Information Processing and Management 58 (2021) 102544

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection

Endang Wahyu Pamungkas [*], Valerio Basile, Viviana Patti

Department of Computer Science, University of Turin, Italy

ARTICLE INFO

Keywords:
Hate speech detection
Cross-lingual classification
Social media
Transfer learning
Zero-shot learning

ABSTRACT

Hate speech is an increasingly important societal issue in the era of digital communication. Hateful expressions often make use of figurative language and, although they represent, in some sense, the dark side of language, they are also often prime examples of creative use of language. While hate speech is a global phenomenon, current studies on automatic hate speech detection are typically framed in a monolingual setting. In this work, 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 experiment 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.

1. Introduction

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 media platforms, governments (Corazza, Menini, Cabrio, Tonelli, & Villata, 2020) and societies, affecting several public debates about immigration, security, and multiculturalism: aggressive and stereotypical statements hinder a constructive dialog between users, thus seriously obstructing the achievement of an equal, cohesive, and inclusive society. Hate Speech (HS) can be defined as any type of communication that is abusive,

[*] Corresponding author.
E-mail addresses: pamungka@di.unito.it (E.W. Pamungkas), valerio.basile@unito.it (V. Basile), viviana.patti@unito.it (V. Patti).

https://doi.org/10.1016/j.ipm.2021.102544
Received 1 November 2020; Received in revised form 23 December 2020; Accepted 7 February 2021
Available online 1 March 2021

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 & Kovačič, 2012). HS is *assaultive*: it verbally attacks, degrades, and persecutes its targets. Psychological research has shown that assaultive speech does not only harm its targets but also triggers prejudice in bystanders. It elicits social exclusion (Leader, Mullen, & Rice, 2009), and higher suicide rates (Mullen & Smyth, 2004) among targets; and prompts prejudice (Soral, Bilewicz, & Winiewski, 2018), intentional bias, and dehumanization in the audience.

Hate speech is becoming a significant problem in online communication on social media, with effects that potentially may result in dangerous criminal acts offline (Müller & Schwarz, 2019; Williams, Burnap, Javed, Liu, & Ozalp, 2020). As an example of an extreme case, in Rohingya, Myanmar, in 2017, hate speech on social media has been heavily implicated in inciting violence against the Rohingya Muslim Minority, including the murder of thousands of civilians, and ICT companies controlling social media platforms had to admit that they failed to prevent such platforms from being used to "foment division and incite offline violence".[1]

With the huge amount of user-generated content on social media, manually analyzing hate speech is impractical. Many studies have been proposed to automate the detection of hate speech in social communication. Most works utilize a supervised approach, and recently deep learning approaches have been applied, achieving state-of-the-art results for some languages (Badjatiya, Gupta, Gupta, & Varma, 2017; Mozafari, Farahbakhsh, & Crespi, 2019). However, most of the proposed models are tested in monolingual settings, mostly in English, while there is an urgent need to develop robust systems to identify online hate speech across multiple languages, considering how is it 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 (EU Commission, 2016).

Similar to other natural language processing (NLP) tasks (Joshi, Santy, Budhiraja, Bali, & Choudhury, 2020), detecting hate speech in less-resourced languages is a prominent and timely challenge. For example, the aforementioned 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.[2] The current availability of datasets in many languages (Poletto, Basile, Sanguinetti, Bosco, & Patti, 2020), makes the time ripe for addressing the multilingual challenge. The main motivation of this work is to overcome the current issues faced by the NLP community in processing data from under-represented languages. This is particularly relevant for tasks dealing with online content, such as hate speech detection, as the Burmese incident demonstrates. Social media platforms need to be able to monitor and support the moderation of user-generated content in many languages, and cross-lingual knowledge transfer could provide an appropriate tool for this purpose.

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, Gupta, Shah, & Lewis, 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 (such as in this work) or one domain is employed to predict samples from a totally unseen language or domain (Goodfellow, Bengio, & Courville, 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).

This work 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 hatefulness 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, Lample, Ranzato, Denoyer, & Jégou, 2017) and Multilingual BERT.[3] 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, Basile, & Patti, 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, 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. In the monolingual setting, the usefulness of external knowledge from HurtLex in abusive language detection tasks has been proven in a previous study (Koufakou, Pamungkas, Basile, & Patti, 2020). 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

[1] https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

[2] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

[3] https://github.com/google-research/bert/blob/master/multilingual.md.

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. Some parts of this work are adopted from our previous study which focuses in the same research direction (Pamungkas & Patti, 2019).

1.1. Research questions

In this paper, we address the following research questions:

RQ1 *What neural architectures are effective for transferring knowledge between languages in a hate speech detection task?*
We review several approaches from related areas such as sentiment analysis, which specifically focus on cross-lingual transfer learning. We propose a joint-learning architecture by adapting several ideas from previous sentiment analysis studies. To provide a deeper analysis, we also implement several baseline models with simpler approaches.

RQ2 *How effective are multilingual pre-trained models for language representation in cross-lingual hate speech detection?*
The availability of multilingual language representations is crucial for the cross-lingual transfer learning task. Therefore, we test and evaluate three publicly available multilingual pre-trained models, namely LASER embeddings (Artetxe & Schwenk, 2019), Facebook MUSE (Conneau et al., 2017), and Multilingual BERT (Devlin, Chang, Lee, & Toutanova, 2019). We also build a model that relies on machine translation and the state-of-the-art pre-trained English BERT model.

RQ3 *What is the role of external multilingual knowledge in improving the model performance in cross-lingual hate speech detection?*
HurtLex is a handcrafted lexicon of offensive, aggressive and hateful words. The original version of this lexicon is in Italian, semi-automatically translated into 67 languages. To answer the third research question, we incorporate features from HurtLex into our models and measure their impact on the performance of the system.

