The University Of Sheffield.

You said what now?!
Hate Speech Detection in Social Media

By:

Zeerak Mustafa Waseem Butt

Interim Panel Report

The University of Sheffield
Faculty of Engineering
Department of Computer Science

June 2018

Table of contents

List of tables **iv**

1 Introduction **1**
- 1.1 Research Questions . . . 3
  - 1.1.1 Abusive Language Detection Using Machine Learning . . . 3
  - 1.1.2 Minimising Human Annotation for Hate Speech Detection . . . 4
  - 1.1.3 Hidden Variables: Critical Analysis of Abusive Language Data Sets 5
  - 1.1.4 Generation of Counter Speech . . . 6
- 1.2 Report Structure . . . 8

2 Background: Social Science **9**
- 2.1 Privilege . . . 9
- 2.2 Oppression . . . 10
- 2.3 Intersectionality . . . 11
- 2.4 Pseudonymity and Anonymity in Social Media . . . 12
  - 2.4.1 Moderation Practises . . . 14
  - 2.4.2 Ethical Considerations . . . 15
- 2.5 Summary . . . 16

3 Literature Review: Natural Language Processing **17**
- 3.1 Natural Language Processing . . . 17
  - 3.1.1 Abusive Language Detection . . . 17
- 3.2 Counter Speech . . . 20
  - 3.2.1 Natural Language Generation . . . 20
- 3.3 Summary . . . 23

4 Work To Date **24**
- 4.1 Typology of Abuse - Published . . . 24
- 4.2 Multitask Learning for Abusive Language Detection - In Press . . . 25

4.3 Summary 27

5 Work Plan and Timetable 28

5.1 Experiments planned 28

5.1.1 Overview 29

5.1.2 RQ2: Distantly Supervised Procedural Algorithm for Abusive Language Detection - Ongoing 30

5.1.3 RQ1: Embedding Spaces for Abusive Language Detection - Ongoing 32

5.1.4 RQ1: Multi-Task Learning for Abusive Language Detection 32

5.1.5 RQ1: Classifying Abusive Documents using Linguistic Information and Meta data 33

5.1.6 RQ3: Critical Analysis of Demographic Profiling in Abusive Language Data Sets 33

5.1.7 RQ4: Generation of Counter Speech for Hate Speech - Early planning 34

5.1.8 RQ3: Computational Analysis of the Influence of Community Guidelines on Politeness on Reddit 35

5.2 Timetable 35

5.3 Contingency Plans 36

5.4 Summary 37

6 (DDP) Progress 38

References 41

Appendix A Published Work 47

Appendix B Applications for Ethical Approval 85

List of tables

1.1 Examples of abusive comments. 2

4.1 Typology of abusive language (Waseem et al., 2017). 25

4.2 Comparison of test-set performance of within-domain and out-of-domain datasets using models trained only on one dataset (first four rows), models trained by concatenating both datasets (middle two rows), and using both datasets in a multi-task learning environment (final four rows). For each training regime, we compare using Bag-of-Words (BoW), the Average of Subword Embeddings (Emb) and both (B+E) as features for each tweet. Key: (R)acism, (S)exism, (H)ate-Speech, (O)ffensive, (N)either. Datasets: Davidson Davidson et al. (2017), WaseemWaseem (2016)/Waseem-HovyWaseem and Hovy (2016) (W/W+H) 26

5.1 Task Timetable 36

Chapter 1

Introduction

As people consume, interact with, and create media online at an ever growing rate, it becomes necessary to consider the content with which they interact and the cultures that are formed in online spaces. As social media platforms have been subject to a growing amount of governmental attention for the abusive content on their platforms (European Commission, 2016; Home Office, 2016), they must find scalable methods for dealing with online abuse which can deal with the large volume of new documents added daily (Bates, 2013; Kuchera, 2014; Sood et al., 2012a). By considering how these cultures are formed, which methods for regulation and self-regulation are effective, and the linguistisc markers we may gain insight into the disticnt forms of culture and speech which is deemed acceptable by a given platform. With this knowledge it is possible to develop methods to encourage healthy conversations and mitigate negative sociatal constructs (Bolukbasi et al., 2016; Hannon, 2016).

Recently, there has been an increase in the amount of research on abusive language and hate speech in the Natural Language Processing (NLP) community. In spite of this uptick, many questions are still remain unanswered and the methods for detection are not generalisable beyond the individual data sets. In addition, while this increase of interest in disparate and abusive treatment of people is quite new in NLP, it has been the subject of long and sustained interest in many other fields such as gender studies (Butler, 1990; de Beauvoir, 1953; McIntosh, 1988), law (Crenshaw, 1989; Han, 2015), media studies (Shah, 2015a,b), sociology (Zaleski et al., 2016), and critical race theory (Crenshaw, 1989; Lawrence III, 1992) to name but a few. In spite of a rich body of work on abuse and discrimination, work in NLP often neglects this history relying solely the community guidelines or terms and conditions documents put forth by social media companies (Nobata et al., 2016; Ross et al., 2016). This lack of consideration of work previous work has the consequence of creating blind spots in annotation guidelines and inconsistent judgements on acceptable language.

Input	System	Output
Example 1	Detection Placeholder	Output placeholder
Example 2	Human Out Of Loop Placeholder	Output placeholder
Example 3	NLG Placeholder	Output placeholder

Table 1.1 Examples of abusive comments.

As abusive language may be expressed along several axis, we focus on abusive language that is directed and either explicit in its potential to be abusive or seeks to use implicit signals to notify abuse Waseem et al. (2017). We limit ourselves to these forms of abuse (please see Table 1.1 for examples) to provide a reasonable scope of the project, and because currently the majority of data sets released deal with directed abuse. Further, due as the relationship between and identities of speaker and listener are important (e.g. black person using the n-word colloqiually with another black person is not necessarily problematic, whereas a non-black person using the n-word directed at a black person is most likely problematic, as they speak to different histories), we will consider directed abuse as it has a clearly identifiable speaker and listener.

In this Ph.D. we will seek to marry the findings and knowledge obtained in fields outside of NLP on the nature of abuse. As machine learning methods rely on the data sets and labels upon which they are trained (Packer et al., 2018; Waseem, 2016), we will seek to influence during annotation of data sets. Further, we seek to encode a specific bias, namely an intersectional feminist bias into the data sets as an explicit bias increases the performance of models trained on the biased data set (Waseem, 2016). Specifically, we will seek to ensure that labels adhere to the guidelines set forth by Waseem and Hovy (2016). During this project, we will also aim to develop methods to limit annotator exposure to abusive documents.

Beyond ensuring an intersectional feminist bias encoded in the data, we will seek to improve the generalisability of machine learning models for hate speech detection.

Furthermore, we will critically assess the data sets that have previously been published for disparate impacts on populations and how this issue may arise as early as in the choice of platform to collect data from.

Finally, we will seek to find methods for automatically generating counter-speech to abusive documents as a way of dealing with abuse online without providing a limit to the freedom of speech.

1.1 Research Questions

1.1.1 Abusive Language Detection Using Machine Learning

Much of the previous work on hate speech and abusive language detection has had a focus of determining the parameters of the task at hand, and while a great deal of decisions and consensus are yet to be reached in the field several issues have already come to light such as the unequal demographic impact stemming from some communities' speech being more prevalent in datasets than other communities, such as the African American community in Davidson et al. (2017). Another issue which has come to light occurs due to the small sizes of data sets available, which results in a risk that models may have extremely poor generalisation often overfitting to innocuous terms such as 'islam" that appear in positive classes (Waseem and Hovy, 2016). Additionally, the data sets are often collected for a specific cultural context meaning that culture-specific understandings of acceptable speech highly influences the data sets, which in turn further decreases the chance of generalisability to other data sets. Another challenge facing the field of abusive language detection is that the vast majority of abusive language data sets display large class imbalances with the majority class being documents that are not abusive, e.g. in Waseem and Hovy (2016), where 31% of the data is labeled as abusive spread over two classes (20% labelled as sexist and 11% labelled as racist). As such classifiers are given the task of identifying abusive language while being given access to very little data and highly variable data in its notions of acceptability and demographic targetting. For a task such as abusive language and hate speech detection, where subjects such as racial and gender minorities are often discussed in both constructive and abusive ways, this means that classifiers are prone to overfit to the majority class thus resulting in poor usability of the models and indeed, poor generalisability beyond the specific data set.

Most data sets that have been for abusive language detection do not allow for generalisation beyond the given data set due to being very small e.g. in Waseem (2016) there are less than 100 documents labelled in the minority class; the data is collected for a limited number of targets (Waseem and Hovy, 2016); or some communities being overrepresented in the data sets (Davidson et al., 2017). In addition, many annotated data sets for abusive language are collected on Twitter (Davidson et al., 2017; Jha and Mamidi, 2017; Waseem, 2016; Waseem and Hovy, 2016) which naturally influences a given model's performance on data obtained from other sources which may have not have length limitations on documents (e.g. comments or posts). Poor generalisability also influences moving from one data set to another Waseem (2016) which greatly reduces the usability of the models that are trained on a given data set. To overcome this issue of poor generalisability stemming from a limited

number of targets, we will be collecting additional data which will be used to explore the utility of neural network models along with linear models. Further, we will seek to abstract away from the words in the documents to avoid overfitting to specific terms. We will to utilise unannotated data from communities that are known to be abusive in order to address poor generalisability issues stemming from a limited number of target groups in the annotated data as well as the issue of documents that exceed Twitter's document length restrictions. Thus, we will seek to address the following research question:

> **RQ1: How can embeddings, weak supervision, linguistic annotation, and meta data help improve platform and topic independent detection of abusive language and hate speech?**

1.1.2 Minimising Human Annotation for Hate Speech Detection

The issue of small data sets that deal with abusive language, and in particular hate speech in part stems from a non-computational issue: Annotation can be psychologically harmful due to the extended and repeated exposure to abuse. For instance, two commercial content moderators for Microsoft are currently suing Microsoft as they allege that their job has resulted in Post-Traumatic Stress Disorder (Levin, 2017). Attempting to limit annotator exposure to a new form of or target of abusive language thus provides issues that have the potential to make work in the field of abusive language detection borderline on ethical appropriateness. Beyond the potential ethical difficulties, another issue with human annotation comes to light.

To alleviate the issues outlined below, we propose building a system that can provide estimates of whether a document contains abuse or not through the use of a collection of NLP methods (such as sentiment analysis, topic modelling, and stance detection amongst others). Thereby, we can control the experience of annotators to a greater degree, which may then allow for less intensive debriefings and reduce the need for mental health profesionals partaking in the debriefing process.

Limited annotation resources As annotation resources may be scarce thus imposing limits on the number of annotators that are recruited or ensuring proper debriefing potentially through a mental health professional.

Limited control on exposure As the documents are not likely to be annotated or the annotations being hidden to prevent bias, it may be hard to employ safeguards ensuring that annotators are not only exposed to abusive documents and that there is a balance in exposure between abusive and non-abusive documents.

Biases Though Bolukbasi et al. (2016); Packer et al. (2018); Zhao et al. (2017) show that computational models contain (unwanted) biases just as humans do. However, unlike humans biased models do not sek to hide these biases, and they can more easily be identified and resolved by addressing the underlying data that gives rise to these biases. Identifying human biases and correcting for them however poses significantly more challenges, as requesting that a human address their biases may produce conflict or the annotator may be unable to address these biases, and in the worst case scenario the annotator may simply stop annotating further.

RQ2: How can NLP techniques be used as proxies for hate speech detection and thus minimise need for human annotation?

1.1.3 Hidden Variables: Critical Analysis of Abusive Language Data Sets

As more data sets for abusive language detection are released it is important to critically assess the hidden variables that exist in the data sets which impact classifiers: i.e. which type of abusive language they can aid in identifying (Jha and Mamidi, 2017; Waseem et al., 2017) as well as which cultural phenomena they speak to.

For instance Waseem et al. (ress) find that the data set published by Davidson et al. (2017) has a great deal of African-American Vernacular English (AAVE) labelled as either hate speech or offensive language. By training a classifier on this data set without ensuring that there are no safeguards to prevent demographic discrimination, the trained model is likely have a preference for detecting that people who use AAVE are being either offensive or abusive, in large parts disregarding whether the contents are truly offensive or abusive. Another issue that comes to light with the fact that AAVE is more likely to be tagged as offensive or hate speech is that the computational models built for removal, moderation, or even generation of counter speech, will explicitly make a judgement about the appropriateness of AAVE as a dialect. An implied role in abusive language detection is the responsibiity of determining which forms of speech are socially acceptable; for this reason it is vital that seek to investigate which demographics, if any, are being singled out by the data sets we employ for training our models.

To the best of our knowledge, only Waseem et al. (ress) begin to touch upon critically assessing abusive language data sets. Therefore, in this project we will seek to identify the types of abusive language that published data sets allow models to detect as well as whihc demographies they might impact negatively.

