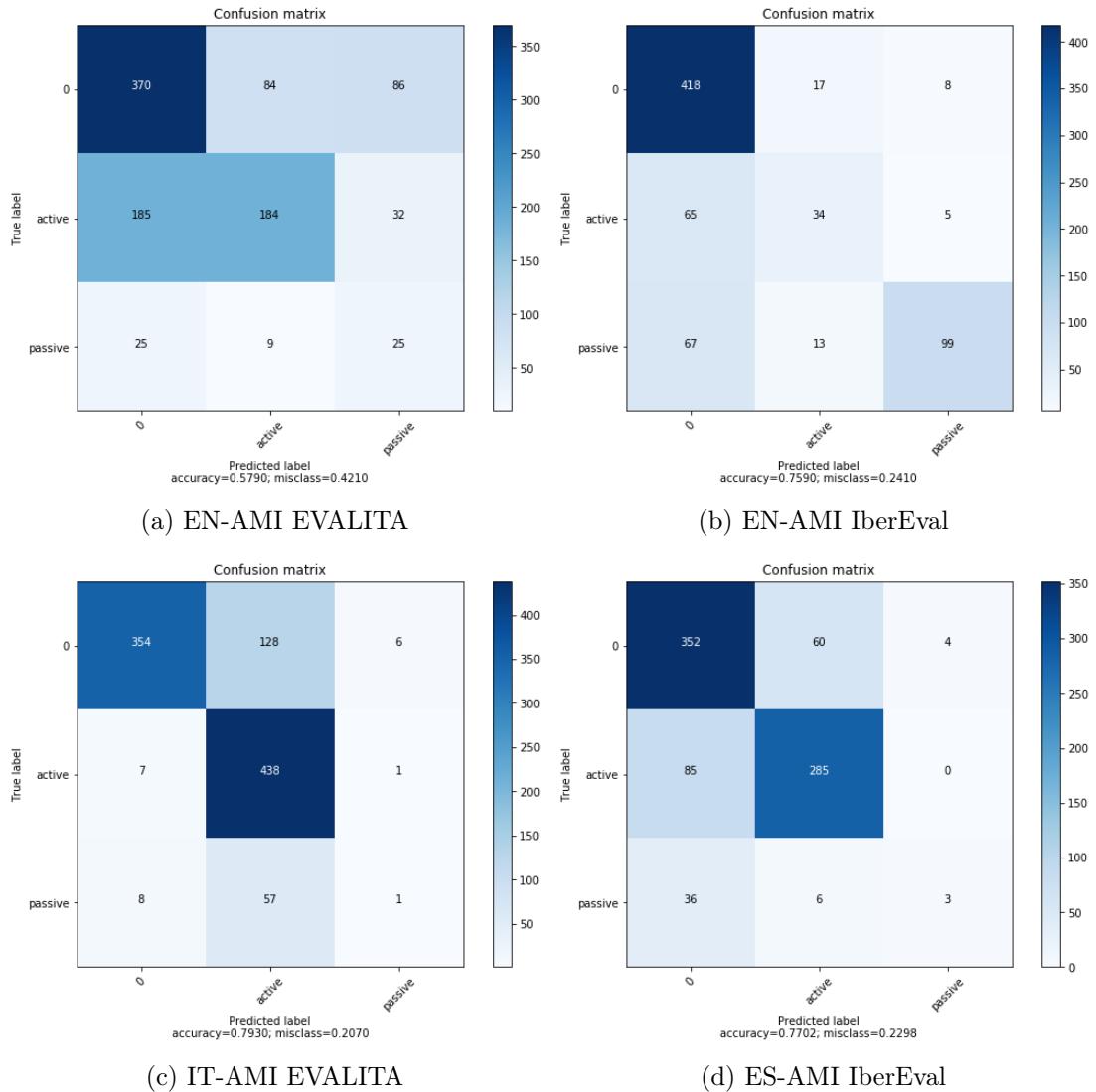


Figure 6.4: Target of Misogyny Classification: Confusion Matrix

### 6.6.2 Relationship between Misogyny and other Abusive Phenomena

The overall results of cross-domain classification of misogyny show that deep learning approaches have a better performance in transferring knowledge between different datasets, including the AMI task datasets. The results also show that BERT is the most robust model in cross-domain setting, achieving a stable result in all the experimental settings. This result is in line with the findings in [Swamy et al., 2019], which experimentally found that BERT has a better capability than other models to generalize over different abusive language detection tasks. In addition, the use of a lexical resource related to abusive language such as HurtLex is able to improve the system performance, providing domain independent information for the systems. The experimental results also show that training a system on a more general abusive language dataset (i.e., OffensEval), and testing it in more specific dataset (i.e., AMI datasets), still obtains a reasonable performance compared to the in-domain setting. On the contrary, when we train the system on more specific datasets such as the AMI datasets, and test it on more general dataset (OffensEval), obtained very poor results. This is consistent with the result obtained by Pamungkas and Patti [2019] on cross-domain classification of abusive language. We also found that system trained on OffensEval is more robust than when training on other datasets, including WaseemS. The results for the WaseemS dataset, which is related to AMI topic-wise, indicate that having similar topic does not guarantee to get competitive results when tested on AMI datasets. We argue that the performance on cross-domain or cross-dataset classification is not only influenced by the topical focus of the datasets, but also, and heavily so, by data collection approaches and annotation procedures. This finding is also supported by Wiegand et al. [2019], who notices how the WaseemS dataset contains biases, which is problematic when the dataset is used for cross-domain classification, since models are not be able to generalize sufficiently. We also found that when the system is trained on misogyny datasets (both EVALITA and IberEval collection) and tested on WaseemS, they experience a bigger drop than when tested on HatEvalM. This result shows how hate speech toward women has a stronger relation to misogyny than sexism. Interestingly, this result is also in line with recent philosophical accounts of misogyny [Manne, 2017, Richardson-Self, 2018], theorizing *misogyny* and *sexism* as related but *distinct* mechanisms that enforce the norms of patriarchy [Manne, 2017], and arguing for considering only misogynistic speech as a specific kind of hate speech [Richardson-Self, 2018].

The dataset augmentation experiment show that augmenting the data training coverage by adding external data only works when the additional data share similar targets or topics, as observed on the AMI, HatEvalM, and WaseemS datasets. Adding a dataset which has different phenomena and topical focus (in this case OffensEval), failed in enhancing the system performance. We argue that additional training data from the loosely related dataset could not be able to extend the coverage of the dataset, which would help to build a robust system, but instead introduces noise which hurts the systems performance. Again, these results confirm that hate speech toward women (modeled in HatEvalM) provide more valuable additional data than sexism (WaseemS), reaffirming that hate speech targeting women is more strongly related to misogyny than sexism.

Notice that we experimented with the *offensive* label provided for tweets in the OffensEval corpus. However, the annotation of target in this corpus (group vs. individual, see Table 6.9) may also be a valuable layer to explore in the future to see how well this information can transfer to and from the AMI dataset, where we have an analogous layer of annotation devoted to identify the nature of the target (active vs. passive).

### 6.6.3 Cross-lingual Automatic Misogyny Identification

Based on the cross-lingual experimental results, we argue that the performance of LSVC models heavily relies on the translation quality, since they only use the token n-grams as their main feature to estimate the probability of a tweet to contain misogyny. In this case, we observe that the machine translation performance is still not good enough for translating to Italian from other languages. Deep learning models have the capability to update their feature representation (word embedding matrix), optimized on the train set during the training phase, giving more flexibility than only relying on the translation result. Therefore, LSTM has a more stable performance across all cross-lingual settings.

We also observe that the vocabulary size of the pre-trained word embedding is a possible cause for the low performance of multilingual embeddings. Indeed, the pre-trained models from FastText contain 2,000,000 vocabulary items for monolingual embeddings and only 200,000 for multilingual embeddings. The low vocabulary coverage leads to a higher number of out of vocabulary (OOV) words, which causes inaccuracies in the word representation matrix.

Both joint-learning models (LSTM-based and BERT-based) performed better than their standard counterparts (LSTM with multilingual embeddings and Multilingual BERT). The better results obtained by our joint-learning model confirms our idea that allowing the network to learn both the original and translated text is able to reduce some of the noise from the translation, while keeping the original structure of the training set. This in turn enables the system to deal with the issues of low vocabulary coverage of multilingual embeddings and quality of the translation result.

Regarding to the performance improvement when exploiting HurtLex, further investigation proved that there is a significant improvement on the recall side when the model includes the HurtLex features, meaning that the system is able to reduce the number of false negatives in the prediction and to detect more misogynistic instances. We argue that HurtLex has a significant impact to inform the models about specific hurtful words, which are possibly not always translated correctly by the machine translation service, or not covered by multilingual word embeddings. For example, offensive words toward women such as “hoe” and “skank” are mis-translated by machine translation (hoe (English) → azada (Spanish)) and (skank (English) → skank (Italian)).

## 6.7 Summary

In this chapter, we present an in-depth study on the phenomena of misogyny in three languages, English, Spanish and Italian, by focusing on three main objectives. Firstly,

we investigate the most important features to detect misogyny and the issues which contribute to the difficulty of misogyny detection, by proposing a novel system and conducting a broad evaluation on this task. Secondly, we study the relationship between misogyny and other abusive language phenomena, by conducting a series of cross-domain classification experiments. Finally, we explore the feasibility of detecting misogyny in a multilingual environment, by carrying out cross-lingual classification experiments. Our system succeeded to outperform all state of the art systems in all benchmark AMI datasets both subtask A and subtask B. We experimentally found that swear words have an impactful role in misogyny phenomena, especially some specific sexist slurs, which also become an important feature to detect misogyny instances. Moreover, intriguing insights emerged from error analysis, in particular about the interaction between different but related abusive phenomena. Based on our cross-domain experiment, we conclude that misogyny is quite a specific kind of abusive language, while we experimentally found that it is different from sexism. Lastly, our cross-lingual experiments show promising results. Our proposed joint-learning architecture obtained a robust performance across languages, worth to be explored in further investigation. All resources and source code developed in this work are publicly available on GitHub.<sup>22</sup>

---

<sup>22</sup><https://github.com/dadangewp/misogyny-project>

# Chapter 7

## Conclusion and Future Works

In this thesis, we addressed several open challenges in abusive language detection by focusing on social media data. The first issue tackled is related to the ambiguity problem related to swear word use, which could have a detrimental effect on abusive language detection models. We explored the phenomenon of swearing in Twitter conversations by automatically predicting the *abusiveness* level of a swear word within the tweet context. We show that resolving the swear word context as either abusive or not abusive helps in improving the models' performance on several downstream abusive language detection tasks. Second, abusive language phenomena are featured by several topical focuses with different vulnerable targets. This characteristic provides another challenge to develop robust models to detect abusive language across different domains. Within this issue, we propose to tackle abusive language detection from a multidomain perspective. We leverage manually annotated datasets to investigate the problem of transferring knowledge from different datasets with different topical focuses and targets. Third, current studies on automatic abusive language detection are typically framed in a monolingual setting, while the abusive language in social media is a global phenomenon. We explore abusive language detection in low-resource languages by transferring knowledge from a resource-rich language, English, in a zero-shot learning fashion. Finally, we also propose to adopt previously implemented approaches, including swear word role analysis, and experiments in cross-domain and cross-lingual settings, focusing on misogyny, a form of online hatred that is widespread across different countries, languages and cultures.