1.2. Contribution

The contributions of this paper can be summarized as follows:

1. We propose a joint-learning architecture for the cross-lingual hate speech detection task. This contribution represents an effort towards creating a more effective model for state-of-the-art neural hate speech detection, therefore answering to RQ1.
2. We experiment with three different multilingual language representation models to transfer knowledge between a pair of languages. We evaluate the effectiveness of these language representation models by using several supervised models to answer RQ2.
3. We investigate the impact of an external resource by experimenting with a multilingual hate lexicon. The result of the analysis provides answers to RQ3 in terms of absolute performance improvement overall.
4. We conduct a deep analysis of the results to gain insight into the main challenges for this task. The results of our qualitative analysis sheds light on the impact of external resource on subsets of instances such as the ones containing idiomatic expressions, providing further evidence for RQ3.

The article is organized as follows. Section 2 introduces related work on this task, starting with several previous studies in hate speech detection. We also present recent work focusing on cross-lingual text classification and more specifically cross-lingual hate speech detection tasks. Section 3 describes the datasets and resources used in our work. In Section 4 we present the models employed in the experiments of this paper, including our proposed joint-learning models and the baseline models. The experimental results and analysis are described in Section 5. Section 6 presents a qualitative analysis based on our experimental results. Finally, Section 7 presents conclusive remarks and ideas for future work.

2. Related works

2.1. Hate speech detection

Automatic hate speech detection is a challenging task. Among the known issues, dataset bias has been pointed out by recent studies (Arango, Pérez, & Poblete, 2020; Wiegand, Ruppenhofer, & Kleinbauer, 2019). Another issue which contributes to the complexity of the task is the light barrier between hate speech and freedom of speech (Pamungkas, Basile, & Patti, 2020; Swamy, Jamatia, & Gambäck, 2019). We only found a few studies that used an unsupervised approach for this task, such as Gitari, Zuping, Damien, and Long (2015), where a manually-built lexicon is used to detect hate speech content. Most works tackle this task by adopting a supervised approach, employing several machine learning models, including but not exclusively neural models. Among the earliest proposed solutions several works relied on machine learning models with manually engineered features, including decision trees (Agarwal & Sureka, 2017; Burnap & Williams, 2015), naive bayes classifiers (Agarwal & Sureka, 2017; Kwok & Wang, 2013), support vector machines (Badjatiya et al., 2017; Burnap & Williams, 2015; Warner & Hirschberg, 2012), logistic regression (Badjatiya et al., 2017; Davidson, Warmsley, Macy, & Weber, 2017; Fehn Unsvåg & Gambäck, 2018; Waseem & Hovy, 2016), and random forest (Agarwal & Sureka, 2017; Badjatiya et al., 2017; Burnap & Williams, 2015). 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, Del Tredici, Yannakoudakis, & Shutova, 2019), and user-related features (Fehn Unsvåg & Gambäck, 2018; Waseem & Hovy, 2016)). Recent works relied on to use of neural-based approaches such as Long Short-Term Memory (LSTM) (Mishra et al., 2019; Vigna, Cimino, Dell'Orletta, Petrocchi, & Tesconi, 2017), Bidirectional Long Short-Term Memory (Bi-LSTM) (Qian, ElSherief, Belding, & Wang, 2018), Gated Recurrent Unit (GRU) (Mossie & Wang, 2019), and Convolutional Neural Network (CNN) (Badjatiya et al., 2017). These models are usually coupled with language representations such as FastText,[4] word2vec,[5] and ELMo (Peters et al., 2018).

Several shared tasks on hate speech detection have been proposed in recent years. Hate Speech and Offensive Content Identification (HASOC) (Mandl et al., 2019) is a hate speech and offensive language identification shared task at FIRE 2019, covering three languages: English, German, and Hindi. Automatic Misogyny Identification (AMI) (Fersini, Nozza and Rosso, 2018; Fersini, Rosso and Anzovino, 2018) focuses on detecting hate speech towards women. HatEval (Basile et al., 2019) was introduced at SemEval 2019 and focused on detecting hateful messages directed towards two specific targets, namely immigrants and women, in English and Spanish. Finally, HaSpeeDe (Bosco, Dell'Orletta, Poletto, Sanguinetti, & Tesconi, 2018) is a shared task in EVALITA 2018 focusing on automatic detection of hate speech towards immigrants in Italian.

2.2. Cross lingual sentiment analysis

We review the cross-lingual study in the sentiment analysis field, a more mature research area related to hate speech detection. The cross-lingual setting has been well-explored in sentiment analysis. The earliest and simplest approach employs automatic tools to translate the data from the source language to the target language, such as in Brooke, Tofiloski, and Taboada (2009) from English to Spanish, Demirtas and Pechenizkiy (2013) from English to Turkish, and Duh, Fujino, and Nagata (2011) from English to Japanese, French, and German. Other studies proposed to use a bilingual sentiment lexicon, which is also obtained via machine translation, to transfer knowledge between languages (Kim & Hovy, 2006; Meng et al., 2012; Mihalcea, Banea, & Wiebe, 2007; Wilson, Wiebe, & Hoffmann, 2005). Despite the limitations of machine translation, these studies found that this simple approach is often quite effective.