Finally, when collecting data the affordances of the platform influence the data e.g. Twitter enforces a character limit, Facebook permits free form text without length boundaries within a closed social network, and Reddit allows for free form text in an open social network. This all influences what the data can be obtained. However, the consideration of these influences on abusive language data has yet to be conducted. Specifically, there has been little consideration to the impact of community guidelines and comunity moderation on the production of abusive language. Therefore it is prudent to investigate how guideliens and moderation impacts a platform. In related work, Crawford and Gillespie (2016) consider how the act of flagging content on platforms may influence cultures, however the impact of guidelines on the amount and type of abuse that is generated on platforms has yet to be studied.

RQ3: Do correlations exist between the individual classes in abusive language data sets and variables such as the race or gender of the speaker, and do they allow us to build high classifiers to detect these in which the predictions have significant correlations with the classes annotated for abuse?

1.1.4 Generation of Counter Speech

On the internet there are several websites specifically created for this purpose such as StormFront[1], an online forum for white nationalsits and Gab[2], a Twitter-like social media created as a response to Twitter removing user accounts for violating Twitter's terms of service. Beyond communities created for the specific purpose of unencumbered free speech or as a meeting place for white nationalists, hate speech and abusive language exists widely on general purpose social media such as Twitter and Facebook (Gorrell et al., 2018; Munger, 2017).

In much of the NLP research conducted on abusive language two goals are stated for the task: Removal of content

ZW: [CITE removal papers]

and filtering for moderators

ZW: [CITE Moderation papers]

. Both of these approaches have noble goals however neither deal with the underlying issue of some users wishing to be abusive or simply not realising that their behaviour is

[1] www.stormfront.org

[2] www.gab.ai

abusive. For instance, the Reddit clone Voat[3] and Twitter clone Gab[4] both of which brand themselves as "free speech" alternatives created in response to Reddit and Twitter enforcing their community guidelines and shutting down subforums and user accounts, respectively. It is therefore prudent to consider methods in which the user may be encouraged to change the non-compliant behaviour to a behaviour that is compliant with the community guidelines and terms and conditions of a given platform without removal of any accounts or shutting down communities. Further, by considering such methods that do not seek to remove the right or ability to speech (e.g. serving more ads to abusive users than to non-abusive users; using bots to generate speech against the abuse; or otherwise instituting a negative cost to being abusive) but rather seeks to model desirable behaviour to deal with abusive language, it is possible to avoid ethical pitfalls and allows us to navigate the tension between the right to protection from persecution and the right to free speech (U.N., 1948). In spite of the benefits of easing ethical consderations around the complex concepts like freedom of speech, to the best of our knowledge no previous work in NLP has attempted to generate counter speech to deal with hate speech or abusive language online.

We propose that generation of counter speech can serve as a productive method for minimising abusive behaviour online. We make this claim on the basis of the findings in Munger (2017), in which the authors find that sending a static message to harassers from an account that shares racial demographic can have limiting effect on the production of abusive content. Therefore, we propose generation of counter speech as an alternate goal for detection of abusive language. We propose that the goal of detection should be different, as the generated texts should be directed to the offending speaker which may only be possible in a public forum and for this reason we must be careful not to address speech which is not abusive or contains hate speech. The need for high precision detection becomes more urgent as our aim is building an automated pipeline without a human in the loop. As we focus on high precision detection models, this further means that there will be some abuse, often more subtle and implicit that will not be identified by our models, thus we give less importance to obtaining high recall scores.

Thus our generation of counter speech becomes the following research question:

RQ4: How can NLG be used as a tool for intervention against online abuse and what is the impact of emulating the stylistic language of the author of the abuse?

[3] www.voat.co
[4] www.gab.ai

1.2 Report Structure

In this chapter, we have introduced the tasks we'll undertake in the thesis along with the research questions and brief outlines of how we will address each research question. For the sake of the confirmation panel, chapters 3, 4, 5, and 6 deal with the requirements of the report, while 2 deals with more background knoweldge that further motivates and provides deeper understanding of the social scientific concepts and how they will be used.[5]

Chapter 2 presents the knowledge from Critical Race Theory, Gender Studies, ethics for social media research, and Psychology upon which we base our methodology and how we will use these concepts.

Chapter 3 presents the literature review on current approaches to abusive language detection and counter speech generation in NLP, and seeks to explain which methods we will use and how they will be used.

Chapter 4 presents the work that has been conducted thus far.

Chapter 5 outlines the tasks that will be carried out during this PhD project, as well as a tentative execution timetable.

Chapter 6 outlines the activities completed regarding the Doctoral Development Program (DDP), as well as other activities that have been carried out.

Appendix A provides the full papers that have been published thus far.

Appendix B provides the successful and pending applications for ethical approval. Successful applications also include the notification of acceptance.

[5] In Chapter 2 we also present our ethical considerations.

Chapter 2

Background: Social Science

In this chapter, we introduce core concepts from social science that we employ in this thesis. This will serve as foundational knowledge for what we interpret as abusive language and hate speech. Subsequently, a consideration of multiple of these may be necessary to determine if speech is derogatory and will inform our recommended solution for handling such speech. Additionally, we seek to introduce how we will use these concepts in the thesis.

2.1 Privilege

The concept of privilege was first introduced by Dubois (1935) as "psychological wage" in his work "Black Reconstruction in America". In this work he describes a deliberate effort to segregate poor whites and blacks to prevent potential alliances between the two. He argued that efforts were made to distinguish working class whites along racial lines by aligning them with their employers. The consequence of this being that working class whites came to feel a supioriority over blacks.

Since this early exploration of psychological wage, or its' more current term: "privilege" a great deal of work has been done exploring the psychological and social differences along racial hierarchies. Further, the concept of privilege has been applied to the concepts of gender (Butler, 1990; de Beauvoir, 1953), race (Crenshaw, 1989; Dubois, 1935), sexuality

ZW: [CITE Focoult]

, religion

ZW: [FInd citations]

, and other aspects of identity. One notable work, that we will rely heavily on in this thesis is McIntosh (1988), in which they explore the differences in privileges afforded along gender and racial lines in her own life. Specifically, she explores the privileges that are

afforded to men, but not to her due to her gender; and privileges that are afforded to her due to her whiteness but not afforded to black women.

In more concrete and operationalisable terms, the concept of privilege seeks to describe that some demographics receive beneficial treatment due to their identities or *intersection* of identities, i.e. gender, race, religion, sexual orientation, and mental health. Many social issues have been explained through the concept of privilege such as gender representations (Butler, 1990), marriage equality for homosexual couples

ZW: [CITE find citation]

, and police treaatment of black people in the United States of America

ZW: [CITE Jurafsky paper]

In recent years, there has been a greater focus on social and economic disparities, including not being financially punished for ones gender expression

ZW: [CITE Find paper from drawer in desk @Sheffield]

; and the increased risk of lethal encounters with law enforcement depending on racial identity Zack (2015).

2.2 Oppression

When speaking of privilege it is necessary to also consider its counter-part: Oppression. Oppression, in the frame of privilege, refers to the (continued) marginalisation of demographics as compared to the privileged groups Abberley (1987). Oppression can thus range from the systematic underpayment of women and trans people

ZW: [CITE Paper in desk + Vox/Propublica reports on workplace gender discrimination.]

, to the underrepresentation of people of colour in popular media

ZW: [CITE Re-find citation for representation in media.]

, to the ability to move safely in public spaces

ZW: [Find citation on trans violence + rape statistics]

.

It is crucial to note that not all marginalisation is oppression Abberley (1987). For instance, should a white man in the United States of America feel unsafe around men of colour, this does not amount to oppression, as while there may be singular instances of feeling unsafe, these instances do not translate to a wider collective experience of their demographic.

As an illustratory example, a white man may experience police brutality at the hands of black police officer, however police brutality and unjust treatment of white men is not a in institutional problem that afflicts most, or even many white men. In comparison, should a black man experience police brutality at the hands of a white police officer, this speaks into a wider history and culture of black men being abused by the police. As such the instutional power that the white and black police officer wield, respectively, speak to two different societal positions and histories of oppression.

However, these imbalances in power do not only occur in physical settings, they also occur in spoken and written word. Consider for instance the case of coded language such as the term "thug". While people of all creeds may use the term, the connotation that occurs with many people is specifically an African American man, thanks to popular media. Thus, while "thug" may be used to refer to a person of any race or gender, the use of the term by politicians invokes a specific imagery. Another such example is with regard to terrorism. In many cases, terrorist acts perpetrated by white people are excused by insanity, frustration, and many other reason

ZW: [CITE Find academic citation for this, otherwise find Vox article speaking to this.]

. However, considering narrative surrounding similar acts perpetrated by brown men, a very different picture arises; one that speaks to terrorist acts and often linked to religious fanaticism, regardless of the piety of the suspect in question.

As we are alluding to, power and privlege are closely related, such that a group which holds institutional power cannot be oppressed (on the axis of the intersection where the group is in power) Eisenstein (1977).

2.3 Intersectionality

In this PhD, we will apply an intersectional approach to analyse biases. Intersectionality is a theoretical framework founded by Crenshaw (1989), the framework seeks to describe how multiple identities can intersect with one another to create unique forms of oppression, e.g. the oppression faced by black homosexual people is different than those faced by straight black people and by homosexual white people. Crenshaw (1989) argues that one cannot separate one identity from another, and that "the intersectional experience is greater than the sum of racism and sexism".

A compelling and clear case of different identities intersecting and influencing is the wage gap (in the United States of America), where there are significant differences between income between black and white women as well as women compared to men Neal (2004).

The justification and need for an intersectional framework for analysing and dealing with oppression is further made clear in the case of Moore v. Hughes Helicopter (Crenshaw, 1989). In this lawsuit Moore, a black woman was not allowed to represent a class of other black women that had been bypassed for promotions due to the fact that neither black men nor white women were discriminated against. Crenshaw (1989) argues hat the most likely interpretation of this decision to deny the class is that not all women were discriminated against and not all black people were discriminated against, only black women and as such Moore could not represent a class of women. Thus, the court ruled that black women, in spite of facing discrimination based on the combination of race and gender, could not constitute their own class with regard to discrimination that only affected them.

Beyond the case of black women, the intersectional framework presented by Crenshaw (1989) has proven useful in applications to sexual orientation, sex work, HIV infected persons (Logie et al., 2011), and other oppressed classes (Erevelles and Minear, 2010). Furthermore, Verloo (2006) analysed the application of intersectionality, or the lack thereof in European Union policies addressing discrimination. In a Natural Language Processing context, Waseem and Hovy (2016) and Waseem (2016) apply research from gender studies and critical race theory to devise their annotation guidelines and their analysis of hate speech.

Intersectionality allows for insight into things that cannot be the object of legislation for instance, seeing members of ones own race in media, speaking time at academic events, and not having to educate their children on systemic racism for the children's physical safety (**?**). The framework also allows for insight into discrimination that can be legislated against, such as the Moore V. Hughes Helicopter case (Crenshaw, 1989).

Banks (2010) suggest that the lack of accountability, the immediacy, and global nature of the internet has allowed for it to become "an ideal tool for extremists and hatemongers to promote hate". In addition, it was the hopes of Facebook that increased accountability would lead to a decrease in hate speech (Levine, 2013).

2.4 Pseudonymity and Anonymity in Social Media

As this Ph.D. will deal with data published by individuals who may not be public figures, it is necessary to consider the topic of anonymity and pseudonymity on social media platforms.

Previous research on anonymity has considered anonymity as a spectrum (Don, 1999; Qia, 2007), going from the completely anonymous to fully named. On the other hand, van der Nagel and Frith (2015) considers anonymity a more complex space, inhabited by pseudonyms, mononyms, stage names, and usernames amongst others. Proponents of named social media sites argue that anonymity encourages anti-social behaviour (Galperin, 2011).

On the other hand van der Nagel and Frith (2015), find this argument to misunderstand privacy and identity both online and offline. van der Nagel and Frith (2015) argue that identity is enacted with territorial and contextual boundaries in the offline world, i.e. experiences shared with ones family may differ from those shared with friends, boundaries that are removed in the online world as users interact with their entire social circle, in what Marwick and danah boyd (2011) refer to as context-collapse. van der Nagel and Frith (2015) suggest that the use of pseudo- and anonymity online can function in a similar fashion to the territorial and contextual boundaries found in the offline world.

To examine how anonymity and pseudonymity is constructed and how it influences, van der Nagel and Frith (2015) and van der Nagel (2013) consider identity on Reddit, a social media site with forums in which users, who are hidden usernames (which may or may not bear reference to their real name) can post and vote on content. Specifically, the forum van der Nagel (2013); van der Nagel and Frith (2015) consider is "Gonewild", a forum which describes itself as "an amateur exhibitionist community" (van der Nagel and Frith, 2015). On Gonewild, users post nude and semi-nude pictures of themselves, however their faces and names are very rarely associated with these posts. In fact, the forum offers guidelines for ensuring safety when posting images:

> The internet is a public place. You are posting naked pictures of yourself on the internet[...] If you want to be as anonymous as possible, take the following precautions. 1. Make a throwaway reddit account.
>
> 2. Don't include your face in your photos. If you must, blur or blackout your features.
>
> 3. Take pictures against difficult-to-identify backgrounds. Plain walls or colours work well. (van der Nagel (2013))

van der Nagel (2013) consider the technological and cultural codes that afford anonymity on Gonewild and the risks of posting sensitive information online in an effort to highlight the importance of nuanced understandings of anonymity online. van der Nagel (2013) argue that anonymity plays a crucial role in online communication, as a way for users to segment and limit their audience according to what they aim to share, i.e. many users on Gonewild create throwaway accounts to limit the risk of other reddit users undermining their credibility because they posted on Gonewild:

> Such attention, when unwanted, can prompt users to disguise themselves on reddit by creating a "throwaway" account. (van der Nagel (2013))

By creating such throwaway, pseudonymous accounts they are afforded the ability to express their sexual selves in while retaining a disconnect with the rest of their online and offline lives without compromising their ability to interact with their audience or face negative repercussions (van der Nagel and Frith, 2015).