This chapter concludes the overall manuscript, by providing the final remarks based on the experimental results, which will be summarized in Section 7.1. Then, we outline the list of publication originated from this research in Section 7.2. Finally, we close this chapter by describing in Section 7.3 possible directions for future works along this research line.

### 7.1 Conclusion

Based on the experiment presented in the previous chapters, we draw several conclusions, which also specifically address the research questions presented in Chapter 1:

- ***What is the role of swear words in abusive language detection task?***

To answer this question, we proposed several experimental steps. First, we developed a new benchmark corpus called SWAD, consisting of English tweets, where abusive swearing is manually annotated at the word-level. Second, we also built models trained on the SWAD corpus to automatically classify abusive and not-abusive swear words and provide an intrinsic evaluation of SWAD. We experimented by modeling this task into three different settings, namely, sequence labeling, simple text classification, and target-based swear word abusiveness prediction. We used BERT for sequence labeling, simpler but more transparent models for text classification, and wide coverage of models, including several state-of-the-art models in aspect-based sentiment analysis for the target-based task. Finally, we explored the usefulness of predicting swear words' abusiveness on several downstream abusive language detection tasks. Based on models built for swear word abusiveness prediction, we introduced a novel feature, namely the swear word abusiveness feature, and infused it to improve current abusive language detection models. Our results confirmed that our annotation is robust based on the sequence labeling performance. On the other hand, text classification results provided new insights on the most predictive features for distinguishing abusive and not-abusive swear words. In particular, we found that a wide range of features can actually improve the models' performance. Meanwhile, our intention to model the task similarly to aspect-based sentiment analysis leads to promising result. Our BERT-based models obtained the best result in this setting, significantly better than simple text classification settings where we implemented more traditional models. Furthermore, we tested our approach of infusing swear word abusiveness features to several abusive language detection tasks, including HatEval, AMI@Evalita, AMI@IberEval, and Davidson dataset, showing consistent and significant performance improvement across topics, except the Davidson dataset. Our further investigation discovered that the different notion of annotation in the Davidson dataset was the main reason why our feature was not impactful. In conclusion, we experimentally found that swear words have an impactful role in abusive language detection tasks. Specifically, we observed that resolving the ambiguity of swear word use as either abusive or not allows to improve the performance of the abusive language detection model. This confirmed our conjecture that, on the one hand, swear words could provide a good signal to spot abusive content, but, on the other hand, it can also lead to false positives when the swear words are used not in an abusive way.

- ***How to build a robust model which facilitates domain transfer for detecting abusive language across different topical focuses and targets?***

We conducted an exploratory experiment on abusive language detection in multiple domains and targets scenario. We focused on social media data, exploiting several datasets across different domains and targets. First, we conducted an exploratory experiment to detect abusive language detection across different datasets with different topical focuses. Second, we carried on a deeper investigation by exploring

the ability of hate speech detection models to capture common properties from generic hate speech datasets and to transfer this knowledge to recognize specific manifestations of hate. We proposed several deep learning models and experiment with binary classification using two generic corpora. We evaluated their ability to detect abusive language in four topically focused datasets: sexism, misogyny, racism, and xenophobia. Lastly, we experimented with the development of models for detecting both the topics (racism, xenophobia, sexism, misogyny) and the targets (gender, ethnicity) of abusive language, going beyond standard binary classification. We relied on multiple topic-specific datasets and develop, in addition to the deep learning models designed to address the first challenge, a multitask architecture that has been shown to be quite effective in cross-domain sentiment analysis task. In addition, we also tested the use of domain-independent feature obtained from language resource called Hurtlex, to enable knowledge transfer between domains.

Based on the cross-domain experiments in abusive language detection, we found that training a system on datasets featured by more general abusive phenomena will produce a more robust system to detect other more specific kinds of abusive languages. Meanwhile, our investigation on more specific hate speech phenomena across different topical focuses and targets obtained two main conclusions. First, we investigated two experimental scenarios: the first one in which a system was trained on a topic-generic dataset and tested on topic-specific data; and a second one in which a given system was trained on a combination of several topic-specific datasets and tested on topic-specific data. The results show that by training a system on a combination of several (training sets from several) topic-specific datasets the system outperforms a system trained on a single topic-generic dataset. This finding partially confirmed the assumption made by [Swamy et al. \[2019\]](#) according to which merging several abusive language datasets could assist in the detection of abusive language in non-generalizable (unseen) problems. Second, combining topically focused datasets enabled the detection of multitar get hate speech even if the topic and/or target are unseen. We proposed a classification setting which allows a given system to detect not only the harmfulness of a tweet, but also its topical focus in the context of a multi-label classification approach. Our findings show that a multitask approach in which the model learns two or more tasks simultaneously, does better, in performance terms, than a single-task system. In the same way, we also proposed a cross-topic and cross-target experimental setting for the task of hate speech detection, where a system is trained on several sets of data with different topical focuses and targets and, then, tested on another dataset where its topical focus and target are unseen during training. We believe that this is an important finding, which will pave the way for targeted hate speech manifestations, stimulated by a triggering event and which will solve the problem of a lack of annotated data for a particular topic/target. Finally, we also observed that HurtLex is able to improve the overall models' performance in cross-domain experiment setting, providing an independent features to allow knowledge transfer between domains. Specifically, we observed that HurtLex features are able to improve the number of true positives of

the prediction results.

- ***How to build a robust model which facilitates language transfer for detecting abusive language across different languages?***

We proposed extensive experiments in hate speech detection in a cross-lingual setting, more specifically by transferring knowledge from a resource-rich language to a number of lower-resource languages in a zero-shot fashion. We proposed a joint-learning architecture to specifically deal with this classification setting, which exploits available multilingual language representations. In addition, we implemented several competitive baseline systems to evaluate the effectiveness of our proposed models. In this direction, we also evaluated the capability of recent multilingual language models in a cross-lingual classification setting. Furthermore, we experimented with the integration of an external source of knowledge, i.e., a multilingual hate lexicon, into our joint-learning models, to test its impact in transferring knowledge between languages. Finally, we conducted a deep analysis on the results, to obtain meaningful insights regarding the main challenges of this task. The zero-shot cross-lingual hate speech classification results show that our joint-learning based models outperform other models in the majority of experimental settings. The joint-learning LSTM with MUSE outperformed joint-learning with Multilingual BERT in most of settings, when trained on both the original and the balanced training set. Surprisingly, we found that a simple model which rely on automatic machine translation and an English BERT pre-trained model achieved a competitive result in this task. Focusing on the use of multilingual language models in our experiments, we found that multilingual BERT obtains a poor performance compared to other models, across all cross-lingual experiment settings. Even when compared to a straightforward logistic regression coupled with LASER embeddings, the Multilingual BERT model obtained a lower result. The overall result indicates that joint-learning LSTM with MUSE is robust across different settings. It is also worth noting that the better performance of joint-learning Multilingual BERT indicates that our joint-learning architecture is able to cope with the Multilingual BERT model issue in cross-lingual classification. The additional features from the multilingual hate lexicon succeeded to improve our joint-learning based models in some experimental scenarios. The most significant improvement was observed with the addition of HurtLex features in the joint-learning BERT model, where they improve the model performance in four out of six experimental settings in terms of macro  $F$ -score and in all experimental settings in terms of  $F_1$ .

- ***How the challenges on swear words use, and on experimenting in cross-domain and cross-lingual settings can be addressed considering a specific abusive phenomenon in social media, namely misogyny, a form of online hatred that is widespread across different countries, languages and cultures?***

We conducted a deep exploration of automatic misogyny identification (AMI). We started from investigating the best approaches to detect misogynistic, exploring

state of the art on several AMI benchmark datasets. We also explored the most predictive features for detecting misogyny, including the analysis of the swear word use in this specific abusive phenomena. We performed a manual error analysis to discover the issues and challenges specific to this kind of classification task. Furthermore, we ran experiments in cross-domain classification, involving some of the AMI datasets, in order to investigate the interaction between misogyny and related phenomena, namely sexism, hate speech, and offensive language. Finally, we conducted an experiment on AMI in a cross-lingual setting, building a joint-learning model based on LSTM and BERT, in order to bridge the gap of AMI in low-resource languages. Our proposed models succeed to outperform the state of the art on all AMI benchmarks, consisting of three different languages: English, Spanish and Italian. We found that traditional models still perform better than more sophisticated, deep learning approaches English and Spanish AMI IberEval. On the other hand, in English and Italian AMI EVALITA, BERT obtains better performance than other models. We also experimentally proved that lexical features such as sexist slurs and woman words (words which are synonyms or related words to “woman”) are among the most predictive features to detect misogyny. We also observed that treating AMI task B as an independent multiclass classification gives a better performance than a pipeline approach with task A. With this approach, we were able to outperform all of the state of the art results on task B with the exact same system used for task A. Our cross-domain classification experiment shows that neural-based models, i.e., LSTM and BERT, facilitate knowledge transfer between different datasets. As expected, our system does not achieve an optimal performance when trained on other abusive phenomena data and tested on AMI data, and vice versa. The experiment with HurtLex have shown that the use of a domain independent resource, such as an abusive language lexicon, was able to boost the cross-domain performance, proving how this approach is capable of facilitating domain transfer between datasets. We also found that augmenting the training set only works when the additional data provide a similar topical focus as the original training dataset. Our cross-domain experimental results confirm that hate speech towards women is a more related phenomenon to misogyny than sexism. The overall results show that BERT is the best model for domain transfer between different datasets, able to obtain robust performance in all experimental settings. Differently from the cross-domain setting, our traditional classifier, i.e. LSVC, still got a better performance than neural architectures in some of our cross-lingual experimental settings. However, further investigations showed that its performance is highly dependant on the quality of the translation result, while deep learning approaches provide a more stable performance across language pairs. Using monolingual word embeddings with translated data with LSTM gives better results than multilingual word embeddings without translating the data. We ascribe this result to the high number of out of vocabulary words resulting from using the multilingual embeddings by FastText. To overcome the translation quality and the out of vocabulary words issues, we proposed a joint-learning model, which was

able to outperform all the other systems. Again, the use of additional knowledge from HurtLex in our joint-learning model improved its performance, mainly on the recall side. Similarly to the cross-domain setting, the overall results exhibit that BERT-based model is the best model in cross-lingual setting experiment, even more robust performance is obtained when we build joint-learning model with multilingual BERT.