With the advancement of neural-based approaches, recent studies focus on exploiting joint-learning architectures that allow the model to learn sentiment information of messages in both languages (source and target) sequentially. Zhou, Wan, and Xiao (2016) proposed an architecture called Bilingual Document Representation Learning (BiDRL), which learns text representations in both languages simultaneously, subsequently passing it to a joint-learning model to predict the sentiment of the text. Feng and Wan (2019) proposed an end-to-end model that uses Long Short-Term Memory (LSTM) as a shared layer to jointly learn sentiment from different languages and multiple domains. Another recent work by Chen, Sun, Athiwaratkun, Cardie, and Weinberger (2018) models cross-lingual sentiment analysis as a multitask learning problem, which allows the model to learn the sentiment and language simultaneously. Other studies focus on investigating the availability of bilingual or multilingual language models for aligning representations between different languages (Jebbara & Cimiano, 2019; Pelicon, Pranjić, Miljković, Škrlj, & Pollak, 2020; Sarkar, Reddy, & Iyengar, 2019). Finally, there have been attempts to exploit social signal information by using emoji as a language-agnostic feature to improve cross-lingual transfer (Chen et al., 2019).

2.3. Multilingual hate speech detection

Recent works evaluated the robustness of the proposed models across multiple languages without experimenting in a cross-lingual setting. Corazza et al. (2020) proposed a robust architecture for detecting hate speech in three languages: Italian, Spanish, and German. Ibrohim and Budi (2019b) focused on evaluating several methods that utilize machine translation to detect HS in Indonesian, English, and Hindi. Ousidhoum, Lin, Zhang, Song, and Yeung (2019) proposed a multilingual HS dataset, which was evaluated with a state-of-the-art multitask learning approach.

Studies on hate speech detection in a cross-lingual setting are still sparse. Stappen, Brunn, and Schuller (2020) 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 of hate speech from English to Spanish and vice versa. Aluru, Mathew, Saha, and Mukherjee (2020) conducted an exploratory work using deep learning approaches to multilingual hate speech detection in nine languages, utilizing several existing models. They found that a logistic regression model performs better on low-resource languages, while BERT is superior in resource-rich languages. Similarly, Rodriguez and Saynova (2020) performed experiments with a few-shot learning approach to detect hate speech in five languages, mainly investigating the dataset scarcity, which contributes to the difficulties of this task. Arango et al. (2020) ran a cross-lingual classification of hate speech in English and Spanish as part of their work on investigating the dataset bias issue in hate speech detection task.

[4] https://fasttext.cc/.

[5] https://code.google.com/archive/p/word2vec/.

Table 1
Size and class distribution of the datasets used in the experiments. HSR is a hate speech instance ratio over all data.

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 hate speech	12,971	0.42
	Founta et al. (2018)	offensive, abusive, aggressive, cyberbullying, spam, and none	58,722	0.08
	Ousidhoum et al. (2019)	hate speech, abusive, offensive, disrespectful, fearful, and normal	1,939	0.66
FR	Ousidhoum et al. (2019)	hate speech, abusive, offensive, disrespectful, fearful, and normal	1,220	0.33
DE	Mandl et al. (2019)	hate speech, not hate speech	4,743	0.03
	Ross et al. (2017)	hate speech, not hate speech	369	0.15
ID	Ibrohim and Budi (2019a)	hate speech, abusive, and neither	13,169	0.42
	Alfina, Mulia, Fanany, and Ekanata (2017)	hateful and normal	713	0.36
IT	Bosco et al. (2018)	hate speech, not hate speech	4,000	0.32
PT	Fortuna, Rocha da Silva, Soler-Company, Wanner, and Nunes (2019)	hate speech, not hate speech	5,670	0.31
ES	Basile et al. (2019)	hate speech, not hate speech	6,599	0.42
	Pereira-Kohatsu, Sánchez, Liberatore, and Camacho-Collados (2019)	hate speech, not hate speech	6,000	0.26

3. Data and resources

3.1. Dataset

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.[6] 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 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*. This corpus was scraped from Twitter by using Twitter Search API based on keywords obtained from HateBase.[7] The collection was manually rated by at least three annotators using the CrowdFlower platform.[8] The final label of each tweet was assigned based on a majority vote, with the inter-annotator agreement of the overall dataset reaching about 0.92. Most instances were labeled as *offensive* (77.4%), while *hate speech* only 5.8%, and the remaining 16.8% were labeled as *neither*.

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*. This dataset was manually annotated by using the Figure Eight platform (now called Appen) with three layers of annotation, including hatefulness (hate speech or not), target range (generic or individual), and aggressiveness (aggressive or not). The annotation process achieved a quite high inter-annotator agreement (0.83, 0.73, and 0.70 respectively). This corpus has been used for HatEval 2019, a shared task at SemEval 2019, which focuses on multilingual hate speech detection.

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*. The initial collection of this dataset contains 30 million tweets gathered by using Twitter Stream API during the period from 30 March 2017 to 9 April 2017. A minimum of five crowd workers annotated each instance, and the final label was decided based on a majority vote.

Ousidhoum et al. Dataset. This dataset (Ousidhoum et al., 2019) contains 13,014 tweets and consists of three different languages: English (5647), French (4014), and Arabic (3353). The dataset was annotated by using a crowdsourcing with the Amazon Mechanical Turk platform.[9] 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 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).[10] 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.

[6] http://hatespeechdata.com/.

[7] http://www.hatebase.org.

[8] Now Appen https://appen.com/.

[9] https://www.mturk.com/.

[10] We combine training and testing data and obtain the number as presented in Table 1.

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 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 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 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 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 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.

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. Below are the description of each model:

LASER Embeddings. Language-Agnostic SEntence Representations (LASER) (Artetxe & 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 pairs of languages to align the embedding representation. As shown on the Github page,[11] 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.[12]

HurtLex. HurtLex is a multilingual lexicon of hate words, originally built from 1082 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.[13] The full description and abbreviation of each category is presented in Table 2. In this work, we only rely on seven languages of HurtLex.

[11] https://github.com/facebookresearch/MUSE.

[12] https://github.com/google-research/bert/blob/master/multilingual.md.