2.4.1 Moderation Practises

To understand how moderation can work, it is necessary to consider the communicative power of reporting frameworks. In Crawford and Gillespie (2016), the analysis of frameworks for reporting is motivated by the fact that reporting content is a method for users to communicate with the platforms. As flags become communicative devices for the users of a site to convey the desired values of the community (Crawford and Gillespie, 2016). At the same time, flagging allows for a company to specifically control the expressiveness of the communication between users and the company and the volume of this communication (Crawford and Gillespie, 2016). Volume can be controlled by the ease of access to flagging documents and expressiveness can be controlled by the level of detail users can provide when reporting. Crawford and Gillespie (2016) argue that a low expressiveness can lead to posts that are reported for a distinct number of reasons can be conflated with one another, if only a few options are available. In addition, a lack of options when flagging can communicate that the company is not interested in distinguishing what users find unacceptable, they simply wish to know that some content is unacceptable (Crawford and Gillespie, 2016).

Oftentimes, flagging is a solitary effort, where single users will communicate that they find content unacceptable, such as when an image of two male actors kissing on the TV show EastEnders was uploaded to Facebook. This image received a great number of reports stating it as sexually graphic and was subsequently removed. Once removed, it created a great deal of outrage as Facebook were accused of homophobia and hypocrisy as gay kisses were being flagged and removed while straight kisses were not. Following this controversy, Facebook undid their removal of the image and apologised for removing it (Crawford and Gillespie, 2016). Considering flags as a communicative device allows for flagging as a strategic means of communication for users to further an agenda. For instance, a group of bloggers angered by the pro-Muslim content on YouTube started “Operation Smackdown” which was launched in 2007 and was active up until 2011. In Operation Smackdown, the bloggers coordinated their supporters to flag certain YouTube videos under the category of “promotes terrorism”. In this coordinated effort, the bloggers created instructions for flagging and created playlists of videos they wanted to targeted that day. They also created a Twitter feed announcing the videos to be flagged. The bloggers would celebrate removed videos and attack YouTube and Google for letting others remain (Crawford and Gillespie, 2016). These

flags can be considered as coordinated attacks which can be perpetrated by a group against both minorities and majorities alike.

2.4.2 Ethical Considerations

Due to the nature of the topics explored in this thesis and the data sources, it is necessary to reflect upon the ethical implication of our work. In this consideration, we focus on participant agency, safety, and anonymity. In this section, we argue the need to balance these in order to ensure the safety of the participants without compromising the research.

Safety

Given that the topic of thesis may touch upon some very sensitive issues, such as gender and racial biases, it is necessary to ensure the safety of participants and researchers involved. Given that much of the controversial data will be personal data, this safety is largely ensured by following the Data Protection Act (DPA) and the General Data Protection Regulation (GDPR) through their focus on ensuring that un-anonymised participant data is only available to the researchers working on the project. Indeed, the DPA suggests that data is anonymised at the earliest possible step, and that it is stored on encrypted drives such that participant, and in extension researcher safety is ensured.

In this project, there will be a need for retaining un-anonymised data sets in regards to the author profiling tasks. To ensure participant safety, we will ensure that any data sets that have not been released publicly are stored in encrypted folders, and encrypted devices where possible. In addition, access to the data will be restricted to the active researchers in the project and potentially the respective supervisors.

Participant Agency and Informed Consent

The social media aspects of this project calls for unobtrusive observation on social media, it is not possible to conduct the research while obtaining informed consent from the users. Rather, we approximate consent by publication. Given the public nature of Twitter, which will be our primary if not only source of social media data, we argue that users are aware of their tweets being published to the world, in addition, should users frequently use “hashtags”, a method for categorising and widening the audience of tweets, it implies that not only are they aware of the public nature of Twitter, they are explicitly seeking to widen the audience of their tweets to a global scope. In our project, we will collect tweets using hashtags and the tweets users of those hashtags publish. In order to ensure that the users are aware and

explicitly utilise the public nature of Twitter, we will not use tweets from users that use hashtags less than at a given ratio[1]

To ensure user agency, we will remove tweets that are deleted and we will only collect tweets from public accounts.

Anonymity

The method that allows for most freedom and safety for users is to fully anonymise data collected on social media. However, due to the phrasing of anonymity in the DPA, which states that a user may not be recoverable obtaining full anonymity on social media data sets is impossible should one seek to work on a user level. In some experiments, namely author profiling, it will not be possible to abstract away from the user level initially. For this reason, at an early stage in our research, it will not be possible to anonymise the data sets fully.

It is our intention to fully anonymise our data sets at the earliest possible point, however, this point will not occur until author profiling tools have been built, such that they can be used in subsequent research. In all subsequent research on social media data, a partial anonymisation will be performed prior to obtaining demographic information by using the author profiling tools. Once the demographic information of a user has been obtained a random hashstring will be generated to identify the user and all user information beyond the demographic knowledge will be dropped. For the sake of reproducibility, a key mapping between user ID's and hashstrings will be stored on a different passport protected encrypted device thus ensuring that the requirements of the DPA are fulfilled.

2.5 Summary

In this chapter, we have introduced several key notions and concepts that will lay a theoretical and philosophical foundation of our work. Specifically, we introduce the concepts of oppression, privilege, and intersectionality. These concepts will be implicitly built into our work, in the cases where it is not made explicit. In addition, we introduce considerations on anonymity and pseudonymity which will be utilised in some projects undertaken in this PhD and how they will be utilised.

[1] We need to determine this ratio based on real use of Twitter.

Chapter 3

Literature Review: Natural Language Processing

In this chapter, we will introduce related work in Natural Language Processing. Additionally, we will introduce the methods which we will employ in this theiss.

ZW: This literature review to be updated with more recent works in NLP, gender studies, policy research, legal research, and abusive language

3.1 Natural Language Processing

In the field of natural language processing (NLP), there has been a recent increase in work focused on abusive language and bias detection. Much of the research is still early-stage work, leaving much room for further inquiry, in particular on considering how biases are employed in written language. In this section, we will provide a detailed overview of recent work and trends in the topics.

3.1.1 Abusive Language Detection

ZW: This needs to be extended with more recent work.

Abusive language detection is a growing field of inquiry. Much off the early work focused on cyberbullying (Chen et al., 2012; Daegon Cho, 2013; Reynolds et al., 2011) and profanity (Sood et al., 2012a,b) with little focus on demographically specified abuse, such as racism, sexism, and anti-semitism (Warner and Hirschberg, 2012). More recently, work on demographically specified has surfaced as an independent task (Agarwal and Sureka, 2016;

Davidson et al., 2017; Park and Fung, 2017; Safi Samghabadi et al., 2017; Silva et al., 2016; Tulkens et al., 2015; Waseem, 2016; Waseem and Hovy, 2016).

A large part of the previous work on hate speech detection has primarily touched upon surface level analysis of abusive language, leaving much room for work to be done. A large effort has been expended in attempting to define annotation schemes. Waseem and Hovy (2016) proposed guidelines derived from gender studies (**?**). A document is labelled as hate speech, if it fails any of the tests.

Waseem and Hovy (2016) a data set for sexist and racist speech on social media which is annotated using their guidelines. In their paper, they investigate the impact of several features on detecting racism and sexism. They find that characters are more discriminative for hate speech detection in line with the findings of Mehdad and Tetreault (2016). In addition, Waseem and Hovy (2016) find that information about a users gender can slightly improve classification performance, however they also find that adding location information slightly harms a classifiers performance. In addition, they find that information on length negatively impacts a classifiers performance.

Ross et al. (2016) investigate annotator agreement for anti-refugee sentiment. They instruct their annotators to follow the Twitter's guidelines for hateful content. They find that on a data set of 541 tweets, they achieve a very poor inter-annotator agreement, suggesting that it is necessary for clear and concise guidelines for annotation of abusive language.

Building on the work of Waseem and Hovy (2016) and Ross et al. (2016), Waseem (2016) consider the impact of annotators' knowledge of hate speech for building models for hate speech detection; they find that employing feminist annotators for labelling data sets allows for more consistent annotations and models as compared to annotators that are not screened for political opinion. Waseem (2016) consider the application of features from sarcasm detection, using Author Historical Salient Terms (AHST) proposed by Bamman and Smith (2015). The feature is generated by computing TF-IDF scores for each user and selecting the 100 highest weighted terms. If a term then occurs both in the document being analysed and in the AHST. Waseem (2016) find that AHST performs extremely poorly, suggesting that hate speech may generally be a one off event, rather than a continuous stream of abuse. It is our contention that another reason AHST might not work is due to the data set employed being highly imbalanced.

Davidson et al. (2017) seek to break down the task of hate speech detection into offensive language and hate speech and obtain labels for a Twitter data set using crowd sourced labour on CrowdFlower.

More recently Badjatiya et al. (2017) trained a deep convolutional neural network (CNN) on the data set annotated by Waseem and Hovy (2016). By using a CNN on the data set

Badjatiya et al. (2017) obtain a significant improvement on Waseem and Hovy's (2016) scores improving the F1 score from 73.89 to 93.00. Given the large increase in scores it is prudent to consider any potential errors. The data set Badjatiya et al. (2017) employ, is highly imbalanced with the positive classes occupying a small minority of the labelled data, there is a risk that their model performs extremely well on the negative class but does not perform well on the positive classes. However, no error analysis is provided in the paper.

In continuation of the results obtained by Badjatiya et al. (2017), Park and Fung (2017) compare using a two-step logistic regression classification, and a single step CNN approach to detecting hate speech. In the single step CNN, the specific form of hate speech is directly predicted, while in the two-step classification scenario, first a classifier is trained to identify whether a document contains abuse followed by predicting the specific type of abuse it is.

A different approach is attempted by Waseem et al. (ress), in which they seek to combine three different datasets for abusive language detection using multi-task learning. With a hypothesis that abuse will differ between geographic and cultural locales, they seek to employ disjoint datasets and train two models, one for each data set that share parameters. We will seek to extend this work to employ more data sets of abusive language as well as related tasks, such as sentiment analysis.

Finally, Jha and Mamidi (2017) break "sexism" down into benevolent and hostile sexism. They apply the ambivalent sexism theory as proposed by (Glick and Fiske, 1996). The ambivalent sexism theory suggests that there are two forms of sexism, benevolent sexism, which on a surface level speaks positively on women, but on a deeper level seeks to assert their inferiority, and hostile sexism, which expresses a strictly negative point of view on women. The following examples illustrate benevolent and hostile sexism respectively: "Women are like flowers who need to be cherished." and "Jus gonna say it..again..DUMB BITCH! #MKR".

As online platforms seek to remove online abuse occurring on their platform, data sets that have been gathered and annotated may have the abusive documents removed, thus requiring several rounds of reannotation of abusive language. In an attempt to deal with this, we will experiment with using documents that are assumed have a higher chance of being abusive as they are posted in forums that are known to abusive. Using these documents we will seek to build different forms of embeddings and evaluating on previously annotated data. Further, in an attempt to mitigate annotator needs, we will build an abusive language potential system which utilises supervised methods for Named Entity Recognition (NER), Gender Identification Sap et al. (2014), and sentiment analysis amongst other methods. The goal of this is to identify the probability that a document has potential to contain abusive

content, in efforts to exact greater control over what documents an annotator is faced with. Additionally, such a system will allow us to identify which documents which are clearly abusive. Thus, we will be able to provide an automated method to create a seed set of positive documents for abusive language detection.

3.2 Counter Speech

In an attempt to mitigate targetted abuse, Munger (2017) attempted an invervention study. In this study, they retrieved tweets containing term *ni**er*, hereafter *n-word*, that were specifically mentioned another Twitter user. For each user using the *n-word*, Munger (2017), collected their user information and the last 1000 tweets. On these tweets, Munger (2017) computed the average number of offensive words per tweet, using a dictionary of offensive words. Munger (2017) discarded some tweets: if the average number of offensive words per tweet fell below a threshold; if the user was not a white man; if the user was a minor; and if the user whom the tweet was directed at appeared to be friends. For the remaining users, Munger (2017) used a bot to send the following message in response to the abusive tweet:

> @[subject] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language (Munger (2017))

Munger (2017) alternated with using a bot with an identifable black name, a bot with an identifiable white name. For each of these bots two versions of the bot existed, one with a high number of followers and one with a low number of followers. Using these bots, they find that only the bot with a high number of followers that belongs to the in-group with the offending user, showed any significant decrease as a result of the tweet being sent.