## 7.2 Research Contribution

In this section, we outline the publications derived from this research, including journal papers and conference papers (reverse chronological order).

- Journal Papers

1. Chiril, P., **Pamungkas, E. W.**, Benamara, F., Moriceau, V., and Patti, V. (2021). Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, 1-31. [Chiril et al. \[2021\]](#)
2. **Pamungkas, E. W.**, Basile, V., and Patti, V. (2021). Towards Multidomain and Multilingual Abusive Language Detection: A Survey. *Personal and Ubiquitous Computing*, 1-27. [Pamungkas et al. \[2021b\]](#)
3. **Pamungkas, E. W.**, Basile, V., and Patti, V. (2021). A Joint Learning Approach with Knowledge Injection for Zero-Shot Cross-lingual Hate Speech Detection. *Information Processing & Management*, 58(4), 102544. [Pamungkas et al. \[2021a\]](#)
4. **Pamungkas, E. W.**, Basile, V., and Patti, V. (2020). Misogyny detection in Twitter: a Multilingual and Cross-domain Study. *Information Processing & Management*, 57(6), 102360. [Pamungkas et al. \[2020b\]](#)

- Conference Papers

1. Koufakou, A., **Pamungkas, E. W.**, Basile, V., and Patti, V. (2020, November). HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In Proceedings of the Fourth Workshop on Online Abuse and Harms, co-located with EMNLP 2020 (pp. 34-43). Association for Computational Linguistics. [Koufakou et al. \[2020\]](#)
2. **Pamungkas, E. W.**, Basile, V., and Patti, V. (2020). Do you really want to hurt me? Predicting abusive swearing in social media. In The 12th Language Resources and Evaluation Conference (pp. 6237-6246). European Language Resources Association. [Pamungkas et al. \[2020a\]](#)
3. **Pamungkas, E. W.**, and Patti, V. (2019, July). Cross-domain and Cross-lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon. In Proceedings of the 57th Annual Meeting of the Association for Computational linguistics: Student Research Workshop (pp.

- 363-370). Association for Computational Linguistics. [Pamungkas and Patti \[2019\]](#)
4. **Pamungkas, E.W.**, Cignarella, A.T., Basile, V., and Patti, V. (2018). Automatic Identification of Misogyny in English and Italian Tweets at Evalita 2018 with a Multilingual Hate Lexicon. In Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018) (Vol. 2263, No. 1, pp. 1-6). CEUR-WS. [Pamungkas et al. \[2018b\]](#)
  5. **Pamungkas, E. W.**, Cignarella, A. T., Basile, V., and Patti, V. (2018). 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. In 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018 (Vol. 2150, pp. 234-241). CEUR-WS. [Pamungkas et al. \[2018c\]](#)

- Other Publications

Below, a list of additional research works carried out is presented. These publications are partially related to the main objectives of this thesis.

1. **Pamungkas, E. W.**, Basile, V., and Patti, V. (2018). Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure. In Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management. CIKM 2018 (Vol. 2482). CEUR-WS. [Pamungkas et al. \[2018a\]](#)
2. Ronzano, F., Barbieri, F., **Pamungkas, E. W.**, Patti, V., and Chiusaroli, F. (2018). Overview of the evalita 2018 italian emoji prediction (itemoji) task. In 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018 (Vol. 2263, pp. 1-9). CEUR-WS. [Ronzano et al. \[2018\]](#)
3. **Pamungkas, E. W.**, and Patti, V. (2018, June). #NonDicevoSulSerio at SemEval-2018 Task 3: Exploiting Emojis and Affective Content for Irony Detection in English Tweets. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 649-654). [Pamungkas and Patti \[2018\]](#)

- Source Code and Corpora

All resources developed and the source code related to experiments described in this thesis are publicly available in the author personal GitHub page.<sup>1</sup>

### 7.3 Future Works

Our investigations in building robust models to detect abusive language content point to several directions for future work:

---

<sup>1</sup><https://github.com/dadangewp>

- ***Swear Words in Abusive Language Detection Tasks.***

We obtained a very encouraging result on our investigation of swear word role in abusive language tasks. However, we believe that there is still a room for improvement for both the corpus and the automatic classification of swearing. We aim to improve the dataset by proposing a fine-grained categorization of swear words, in line with the ones introduced by [Pinker \[2007\]](#) and [McEnery \[2006\]](#)). We also plan to apply our swear word abusiveness feature to other tasks and datasets [[Poletto et al., 2020](#)], in order to obtain the full picture of its impact in abusive language detection tasks. Applying our methodology to other languages, like Italian, is also an interesting matter of future work, even if it is not trivial, as it depends on the availability of language resources and robust NLP tools for them. Fortunately, full-fledged NLP pipelines do exist for many languages, thanks for instance to large-scale initiatives such as Universal Dependencies, which provides among its deliverables the UDpipe software library and a broad set of trained models in more than 70 languages [[Nivre et al., 2016](#), [Straka et al., 2016](#)]. Deep learning models, including transformer-based networks are also surfacing for languages less resources than English — see for instance the Italian BERT model ALBERTo [[Polignano et al., 2019](#)]. Finally, the multilingual lexicon of offensive words HurtLex [[Bassignana et al., 2018](#)] could provide a solid basis to compile lists of swear words in its 53 covered languages.

- ***Abusive Language Detection Across Domains.***

Based on the recent studies, combining several datasets to enhance topic coverage of training data is not always leading to positive results. As future work in this direction, we aim at exploring more deeply the issue related to different coverage, topical focuses and abusive phenomena in characterizing the datasets in this field, taking a semantic ontology-based approach to clearly represent the relations between concepts and linguistic phenomena involved. This will allow us to further explore and refine the idea that combining some datasets can produce a more robust system to detect abusive language across different domains. We also observed that previous approaches struggle to transfer knowledge between different domains, especially to deal with some specific taboo words which are linked to different (abusive) contexts in different domains [[Nozza, 2021](#)]. We also found that the use of domain-independent resources to transfer knowledge between domains was able to partially solve this issue. We plan to explore the use of other domain-independent feature in multidomain abusive language detection task. Another path to explore is the impact of bias in multitarget hate speech detection. Bias in abusive language datasets is an open problem already observed by several previous studies [[Wiegand et al., 2019](#), [Davidson et al., 2019](#), [Park et al., 2018](#), [Mozafari et al., 2019](#)], in which different variants of bias, such as topic bias, author bias, gender and racial bias were explored. As no further investigation on developing an approach in debiasing abusive language datasets has been offered, we also plan to examine this direction in the future in the interests of keeping hate speech detection fair and compliant. On the theoretical

counterpart, a careful study of the notion of every abusive behavior online which is modeled with the purpose of automatic detection is important, to obtain a clearer terminology and understanding of the abusive phenomena we want to capture in language. The study by [Vidgen et al. \[2019\]](#) proposed several possible solutions to address this issue, which can be considered for future works.

- ***Abusive Language Detection Across Languages.***

Based on the overall results in our cross lingual abusive language experiment and the experimental results of previous studies, we observed several issues and difficulties of this task. Biased datasets and the lack of ability of the current multilingual language models to transfer knowledge between different languages remain open problems. However, the results of our experiments show that zero-shot methods are promising in this scenario, paving the way for extended work in this direction. This is particularly relevant to the task of hate speech detection, where the aforementioned issues are ubiquitous. In future work, we plan to have a deeper analysis of the impact of automatic translation results on our joint-learning based models. We also plan to better investigate the issue related to the use of derogatory words in different languages, which we believe may have a significant impact on cross-lingual classification of hate speech. Some recent studies also found that some specific taboo words have different meaning and are linked to different abusive contexts in different languages, and this could be detrimental for the classification models [Pamungkas et al. \[2021a\]](#), [Nozza \[2021\]](#). Our experimental evaluation and subsequent qualitative analysis suggests that the explicit integration of linguistic knowledge from a multilingual abusive language lexicon helps to provide a better representation of the words, in particular by accounting for creative language use such as metaphors and figurative language. Furthermore, we also notice the unstable performance of models in different target languages, specifically models' performance in more resource-rich languages is higher than in lower-resource languages. This may be related to multilingual language representation models being trained on different amounts of data in different languages [\[Wu and Dredze, 2020\]](#). Therefore, we argue that tackling the lesser performance of multilingual language representations in low-resource languages also needs to be considered for future work. Finally, we also plan to extend our experiment into more low resource languages such as Arabic [\[Ousidhoum et al., 2019\]](#), Russian [\[Pronoza et al., 2021\]](#), Polish [\[Ptaszynski et al., 2019\]](#), Vietnamese [\[Vu et al., 2020\]](#) and etc.

- ***Automatic Misogyny Identification Task***

We found that experimenting on cross-domain and cross-lingual settings is beneficial also focusinng on a specific kind of abusive phenomena, namely misogyny. In future work, we plan to implement a transfer learning approach for improving the performance of the system in AMI Task A, by propagating information from the AMI task B classification. Transfer learning is also a potential solution for the domain adaptation issue in both cross-domain and cross-lingual settings. There is also a newer edition of the AMI shared task proposed at Evalita 2020 [\[Fersini et al.,](#)

[2020], which provide a new collection of dataset, including synthetical data which were added to address bias mitigation in the dataset. It would be also interesting to validate the state of the art approaches proposed in this study to the new available dataset and provide a further analysis of this task. A new shared task called ARMI (Arabic Misogyny Identification Shared Task)<sup>2</sup> will be also held this year, which focuses to the detection of misogyny in Arabic. This new Arabic dataset could also provide an interesting benchmark enabling new investigations towards robust models to detect misogyny in multilingual environments.