[13] http://hatespeech.di.unito.it/resources.html.

Table 2
HurtLex categories.

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

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 (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 three novel models based on a joint-learning approach. The rest of the models were 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 & 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.[14]

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,[15] which allows us to integrate BERT with the Keras functional layer.[16] 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.[17] 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).[18] Fig. 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 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

[14] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[15] https://www.tensorflow.org/hub.

[16] https://keras.io/.

[17] https://translate.google.com/.

[18] https://github.com/facebookresearch/MUSE.

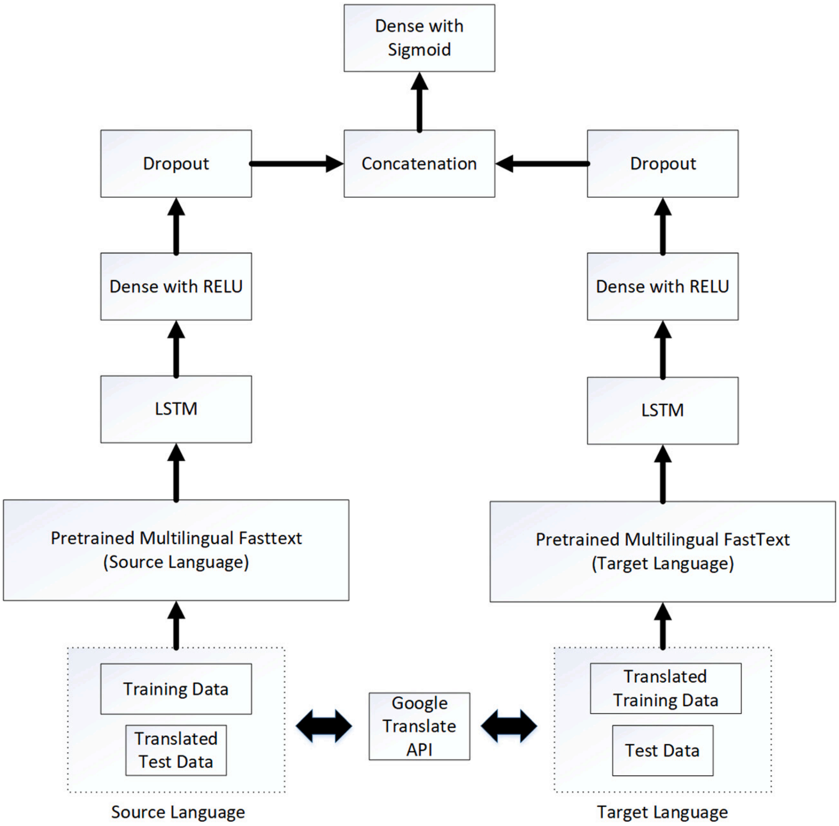

Fig. 1. Joint-Learning LSTM Model Architecture.

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 Fig. 4. We use two HurtLex embeddings in each architecture to accommodate the input from the source and target languages (see Fig. 2).

Joint-Learning Model Based on Multilingual BERT. We incorporated Multilingual BERT (`bert-multi-cased`) into our joint-learning architecture — see Fig. 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 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 Fig. 4.

5. Results and analysis

Table 3 shows the results of our experiment. First, we will 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 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, Basile, & Patti, 2020b), resulting in poor performance.

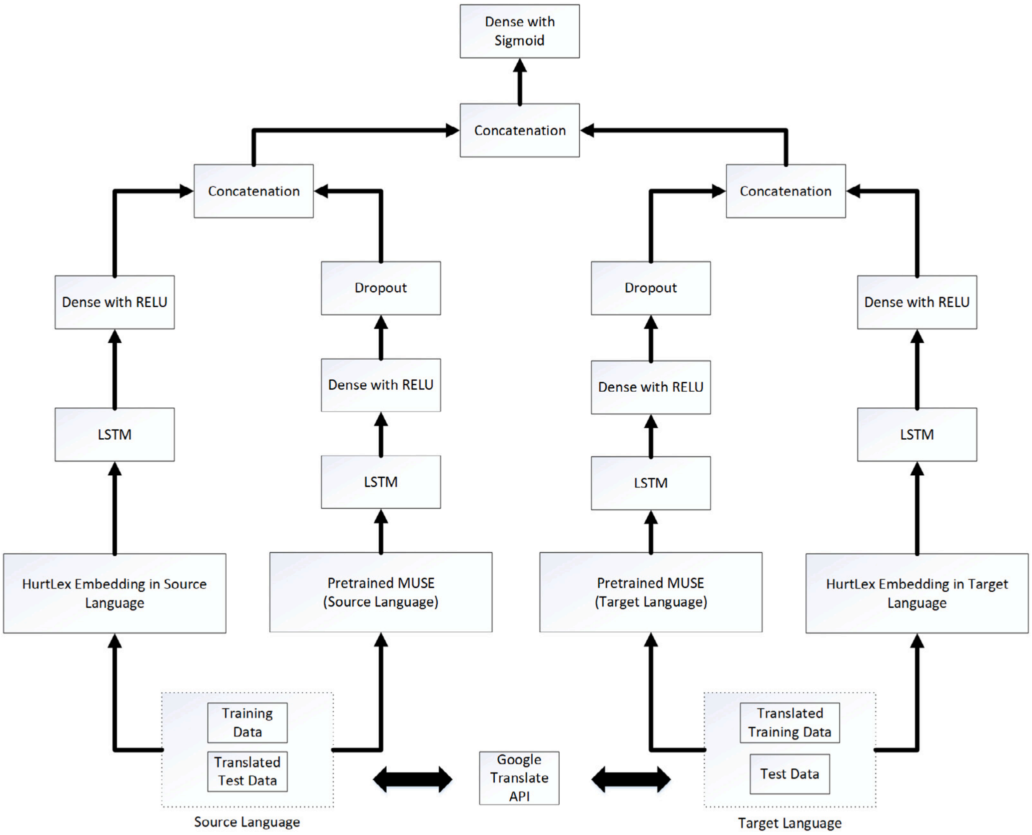

Fig. 2. Joint-Learning LSTM-HurtLex Model Architecture.