ZW: Check if Susan Benesch work hasn't used counter speech.

To the best of our knowledge, this is the only paper that has attempted using automated counter speech as a means to react to abusive speech.

3.2.1 Natural Language Generation

As Munger (2017) shows, there is some potential in using alternative methods to moderation when dealing with abusive language online; particularly as the issues surrounding freedom of speech may be circumvented. For this reason we will look into Natural language generation (NLG). NLG deals with automatically generating language. While previous work has

attempted to generate language for social media (Sordoni et al., 2015), no previous work has dealt with generating language to counter abusive language online.[1]

NLG can be split into 6 tasks:

1. *C*ontent Determination - Deciding which information to include in the text under construction,
2. *T*ext Structuring - Determining in which order information will be presented in the text,
3. *S*entence Aggregation - Deciding which information to present in individual sentences,
4. *L*exicalisation - Finding the right words and phrases to express information,
5. *R*eferring Expression - Selecting the words and phrases to identify domain objects,
6. *L*inguistic Realisation - Combining all words and phrases into well-formed sentences.

(gatt and krahmer (2018); Reiter and Dale (1997, 2000))

Traditional NLG approaches have stuck closely to the above 6 tasks, in varying degrees of modularity and with varying degrees of manual labour required gatt and krahmer (2018). More recently, statistical approaches have been attempted for each of these tasks **??????**.

As no previous work has dealt with generating language, we must look at NLG for social media as a source of our methods along with the social scientific research on counter speech for hate speech. Further, we suggest that the generated language will be most effective if it has the same writing style as the abusive tweet. Munger (2017) found that racial identity and a high follower count influenced the use of racial slurs, specifically, this suggests that it may be possible to prolong the inhibiting effects freported by Munger (2017). As we will work with generating documents for Twitter, we identify the following two constraints for our NLG system:

1. Max Document Length: 280 Characters - Limitation on Tweets,
2. Stylistic mimicry of abusing user by approximate demographic,

As gatt and krahmer (2018) point out, generation for stylistic variation is dominated by deep neural network architectures, as such we will be employing deep neural networks.

[1] Note that Munger (2017) does not generate language but uses a single static message.

Deep Neural Networks (NN) have previously been used for languag generation built as either encoder-decoder architectures or as conditioned language models, we will focus on encoder-decoder architectures.

Encoder-decoder frameworks are neural network architectures, in whihc a Recurrent Neural Network (RNN) is used to encode the input into a vector representation; the vector output of the encoder then serves as the auxiliary input to a decoder RNN (**?**). This framework has been applied in the context of response generation, for instance by Sordoni et al. (2015) in which they build two Recurrent Neural Network Language Models (RLM) (**?**), in which one of the RLMs acts as an encoder and the other as a decoder. They utilise a negative log likelihood as their loss function, given by

$$L(s) = -\sum_{t=1}^{T} \log(P)(s_t|s_1, \ldots s_T) \tag{3.1}$$

The architectures proposed by Sordoni et al. (2015) consider three linguistic entities in a conversation between two users: the (linguistic) context c, the message m, and the response r, where c represents a sequence of past dialogue exchanges of any length and m is the message which the response r is responding to. Using this information they seek to train context based models to generate $r, r = r_1, \ldots, r_T$ using context based models. In their first model Sordoni et al. (2015) simply concatenate *c, m, r* into a single sentence s and seek to minimise $L(s)$. This model does not deal with long range dependencies, which may become an issue as s grows.

In the second model c and m are encoded into fixed-length vector representations which is used by the RLM to decode the response. Thus c and m are considered as as single sentence for which a single bag-of-words representation, b_{cm} is computed. **?** then provide b_{cm} as input to a non-linear feed-forward architecture that produces a fixed-length representation that is used to bias the recurrent state of the RLM. Their third model is similar to the second with one notable difference, rather than generating a single bag-of-words representation b_{cm}, two representations b_c and b_m are created and concatenated. They find that adding contextual information allows for their models to improve on the baselines.

As we will seek to generate text that mimics the style of the person to whom our system would be responding, we will need to handle stylistic variation such that authors can be mimicked. We will build on the work by Sordoni et al. (2015), extending their model to take in multiple contexts rather than a single context. We will provide three different contexts:

1. A history of tweets from the users c_u,

2. a history of abusive tweets from multiple users c_h,

3. and a history of counter-speech documents c_c.

The message m will still remain the tweet that we are generating a response to. We will be providing these histories as additional context, as we will seek to build a model that is informed by the user's general style, the abusive behaviour of many users across contexts, and how counter-speech has occurred on Twitter. In such a manner, we hope to generate responses that take contextual information about how abuse is formed, are modelled according to the style of the author, and take into account how counter-speech dealing with sexism and racism online has been performed by humans.

Having generated content which simulates the style of the abusive comment, we will seek to evaluate our generated texts using perplexity to measure its naturalness as well as cosine similarity and BLEU with human generated counter speech.

3.3 Summary

In this chapter, we have introduced the NLP work related to thesis and sought to show how we will build and expand on this work.

Chapter 4

Work To Date

In this chapter I will briefly introduce the work I have concluded for the PhD so far (see Appendix A for full papers).

4.1 Typology of Abuse - Published

In Waseem et al. (2017), we consider the implications of the current state of abusive language detection. in Davidson et al. (2017) and Schmidt and Wiegand (2017), clear misunderstandings of the term "hate speech" occur, considering both a legal, gender studies and critical race studies point of view. To effect this, we worked on a paper to emphasise and encourage researchers to focus on the manner in which the abuse occurs rather than the kind of abuse occurring, as focusing overly on the type of abuse will inevitably lead to potential overlaps between multiple tasks in the general topic of abusive language detection.

In the work, we propose that abusive language will exist in the following forms (please see table 4.1 for more examples):

- Directed Abuse
  In this case the abuse is aimed a particular individual or entity.
- Generalised Abuse
  Is the abuse generalised, it seeks to attack a large group of people, for instance by use of stereotyping an oppressed minority.
- Explicit Abuse
  Explicit Abuse may be explicit for instance by using bad words (slurs amongst other things).

- Implicit Abuse
  Implicit Abuse is abusive comments that seek to utilise alliterations, similes, and euphemisms amongst others to communicate abusive content. An example of this, is referring to refugees from Syria in terms of waves and streams, where the implicit connotations are that they are overrunning Europeans and must be stopped, lest we all "drown".

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ‸ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "Gas the skypes" (Magu et al., 2017), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011)

Table 4.1 Typology of abusive language (Waseem et al., 2017).

in Table 4.1 we see a mapping of the categories proposed and how they may interact.

4.2 Multitask Learning for Abusive Language Detection - In Press

This project was a collaboration between Zeerak Waseem, James Thorne (University of Sheffield), and Joachim Bingel (University of Copenhagen)[1].

This project sought to investigate the applicability of Multi-Task Learning for abusive language detection. The overarching aim of the project was to investigate the hypothesis that disparate data sets annotated for different kinds of abusive language in which the demographics writing the abuse can are different can be utilised for a unified model for abusive language detection.

[1]This project was the result of equal contribution from all authors.

We applied data from Waseem (2016); Waseem and Hovy (2016) and Davidson et al. (2017). The tweets collected by Waseem and Hovy (2016) and Waseem (2016) were collected without a specific demographic selected whereas Davidson et al. (2017) collected tweets that originated from the United States of America. Given that tehse differences meant that the datasets contained documents from several different demographic groups, each with their own understandings of bias and abuse we sought to investigate whether it was possible to utilise the distinct groups to improve detection of abusive language.

To investigate this we built classifiers trained on a single data set[2], the concatenation of all data sets, and a multi-task learning configuration in which multiple models are trained simultaneously.

Table 4.2 Comparison of test-set performance of within-domain and out-of-domain datasets using models trained only on one dataset (first four rows), models trained by concatenating both datasets (middle two rows), and using both datasets in a multi-task learning environment (final four rows). For each training regime, we compare using Bag-of-Words (BoW), the Average of Subword Embeddings (Emb) and both (B+E) as features for each tweet. Key: (R)acism, (S)exism, (H)ate-Speech, (O)ffensive, (N)either. Datasets: Davidson Davidson et al. (2017), WaseemWaseem (2016)/WaseemHovyWaseem and Hovy (2016) (W/W+H)

Training Objective		**Feats**	**F_1-Scores of Predictions on Test Sets**							
Primary	**Aux**		**W/W+H**				**Davidson**			
			R	S	N	Avg	H	O	N	Avg
W/W+H	-	BoW	0.70	0.65	0.88	0.82	0.00	0.64	0.42	0.57
W/W+H	-	Emb	0.30	0.42	0.85	0.71	0.01	0.04	0.29	0.08
W/W+H	-	B+E	0.00	0.00	0.82	0.57	0.00	0.00	0.29	0.05
Davidson	-	BoW	0.22	0.29	0.69	0.56	0.32	0.94	0.84	0.89
Davidson	-	Emb	0.00	0.32	0.60	0.48	0.19	0.92	0.69	0.84
Davidson	-	B+E	0.25	0.33	0.70	0.58	0.39	0.82	0.94	0.89
Both	-	BoW	0.21	0.54	0.81	0.70	0.20	0.92	0.77	0.86
Both	-	Emb	0.21	0.45	0.76	0.64	0.05	0.90	0.64	0.80
Both	-	B+E	0.17	0.53	0.81	0.69	0.31	0.92	0.77	0.86
W/W+H	Davidson	BoW	0.64	0.63	0.87	0.80	0.39	0.94	0.84	0.89
W/W+H	Davidson	Emb	0.32	0.50	0.84	0.72	0.10	0.91	0.64	0.82
W/W+H	Davidson	B+E	0.51	0.53	0.86	0.75	0.16	0.93	0.78	0.86
Davidson	W/W+H	BoW	0.66	0.62	0.86	0.79	0.37	0.94	0.83	0.89
Davidson	W/W+H	Emb	0.39	0.49	0.84	0.73	0.09	0.91	0.62	0.81
Davidson	W/W+H	B+E	0.60	0.57	0.85	0.77	0.14	0.93	0.78	0.86

We find that while we did not see improvements over the concatenated models, strong improvements were made over simply using a single training set. Importantly, we see that

[2]Waseem and Hovy (2016) and Waseem (2016) are concatenated into a single data set as they are collected from the same initial scrape

using multi-task learning we achieve similar performance to concatenating a data set, which allows for further research to continue with this setting using more data sets that may give related signals to abusive language detection.

4.3 Summary

In this chapter, we have briefly introduced the projects which have lead to publication at this point. Please see Appendix A for more detail.

Chapter 5

Work Plan and Timetable

In this chapter, we detail the experiments we will be working on, their components, and provide a timeline for the completion of each experiment. Recall from Chapter 1, that the research questions we'll be investigating are:

RQ1 How can embeddings, weak supervision, linguistic annotation, and meta data help improve platform and topic independent detection of abusive language and hate speech?

RQ2 How can NLP techniques be used as proxies for hate speech detection and thus minimise need for human annotation?

RQ3 Do correlations exist between the individual classes in abusive language data sets and variables such as the race or gender of the speaker, and do they allow us to build high classifiers to detect these in which the predictions have significant correlations with the classes annotated for abuse?

RQ4 How can NLG be used as a tool for intervention against online abuse and what is the impact of emulating the stylistic language of the author of the abuse?

In each subsection of the experiments planned, we provide an explanation as to how each experiment and its components build towards these aims.

5.1 Experiments planned

To answer the four research questions, we will conduct experiments which we believe will give insight into the answers to each question.

5.1.1 Overview

Here we provide an overview of how the individual experiments connect together to a cohesive answer to the overall questions of how to detect abusive language using machine learning and how we can seek to detect and address the potential of oppressed groups being overrepresented in the positive classes in abusive language data sets.

Research Questions and Experiments

Considering first the experiments that seek to address **RQ1**, we have the following experiments outlined:

1. Embedding Spaces for Abusive Language Detection (see section 5.1.3).
2. Multi-Task Learning for Abusive Language Detection (see section 5.1.4).
3. Classifying Abusive Documents using Linguistic Information and Meta data (see section 5.1.5).

These three experiment address three different ways of classifying abusive language documents such that there is a greater chance of generalisability of the machine learning classifiers built.

Next, looking at the experiments addressing **RQ2**, we have outlined the following experiments:

1. Distantly Supervised Procedural Algorithm for Abusive Language Detection (see section 5.1.2).

In this experiment, we specifically seek to investigate whether other NLP tasks such as sentiment analysis can be used as proxies for abusive language detection, thus allowing us to assign confidence of a document being abusive without human consideration of whether it is abusive. This research question, and experiment specifically seeks to investigate automated methods for generating labels for abusive language detection and thereby allowing for larger data sets for abusive language detection to be built.