---

<sup>2</sup><https://sites.google.com/view/armi2021/>

# Bibliography

- S. Agarwal and A. Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *CoRR*, abs/1701.04931, 2017. URL <http://arxiv.org/abs/1701.04931>.
- R. Ahluwalia, H. Soni, E. Callow, A. C. A. Nascimento, and M. D. Cock. Detecting hate speech against women in english tweets. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2263/paper032.pdf>.
- H. Ahn, J. Sun, C. Y. Park, and J. Seo. NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1576–1586, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.semeval-1.206>.
- S. Akhtar, V. Basile, and V. Patti. A new measure of polarization in the annotation of hate speech. In M. Alviano, G. Greco, and F. Scarcello, editors, *AI\*IA 2019 - Advances in Artificial Intelligence - XVIIIfth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19-22, 2019, Proceedings*, volume 11946 of *Lecture Notes in Computer Science*, pages 588–603. Springer, 2019. doi: 10.1007/978-3-030-35166-3\\_41. URL [https://doi.org/10.1007/978-3-030-35166-3\\_41](https://doi.org/10.1007/978-3-030-35166-3_41).
- A. Alakrot, L. Murray, and N. S. Nikolov. Dataset construction for the detection of anti-social behaviour in online communication in arabic. In K. Shaalan and S. R. El-Beltagy, editors, *Fourth International Conference On Arabic Computational Linguistics, ACLING 2018, November 17-19, 2018, Dubai, United Arab Emirates*, volume 142 of *Procedia Computer Science*, pages 174–181. Elsevier, 2018. doi: 10.1016/j.procs.2018.10.473. URL <https://doi.org/10.1016/j.procs.2018.10.473>.
- N. Albadi, M. Kurdi, and S. Mishra. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In U. Brandes, C. Reddy, and A. Tagarelli, editors, *IEEE/ACM 2018 International Conference on Advances in Social*

*Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 69–76. IEEE Computer Society, 2018. doi: 10.1109/ASONAM.2018.8508247. URL <https://doi.org/10.1109/ASONAM.2018.8508247>.

- I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE, 2017.
- K. Allan and K. Burridge. *Forbidden words: Taboo and the censoring of language*. Cambridge University Press, 2006.
- S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465, 2020. URL <https://arxiv.org/abs/2004.06465>.
- M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain*, volume 6, page 23, 2018.
- M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of misogynistic language on Twitter. In M. Silberstein, F. Atigui, E. Kornysheva, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, volume 10859 of *Lecture Notes in Computer Science*, pages 57–64. Springer, 2018. doi: 10.1007/978-3-319-91947-8\\_6. URL [https://doi.org/10.1007/978-3-319-91947-8\\_6](https://doi.org/10.1007/978-3-319-91947-8_6).
- A. Arango, J. Pérez, and B. Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, page 101584, 2020.
- M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610, 2019. URL <https://transacl.org/ojs/index.php/tacl/article/view/1742>.
- P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 759–760. ACM, 2017. doi: 10.1145/3041021.3054223. URL <https://doi.org/10.1145/3041021.3054223>.
- J. Y. Bak, S. Kim, and A. Oh. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 103–112. Association for Computational Linguistics, 2022.

*Linguistics: Short Papers- Volume 2*, pages 60–64. Association for Computational Linguistics, 2012.

- A. Bakarov. Vector space models for automatic misogyny identification (short paper). In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–3. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2263/paper035.pdf>.
- T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, Oct. 2013. Asian Federation of Natural Language Processing. URL <https://www.aclweb.org/anthology/I13-1041>.
- J. Bartlett, R. Norrie, S. Patel, R. Rumpel, and S. Wibberley. Misogyny on twitter. *Demos*, 2014.
- A. Basile and C. Rubagotti. Crotonemilano for AMI at evalita2018. A performant, cross-lingual misogyny detection system. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–5. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2263/paper034.pdf>.
- V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL <https://www.aclweb.org/anthology/S19-2007>.
- E. Bassignana, V. Basile, and V. Patti. Hurtlex: A multilingual lexicon of words to hurt. In E. Cabrio, A. Mazzei, and F. Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2253/paper49.pdf>.
- C. Baziotsis, N. Pelekis, and C. Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

- C. Bianchi. Slurs and appropriation: An echoic account. *Journal of Pragmatics*, 66:35 – 44, 2014. ISSN 0378-2166. doi: <https://doi.org/10.1016/j.pragma.2014.02.009>. URL <http://www.sciencedirect.com/science/article/pii/S0378216614000526>.
- S. Bodapati, S. Gella, K. Bhattacharjee, and Y. Al-Onaizan. Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3515. URL <https://www.aclweb.org/anthology/W19-3515>.
- A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1105. URL <https://www.aclweb.org/anthology/W18-1105>.
- C. Bosco, F. Dell’Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–9. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2263/paper010.pdf>.
- L. Boualili, J. G. Moreno, and M. Boughanem. Markedbert: Integrating traditional IR cues in pre-trained language models for passage retrieval. In J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1977–1980. ACM, 2020. doi: 10.1145/3397271.3401194. URL <https://doi.org/10.1145/3397271.3401194>.
- L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1176. URL <https://www.aclweb.org/anthology/D19-1176>.
- U. Bretschneider and R. Peters. Detecting cyberbullying in online communities. In *24th European Conference on Information Systems, ECIS 2016, Istanbul, Turkey, June 12-15, 2016*, page Research Paper 61, 2016. URL [http://aisel.aisnet.org/ecis2016\\_rp/61](http://aisel.aisnet.org/ecis2016_rp/61).
- U. Bretschneider and R. Peters. Detecting offensive statements towards foreigners in social media. In T. Bui, editor, *50th Hawaii International Conference on System Sciences*,

*HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, pages 1–10. ScholarSpace / AIS Electronic Library (AISel), 2017. URL <http://hdl.handle.net/10125/41423>.

- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>.
- P. Burnap and M. L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.
- D. Buscaldi. Tweetaneuse @ AMI EVALITA2018: character-based models for the automatic misogyny identification task (short paper). In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–4. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2263/paper036.pdf>.
- I. Cachola, E. Holgate, D. Preoțiuc-Pietro, and J. J. Li. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938, 2018.
- Y. Cai and X. Wan. Multi-domain sentiment classification based on domain-aware embedding and attention. In S. Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4904–4910. ijcai.org, 2019. doi: 10.24963/ijcai.2019/681. URL <https://doi.org/10.24963/ijcai.2019/681>.
- E. Cambria, P. Chandra, A. Sharma, and A. Hussain. Do not feel the trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web*, volume 664 of *CEUR Workshop Proceedings*, pages 1–12, Shanghai, China, 2010. CEUR-WS.org.
- E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80, 2017.
- J. S. Canós. Misogyny identification through SVM at ibereval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Alborno, editors, *Proceedings*

*of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 229–233. CEUR-WS.org, 2018.  
URL [http://ceur-ws.org/Vol-2150/AMI\\_paper1.pdf](http://ceur-ws.org/Vol-2150/AMI_paper1.pdf).

- T. Caselli, V. Basile, J. Mitrovic, and M. Granitzer. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472, 2020a. URL <https://arxiv.org/abs/2010.12472>.
- T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France, May 2020b. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://www.aclweb.org/anthology/2020.lrec-1.760>.
- C. Casula. Transfer learning for multilingual offensive language detection with bert. Master's thesis, Uppsala University, 2020.
- D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. Mean Birds: Detecting Aggression and Bullying on Twitter. In P. Fox, D. L. McGuinness, L. Poirier, P. Boldi, and K. Kinder-Kurlanda, editors, *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 13–22. ACM, 2017. doi: 10.1145/3091478.3091487. URL <https://doi.org/10.1145/3091478.3091487>.
- H. Chen, S. McKeever, and S. J. Delany. Abusive text detection using neural networks. In J. McAuley and S. McKeever, editors, *Proceedings of the 25th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, December 7 - 8, 2017*, volume 2086 of *CEUR Workshop Proceedings*, pages 258–260. CEUR-WS.org, 2017a. URL [http://ceur-ws.org/Vol-2086/AICS2017\\_paper\\_44.pdf](http://ceur-ws.org/Vol-2086/AICS2017_paper_44.pdf).
- P. Chen, Z. Sun, L. Bing, and W. Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark, Sept. 2017b. Association for Computational Linguistics. doi: 10.18653/v1/D17-1047. URL <https://www.aclweb.org/anthology/D17-1047>.
- Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.
- P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully. He said “who’s gonna take care of your children when you are at ACL?”: Reported Sexist Acts are Not Sexist. In *Proceedings of the 58th Annual Meeting of the Association*

*for Computational Linguistics*, pages 4055–4066, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.373. URL <https://www.aclweb.org/anthology/2020.acl-main.373>.

P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31, 2021. URL <https://link.springer.com/article/10.1007/s12559-021-09862-5>. Published online: 28 June 2021.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL <https://www.aclweb.org/anthology/D14-1179>.

S. A. Chowdhury, H. Mubarak, A. Abdelali, S.-g. Jung, B. J. Jansen, and J. Salminen. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://www.aclweb.org/anthology/2020.lrec-1.761>.

Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829. Association for Computational Linguistics, 2019. URL <https://aclanthology.org/P19-1271>.

L. Code. *Encyclopedia of feminist theories*. Routledge, 2002.

Ç. Çöltekin. A corpus of turkish offensive language on social media. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6174–6184. European Language Resources Association, 2020. URL <https://www.aclweb.org/anthology/2020.lrec-1.758/>.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. Cross-platform evaluation for italian hate speech detection. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL <http://ceur-ws.org/Vol-2481/paper22.pdf>.

- M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Techn.*, 20(2):10:1–10:22, 2020a. doi: 10.1145/3377323. URL <https://doi.org/10.1145/3377323>.
- M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. Hybrid emoji-based masked language models for zero-shot abusive language detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 943–949, Online, Nov. 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.84. URL <https://www.aclweb.org/anthology/2020.findings-emnlp.84>.
- K. Crenshaw. *On Intersectionality: The Essential Writings of Kimberle Crenshaw*. New Press, 2015. ISBN 9781595587046. URL <https://books.google.it/books?id=xlfctgAACAAJ>.
- T. Dadu and K. Pant. Towards code-switched classification exploiting constituent language resources. *CoRR*, abs/2011.01913, 2020a. URL <https://arxiv.org/abs/2011.01913>.
- T. Dadu and K. Pant. Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2183–2189, Barcelona (online), Dec. 2020b. International Committee for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.semeval-1.290>.
- H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P07-1033>.
- T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press, 2017. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665>.
- T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3504. URL <https://www.aclweb.org/anthology/W19-3504>.
- O. de Gibert, N. Pérez, A. G. Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. In D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, editors, *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 11–20. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-5102. URL <https://doi.org/10.18653/v1/w18-5102>.
- T. De Mauro. Le parole per ferire. *Internazionale*, 2016. 27 settembre 2016.