Table 3
Results of cross-lingual hate speech detection on the original distribution of training sets.

	LR + Laser								BERT Multilingual							
	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	**.634**	.672	.740	.657	.696	.427	.526	.471	.584	.614
DE	.973	.903	.937	.178	.456	.254	.595	.883	.971	.928	.949	.201	.398	.268	**.608**	.905
ID	.625	.858	.723	.593	.285	.385	.554	.618	.661	.645	.653	.524	.543	.533	**.593**	.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	**.598**	.723
ES	.748	.845	.794	.602	.453	.517	**.655**	.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	**.731**	.773								
PT	.732	.917	.814	.600	.272	.375	.594	.713								
ES	.751	.771	.761	.535	.508	.521	.641	.681								

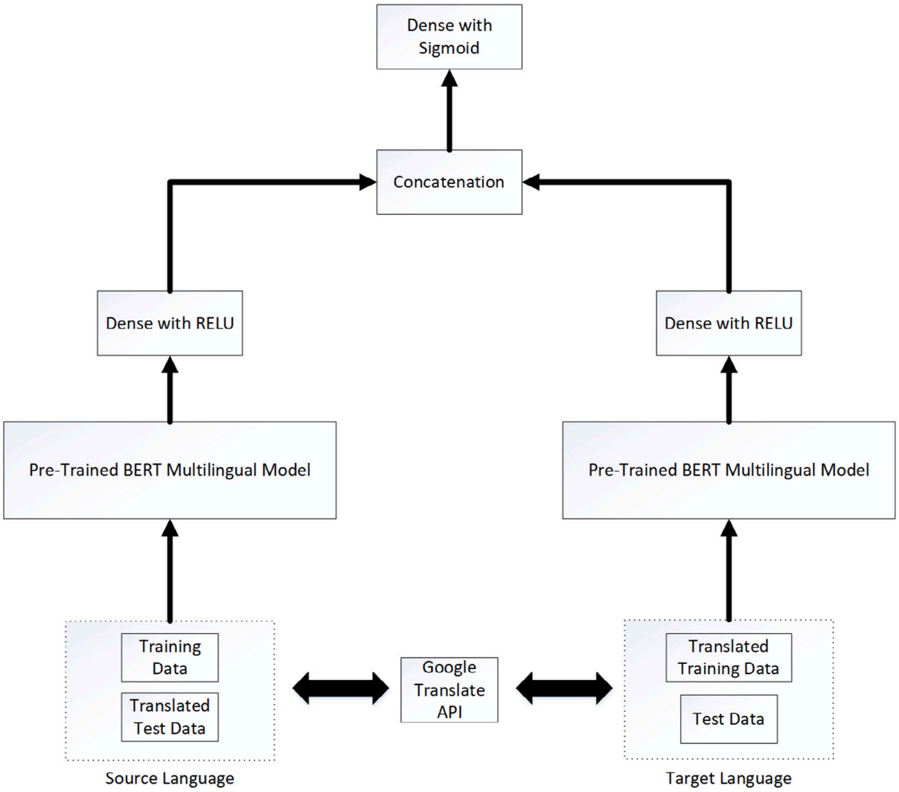

Fig. 3. Joint-Learning BERT Model Architecture.

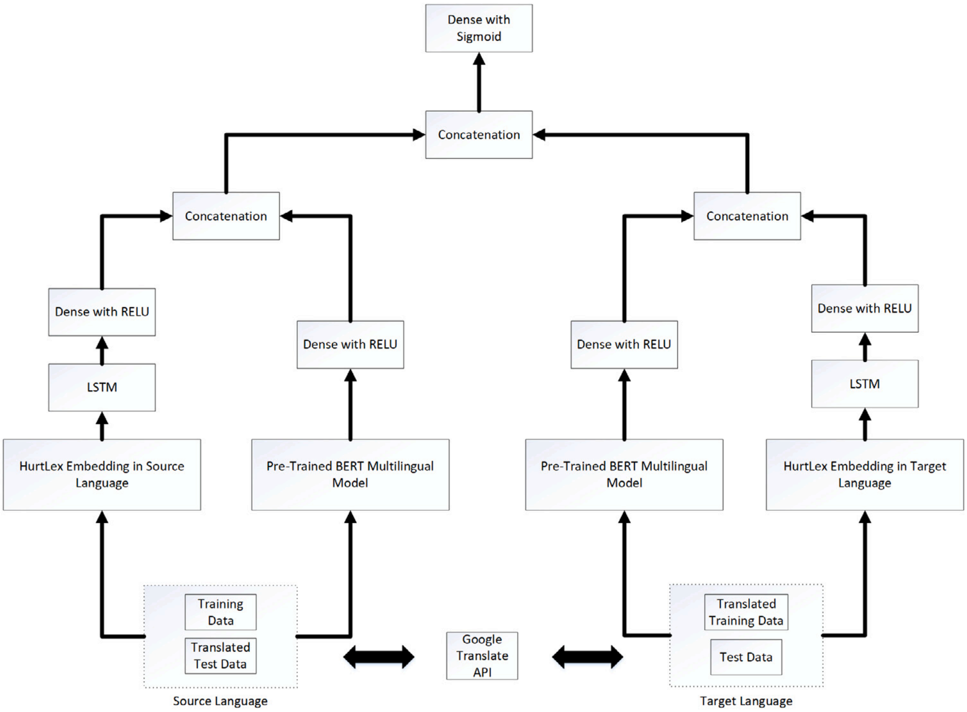

Fig. 4. Joint-Learning BERT-HurtLex Model Architecture.