Moving on to **RQ3**, this research question seeks to address the issue of confounding variables, demographic overrepresentation, and the influence of enforced community guidelines on abusive comments through the following experiments:

1. Computational Analysis of the Influence of Community Guidelines on Politeness on Reddit (see section 5.1.8).

2. Critical Analysis of Demographic Profiling in Abusive Language Data Sets (see section 5.1.6).

Through these experiemnts we seek to highlight how enforced community guidelines influence the abuse generated and how some abusive language data sets that have currently been released are biased towards confounding variables and certain demographics. These experiemnts will go to show that not only do correlations exist, but the individual communities have the power and agency to influence the abuse that is generated. Further, they will seek to highlight issues in creating abusive language data sets and how deciding which communties to scrape from can influence the subsequent abuse that is labelled.

Finally, considering **RQ4** we have outlined the following experiments:

1. Generation of Counter Speech for Hate Speech (see section 5.1.7)

This experiment stands as our final experiment and will be the last part which is executed due to the necessity of abusive language classifiers that can detect abuse with high precision, as such this experiment relies on **RQ1** and as a large data set of abusive documents and the counter speech to them may be needed it also relies on **RQ2**.

Cohesiveness of Experiments and Research Questions

Considering the experiments for each research question, we show that this thesis has 3 central focus areas: How to generate data for abusive language, how to build generalisable abusive language classifiers, and how to generate responses to abusive language. These aims, while independent of one another, are planned such that they each rely on the previous steps and the answer to each question leads to the next.

5.1.2 RQ2: Distantly Supervised Procedural Algorithm for Abusive Language Detection - Ongoing

This experiment is only beginning as ethical approval has only recently been granted. The aim of this experiment is to identify a method for detecting hate speech without providing a system with documents labelled for hate speech. Rather, we will aim to consider proxies to hate speech, such as impoliteness, negativity, the use of slurs, correlations with documents

known to be abusive, and so forth. Additionally, this experiment aims to minimise the need for human annotators for hate speech detection, as human annotators often introduce bias into the data set. This experiment is a collaborative effort with Factmata and the Univeristy of Michigan. As this experiment seeks to minimise annotator exposure to abuse, it is central to the completion of this thesis.

This experiment seeks to minimise need for human annotators, therefore we will use knowledge that some communities are known to be highly toxic and abusive. By using the comments and posts from such communities we can begin to determine features of abusive and discriminatory language. Additionally, by using these communities, we can generate embeddings that specifically address the semantic similarities of words in abusive contexts. Additionally, we will seek to find a way to discover new abusive terms by identifying semantically similar terms to known slurs and references to demographic groups. This information will then be used to create additional rules for a Named Entitity Recognition (NER) system such to help identify instances of abuse that would otherwise have alluded us. Beyond named entity recognition, we will also be computing the sentiment of a document, the sentiment towards a target, and stance towards a target person or subject. Further, we will compute the distances between a given document and documents that occur in abusive and non-abusive communities. The data sets which we use for this will be the datasets provided by Waseem (2016); Waseem and Hovy (2016) and Davidson et al. (2017). Additionally, we use data collected from Reddit.

The intuition behind this system is that while we do not have an automatic hate speech classification system, we do have proxies for hate speech that, when used in combination with one another can give us an idea of whether a document is likely to be abusive. While this may allow for an unsupervised collection of positive labels, there is a high risk that many of the labels would be highly explicit in their abuse. For this reason we propose that we build this system such that the aforementioned proxies are consulted individually, following which we are able to generate probabilities to assign a document of its potential to be abusive given the results from the respective NLP tools, and whether the individual tool is focused on high recall (NER) or high precision (community vectors).

As this experiment will require a large set of labelled and unlabelled data, it will be necessary to first undertake the other data generation experiments, while we collect data that has distant labels.

AV: Is the data generated going to be used to train a hate speech detector? In a way, the data generation approach would already qualify as one.

ZW: It will, but first we'll have it annotated by humans to see how well the method performs. Ideally, this method should allow you to generate labels for an unknown domain and unknown target by leveraging knowledge of how abusive communities speak.

5.1.3 RQ1: Embedding Spaces for Abusive Language Detection - Ongoing

In this experiment, we will investigate embeddings for abusive language detection. As this experiment deals with improving the generalisability of abusive language detection models, it is central to this thesis.

We will initially seek to analyse the usefulness of off-the-shelf embeddings for hate speech detection. Following this, we will seek to identify whether updating off-the-shelf embeddings with embeddings trained on data sets with large amounts of abusive language allow for improving model performance. We will also try with concatenation of off-the-shelf and pre-trained embeddings. Finally, we will investigate the applicability of various forms of embeddings for detecting abuse occurring on other platforms than has been used to generate the embeddings. For this experiment, we plan to use a mix of linear models, particularly Support Vector Machines (SVM), Decision Trees, and Logistic Regression, and neural network models in particular Long-Short Term Memory (LSTM) models and CNNs (Convolutional Neural Networks). Further, we will be using different forms of embeddings both pre-trained embeddings as well as well as embeddings trained on datasets of abusive language.

This experiment relies on the experiments that create data sets, as the data collected will serve as the foundation for training embeddings.

5.1.4 RQ1: Multi-Task Learning for Abusive Language Detection

In this experiment we follow up on the work of Waseem et al. (ress), in which they seek to utilise multi-task learning for abusive language detection. In Waseem et al. (ress) they utilise multiple abusive language data sets, in this experiment we will seek to utilise data sets for other tasks such as sentiment analysis, dependency parsing, cyber bullying, and named entity recognition as our auxiliary tasks in order to investigate the utility of related tasks for improving abusive language detection. Our main task will be detecting abusive language. As we will use tasks where large scale data is available, we hope to improve generalisability

of abusive language detection methods without the use of large scale data sets for abusive language. As such, this experiment is core to the thesis.

As there is no not a need for collecting and annotating data sets, this experiment does not rely on others.

5.1.5 RQ1: Classifying Abusive Documents using Linguistic Information and Meta data

In this experiment, we will seek to address the issue of overfitting to the small positive classes in abusive language data sets and thus having models have confounding variables. We will aim to avoid confounding variables in our models by seeking to utilise linguistic knowledge such as the part of speech tags and dependency trees. Further, we will seek to identify whether meta data such as the creation time of the user account, the frequency of tweets, and the number of followers and followees is useful for abusive language detection. In this experimeinet we will be using previously published data sets (Davidson et al., 2017; Waseem, 2016; Waseem and Hovy, 2016) in addition to data obtained from abusive communities on Reddit. We will be applying linear models (logistic regression and SVMs) and neural network methods (LSTMs and CNNs). As this experiment seeks to improve generalisability of abusive language classifiers through the use of linguistic annotation and user meta data, it is central to the thesis.

This experiment does not explicitly require more data, however we will let this experiment depend on the data obtained through the collection of abusive communities on Reddit.

5.1.6 RQ3: Critical Analysis of Demographic Profiling in Abusive Language Data Sets

In Davidson et al. (2017), a great deal of the abusive language, tagged as offensive or hate speech, consists of African American Vernacular English (AAVE). In this paper, will seek to show that by using a classifier trained on this dataset, the African American diasporas are disproportionately targeted for the use of their dialect which may be construed as offensive. We will be using Waseem (2016); Waseem and Hovy (2016) and Davidson et al. (2017) as our sources of data. To provide fair comparison, we will be using the same models that have been used in the papers introducing the data sets. As this experiment deals with identifying the impact of a data sets on the oppressed populations it is central to this thesis.

We will investigate the impact on the African-American community by reimplementing the classifier trained in Davidson et al. (2017) to obtain the labels. Following this, we will

correlate the labels predicted on the test set with our annotations of the test set using AAVE. Additionally, we will compute correlations between common terms in AAVE such as “finna”, “n**ga(h)” with their likelihood of being tagged as offensive or hate speech. The majority of the work will seek to problematise data sampling solely based on slurs and call for data collection to be done with respect to the oppression of certain demographics and with respect to influencers such as dialect. Finally, the paper will seek to highlight the importance of annotators with linguistic knowledge for a task such as hate speech detection.

This experiment does not rely on other experiments, as it seeks to deal with considerations of flaws and weaknesses in previously published work.

5.1.7 RQ4: Generation of Counter Speech for Hate Speech - Early planning

In this experiment, we will seek to build on the work of Munger (2017) in which they experiment with the use of bots to mitigate abusive behaviours on-line. As this experiment seeks to explore the possibility of dealing with abusive language without prompting for the document or user to be deleted, this will be central to the thesis.

We pursue this, as a fundamental question of how to deal with abuse once it is discovered becomes key to how the task is handled. While removal, or improved moderation, is often argued as the motivation for abusive language detection research, it is clear that simply removing avenues for speech will not prevent the speech - it may simply move to another context, whether online or offline where it is deemed acceptable. To consider alternative methods to removal, we investigate the potential of generating content that can act as counter speech. To this end, we propose generation of two forms of responses:

1. Generation of dialogue promoting speech.
2. Generation of Non-Violent Communication (De-escalating Speech)

The tasks differ in the functional use of language. In task 1, the primary concern is to pose questions that invite for dialogue yet are critical of the outset. In task 2, the primary objective is to attempt to de-escelate communications from a given user to a target. Thus, while task 1 seeks to have an abusing user explore the roots of their abuse, task 2 seeks to simply de-escalate a given situation. While a number of generation experiments could be interesting to pursue, we have chosen these as they frame and highlight the offending user without an attempt to obviously any behaviour, but seek for the offender to explain or to

communicate in a calmer, less aggressive form. As such, we hope to show one viable option to removal of abusive documents.

This experiment relies on all of the detection experiments, as identifying which users should be addressed will rely on high-precision methods for detection. Further, it will rely on the critical analysis work, as we will seek to ensure that oppressed demographics are not targetted higher due to biases in the detection methods and data sets upon which they are built.

5.1.8 RQ3: Computational Analysis of the Influence of Community Guidelines on Politeness on Reddit

In this experiment, we will identify two reddit communities that are highly similar in the topics discussed but with differences in community guidelines on what is acceptable speech. We will then seek to see how these communities differ in their rhetorical styles and the amount of abuse that is in the communities. This experiment is being undertaken with Dennis Tenen (University of Columbia). As this will shed light on the importance of guidelines and moderation on abuse, this will be central to the thesis.

This experiment does not depend on any other experiment, however it relies on acquiring data from Reddit.

5.2 Timetable

In order to answer the research questions described in Section **??**, the following tasks will be carried out: These are the experiments which we are aiming to work on throughout the remainder of this Ph.D. studentship. Items marked with an "*" have already received ethical approval, while items marked with "$^{+}$" have yet to have an application for ethical approval filed. Items marked with "$^{-}$" do not need an ethical approval.

ZW: @KB,@AV Please let me know if the items marked as not needing approval actually won't need approval.

1. Critical Work
   (a) $^{-}$ Critical Analysis of Demographic Profiling in Abusive Language Data Sets
   (b) $^{-}$ Computational Analysis of the Influence of Community Guidelines on Reddit
2. Generalisability of and Improved Classification Models for Abusive Language Models

		2018				2019				2020	
Overall Task	Subtask	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
Literature Review		•	•	•	•						
Data Generation and Annotation						•					
	Minimise Annotator Interaction		•	•							
Critical Work										•	
	Demographic Profiling		•	•			•		•		
	Analysis of Community Guidelines			•	•						
Improved Abusive Language Detection									•		
	Embeddings for Detection			•	•	•	•	•			
	Multi-Task Learning Using Out-Of-Domain data				•	•	•	•			
	Linguistic Annotation & Meta data			•	•	•					
Natural Language Generation											•
	Dialogue Promoting Speech							•	•	•	
	Non-violent communication							•	•	•	

Table 5.1 Task Timetable

(a) $^{-}$ Exploration of (Pre-trained) Embeddings for Abusive Language Detection

(b) * Multi-Task Learning for Abusive Language Detection Using Out of Domain data sets

(c) $^{-}$ Classifying Abusive Documents using Linguistic Annotation and Meta data

3. Abusive Data Generation

   (a) * Minimise User Interactions with Abuse in Annotation Processes

4. Natural Language Generation

   (a) $^{+}$ Generating Dialogue Promoting Speech

   (b) $^{+}$ Generating non-violent communication (de-escalating speech)

In Table 5.1, we present a time table for the completion of each subtask.

Table 5.1 presents the timetable for the activities previously outlined. Each task is scheduled on a quarterly basis to provide an overview of the expected time necessary to complete it. For each group of tasks, a quarter is given as the period in which we will be writing the respective chapters of the thesis. At the end of each individual experiment component, we schedule a month for submission of papers to conferences and journals.

5.3 Contingency Plans

Should a experiment overrun its allotted time by a quarter or more, we will drop a subsequent experiment. For instance, should multi-task learning for abuse detection take longer to implement and test than expected, then we will drop working on the experiment working on detecting hate speech without the use n-grams of words. Additioanlly, we leave 3 months

for finalising the final chapters of the thesis. These three months also serve as a buffer for experiments and time running over, as we will have another year beyond our experiment plan in which we may finalise experiments and the thesis writing.