- R. P. de Pelle and V. P. Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, page 10. SBC, 2017.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://www.aclweb.org/anthology/N19-1423>.
- K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*, volume WS-11-02 of *AAAI Workshops*. AAAI, 2011. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841>.
- A. Downs. Up and down with ecology. *The “issue attention” cycle*, 1973.
- C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.370. URL <https://www.aclweb.org/anthology/2020.acl-main.370>.
- M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, pages 42–51. AAAI Press, 2018a. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17910>.
- M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. M. Belding. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 52–61. AAAI Press, 2018b. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17905>.
- K. Erjavec and M. P. Kovačič. “You Don’t Understand, This is a New War!” Analysis of Hate Speech in News Web Sites’ Comments. *Mass Communication and Society*, 15(6): 899–920, 2012. doi: 10.1080/15205436.2011.619679. URL <https://doi.org/10.1080/15205436.2011.619679>.
- EU Commission. Code of conduct on countering illegal hate speech online, 2016. URL [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online\\_en#theeucodeofconduct](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en#theeucodeofconduct).

- K. B. Fägersten. *Who's swearing now? The social aspects of conversational swearing.* Cambridge Scholars Publishing, 2012.
- F. Fasoli, A. Carnaghi, and M. P. Paladino. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, 52:98–107, 2015.
- E. Fehn Unsvåg and B. Gambäck. The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5110. URL <https://www.aclweb.org/anthology/W18-5110>.
- J. Fernquist, O. Lindholm, L. Kaati, and N. Akrami. A study on the feasibility to detect hate speech in swedish. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*, pages 4724–4729. IEEE, 2019. doi: 10.1109/BigData47090.2019.9005534. URL <https://doi.org/10.1109/BigData47090.2019.9005534>.
- E. Fersini, D. Nozza, and P. Rosso. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–9. CEUR-WS.org, 2018a. URL <http://ceur-ws.org/Vol-2263/paper009.pdf>.
- E. Fersini, D. Nozza, and P. Rosso. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR.org*, 2018b.
- E. Fersini, P. Rosso, and M. Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org, 2018c. URL <http://ceur-ws.org/Vol-2150/overview-AMI.pdf>.
- E. Fersini, P. Rosso, and M. Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@ SEPLN*, pages 214–228, 2018d.
- E. Fersini, D. Nozza, and P. Rosso. AMI @ EVALITA2020: automatic misogyny identification. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL <http://ceur-ws.org/Vol-2765/paper161.pdf>.

- D. Fiser, T. Erjavec, and N. Ljubesic. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In Z. Waseem, W. H. K. Chung, D. Hovy, and J. R. Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017*, pages 46–51. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-3007. URL <https://doi.org/10.18653/v1/w17-3007>.
- D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont. Proceedings of the 2nd workshop on abusive language online (alw2). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/W18-5100>.
- K. Florio, V. Basile, M. Polignano, P. Basile, and V. Patti. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180, 2020.
- P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30, 2018. doi: 10.1145/3232676. URL <https://doi.org/10.1145/3232676>.
- P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3510. URL <https://www.aclweb.org/anthology/W19-3510>.
- A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press, 2018. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909>.
- S. Frenda, B. Ghanem, E. Guzmán-Falcón, M. Montes-y-Gómez, and L. V. Pineda. Automatic expansion of lexicons for multilingual misogyny detection. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org, 2018a. URL <http://ceur-ws.org/Vol-2263/paper031.pdf>.
- S. Frenda, B. Ghanem, and M. Montes-y-Gómez. Exploration of misogyny in spanish and english tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th*

*Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 260–267. CEUR-WS.org, 2018b. URL [http://ceur-ws.org/Vol-2150/AMI\\_paper6.pdf](http://ceur-ws.org/Vol-2150/AMI_paper6.pdf).

- S. Frenda, B. Ghanem, M. Montes-y-Gómez, and P. Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent and Fuzzy Systems*, 36(5):4743–4752, 2019. doi: 10.3233/JIFS-179023. URL <https://doi.org/10.3233/JIFS-179023>.
- R. Fulper, G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe. Misogynistic language on Twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, pages 1–4, 2014.
- B. Gambäck and U. K. Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3013. URL <https://www.aclweb.org/anthology/W17-3013>.
- Y. Ganin and V. S. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org, 2015. URL <http://proceedings.mlr.press/v37/ganin15.html>.
- L. Gao, A. Kuppersmith, and R. Huang. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In G. Kondrak and T. Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 774–782. Asian Federation of Natural Language Processing, 2017. URL <https://www.aclweb.org/anthology/I17-1078/>.
- M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayanan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference, WWW ’19*, page 514–525, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313698. URL <https://doi.org/10.1145/3308558.3313698>.
- M. Gauthier, A. Guille, A. Deseille, and F. Rico. Text mining and twitter to analyze British swearing habits. *Handbook of Twitter for Research*, 2015.
- N. D. Gitari, Z. Zuping, H. Damien, and J. Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

- G. Glavaš, M. Karan, and I. Vulić. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.coling-main.559>.
- I. Goenaga, A. Atutxa, K. Gojenola, A. Casillas, A. D. de Ilarrazá, N. Ezeiza, M. Oronoz, A. Pérez, and O. Perez-de-Viñaspre. Automatic misogyny identification using neural networks. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 249–254. CEUR-WS.org, 2018. URL [http://ceur-ws.org/Vol-2150/AMI\\_paper4.pdf](http://ceur-ws.org/Vol-2150/AMI_paper4.pdf).
- J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, Q. Gergory, R. K. Gnanasekaran, R. R. Gunasekaran, K. M. Hoffman, J. Hottle, V. Jienjiltlert, S. Khare, R. Lau, M. J. Martindale, S. Naik, H. L. Nixon, P. Ramachandran, K. M. Rogers, L. Rogers, M. S. Sarin, G. Shahane, J. Thanki, P. Vengataraman, Z. Wan, and D. M. Wu. A large labeled corpus for online harassment research. In P. Fox, D. L. McGuinness, L. Poirier, P. Boldi, and K. Kinder-Kurlanda, editors, *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 229–233. ACM, 2017. doi: 10.1145/3091478.3091509. URL <https://doi.org/10.1145/3091478.3091509>.
- R. Gomez, J. Gibert, L. Gómez, and D. Karatzas. Exploring hate speech detection in multimodal publications. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1459–1467. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093414. URL <https://doi.org/10.1109/WACV45572.2020.9093414>.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- I. J. Goodfellow, Y. Bengio, and A. C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL <http://www.deeplearningbook.org/>.
- E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- J. Guberman and L. Hemphill. Challenges in modifying existing scales for detecting harassment in individual tweets. In T. Bui, editor, *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January*

4-7, 2017, pages 1–10. ScholarSpace / AIS Electronic Library (AISeL), 2017. URL <http://hdl.handle.net/10125/41422>.

- H. Haddad, H. Mulki, and A. Oueslati. T-HSAB: A tunisian hate speech and abusive dataset. In K. Smaïli, editor, *Arabic Language Processing: From Theory to Practice - 7th International Conference, ICALP 2019, Nancy, France, October 16-17, 2019, Proceedings*, volume 1108 of *Communications in Computer and Information Science*, pages 251–263. Springer, 2019. doi: 10.1007/978-3-030-32959-4\18. URL [https://doi.org/10.1007/978-3-030-32959-4\\_18](https://doi.org/10.1007/978-3-030-32959-4_18).
- H. L. Hammer. Automatic detection of hateful comments in online discussion. In L. A. Maglaras, H. Janicke, and K. I. Jones, editors, *Industrial Networks and Intelligent Systems - Second International Conference, INISCOM 2016, Leicester, UK, October 31 - November 1, 2016, Revised Selected Papers*, volume 188 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 164–173, 2016. doi: 10.1007/978-3-319-52569-3\15. URL [https://doi.org/10.1007/978-3-319-52569-3\\_15](https://doi.org/10.1007/978-3-319-52569-3_15).
- C. V. Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. D. Pauw, W. Daelemans, and V. Hoste. Detection and fine-grained classification of cyberbullying events. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 672–680. RANLP 2015 Organising Committee / ACL, 2015. URL <https://www.aclweb.org/anthology/R15-1086/>.
- S. Hewitt, T. Tiropanis, and C. Bokhove. The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335. ACM, 2016.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL <http://dx.doi.org/10.1162/neco.1997.9.8.1735>.
- E. Holgate, I. Cachola, D. Preoțiuc-Pietro, and J. J. Li. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1471. URL <https://www.aclweb.org/anthology/D18-1471>.
- C. J. Hutto and E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, editors, *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press, 2014. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109>.

- M. O. Ibrohim and I. Budi. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229, 2018.
- M. O. Ibrohim and I. Budi. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy, Aug. 2019a. Association for Computational Linguistics. doi: 10.18653/v1/W19-3506. URL <https://www.aclweb.org/anthology/W19-3506>.
- M. O. Ibrohim and I. Budi. Translated vs Non-Translated Method for Multilingual Hate Speech Identification in Twitter. *Int. J. Adv. Sci. Eng. Inf. Technol.*, 9(4):1116–1123, 2019b.
- A. M. Ishmam and S. Sharmin. Hateful speech detection in public facebook pages for the bengali language. In M. A. Wani, T. M. Khoshgoftaar, D. Wang, H. Wang, and N. Seliya, editors, *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*, pages 555–560. IEEE, 2019. doi: 10.1109/ICMLA.2019.00104. URL <https://doi.org/10.1109/ICMLA.2019.00104>.
- T. Jay. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets*. John Benjamins Publishing, 1992.
- T. Jay. *Why we curse: A neuro-psycho-social theory of speech*. John Benjamins Publishing, 1999.
- T. Jay. Do offensive words harm people? *Psychology, public policy, and law*, 15(2):81, 2009a.
- T. Jay. The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, 4(2):153–161, 2009b.
- T. Jay and K. Janschewitz. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288, 2008.
- T. Jay, K. King, and T. Duncan. Memories of punishment for cursing. *Sex Roles*, 55(1-2):123–133, 2006.
- A. Jha and R. Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2902. URL <https://www.aclweb.org/anthology/W17-2902>.
- L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL <https://aclanthology.org/P11-1016>.