Table 4
Results of cross-lingual hate speech detection on the balanced training set.

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

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 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 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** 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, McKeever, & Delany, 2017).

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, while Table 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 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 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 (Arango et al., 2020; Wiegand et al., 2019).

Table 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.

Table 5
Results of cross-lingual hate speech detection on individual datasets with the original training set distribution.

	LR + Laser								BERT Multilingual							
	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 6
Results of cross-lingual hate speech detection of each dataset on the balanced training set.

	LR + Laser								BERT Multilingual							
	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	.574	.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								

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

Table 7
Results of cross-lingual hate speech detection with additional external resource on the balanced training set.

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

6.1. External knowledge on hate words

Based on the results in Table 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-dependent 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. 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 :

🐦 Maju lu sini **anjing** URL

🇬🇧 *Come on here **dog** URL*

Example 2 :

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

🇬🇧 *Ahok: choosing a leader based on religion violates the constitution. It 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 northern Italians. The word belongs to the group of HurtLex terms evoking a negative stereotype, where the offense does not target a single individual but rather an entire category (in this case geographically connoted). 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.

More in general, the quality of translation is sometimes an issue. In particular, all instances containing idiomatic terms (see examples above such as “dodol”) or expressions that are otherwise not directly translatable, such as metaphors, can produce mismatches between the data translated from English and the original test sets. This finding is consistent with the study of Glavaš, Karan, and Vulić (2020), where translation implies a loss of abusiveness of several words in the target languages (Albanian, Croatian, German, Russian, and Turkey). In fact, the expressions that offend may vary in a very imaginative way from one language to another, giving rise to expressive formulas that often sound bizarre or incomprehensible when viewed through the lenses of a speaker of another mother tongue. Often the words of a language are not offensive *per se*: it is the negative cultural values associated with a concept that makes a word or phrase insulting. For example, “fox” can be used to convey a positive stereotype in English and to refer to a woman who is very intelligent and attractive, while its translation in another language can be used as a serious insult, as in Spanish, where “zorra” is often used as a synonym for whore. Moreover, there are English specific expressions that do not have any equivalent in the target language (i.e., “piece of shit” and “scumbag”), and also English puns or word plays merging personal names with animal names to mock a person (i.e., the “Hildebeest”, a nickname that is used to mock Hillary Clinton, via fusing “Hillary” and “Wildebeest”) for which machine translation does not return an equivalent in the target language.

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. Indeed, in our analysis, we found a number of such instances that were translated incorrectly.

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 8 summarized the datasets, including their topical focus and collection period. As shown in Tables 5 and 6, our model obtained 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, Basile, Polignano, Basile, and Patti (2020). The datasets in the same language shown in Table 8 were collected at different times, which we believe affects their topical focus and therefore the cross-dataset classification results.

Table 8
Dataset topical focuses and its collection time.

Lang.	Dataset	Topical focus	Collection
EN	Davidson et. al.	Topic generic	–
	Basile et. al.	Hate speech towards immigrants and women	Jul–Sept 2018
	Founta et. al.	Topic generic	Mar–Apr 2017
	Ousidhoum et. al.	Some controversial topics including feminism, immigrants and islamic-leftism	–
FR	Ousidhoum et. al.	Some controversial topics including feminism, immigrants 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 events	Feb–Apr 2017
IT	Bosco et. al.	HS related to immigrants, muslim, and roma	–
PT	Fortuna et. al.	Generic topic	Jan–Mar 2017
ES	Basile et. al.	Hate speech towards immigrants and women	Jul–Sept 2018
	Pereira et. al.	Generic topic	Feb–Dec 2017

7. Conclusion and future works

In this work, we presented the results of 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 approach. 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 evaluate the capability of recent multilingual language models in a cross-lingual classification setting. Furthermore, we experiment 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 conduct 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 is 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.

In summary, we report that our novel method outperformed existing models, including **LR + LASER Embedding**, **BERT Multilingual**, and **BERT + Translation**, which were all tested on recent benchmarks (Aluru et al., 2020). We also conduct a more in-depth analysis of the results to provide insightful findings to shed some light for the future works. Overall, our work is complementing other works that focused on the zero-shot cross-lingual hate speech detection task, which is still very limited.

The research we propose responds to a societal need for change and will be impactful from many perspectives. Several actors have an increasing need for automatic support to moderate or monitor and map the dynamics and the diffusion of HS dynamics over different territories (Capozzi et al., 2019; Paschalides et al., 2020). These include institutions, NGOs, and ICT companies that have to comply with governments' demand for counteracting the HS phenomenon.[19] This is only possible at a large scale by employing robust computational methods, and the method we propose here can have a practical counterpart in the implementation of tools supporting content moderation in social media platforms, where the multilinguality issue is crucial to scaling up. This will hopefully contribute to making social media a safer environment for vulnerable or marginalized groups, preventing the rise of hate speech online with a broader coverage in terms of languages and territories, countering the effects of toxic online discourse on society, and favoring the diffusion of voices speaking up to empower discriminated people in expressing themselves online and offline.

Based on the overall results in all experiments, we observed several issues and difficulties of this task. Among the issues raised by the present work, we observed that dataset bias 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

[19] See for instance, the recently issued EU commission Code of Conduct on countering illegal hate speech online (EU Commission, 2016).