5.4 Summary

In this chapter, we have broken each research question into the experiments and provide a timeline for each experiment. While there may be deviations from this timeline, we will seek to adhere to it as best possible.

Chapter 6

(DDP) Progress

In this chapter we present a list of activities that have been performed in order to comply with the Doctoral Development Program (DDP) requirements:

1. DDP
   (a) The Training Need Analysis (TNA) for the first year was elaborated.
   (b) I have taken the module FCE6100 “Professional Behaviour and Ethical Conduct”.
   (c) I have participated in and lead several sessions of the NLP Group Reading Group.
   (d) I have attended several and lead one of the NLP Group Seminars.
2. Conference Participation
   (a) Participation at ACL, 2017
   (b) Participation at EMNLP, 2017
3. Organisational Work
   (a) Organise the First Workshop on Abusive Language Online at ACL 2017
   (b) Co-organiser of the Widening NLP Workshop at NAACL 2018.
   (c) Co-organiser of the 2nd Workshop on Abusive Language Online at EMNLP 2018.
4. Summer/Winter Schools & Workshops
   (a) Participation in The 7th Annual Lisbon Machine Learning Summer School (LxMLS 2017)

(b) Participated in workshop: Understanding Euroscepticism through the Lens of Big-Data, Villa Vigoni, 04/12/2017-07/12/2017

5. Published Papers

   (a) Published a paper at The First Workshop on Abusive Language Online (first author)

   (b) Participated in the PAN Author Profiling Task with Adam Poulston and Mark Stevenson. We obtained the 7th place out of 22. Published working notes for our approach to the task (second author)

   (c) Published Book Chapter on Multi-Task Learning for Hate Speech Detection

6. Invited Talks

   (a) Given an invited talk at Potsdam University, 08/05/2017

   (b) Given an invited talk at Leuphana University, 23/11/2017

   (c) Given an invited talk at Sapienza University of Rome, 01/12/2017

   (d) Given an invited talk at the Lorentz Workshop on Intersectionality and Algorithmic Discrimination, 18/12/2017 - 22/12/2017

   (e) Given an invited talk at Heidelberg University, 23/01/2018

   (f) Participated workshop: The turn to artificial intelligence in governing communication online, 20/03/2017

   (g) Given an invited talk at the Alan Turing Institute, 09/04/2018

7. Reviewing

   (a) ConLL, 2017

   (b) NLP and CSS, 2017

   (c) COLING 2018

   (d) EMNLP 2018

   (e) Big Data and Society

   (f) Transactions on Asian and Low-Resource Language Information Processing

   (g) Transactions on the Web

   (h) GeoJournal

In addition to the above, I have kept a focus and will continue to focus on improving my skills and understanding of relevant topics

1. Core Computer Science topics
2. German Language Skills
3. Linear Algebra
4. Probability Theory
5. Statistics
6. English Punctuation and grammar
7. Machine Learning Methods
    - (Deep) Neural Networks
    - Hidden Markov Models
    - Sequence-to-Sequence models

Finally, there is a chance that I may be editing a book (on abusive language online), and contributing to another book on uncertainty.

References

(1999). Identity and deception in the virtual community. In Smith, M. A. and Kollock, P., editors, *Communities in cyberspace*, pages 29–59. Routledge.

(2007). Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication*, 12(4).

Abberley, P. (1987). The concept of oppression and the development of a social theory of disability. *Disability, Handicap & Society*, 2(1):5–19.

Agarwal, S. and Sureka, A. (2016). Spider and the flies : Focused crawling on tumblr to detect hate promoting communities. *CoRR*, abs/1603.09164.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Bamman, D. and Smith, N. (2015). Contextualized sarcasm detection on twitter.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3).

Bates, L. (2013). Does facebook have a problem with women? `https://www.theguardian.com/lifeandstyle/2013/feb/19/facebook-images-rape-domestic-violence`.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Burnap, P. and 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.

Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York.

Chen, Y., Zhang, L., Michelony, A., and Zhang, Y. (2012). 4is of social bully filtering: Identity, inference, influence, and intervention. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2677–2679, New York, NY, USA. ACM.

Crawford, K. and Gillespie, T. (2016). What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist eory and antiracist politics. *University of Chicago Legal Forum*, 1989(1).

Daegon Cho, A. A. (2013). The more social cues, the less trolling? an empirical study of online commenting behavior. In *Proceedings of the Twel h Workshop on the Economics of Information Security (WEIS 2013)*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

de Beauvoir, S. (1953). *The Second Sex*. Knopf, New York.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02).

Dubois, W. (1935). *Black Reconstruction in America: an essay toward a history of the part which black folk played in the attempt to reconstruct democracy in America, 1860-1880*. Philadelphia, Penn. : Albert Saifer.

Eisenstein, Z. (1977). Constructing a theory of capitalist patriarchy and socialist feminism. *Insurgent Sociologist*, 7(3):3–17.

Erevelles, N. and Minear, A. (2010). Unspeakable offenses: Untangling race and disability in discourses of intersectionality. *Journal of Literary & Cultural Disability Studies*, 4(2).

European Commission (2016). Code of conduct on countering illegal hate speech online. Technical report.

Galperin, E. (2011). Randi zuckerberg runs in the wrong direction on pseudonymity online.

gatt, a. and krahmer, e. (2018). survey of the state of the art in natural language generation: core tasks, applications and evaluation. *journal of artificial intelligence research (jair)*, 61:65–170.

Glick, P. and Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491.

Gorrell, G., Greenwood, M., Roberts, I., Maynard, D., and Bontcheva, K. (2018). Online abuse of uk mps in 2015 and 2017: Perpetrators, targets, and topics. *arXiv*. © 2018 The Author(s). For reuse permissions, please contact the Author(s).

Han, S. (2015). *Letters of the Law: Race and the Fantasy of Colorblindness in American Law*. Stanford University Press, Stanford, CA.

Hannon, C. (2016). Gender and status in voice user interfaces. *interactions*, 23(3):34–37.

Hine, G. E., Onaolapo, J., Cristofaro, E. D., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., and Blackburn, J. (2017). A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. In *ICWSM*, volume abs/1610.03452.

Home Office (2016). Action against hate the uk government's plan for tackling hate crime. Technical report.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. pages 7–16. Association for Computational Linguistics.

Kontostathis, A., Reynolds, K., Garron, A., and Edwards, L. (2013). Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 195–204, New York, NY, USA. ACM.

Kuchera, B. (2014). Twitter can fix its harassment problem, but why mess with success? `http://www.polygon.com/2014/7/30/5952135/twitter-harassment-problems`.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 1621–1622. AAAI Press.

Lawrence III, C. R. (1992). Crossburning and the sound of silence: Antisubordination theory and the first amendment. *Villanova Law Review*, 37(4).

Levin, S. (2017). Moderators who had to view child abuse content sue microsoft, claiming ptsd. Last Accessed: May 22nd, 2018.

Levine, M. (2013). Controversial, harmful and hateful speech on facebook. `https://www.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054/`.

Logie, C. H., James, L., Tharao, W., and Loutfy, M. R. (2011). Hiv, gender, race, sexual orientation, and sex work: A qualitative study of intersectional stigma experienced by hiv-positive women in ontario, canada. *PLoS:med*, 11(8).

Magu, R., Joshi, K., and Luo, J. (2017). Detecting the hate code on social media. *arXiv preprint arXiv:1703.05443*.

Marwick, A. E. and danah boyd (2011). I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.

McIntosh, P. (1988). White privilege and male privilege: A personal account of coming to see correpondences through work in women's studies.

Mehdad, Y. and Tetreault, J. R. (2016). Do characters abuse more than words? In *SIGDIAL Conference*.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

Neal, D. (2004). The measured black-white wage gap among women is too small. *Journal of Political Economy*, 112(S1):S1–S28.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Packer, B., Halpern, Y., Guajardo-Céspedes, M., and Mitchell, M. (2018). Text embedding models contain bias. here's why that matters. Last Accessed: May 23rd, 2018.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. pages 41–45. Association for Computational Linguistics.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M., Wojatzki, M., and Zesch, T., editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.

Safi Samghabadi, N., Maharjan, S., Sprague, A., Diaz-Sprague, R., and Solorio, T. (2017). Detecting nastiness in social media. pages 63–72. Association for Computational Linguistics.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., and Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *EMNLP*, pages 1146–1151. Association for Computational Linguistics.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Shah, N. (2015a). Exposed net porn: Penetrating regulation, bodies, and sexuality in the age of the internet. In Wendy Hui Kyong Chun, Anna Watkins Fisher, T. K., editor, *New Media, Old Media: A History and Theory Reader*, pages 539 – 551. Routledge.

Shah, N. (2015b). Sluts 'r' us: Intersections of gender, protocol and agency in the digital age. *First Monday*, 20(4).

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). Analyzing the targets of hate in online social media. *CoRR*, abs/1603.07709.

Sood, S., Antin, J., and Churchill, E. (2012a). Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490. ACM.

Sood, S. O., Antin, J., and Churchill, E. F. (2012b). Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2015). Detecting racism in dutch social media posts.

U.N., U. N. G. A. (1948). The universal declaration of human rights. `http://www.un.org/en/universal-declaration-human-rights/`.

van der Nagel, E. (2013). Faceless bodies: Negotiating technological and cultural codes on reddit gonewild. *Scan: Journal of Media Arts Culture*, 10(2).

van der Nagel, E. and Frith, J. (2015). Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of r/gonewild. *First Monday*, 20(3).

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680. INCOMA Ltd. Shoumen, BULGARIA.

Verloo, M. (2006). Multiple inequalities, intersectionality and the european union. *European Journal of Women's Studies*, 13(3).

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, San Diego, California. Association for Computational Linguistics.

Waseem, Z., Thorne, J., and Bingel, J. (In Press). Bridging the gaps: Multi-task learning for domain transfer of hate speech detection. In Golbeck, J., editor, *Online Harassment*. Springer.

Zack, N. (2015). *White Privilege and Black Rights: The Injustice of U.S. Police Racial Profiling and Homicide*. Rowman & Littlefield.

Zaleski, K. L., Gundersen, K. K., Baes, J., Estupinian, E., and Vergara, A. (2016). Exploring rape culture in social media forums. *Computers in Human Behavior*, 63:922 – 927.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association of Computational Linguistics.

Appendix A

Published Work

Understanding Abuse: A Typology of Abusive Language Detection Subtasks

Zeerak Waseem
Department of Computer Science
University of Sheffield
United Kingdom
z.w.butt@sheffield.ac.uk

Thomas Davidson
Department of Sociology
Cornell University
Ithica, NY
trd54@cornell.edu

Dana Warmsley
Department for Applied Mathematics
Cornell University
Ithica, NY
dw457@cornell.edu

Ingmar Weber
Qatar Computing Research Institute
HBKU
Doha, Qatar
iweber@hbku.edu.qa

Abstract

As the body of research on abusive language detection and analysis grows, there is a need for critical consideration of the relationships between different subtasks that have been grouped under this label. Based on work on hate speech, cyberbullying, and online abuse we propose a typology that captures central similarities and differences between subtasks and we discuss its implications for data annotation and feature construction. We emphasize the practical actions that can be taken by researchers to best approach their abusive language detection subtask of interest.

1 Introduction

There has been a surge in interest in the detection of abusive language, hate speech, cyberbullying, and trolling in the past several years (Schmidt and Wiegand, 2017). Social media sites have also come under increasing pressure to tackle these issues. Similarities between these subtasks have led scholars to group them together under the umbrella terms of "abusive language", "harmful speech", and "hate speech" (Nobata et al., 2016; Faris et al., 2016; Schmidt and Wiegand, 2017) but little work has been done to examine the relationship between them. As each of these subtasks seeks to address a specific yet partially overlapping phenomenon, we believe that there is much to gain by studying how they are related.

The overlap between subtasks is illustrated by the variety of labels used in prior work. For example, in annotating for cyberbullying events, Van Hee et al. (2015b) identifies discriminative remarks (racist, sexist) as a subset of "insults", whereas Nobata et al. (2016) classifies similar remarks as "hate speech" or "derogatory language". Waseem and Hovy (2016) only consider "hate speech" without regard to any potential overlap with bullying or otherwise offensive language, while Davidson et al. (2017) distinguish hate speech from generally offensive language. Wulczyn et al. (2017) annotates for personal attacks, which likely encompasses identifying cyberbullying, hate speech, and offensive language. The lack of consensus has resulted in contradictory annotation guidelines - some messages considered as hate speech by Waseem and Hovy (2016) are only considered derogatory and offensive by Nobata et al. (2016) and Davidson et al. (2017).

To help to bring together these literatures and to avoid these contradictions, we propose a typology that synthesizes these different subtasks. We argue that the differences between subtasks within abusive language can be reduced to two primary factors:

1. *Is the language directed towards a specific individual or entity or is it directed towards a generalized group?*

2. *Is the abusive content explicit or implicit?*

Each of the different subtasks related to abu-

sive language occupies one or more segments of this typology. Our aim is to clarify the similarities and differences between subtasks in abusive language detection to help researchers select appropriate strategies for data annotation and modeling.