- T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In C. Nedellec and C. Rouveiro, editors, *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer, 1998. doi: 10.1007/BFb0026683. URL <https://doi.org/10.1007/BFb0026683>.
- D. I. Johnson. Swearing by peers in the work setting: Expectancy violation valence, perceptions of message, and perceptions of speaker. *Communication Studies*, 63(2): 136–151, 2012.
- P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL <https://www.aclweb.org/anthology/2020.acl-main.560>.
- D. Jurgens, E. Chandrasekharan, and L. Hemphill. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666. Association for Computational Linguistics (ACL), 2019a.
- D. Jurgens, L. Hemphill, and E. Chandrasekharan. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1357. URL <https://www.aclweb.org/anthology/P19-1357>.
- M. Karan and J. Šnajder. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5117. URL <https://www.aclweb.org/anthology/W18-5117>.
- A. Khatua, C. E., and A. Khatua. Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400, 2018.
- A. Khatua, E. Cambria, K. Ghosh, N. Chaki, and A. Khatua. Tweeting in support of lgbt? a deep learning approach. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD ’19, page 342–345, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362078. doi: 10.1145/3297001.3297057. URL <https://doi.org/10.1145/3297001.3297057>.
- S. Kiritchenko and I. Nejadgholi. Towards ethics by design in online abusive content detection. *CoRR*, abs/2010.14952, 2020. URL <https://arxiv.org/abs/2010.14952>.

- V. Kolhatkar, H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, pages 1–36, 2019.
- A. Koufakou, E. W. Pamungkas, V. Basile, and V. Patti. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.5. URL <https://www.aclweb.org/anthology/2020.alw-1.5>.
- C. Kramerae and D. Spender. *Routledge international encyclopedia of women*. New York, London: Routledge, 2000.
- R. Kshirsagar, T. Cukuvac, K. McKeown, and S. McGregor. Predictive embeddings for hate speech detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5104. URL <https://www.aclweb.org/anthology/W18-5104>.
- R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W18-4401>.
- R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France, May 2020. European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/2020.trac2-1.1>.
- J. Kurrek, H. M. Saleem, and D. Ruths. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.17. URL <https://www.aclweb.org/anthology/2020.alw-1.17>.
- I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In M. desJardins and M. L. Littman, editors, *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*, pages 1621–1622. AAAI Press, 2013. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6419>.
- K. H. Kwon and A. Gruzd. Is offensive commenting contagious online? examining public vs interpersonal swearing in response to donald trump’s youtube campaign videos. *Internet Research*, 27(4):991–1010, 2017.

- G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, pages 1–14. OpenReview.net, 2018. URL <https://openreview.net/forum?id=H196sainb>.
- J. Langham and K. Gosha. The classification of aggressive dialogue in social media platforms. In R. Kishore, D. Beimborn, R. K. Bandi, B. Aubert, D. Compeau, and M. Tarafdar, editors, *Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research, SIGMIS-CPR 2018, Buffalo-Niagara Falls, NY, USA, June 18-20, 2018*, pages 60–63. ACM, 2018. doi: 10.1145/3209626.3209720. URL <https://doi.org/10.1145/3209626.3209720>.
- Y.-H. Lin, C.-Y. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, S. Rijhwani, J. He, Z. Zhang, X. Ma, A. Anastasopoulos, P. Littell, and G. Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1301. URL <https://www.aclweb.org/anthology/P19-1301>.
- H. Liu, F. Chiroma, and M. Cocea. Identification and classification of misogynous tweets using multi-classifier fusion. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 268–273. CEUR-WS.org, 2018a. URL [http://ceur-ws.org/Vol-2150/AMI\\_paper7.pdf](http://ceur-ws.org/Vol-2150/AMI_paper7.pdf).
- P. Liu, X. Qiu, and X. Huang. Adversarial Multi-task Learning for Text Classification. In R. Barzilay and M. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1–10. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1001. URL <https://doi.org/10.18653/v1/P17-1001>.
- Q. Liu, H. Zhang, Y. Zeng, Z. Huang, and Z. Wu. Content attention model for aspect based sentiment analysis. In P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1023–1032. ACM, 2018b. doi: 10.1145/3178876.3186001. URL <https://doi.org/10.1145/3178876.3186001>.
- Q. Liu, Y. Zhang, and J. Liu. Learning Domain Representation for Multi-Domain Sentiment Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 541–550, New Orleans, Louisiana, June 2018c. Association for Computational Linguistics. doi: 10.18653/v1/N18-1050. URL <https://www.aclweb.org/anthology/N18-1050>.

- X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.
- N. Ljubešić, T. Erjavec, and D. Fišer. Datasets of Slovene and Croatian moderated news comments. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5116. URL <https://www.aclweb.org/anthology/W18-5116>.
- D. Ma, S. Li, X. Zhang, and H. Wang. Interactive attention networks for aspect-level sentiment classification. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4068–4074. ijcai.org, 2017. doi: 10.24963/ijcai.2017/568. URL <https://doi.org/10.24963/ijcai.2017/568>.
- N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43, 2019.
- S. Malmasi and M. Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.
- T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandalia, and A. Patel. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In P. Majumder, M. Mitra, S. Gangopadhyay, and P. Mehta, editors, *FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019*, pages 14–17. ACM, 2019. doi: 10.1145/3368567.3368584. URL <https://doi.org/10.1145/3368567.3368584>.
- K. Manne. *Down girl: The logic of misogyny*. Oxford University Press, 2017.
- T. Marwa, O. Salima, and M. Souham. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE, 2018.
- B. Mathew, N. Kumar, P. Goyal, A. Mukherjee, et al. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712*, 2018.
- B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, and A. Mukherjee. Thou shalt not hate: Countering online hate speech. In J. Pfeffer, C. Budak, Y. Lin, and F. Morstatter, editors, *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 369–380. AAAI Press, 2019. URL <https://aaai.org/ojs/index.php/ICWSM/article/view/3237>.
- P. Mathur, R. R. Shah, R. Sawhney, and D. Mahata. Detecting offensive tweets in hindi-english code-switched language. In L. Ku and C. Li, editors, *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*,

*SocialNLP@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 18–26. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-3504. URL <https://doi.org/10.18653/v1/w18-3504>.

- A. McEnery. *Swearing in English: blasphemy, purity and power from 1586 to the present*. London: Routledge, 2006.
- M. R. Mehl and J. W. Pennebaker. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, 84(4):857, 2003.
- S. Menini, G. Moretti, M. Corazza, E. Cabrio, S. Tonelli, and S. Villata. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3511. URL <https://www.aclweb.org/anthology/W19-3511>.
- J. S. Meyer and B. Gambäck. A platform agnostic dual-strand hate speech detector. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 146–156, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3516. URL <https://www.aclweb.org/anthology/W19-3516>.
- P. Michal, D. Pawel, M. Tatsuaki, M. Fumito, R. Rafal, A. Kenji, and M. Yoshio. In the service of online order: Tackling cyber-bullying with machine learning and affect analysis. *International Journal of Computational Linguistics Research*, 1(3):135–154, 2010.
- P. Mishra, M. D. Tredici, H. Yannakoudakis, and E. Shutova. Author profiling for abuse detection. In E. M. Bender, L. Derczynski, and P. Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1088–1098. Association for Computational Linguistics, 2018. URL <https://www.aclweb.org/anthology/C18-1093/>.
- Z. Mossie and J.-H. Wang. Social network hate speech detection for amharic language. *Computer Science & Information Technology*, pages 41–55, 2018.
- Z. Mossie and J.-H. Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, page 102087, 2019.
- M. Mozafari, R. Farahbakhsh, and N. Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, editors, *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 928–940. Springer, 2019. doi: 10.1007/978-3-030-36687-2\\_77. URL [https://doi.org/10.1007/978-3-030-36687-2\\_77](https://doi.org/10.1007/978-3-030-36687-2_77).

- M. Mozafari, R. Farahbakhsh, and N. Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.
- H. Mubarak, K. Darwish, and W. Magdy. Abusive language detection on arabic social media. In Z. Waseem, W. H. K. Chung, D. Hovy, and J. R. Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017*, pages 52–56. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-3008. URL <https://doi.org/10.18653/v1/w17-3008>.
- H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3512. URL <https://www.aclweb.org/anthology/W19-3512>.
- K. Müller and C. Schwarz. Fanning the flames of hate: Social media and hate crime. Available at SSRN 3082972, 2019.
- K. Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.
- G. Nascimento, F. Carvalho, A. M. da Cunha, C. R. Viana, and G. P. Guedes. Hate speech detection using brazilian imageboards. In J. A. F. dos Santos and D. C. Muchaluat-Saade, editors, *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia 2019, Rio de Janeiro, Brazil, October 29 - November 01, 2019*, pages 325–328. ACM, 2019. doi: 10.1145/3323503.3360619. URL <https://doi.org/10.1145/3323503.3360619>.
- R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. URL <https://doi.org/10.1016/j.artint.2012.07.001>.
- I. Nejadgholi and S. Kiritchenko. On cross-dataset generalization in automatic detection of online abuse. *CoRR*, abs/2010.07414, 2020. URL <https://arxiv.org/abs/2010.07414>.
- V. Nina-Alcocer. AMI at ibereval2018 automatic misogyny identification in spanish and english tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 274–279. CEUR-WS.org, 2018. URL [http://ceur-ws.org/Vol-2150/AMI\\_paper8.pdf](http://ceur-ws.org/Vol-2150/AMI_paper8.pdf).
- R. Nithyanand, B. Schaffner, and P. Gill. Measuring offensive speech in online political discourse. In J. Penney and N. Weaver, editors, *7th USENIX Workshop on Free and Open Communications on the Internet, FOCI 2017, Vancouver, BC, Canada, August*

- 14, 2017. USENIX Association, 2017. URL <https://www.usenix.org/conference/foci17/workshop-program/presentation/nithyanand>.
- J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/L16-1262>.
- C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.
- D. Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.114. URL <https://aclanthology.org/2021.acl-short.114>.
- A. Olteanu, C. Castillo, J. Boy, and K. R. Varshney. The effect of extremist violence on hateful speech online. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25–28, 2018*, pages 221–230. AAAI Press, 2018. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17908>.
- E. Ombui, L. Muchemi, and P. Wagacha. Hate speech detection in code-switched text messages. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6. IEEE, 2019.
- O. Oriola and E. Kotzé. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509, 2020. doi: 10.1109/ACCESS.2020.2968173. URL <https://doi.org/10.1109/ACCESS.2020.2968173>.
- N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1474. URL <https://www.aclweb.org/anthology/D19-1474>.
- K. B. Ozler, K. Kenski, S. Rains, Y. Shmargad, K. Coe, and S. Bethard. Fine-tuning for multi-domain and multi-label uncivil language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 28–33, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.4. URL <https://www.aclweb.org/anthology/2020.alw-1.4>.