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 conduct a deeper analysis of the impact of automatic translation on joint-learning models. Manual error analysis highlighted that some key terms are not translated accurately, especially derogatory words, usually related to cultural values in different languages. Topic bias is also an issue we found relevant to this work. We plan to investigate how topic bias behaves in hate speech datasets across languages and how to alleviate it with our 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. 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.

CRediT authorship contribution statement

Endang Wahyu Pamungkas: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original, Writing - review & editing. **Valerio Basile:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing - original, Writing - review & editing. **Viviana Patti:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing - original, Writing - review & editing.

Acknowledgments

This work is partially funded by the project “Be Positive!” (under the 2019 “Google.org Impact Challenge on Safety” call).

References

Agarwal, S., & Sureka, A. (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. CoRR, abs/1701.04931, arXiv:1701.04931.

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 233–238). IEEE.

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. CoRR, abs/2004.06465, URL arXiv:2004.06465.

Arango, A., Pérez, J., & Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, Article 101584.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610, URL https://transacl.org/ojs/index.php/tacl/article/view/1742.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In R. Barrett, R. Cummings, E. Agichtein, & E. Gabrilovich (Eds.), *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017* (pp. 759–760). ACM, http://dx.doi.org/10.1145/3041021.3054223.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019* (pp. 54–63). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/s19-2007.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In E. Cabrio, A. Mazzei, & F. Tamburini (Eds.), *CEUR Workshop Proceedings*: *vol. 2253*, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018), Torino, Italy, December 10-12, 2018* (pp. 1–6). CEUR-WS.org, URL http://ceur-ws.org/Vol-2253/paper49.pdf.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *CEUR Workshop Proceedings*: *vol. 2263*, *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* (pp. 1–9). CEUR-WS.org, URL http://ceur-ws.org/Vol-2263/paper010.pdf.

Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, & N. Nikolov (Eds.), *Recent Advances in Natural Language Processing, RANLP 2009, 14-16 September, 2009, Borovets, Bulgaria* (pp. 50–54). RANLP 2009 Organising Committee / ACL, URL https://www.aclweb.org/anthology/R09-1010/.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Capozzi, A. T., Lai, M., Basile, V., Poletto, F., Sanguinetti, M., Bosco, C., et al. (2019). Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *CEUR Workshop Proceedings*: *vol. 2481*, *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-It 2019)* (pp. 1–6). Bari, Italy: CEUR-WS.org.

Chen, H., McKeever, S., & Delany, S. J. (2017). Abusive text detection using neural networks. In J. McAuley, & S. McKeever (Eds.), *CEUR Workshop Proceedings*: *vol. 2086*, *Proceedings of the 25th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, December 7 - 8, 2017* (pp. 258–260). CEUR-WS.org, URL http://ceur-ws.org/Vol-2086/AICS2017_paper_44.pdf.

Chen, Z., Shen, S., Hu, Z., Lu, X., Mei, Q., & Liu, X. (2019). Emoji-powered representation learning for cross-lingual sentiment classification. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, & L. Zia (Eds.), *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019* (pp. 251–262). ACM, http://dx.doi.org/10.1145/3308558.3313600.

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., & Weinberger, K. Q. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, *6*, 557–570, URL https://transacl.org/ojs/index.php/tacl/article/view/1413.

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

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology*, *20*(2), 10:1–10:22. http://dx.doi.org/10.1145/3377323.

Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). 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* (pp. 512–515). AAAI Press, URL https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665.

De Mauro, T. (2016). Le parole per ferire, internazionale. 27 settembre 2016.

Demirtas, E., & Pechenizkiy, M. (2013). Cross-lingual polarity detection with machine translation. In E. Cambria, B. Liu, Y. Zhang, & Y. Xia (Eds.), *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013* (pp. 9:1–9:8). ACM, http://dx.doi.org/10.1145/2502069.2502078.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1423.

Downs, A. (1973). Up and down with ecology, the ‘‘issue attention’’ cycle.

Duh, K., Fujino, A., & Nagata, M. (2011). Is machine translation ripe for cross-lingual sentiment classification?. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers* (pp. 429–433). The Association for Computer Linguistics, URL https://www.aclweb.org/anthology/P11-2075/.

Erjavec, K., & Kovačič, M. P. (2012). ‘‘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. http://dx.doi.org/10.1080/15205436.2011.619679.

EU Commission (2016). Code of conduct on countering illegal hate speech online. URL https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en#theeucodeofconduct.

Fehn Unsvåg, E., & Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 75–85). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W18-5110, URL https://www.aclweb.org/anthology/W18-5110.

Feng, Y., & Wan, X. (2019). Towards a unified end-to-end approach for fully unsupervised cross-lingual sentiment analysis. In M. Bansal, & A. Villavicencio (Eds.), *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019* (pp. 1035–1044). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/K19-1097.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *CEUR Workshop Proceedings*: *vol. 2263*, *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* (p. 59). CEUR-WS.org, URL http://ceur-ws.org/Vol-2263/paper009.pdf.

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Eds.), *CEUR Workshop Proceedings*: *vol. 2150*, *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* (pp. 214–228). CEUR-WS.org, URL http://ceur-ws.org/Vol-2150/overview-AMI.pdf.

Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, *10*(12), 4180.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 94–104). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-3510, URL https://www.aclweb.org/anthology/W19-3510.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., et al. (2018). 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* (pp. 491–500). AAAI Press, URL https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Glavaš, G., Karan, M., & Vulić, I. (2020). XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6350–6365). Barcelona, Spain (Online): International Committee on Computational Linguistics, URL https://www.aclweb.org/anthology/2020.coling-main.559.

Goodfellow, I. J., Bengio, Y., & Courville, A. C. (2016). *Deep Learning, Adaptive computation and machine learning*. MIT Press, URL http://www.deeplearningbook.org/.