2 A typology of abusive language

Much of the work on abusive language subtasks can be synthesized in a two-fold typology that considers whether (i) the abuse is directed at a specific target, and (ii) the degree to which it is explicit.

Starting with the targets, abuse can either be directed towards a specific individual or entity, or it can be used towards a generalized *Other*, for example people with a certain ethnicity or sexual orientation. This is an important sociological distinction as the latter references a whole category of people rather than a specific individual, group, or organization (see Brubaker 2004, Wimmer 2013) and, as we discuss below, entails a linguistic distinction that can be productively used by researchers. To better illustrate this, the first row of Table 1 shows examples from the literature of directed abuse, where someone is either mentioned by name, tagged by a username, or referenced by a pronoun.[1] Cyberbullying and trolling are instances of directed abuse, aimed at individuals and online communities respectively. The second row shows cases with abusive expressions towards generalized groups such as racial categories and sexual orientations. Previous work has identified instances of hate speech that are both directed and generalized (Burnap and Williams, 2015; Waseem and Hovy, 2016; Davidson et al., 2017), although Nobata et al. (2016) come closest to making a distinction between directed and generalized hate.

The other dimension is the extent to which abusive language is explicit or implicit. This is roughly analogous to the distinction in linguistics and semiotics between *denotation*, the literal meaning of a term or symbol, and *connotation*, its sociocultural associations, famously articulated by Barthes (1957). Explicit abusive language is that which is unambiguous in its *potential* to be abusive, for example language that contains racial or homophobic slurs. Previous research has indicated a great deal of variation within such language (Warner and Hirschberg, 2012; Davidson et al., 2017), with abusive terms being used in a colloquial manner or by people who are victims of abuse. Implicit abusive language is that which does not immediately imply or denote abuse. Here, the true nature is often obscured by the use of ambiguous terms, sarcasm, lack of profanity or hateful terms, and other means, generally making it more difficult to detect by both annotators and machine learning approaches (Dinakar et al., 2011; Dadvar et al., 2013; Justo et al., 2014). Social scientists and activists have recently been paying more attention to implicit, and even unconscious, instances of abuse that have been termed "micro-aggressions" (Sue et al., 2007). As the examples show, such language may nonetheless have extremely abusive connotations. The first column of Table 1 shows instances of explicit abuse, where it should be apparent to the reader that the content is abusive. The messages in the second column are implicit and it is harder to determine whether they are abusive without knowing the context. For example, the word "them" in the first two examples in the generalized and implicit cell refers to an ethnic group, and the words "skypes" and "Google" are used as euphemisms for slurs about Jews and African-Americans respectively. Abuse using sarcasm can be even more elusive for detection systems, for instance the seemingly harmless comment praising someone's intelligence was a sarcastic response to a beauty pageant contestants unsatisfactory answer to a question (Dinakar et al., 2011).

3 Implications for future research

In the following section we outline the implications of this typology, highlighting where the existing literatures indicate how we can understand, measure, and model each subtype of abuse.

3.1 Implications for annotation

In the task of annotating documents that contain bullying, it appears that there is a common understanding of what cyberbullying entails: an intentionally harmful electronic attack by an individual or group against a victim, usually repetitive in nature (Dadvar et al., 2013). This consensus allows for a relatively consistent set of annotation guidelines across studies, most of which simply ask annotators to determine if a post contains bullying or harassment (Dadvar et al., 2014; Kontostathis et al., 2013; Bretschneider et al., 2014).

[1] All punctuation is as reported in original papers. We have added all the * symbols.

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

Table 1: **Typology of abusive language.**

High inter-annotator agreement on cyberbullying tasks (93%) (Dadvar et al., 2013) further indicates a general consensus around the features of cyberbullying (Van Hee et al., 2015b). After bullying has been identified annotators are typically asked more detailed questions about the extremity of the bullying, the identification of phrases that indicate bullying, and the roles of users as bully/victim (Dadvar et al., 2014; Van Hee et al., 2015b; Kontostathis et al., 2013).

We expect that consensus may be due to the directed nature of the phenomenon. Cyberbullying involves a victim whom annotators can identify and relatively easily discern whether statements directed towards the victim should be considered abusive. In contrast, in work on annotating harassment, offensive language, and hate speech there appears to be little consensus on definitions and lower inter-annotator agreement ($\kappa \approx 0.60 - 0.80$) (Ross et al., 2016; Waseem, 2016a; Tulkens et al., 2016; Bretschneider and Peters, 2017) are obtained. Given that these tasks are often broadly defined and the target is often generalized, all else being equal, it is more difficult for annotators to determine whether statements should be considered abusive. Future work in these subtasks should aim to have annotators distinguish between targeted and generalized abuse so that each subtype can be modeled more effectively.

Annotation (via crowd-sourcing and other methods) tends to be more straightforward when explicit instances of abusive language can be identified and agreed upon (Waseem, 2016b), but is considerably more difficult when implicit abuse is considered (Dadvar et al., 2013; Justo et al., 2014; Dinakar et al., 2011). The connotations of language can be difficult to classify without domain-specific knowledge. Furthermore, while some argue that detailed guidelines can help annotators to make more subtle distinctions (Davidson et al., 2017), others find that they do not improve the reliability of non-expert classifications (Ross et al., 2016). In such cases, expert annotators with domain specific knowledge are preferred as they tend to produce more accurate classifications (Waseem, 2016a).

Ultimately, the nature of abusive language can be extremely subjective, and researchers must endeavor to take this into account when using human annotators. Davidson et al. (2017), for instance, show that annotators tend to code racism as hate speech at a higher rate than sexism. As such, it is important that researchers consider the social biases that may lead people to disregard certain types of abuse.

The type of abuse that researchers are seeking to identify should guide the annotation strategy. Where subtasks occupy multiple cells in our typology, annotators should be allowed to make nuanced distinctions that differentiate between different types of abuse. In highlighting the major differences between different abusive language detection subtasks, our typology indicates that different annotation strategies are appropriate depending on the type of abuse.

3.2 Implications for modeling

Existing research on abusive language online has used a diverse set of features. Moving forward, it is important that researchers clarify which features are most useful for which subtasks and which subtasks present the greatest challenges. We do not attempt to review all the features used (see Schmidt and Wiegand 2017 for a detailed review)

but make suggestions for which features could be most helpful for the different subtasks. For each aspect of the typology, we suggest features that have been shown to be successful predictors in prior work. Many features occur in more than one form of abuse. As such, we do not propose that particular features are necessarily unique to each phenomenon, rather that they provide different insights and should be employed depending on what the researcher is attempting to measure.

Directed abuse. Features that help to identify the target of abuse are crucial to directed abuse detection. Mentions, proper nouns, named entities, and co-reference resolution can all be used in different contexts to identify targets. Bretschneider and Peters (2017) use a multi-tiered system, first identifying offensive statements, then their severity, and finally the target. Syntactical features have also proven to be successful in identifying abusive language. A number of studies on hate speech use part-of-speech sequences to model the expression of hatred (Warner and Hirschberg, 2012; Gitari et al., 2015; Davidson et al., 2017). Typed dependencies offer a more sophisticated way to capture the relationship between terms (Burnap and Williams, 2015). Overall, there are many tools that researchers can use to model the relationship between abusive language and targets, although many of these require high-quality annotations to use as training data.

Generalized abuse. Generalized abuse online tends to target people belonging to a small set of categories, primarily racial, religious, and sexual minorities (Silva et al., 2016). Researchers should consider identifying forms of abuse unique to each target group addressed, as vocabularies may depend on the groups targeted. For example, the language used to abuse trans-people and that used against Latin American people are likely to differ, both in the nouns used to denote the target group and the other terms associated with them. In some cases a lexical method may therefore be an appropriate strategy. Further research is necessary to determine if there are underlying syntactic structures associated with generalized abusive language.

Explicit abuse Explicit abuse, whether directed or generalized, is often indicated by specific keywords. Hence, dictionary-based approaches may be well suited to identify this type of abuse (Warner and Hirschberg, 2012; Nobata et al., 2016), although the presence of particular words should not be the only criteria, even terms that denote abuse may be used in a variety of different ways (Kwok and Wang, 2013; Davidson et al., 2017). Negative polarity and sentiment of the text are also likely indicators of explicit abuse that can be leveraged by researchers (Gitari et al., 2015).

Implicit abuse. Building a specific lexicon may prove impractical, as in the case of the appropriation of the term "skype" in some forums (Magu et al., 2017). Still, even partial lexicons may be used as seeds to inductively discover other keywords by use of a semi-supervised method proposed by King et al. (2017). Additionally, character n-grams have been shown to be apt for abusive language tasks due to their ability to capture variation of words associated with abuse (Nobata et al., 2016; Waseem, 2016a). Word embeddings are also promising ways to capture terms associated with abuse (Djuric et al., 2015; Badjatiya et al., 2017), although they may still be insufficient for cases like 4Chan's connotation of "skype" where a word has a dominant meaning and a more subversive one. Furthermore, as some of the above examples show, implicit abuse often takes on complex linguistic forms like sarcasm, metonymy, and humor. Without high quality labeled data to learn these representations, it may be difficult for researchers to come up with models of syntactic structure that can help to identify implicit abuse. To overcome these limitations researchers may find it prudent to incorporate features beyond just textual analysis, including the characteristics of the individuals involved (Dadvar et al., 2013) and other extra-textual features.

4 Discussion

This typology has a number of implications for future work in the area.

First, we want to encourage researchers working on these subtasks to learn from advances in other areas. Researchers working on purportedly distinct subtasks are often working on the same problems in parallel. For example, the field of hate speech detection can be strengthened by interactions with work on cyberbullying, and vice versa, since a large part of both subtasks consists of identifying targeted abuse.

Second, we aim to highlight the important distinctions within subtasks that have hitherto been ignored. For example, in much hate speech research, diverse types of abuse have been lumped

together under a single label, forcing models to account for a large amount of within-class variation. We suggest that fine-grained distinctions along the axes allows for more focused systems that may be more effective at identifying particular types of abuse.

Third, we call for closer consideration of how annotation guidelines are related to the phenomenon of interest. The type of annotation and even the choice of annotators should be motivated by the nature of the abuse. Further, we welcome discussion of annotation guidelines and the annotation process in published work. Many existing studies only tangentially mention these, sometimes never explaining how the data were annotated.

Fourth, we encourage researchers to consider which features are most appropriate for each subtask. Prior work has found a diverse array of features to be useful in understanding and identifying abuse, but we argue that different feature sets will be relevant to different subtasks. Future work should aim to build a more robust understanding of when to use which types of features.

Fifth, it is important to emphasize that not all abuse is equal, both in terms of its effects and its detection. We expect that social media and website operators will be more interested in identifying and dealing with explicit abuse, while activists, campaigners, and journalists may have more incentive to also identify implicit abuse. Targeted abuse such as cyberbullying may be more likely to be reported by victims and thus acted upon than generalized abuse. We also expect that implicit abuse will be more difficult to detect and model, although methodological advances may make such tasks more feasible.

5 Conclusion

We have presented a typology that synthesizes the different subtasks in abusive language detection. Our aim is to bring together findings in these different areas and to clarify the key aspects of abusive language detection. There are important analytical distinctions that have been largely overlooked in prior work and through acknowledging these and their implications we hope to improve abuse detection systems and our understanding of abusive language.

Rather than attempting to resolve the "definitional quagmire" (Faris et al., 2016) involved in neatly bounding and defining each subtask we encourage researchers to think carefully about the phenomena they want to measure and the appropriate research design. We intend for our typology to be used both at the stage of data collection and annotation and the stage of feature creation and modeling. We hope that future work will be more transparent in discussing the annotation and modeling strategies used, and will closely examine the similarities and differences between these subtasks through empirical analyses.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. pages 759–760.

Roland Barthes. 1957. *Mythologies*. Seuil.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Uwe Bretschneider, Thomas Whner, and Ralf Peters. 2014. Detecting online harassment in social networks. In *ICIS 2014 Proceedings: Conference Theme Track: Building a Better World through IS*.

Rogers Brubaker. 2004. *Ethnicity without groups*. Harvard University Press.

Pete Burnap and Matthew L Williams. 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.

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Conference on Artificial Intelligence*. Springer International Publishing.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, pages 693–696.

Thomas Davidson, Dana Warmsley, Micheel Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media*. Montreal, Canada, pages 512–515.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and

mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):18.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web* 11(02).

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, pages 29–30.

Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo. 2016. Understanding harmful speech online. *Berkman Klein Center Research Publication* 21.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10(4):215–230.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. In *Proceedings of the Eleventh International Conference on Web and Social Media*. Montreal, Canada, pages 92–101.

Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ins Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems* 69:124 – 133.

Gary King, Patrick Lam, and Margaret E Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science* .

April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, New York, NY, USA, WebSci '13, pages 195–204.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'13, pages 1621–1622.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the Eleventh International Conference on Web and Social Media*. Montreal, Canada, pages 608–612.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. pages 145–153.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. pages 6–9.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, pages 1–10.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media*. Cologne, Germany, pages 687–690.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American Psychologist* 62(4):271–286.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. The automated detection of racist discourse in dutch social media. *CLIN Journal* 6:3–20.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015a. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria, pages 672–680.