- E. W. Pamungkas and V. Patti. #NonDicevoSulSerio at SemEval-2018 task 3: Exploiting emojis and affective content for irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 649–654, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1106. URL <https://aclanthology.org/S18-1106>.
- E. W. Pamungkas and V. Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In F. Alva-Manchego, E. Choi, and D. Khashabi, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 363–370. Association for Computational Linguistics, 2019. URL <https://www.aclweb.org/anthology/P19-2051/>.
- E. W. Pamungkas, V. Basile, and V. Patti. Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. In A. Cuzzocrea, F. Bonchi, and D. Gunopoulos, editors, *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018*, volume 2482 of *CEUR Workshop Proceedings*, pages 1–7. CEUR-WS.org, 2018a. URL <http://ceur-ws.org/Vol-2482/paper37.pdf>.
- E. W. Pamungkas, A. T. Cignarella, V. Basile, and V. Patti. Automatic identification of misogyny in english and italian tweets at EVALITA 2018 with a multilingual hate lexicon. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–9. CEUR-WS.org, 2018b. URL <http://ceur-ws.org/Vol-2263/paper033.pdf>.
- E. W. Pamungkas, A. T. Cignarella, V. Basile, and V. Patti. 14-ExLab@UniTo for AMI at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 234–241. CEUR-WS.org, 2018c. URL [http://ceur-ws.org/Vol-2150/AMI\\_paper2.pdf](http://ceur-ws.org/Vol-2150/AMI_paper2.pdf).
- E. W. Pamungkas, V. Basile, and V. Patti. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://www.aclweb.org/anthology/2020.lrec-1.765>.

- E. W. Pamungkas, V. Basile, and V. Patti. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management*, 57(6):102360, 2020b. URL <https://www.sciencedirect.com/science/article/pii/S0306457320308554>.
- E. W. Pamungkas, V. Basile, and V. Patti. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544, 2021a. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2021.102544>. URL <https://www.sciencedirect.com/science/article/pii/S0306457321000510>.
- E. W. Pamungkas, V. Basile, and V. Patti. Towards Multidomain and Multilingual Abusive Language Detection: A Survey. *Personal and Ubiquitous Computing*, 2021b. URL <https://link.springer.com/article/10.1007/s00779-021-01609-1>. Published online: 11 August 2021.
- S. J. Pan, X. Ni, J. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 751–760. ACM, 2010. doi: 10.1145/1772690.1772767. URL <https://doi.org/10.1145/1772690.1772767>.
- J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL <https://www.aclweb.org/anthology/D18-1302>.
- J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In S. G. Lukosch, A. Sarcevic, M. Lewkowicz, and M. J. Muller, editors, *Proceedings of the 19th International Conference on Supporting Group Work, Sanibel Island, FL, USA, November 13 - 16, 2016*, pages 369–374. ACM, 2016. doi: 10.1145/2957276.2957297. URL <https://doi.org/10.1145/2957276.2957297>.
- J. Pavlopoulos, P. Malakasiotis, J. Bakagianni, and I. Androutsopoulos. Improved abusive comment moderation with user embeddings. In O. Popescu and C. Strapparava, editors, *Proceedings of the 2017 Workshop: Natural Language Processing meets Journalism, NLPmJ@EMNLP, Copenhagen, Denmark, September 7, 2017*, pages 51–55. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-4209. URL <https://doi.org/10.18653/v1/w17-4209>.
- M. Peng, Q. Zhang, Y.-g. Jiang, and X. Huang. Cross-Domain Sentiment Classification with Target Domain Specific Information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1233. URL <https://www.aclweb.org/anthology/P18-1233>.

- J. C. Pereira-Kohatsu, L. Q. Sánchez, F. Liberatore, and M. Camacho-Collados. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19(21):4654, 2019. doi: 10.3390/s19214654. URL <https://doi.org/10.3390/s19214654>.
- J. M. Pérez, A. Arango, and F. Luque. ANDES at SemEval-2020 task 12: A jointly-trained BERT multilingual model for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1524–1531, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.semeval-1.199>.
- M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In M. A. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1202. URL <https://doi.org/10.18653/v1/n18-1202>.
- S. Pinker. *The stuff of thought: Language as a window into human nature*. Penguin, 2007.
- Z. Pitenis, M. Zampieri, and T. Ranasinghe. Offensive language identification in greek. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5113–5119. European Language Resources Association, 2020. URL <https://www.aclweb.org/anthology/2020.lrec-1.629/>.
- B. Poland. *Haters: Harassment, abuse, and violence online*. U of Nebraska Press, 2016.
- F. Poletto, V. Basile, C. Bosco, V. Patti, and M. Stranisci. Annotating hate speech: Three schemes at comparison. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL <http://ceur-ws.org/Vol-2481/paper56.pdf>.
- F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and Benchmark Corpora for Hate Speech Detection: a Systematic Review. *Language Resources and Evaluation*, 2020. URL <https://link.springer.com/article/10.1007/s10579-020-09502-8>.
- M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, 2019. URL <http://ceur-ws.org/Vol-2481/paper57.pdf>.
- S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

- E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso. Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management*, 58(6):102674, 2021.
- M. Ptaszynski, A. Pieciukiewicz, and P. Dybała. Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter. In *Proceedings of the PolEval 2019 Workshop*, page 89, 2019.
- J. Qian, M. ElSherief, E. Belding, and W. Y. Wang. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2019. URL <https://www.aclweb.org/anthology/N18-2019>.
- J. Qian, M. ElSherief, E. M. Belding, and W. Y. Wang. Hierarchical CVAE for fine-grained hate speech classification. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3550–3559. Association for Computational Linguistics, 2018b. doi: 10.18653/v1/d18-1391. URL <https://doi.org/10.18653/v1/d18-1391>.
- J. Qian, A. Bethke, Y. Liu, E. M. Belding, and W. Y. Wang. A benchmark dataset for learning to intervene in online hate speech. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4754–4763. Association for Computational Linguistics, 2019a. doi: 10.18653/v1/D19-1482. URL <https://doi.org/10.18653/v1/D19-1482>.
- J. Qian, M. ElSherief, E. M. Belding, and W. Y. Wang. Learning to decipher hate symbols. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3006–3015. Association for Computational Linguistics, 2019b. doi: 10.18653/v1/n19-1305. URL <https://doi.org/10.18653/v1/n19-1305>.
- B. Radfar, K. Shivaram, and A. Culotta. Characterizing variation in toxic language by social context. In M. D. Choudhury, R. Chunara, A. Culotta, and B. F. Welles, editors, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 959–963. AAAI Press, 2020. URL <https://aaai.org/ojs/index.php/ICWSM/article/view/7366>.
- H. Rainie, J. Q. Anderson, and J. Albright. *The future of free speech, trolls, anonymity and fake news online*. Pew Research Center Washington, DC, 2017.

- S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova. Joint modelling of emotion and abusive language detection. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4270–4279. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.394. URL <https://doi.org/10.18653/v1/2020.acl-main.394>.
- T. Ranasinghe and M. Zampieri. Multilingual offensive language identification with cross-lingual embeddings. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5838–5844. Association for Computational Linguistics, 2020. URL <https://www.aclweb.org/anthology/2020.emnlp-main.470/>.
- P. Rani, S. Suryawanshi, K. Goswami, B. R. Chakravarthi, T. Fransen, and J. P. McCrae. A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 42–48, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL <https://www.aclweb.org/anthology/2020.trac-1.7>.
- A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer, 2010.
- N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. *CoRR*, abs/2004.09813, 2020. URL <https://arxiv.org/abs/2004.09813>.
- M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. P. Sheth. A quality type-aware annotated corpus and lexicon for harassment research. In H. Akkermans, K. Fontaine, I. Vermeulen, G. Houben, and M. S. Weber, editors, *Proceedings of the 10th ACM Conference on Web Science, WebSci 2018, Amsterdam, The Netherlands, May 27-30, 2018*, pages 33–36. ACM, 2018. doi: 10.1145/3201064.3201103. URL <https://doi.org/10.1145/3201064.3201103>.
- M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. M. Jr. Characterizing and detecting hateful users on twitter. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 676–679. AAAI Press, 2018. URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17837>.
- L. Richardson-Self. Woman-hating: On misogyny, sexism, and hate speech. *Hypatia*, 33(2):256–272, 2018. doi: 10.1111/hypa.12398. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/hypa.12398>.

- R. W. Rieber, C. Wiedemann, and J. D’Amato. Obscenity: Its frequency and context of usage as compared in males, nonfeminist females, and feminist females. *Journal of Psycholinguistic Research*, 8(3):201–223, 1979.
- M. Rizoiu, T. Wang, G. Ferraro, and H. Suominen. Transfer learning for hate speech detection in social media. *CoRR*, abs/1906.03829, 2019. URL <http://arxiv.org/abs/1906.03829>.
- S. Rojas-Galeano. On obstructing obscenity obfuscation. *ACM Transactions on the Web (TWEB)*, 11(2):12, 2017.
- F. Ronzano, F. Barbieri, E. Wahyu Pamungkas, V. Patti, F. Chiusaroli, et al. Overview of the evalita 2018 italian emoji prediction (itamoji) task. In *6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018*, volume 2263, pages 1–9. CEUR-WS, 2018. URL <http://ceur-ws.org/Vol-2263/paper004.pdf>.
- H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur. Using fuzzy fingerprints for cyberbullying detection in social networks. In *2018 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–7. IEEE, 2018. doi: 10.1109/FUZZ-IEEE.2018.8491557. URL <https://doi.org/10.1109/FUZZ-IEEE.2018.8491557>.
- B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118, 2017. URL <http://arxiv.org/abs/1701.08118>.
- H. Ross. Patterns of swearing. *Discovery: The Popular Journal of Knowledge*, pages 479–481, 1969.
- S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard. Sluice networks: Learning what to share between loosely related tasks. *stat*, 1050:23, 2017.
- N. Safi Samghabadi, A. Hatami, M. Shafaei, S. Kar, and T. Solorio. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.10. URL <https://www.aclweb.org/anthology/2020.alw-1.10>.
- P. Saha, B. Mathew, P. Goyal, and A. Mukherjee. Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*, 2018.
- P. Saha, B. Mathew, P. Goyal, and A. Mukherjee. Hatemonitors: Language agnostic abuse detection in social media. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, pages 246–253, Kolkata, India, December 2019.