Ibrohim, M. O., & Budi, I. (2019a). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 46–57). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-3506, URL https://www.aclweb.org/anthology/W19-3506.

Ibrohim, M. O., & Budi, I. (2019b). Translated vs non-translated method for multilingual hate speech identification in Twitter. *International Journal on Advanced Science, Engineering and Information Technology*, *9*(4), 1116–1123.

Jebbara, S., & Cimiano, P. (2019). Zero-shot cross-lingual opinion target extraction. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 2486–2495). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1257.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). 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* (pp. 6282–6293). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.560, URL https://www.aclweb.org/anthology/2020.acl-main.560.

Kim, S., & Hovy, E. H. (2006). Identifying and analyzing judgment opinions. In R. C. Moore, J. A. Bilmes, J. Chu-Carroll, & M. Sanderson (Eds.), *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA* (pp. 200–207). The Association for Computational Linguistics, URL https://www.aclweb.org/anthology/N06-1026/.

Koufakou, A., Pamungkas, E. W., Basile, V., & Patti, V. (2020). HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 34–43). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.alw-1.5.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In M. desJardins, & M. L. Littman (Eds.), *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA* (pp. 1621–1622). AAAI Press, URL http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6419.

Leader, T., Mullen, B., & Rice, D. (2009). Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: what puts the "hate" into hate speech?. *Journal of Personality and Social Psychology*, *96 1*, 170–182.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., et al. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3125–3135). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1301, URL https://www.aclweb.org/anthology/P19-1301.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandalia, C., et al. (2019). 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, P. Mehta (Eds.), *FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019* (pp. 14–17). ACM, http://dx.doi.org/10.1145/3368567.3368584.

Meng, X., Wei, F., Xu, G., Zhang, L., Liu, X., Zhou, M., et al. (2012). Lost in translations? Building sentiment lexicons using context based machine translation. In M. Kay, C. Boitet (Eds.), *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India* (pp. 829–838). Indian Institute of Technology Bombay, URL https://www.aclweb.org/anthology/C12-2081/.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In J. A. Carroll, A. van den Bosch, & A. Zaenen (Eds.), *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic* (pp. 976–983). The Association for Computational Linguistics, URL https://www.aclweb.org/anthology/P07-1123/.

Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2019). Author profiling for hate speech detection. arXiv preprint arXiv:1902.06734.

Mossie, Z., & Wang, J.-H. (2019). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, Article 102087.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), *Studies in Computational Intelligence*: *vol. 881*, *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* (pp. 928–940). Springer, http://dx.doi.org/10.1007/978-3-030-36687-2_77.

Mullen, B., & Smyth, J. (2004). Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine*, *66*, 343–348.

Müller, K., & Schwarz, C. (2019). Fanning the flames of hate: Social media and hate crime. Available at SSRN 3082972.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. (2019). Multilingual and multi-aspect hate speech analysis. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *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* (pp. 4674–4683). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1474.

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6237–6246). Marseille, France: European Language Resources Association, ISBN: 979-10-95546-34-4, URL https://www.aclweb.org/anthology/2020.lrec-1.765.

Pamungkas, E. W., Basile, V., & Patti, V. (2020b). Misogyny detection in Twitter: a multilingual and cross-domain study. *Information Processing & Management*, *57*(6), Article 102360, URL https://www.sciencedirect.com/science/article/pii/S0306457320308554.

Pamungkas, E. W., & Patti, V. (2019). 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). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-2051.

Paschalides, D., Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., et al. (2020). MANDOLA: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Transactions on Internet Technology*, *20*(2), http://dx.doi.org/10.1145/3371276.

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., & Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, *10*(17), 5993.

Pereira-Kohatsu, J. C., Sánchez, L. Q., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors*, *19*(21), 4654. http://dx.doi.org/10.3390/s19214654.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. In M. A. Walker, H. Ji, A. Stent (Eds.), *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)* (pp. 2227–2237). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n18-1202.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, URL https://link.springer.com/article/10.1007/s10579-020-09502-8.

Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2018). 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)* (pp. 118–123). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-2019, URL https://www.aclweb.org/anthology/N18-2019.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. CoRR, abs/2004.09813, URL arXiv:2004.09813.

Rodriguez, D., & Saynova, D. (2020). *Machine Learning for Detecting Hate Speech in Low Resource Languages* (Master's thesis), Göteborgs Universitet.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. CoRR, abs/1701.08118, URL arXiv:1701.08118.

Sarkar, A., Reddy, S., & Iyengar, R. S. (2019). Zero-shot multilingual sentiment analysis using hierarchical attentive network and BERT. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval* (pp. 49–56).

Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 3795–3805). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1380.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. http://dx.doi.org/10.1002/ab.21737, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ab.21737, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21737.

Stappen, L., Brunn, F., & Schuller, B. W. (2020). Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. CoRR, abs/2004.13850, arXiv:2004.13850.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 940–950). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/K19-1088, URL https://www.aclweb.org/anthology/K19-1088.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In A. Armando, R. Baldoni, & R. Focardi (Eds.), *CEUR Workshop Proceedings*: *vol. 1816*, *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017* (pp. 86–95). CEUR-WS.org, URL http://ceur-ws.org/Vol-1816/paper-09.pdf.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media* (pp. 19–26). Montréal, Canada: Association for Computational Linguistics, URL https://www.aclweb.org/anthology/W12-2103.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016* (pp. 88–93). The Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n16-2013.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 602–608). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1060.

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). 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.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada* (pp. 347–354). The Association for Computational Linguistics, URL https://www.aclweb.org/anthology/H05-1044/.

Zhou, X., Wan, X., & Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers* (pp. 1403–1412). The Association for Computer Linguistics, http://dx.doi.org/10.18653/v1/p16-1133.