- J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, and B. J. Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1, 2020.
- M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci. An italian Twitter corpus of hate speech against immigrants. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, pages 2798–2805. European Language Resources Association (ELRA), 2018. URL <http://www.lrec-conf.org/proceedings/lrec2018/summaries/710.html>.
- M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL <http://ceur-ws.org/Vol-2765/paper162.pdf>.
- J. Schäfer and B. Burtenshaw. Offence in dialogues: A corpus-based study. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019*, pages 1085–1093. INCOMA Ltd., 2019. doi: 10.26615/978-954-452-056-4\\_125. URL [https://doi.org/10.26615/978-954-452-056-4\\_125](https://doi.org/10.26615/978-954-452-056-4_125).
- A. Schmidt and M. Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL <https://www.aclweb.org/anthology/W17-1101>.
- J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele, and G. Rehm. Towards the automatic classification of offensive language and related phenomena in german tweets. In *14th Conference on Natural Language Processing KONVENS 2018*, page 95, 2018.
- S. Schuster, S. Gupta, R. Shah, and M. Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3795–3805. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1380. URL <https://doi.org/10.18653/v1/n19-1380>.
- S. Sharifirad and A. Jacovi. Learning and understanding different categories of sexism using convolutional neural network's filters. In *Proceedings of the 2019 Workshop on*

- Widening NLP*, pages 21–23, Florence, Italy, Aug. 2019. Association for Computational Linguistics. URL <https://aclanthology.org/W19-3609>.
- H. K. Sharma, K. Kshitiz, et al. Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 265–272. IEEE, 2018.
- A. Shirbandi and B. Moradi. Comparative Study of Combination of Convolutional and Recurrent Neural Network for Natural Language Processing. Technical report, EasyChair, 2019.
- E. Shushkevich and J. Cardiff. Misogyny detection and classification in english tweets: The experience of the ITT team. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org, 2018a. URL <http://ceur-ws.org/Vol-2263/paper030.pdf>.
- E. Shushkevich and J. Cardiff. Classifying misogynistic tweets using a blended model: The AMI shared task in IBEREVAL 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 255–259. CEUR-WS.org, 2018b. URL [http://ceur-ws.org/Vol-2150/AMI\\_paper5.pdf](http://ceur-ws.org/Vol-2150/AMI_paper5.pdf).
- G. I. Sigurbergsson and L. Derczynski. Offensive language and hate speech detection for danish. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3498–3508. European Language Resources Association, 2020. URL <https://www.aclweb.org/anthology/2020.lrec-1.430/>.
- A. Singh, E. Blanco, and W. Jin. Incorporating Emoji Descriptions Improves Tweet Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2096–2101, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1214. URL <https://aclanthology.org/N19-1214>.
- S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International*

*Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.* OpenReview.net, 2017. URL <https://openreview.net/forum?id=r1Aab85gg>.

- H. Sohn and H. Lee. MC-BERT4HATE: hate speech detection using multi-channel BERT for different languages and translations. In P. Papapetrou, X. Cheng, and Q. He, editors, *2019 International Conference on Data Mining Workshops, ICDM Workshops 2019, Beijing, China, November 8-11, 2019*, pages 551–559. IEEE, 2019. doi: 10.1109/ICDMW.2019.00084. URL <https://doi.org/10.1109/ICDMW.2019.00084>.
- S. Sood, J. Antin, and E. Churchill. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490. ACM, 2012.
- R. Sprugnoli, S. Menini, S. Tonelli, F. Oncini, and E. Piras. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5107. URL <https://www.aclweb.org/anthology/W18-5107>.
- L. Stappen, F. Brunn, and B. W. Schuller. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850, 2020. URL <https://arxiv.org/abs/2004.13850>.
- J. Steinberger, T. Brychcín, T. Hercig, and P. Krejzl. Cross-lingual flames detection in news discussions. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 694–700. INCOMA Ltd., 2017. doi: 10.26615/978-954-452-049-6\\_089. URL [https://doi.org/10.26615/978-954-452-049-6\\_089](https://doi.org/10.26615/978-954-452-049-6_089).
- R. Stephens and C. Umland. Swearing as a response to pain-effect of daily swearing frequency. *The Journal of Pain*, 12(12):1274–1281, 2011.
- M. Straka, J. Hajič, and J. Straková. UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/L16-1680>.
- E. Sulis, D. I. H. Farías, P. Rosso, V. Patti, and G. Ruffo. Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge Based Systems*, 108:132–143, 2016. doi: 10.1016/j.knosys.2016.05.035. URL <https://doi.org/10.1016/j.knosys.2016.05.035>.

- S. D. Swamy, A. Jamatia, and B. Gambäck. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1088. URL <https://www.aclweb.org/anthology/K19-1088>.
- D. Tang, B. Qin, X. Feng, and T. Liu. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL <https://www.aclweb.org/anthology/C16-1311>.
- M. Thelwall. Fk yea i swear: cursing and gender in myspace. *Corpora*, 3(1):83–107, 2008.
- C. Themeli, G. Giannakopoulos, and N. Pittaras. A study of text representations in hate speech detection. *CoRR*, abs/2102.04521, 2021. URL <https://arxiv.org/abs/2102.04521>.
- S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans. The automated detection of racist discourse in Dutch social media. *Computational Linguistics in the Netherlands Journal*, 6:3–20, 2016.
- R. van der Goot, N. Ljubesic, I. Matroos, M. Nissim, and B. Plank. Bleaching text: Abstract features for cross-lingual gender prediction. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 383–389. Association for Computational Linguistics, 2018. URL <https://aclanthology.info/papers/P18-2061/p18-2061>.
- C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria, Sept. 2015. INCOMA Ltd. Shoumen, BULGARIA. URL <https://www.aclweb.org/anthology/R15-1086>.
- C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10), 2018.
- N. Vashistha and A. Zubiaga. Online multilingual hate speech detection: experimenting with hindi and english social media. *Information*, 12(1):5, 2021.
- B. Vidgen and L. Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12):e0243300, 2020.
- B. Vidgen and T. Yasseri. Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78, 2020.

- B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3509. URL <https://www.aclweb.org/anthology/W19-3509>.
- F. D. Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In A. Armando, R. Baldoni, and R. Focardi, editors, *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017*, volume 1816 of *CEUR Workshop Proceedings*, pages 86–95. CEUR-WS.org, 2017. URL <http://ceur-ws.org/Vol-1816/paper-09.pdf>.
- X. Vu, T. Vu, M. Tran, T. Le-Cong, and H. T. M. Nguyen. HSD shared task in VLSP campaign 2019: Hate speech detection for social good. *CoRR*, abs/2007.06493, 2020. URL <https://arxiv.org/abs/2007.06493>.
- B. C. Wallace, L. Kertz, E. Charniak, et al. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–516, 2014.
- K. Wang, D. Lu, S. C. Han, S. Long, and J. Poon. Detect all abuse! toward universal abusive language detection models. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6366–6376. International Committee on Computational Linguistics, 2020. URL <https://www.aclweb.org/anthology/2020.coling-main.560/>.
- W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. Cursing in English on twitter. In S. R. Fussell, W. G. Lutters, M. R. Morris, and M. Reddy, editors, *Computer Supported Cooperative Work, CSCW ’14, Baltimore, MD, USA, February 15-19, 2014*, pages 415–425. ACM, 2014a. doi: 10.1145/2531602.2531734. URL <https://doi.org/10.1145/2531602.2531734>.
- W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 415–425. ACM, 2014b.
- Y. Wang, M. Huang, X. Zhu, and L. Zhao. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1058. URL <https://www.aclweb.org/anthology/D16-1058>.
- W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada,

- June 2012. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W12-2103>.
- Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL <https://www.aclweb.org/anthology/N16-2013>.
- Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3012. URL <https://www.aclweb.org/anthology/W17-3012>.
- Z. Waseem, J. Thorne, and J. Bingel. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online Harassment*, pages 29–55. Springer, 2018.
- M. Wiegand and J. Ruppenhofer. Exploiting emojis for abusive language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 369–380, Online, Apr. 2021. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2021.eacl-main.28>.
- M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1095. URL <https://www.aclweb.org/anthology/N18-1095>.
- M. Wiegand, M. Siegel, and J. Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*, page 1, 2018b.
- M. Wiegand, J. Ruppenhofer, and T. Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL <https://www.aclweb.org/anthology/N19-1060>.
- M. Wiegand, J. Ruppenhofer, and E. Eder. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

*Linguistics: Human Language Technologies*, pages 576–587, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nacl-main.48. URL <https://aclanthology.org/2021.nacl-main.48>.

- M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp. Hate in the machine: anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117, 2020.
- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771, 2019.
- L. Wright, D. Ruths, K. P. Dillon, H. M. Saleem, and S. Benesch. Vectors for counterspeech on twitter. In Z. Waseem, W. H. K. Chung, D. Hovy, and J. R. Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017*, pages 57–62. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-3009. URL <https://doi.org/10.18653/v1/w17-3009>.
- S. Wu and M. Dredze. Are all languages created equal in multilingual bert? In S. Gella, J. Welbl, M. Rei, F. Petroni, P. S. H. Lewis, E. Strubell, M. J. Seo, and H. Hajishirzi, editors, *Proceedings of the 5th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2020, Online, July 9, 2020*, pages 120–130. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.repl4nlp-1.16. URL <https://doi.org/10.18653/v1/2020.repl4nlp-1.16>.
- E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM, 2017. doi: 10.1145/3038912.3052591. URL <https://doi.org/10.1145/3038912.3052591>.
- D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB*, volume 2, pages 1–7, 2009.
- Z. Yuan, S. Wu, F. Wu, J. Liu, and Y. Huang. Domain attention model for multi-domain sentiment classification. *Knowl. Based Syst.*, 155:1–10, 2018. doi: 10.1016/j.knosys.2018.05.004. URL <https://doi.org/10.1016/j.knosys.2018.05.004>.
- M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1144. URL <https://www.aclweb.org/anthology/N19-1144>.

- M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010. URL <https://www.aclweb.org/anthology/S19-2010>.
- M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.semeval-1.188>.
- S. Zannettou, J. Finkelstein, B. Bradlyn, and J. Blackburn. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 786–797, 2020.
- K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, and E. Chen. Interactive Attention Transfer Network for Cross-Domain Sentiment Classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5773–5780. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33015773. URL <https://doi.org/10.1609/aaai.v33i01.33015773>.
- X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In *15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016, Anaheim, CA, USA, December 18-20, 2016*, pages 740–745. IEEE Computer Society, 2016. doi: 10.1109/ICMLA.2016.0132. URL <https://doi.org/10.1109/ICMLA.2016.0132>.