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Véronique Hoste, and Walter Daelemans. 2015b. Guidelines for the fine-grained analysis of cyberbullying. Technical report, LT3, Ghent University, Belgium.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, LSM '12, pages 19–26.

Zeerak Waseem. 2016a. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, pages 138–142.

Zeerak Waseem. 2016b. *Automatic Hate Speech Detection*. Master's thesis, University of Copenhagen.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the*

NAACL Student Research Workshop. Association for Computational Linguistics, San Diego, California, pages 88–93.

Andreas Wimmer. 2013. *Ethnic boundary making: Institutions, power, networks*. Oxford University Press.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*.

Noname manuscript No.
(will be inserted by the editor)

Bridging the Gaps

Multi-Task Learning for Domain Transfer of Hate Speech Detection

Zeerak Waseem · James Thorne · Joachim Bingel

Received: date / Accepted: date

Abstract Accurately detecting hate speech using supervised classification is dependent on data that is annotated by humans. Attaining high agreement amongst annotators though is difficult due to the subjective nature of the task, and different cultural, geographic and social backgrounds of the annotators. Furthermore, existing datasets capture only single types of hate speech such as sexism or racism; or single demographics such as people living in the United States, which negatively affects the recall when classifying data that are not captured in the training examples. End users of websites where hate speech may occur are exposed to risk of being exposed to explicit content due to the shortcomings in the training of automatic hate speech detection systems where unseen forms of hate speech or hate speech towards unseen groups are not captured.

In this paper, we investigate methods for bridging differences in annotation and data collection of abusive language tweets such as different annotation schemes, labels, or geographic and cultural influences from data sampling. We consider three distinct sets of annotations, namely the annotations provided by [43], [45], and [16]. Specifically, we train a machine learning model using a multi-task learning (MTL) framework, where typically some auxiliary task is learned alongside a main task in order to gain better performance on the latter. Our approach distinguishes itself from most previous work in that we aim to train a model that is robust across data originating from different distributions and labeled under differing annotation guidelines, and that we understand these different datasets as different learning

All authors contributed equally.

Z. Waseem
University of Sheffield E-mail: z.w.butt@sheffield.ac.uk

J. Thorne
University of Sheffield E-mail: j.thorne@sheffield.ac.uk

J. Bingel
University of Copenhagen, E-mail: bingel@di.ku.dk

objectives in the way that classical work in multi-task learning does with different tasks.

Here, we experiment with using fine-grained tags for annotation. Aided by the predictions in our models as well as the baseline models, we seek to show that it is possible to utilize distinct domains for classification as well as showing how cultural contexts influence classifier performance as the datasets we use are collected either exclusively from the U.S. [16] or collected globally with no geographic restriction [43, 45].

Our choice for a multi-task learning set-up is motivated by a number of factors. Most importantly, MTL allows us to share knowledge between two or more objectives, such that we can leverage information encoded in one dataset to better fit another. As shown by [3] and [30], this is particularly promising when the auxiliary task has a more coarse-grained set of labels in comparison to the main task. Another benefit of MTL is that it lets us learn lower-level representations from greater amounts of data when compared to a single-task setup. This, in connection with MTL being known to work as a regularizer, is not only promising when it comes to fitting the training data, but also helps to prevent overfitting, especially when we have to deal with small datasets.

Keywords Multi-Task Learning · Abusive Language Detection · Social Media Analysis · Domain Transfer

1 Introduction

With the growing amount of user-generated content online, issues such as online abuse become more important to tackle as they affect a great number of people. A recent study undertaken by the Pew Research Center found that 73% of online adult users had witnessed online harassment, and 40% had been personally targeted [35]. Given staggering numbers such as these, it is clear that current methods of detecting abuse deployed on internet platforms are not effective in shielding users from witnessing or experiencing these forms of violence or harassment.

Alongside the pressure generated by public outcry [13], multiple government agencies have applied political pressure on social media companies to tackle the threat of online hate speech and abuse [42]. For example: the British Home Office created an action plan to deal with hate crime, in which online hate speech is explicitly mentioned [23]; Germany has introduced a €50 million fine for social media companies systematically failing to remove hate speech within 24 hours [42]; and the European Commission has set released a code of conduct for dealing with hate speech online [19].

As it stands, a vast majority of the moderation of abusive language and hate speech for these platforms is performed by human moderators, in spite of the exposure to online abuse having a profound impact on mental wellbeing [28, 40, 6]. Notably, two previous moderators have sued Microsoft for negligent infliction of emotional distress resulting in post traumatic stress disorder for being tasked with moderating child abuse [28]. Studies have shown the adverse effects of cyberbullying on youth [40] and negative effects on self-esteem of adults that were exposed to online hate speech [6]. Exposing moderation staff to every abusive post reported on an online platform has the potential to cause harm. Not only is it possible to mitigate this risk through detection of the explicit materials using automated means,

automatically detecting and auto-moderating content containing hate speech will limit exposure of offensive materials to the end users, thereby reducing risk the negative consequences. For example, it could be possible to shield children from cyberbullying.

Furthermore, correlations between increases in hate speech online and increases in hate crime have been shown [32]. Thus, not only are online safety and mental health compromised by hate speech, but offline safety can be ensured by being able to detect such increases and alerting authorities of potential risks of increase in hate crime.

1.1 Hate speech detection

Considering the task of detecting hate speech, it is important to recall that word senses may change as the dialect, sociolect, language, and culture changes [36,8]. Currently, computational methods for hate speech detection cannot and do not try to consider the influence of socio-demographic variables. Furthermore, the issue of cultural and sociodemographic influences on the data sample are not considered, nor has the consideration of how to overcome these cultural differences in datasets collection.

The influence of these issues culminate in models that are guaranteed to have poor generalization when they are applied to different socio-demographic or cultural contexts. For example, a model trained on Standard American English (SAE) on a Twitter data sample will be likely to evaluate the use of the *n-word* as being offensive when applied to a sample of tweets that are written in African American Vernacular English (AAVE) in spite of the cultural context and acceptability of the use of the word being completely different as it is likely to primarily be African Americans writing in this dialect.

A further issue that affects generalization is the data sampling and annotation methodology. When considering data samples that are collected from two similar but different cultures, hate speech directed to one particular demographic may contain targeted locutions that may only appear offensive to one community. Distinct sets of annotation guidelines may be generated that are specific to the task [43,45,16,46]. This in turn increases the barrier for combining the datasets and using them to train models on hate speech that can detect multiple forms of abuse.

Finally, given a number of datasets sampled from distinct cultural contexts, a possible approach for inducing a joint model from these might be to concatenate the datasets.[1] However, differences in size between these datasets may lead to a bias for the larger dataset, and by extension a bias for the culture captured within it. In this case, the detection of hate speech would be biased towards the cultural assumptions that this dataset makes. In contrast to simply merging datasets, multi-task learning allows us to differentiate between them while still training a common model that exploits their commonalities. Careful optimization of hyperparameters, e.g. pertaining to model topology or differing learning rates for the individual datasets, further allows us to explicitly control and correct for a potential bias, or to introduce a certain bias if we deem this desirable. We discuss multi-task learning in more detail below.

[1] After re-annotation to unify class labels, if necessary.

1.2 Multi-task learning

In this work, we seek to address the shortcoming of the previous work by considering issues of generalizability of models to accurately classify hate speech and offensive language across datasets and cultural contexts. To tackle these problems, we make use of multi-task learning (MTL), a machine learning framework that seeks to utilize the similarities and subtle differences in annotations and datasets to improve performance on and regularize against another. To the best of our knowledge, this is the first work exploring the utility of multi-task learning for abusive language.

1.2.1 Motivation

Multi-task learning has its origins in the seminal works by Caruana [10, 9] and has since been applied to a wide range of areas in machine learning, including computer vision [21], bio-informatics [37] and numerous subfields of natural language processing [27, 3, 30]. The core idea in multi-task learning is to train a model that generates outputs for several related tasks from a single common input. We contrast this against classical machine learning approaches where typically a model is a function from one input to a single output space.

The rationale behind this idea is that certain information, which is encoded in the training data of some task, may help the model generalize better when learning how to make predictions for another related task. We can draw parallels to intuitions and observations we can make about human learning: whenever we learn a new skill, we build on other skills that we may have gained earlier. For example, when learning a foreign language, we benefit from other languages that we have learned in the past. This benefit is particularly strong when the languages in question are closely related, i.e. when they share a lot of their vocabulary or structure.[2]

From a more theoretical point of view, multi-task learning has the benefit of serving as a regularizer to a certain task which allows models to be constructed that can generalize better to unseen data. More specifically, because we simultaneously optimize parameters for several tasks, the additional information that is encoded in the auxiliary tasks acts as a mechanism which prevents the model from overfitting to the training data and becoming so specific that new data, while from the same domain and general distribution, cannot be modeled well. Previous work [3] also suggests that MTL can help a model escape from local optima, i.e. suboptimal solutions, in which it would get stuck in a single-task scenario. It has also been observed that in sequence-to-sequence architectures, the inductive bias introduced by MTL tends to have strikingly similar effects to an attention mechanism typically found in neural decoders [7], suggesting that MTL helps to focus attention on relevant parts of the input.

Another advantage of MTL that we exploit in this work is the ability to learn from multiple disjoint datasets. This means that we can combine datasets from more or less different tasks without the need for re-annotating the other data so

[2] The fact that in MTL we tend to learn both tasks simultaneously rather than in succession weakens this analogy to some degree. In fact, the simultaneous learning of two languages could actually make learning harder for humans. For a machine, however, the temporal order is less critical given its far superior memory when compared to humans.

that the label spaces are the same. This is because, as explained in Section 3.3, we can alternate between optimizing for different tasks during the training process. A consequence of this is that we can benefit from both an augmented data source while ensuring the model to generalize better across different kinds of input (e.g. tweets which originate from different domains and demographies).

1.2.2 Task choices

While the simultaneous language learning analogy above illustrates an approach to MTL that has received relatively large popularity in natural language processing, one is often only concerned with one particular task while leveraging other tasks to help this process. In such a scenario, it is common practice to distinguish between these as a *primary task* and one or more *auxiliary tasks*.

The relative importance of the tasks that we specify influences some of the design decisions of the modeling and training our data in a multi-task environment.[3] If, for example, we are ultimately only interested in a single task, we obviously only want to optimize our model architecture (and selection of auxiliary tasks) to yield the best possible performance for that primary task. If, however, we are equally interested in good performance across all tasks, our job becomes considerably harder, as we potentially need to find a compromise between performance scores across all tasks.

As discussed in [4], there are two distinct approaches to choosing an auxiliary task in the language processing architecture. The first is to select one or several tasks that are similar in their linguistic annotations to the main task (e.g. to induce better dependency parsing models by also letting the model learn syntactic categories such as parts-of-speech). The second approach is to use some non-linguistic auxiliary task whose annotations encode some signal that could be useful for the main task. A particularly interesting example is that of [27], where eye-tracking data is used to inform sentence compression.

1.2.3 Limits of multi-task learning

Multi-task learning may not always be beneficial in improving the accuracy of a classifier. Besides increasing model complexity and training time, the relation between the tasks and the respective datasets are critical for the success of MTL. Previous studies [30,3,5] have shown systematically that MTL may lead to detrimental performance on the main task compared to training a single-task model, and have explored the conditions under which some task may aid another. While those findings are not always compatible, a common denominator of these studies is that a high entropy in the label distribution of the auxiliary task is beneficial for the main task. In other words, if the auxiliary task has very predictable labels, performance gains on the main task become less likely.

[3] Such choices include the number and width of the hidden layers, input representations, task-specific learning rates, training schedules, among others.

1.3 Utility of multi-task learning for hate speech detection

Training a classifier to detect hate speech in a supervised setting requires training data that has been annotated by humans. Currently available resources (e.g. [43, 45, 16]) only capture types of hate speech, or single geographies meaning that a system to detect hate speech based on these data may not correctly identify hate speech outside of this domain. Generating new training data is expensive and exposes the annotator to explicit content. By applying a multi-task learning framework, we aim to provide a method which can easily be extended and allow for generalization onto unseen forms and targets of hate speech minimizing the cost of generating new datasets.

Considering classification confidence, our approach may be used for a automated content approval system which relies on detecting multiple forms of hate speech and abuse. In such a system, documents which are predicted to be hate speech with a high confidence may be automatically rejected, whereas comments for which the prediction has low confidence may be subject to human moderation. In this way, such a system would allow for human moderators to focus on borderline cases where human cognition and ability to consider context is required, exposing the moderators to explicit materials only when absolutely necessary.

2 Data

In this work, we utilize three previously published datasets for hate speech detection on Twitter data [43, 45, 16]. As [43] and [45] are annotated using the same definition of hate speech and are in fact partially overlapping datasets, we collapse these into a single composite dataset. Below, we give a comprehensive comparison of the three datasets and their annotation methods.

Intersectionality Before we begin with our introduction to the datasets, it is important that we define a key concept: "intersectionality". Intersectionality was originally coined by [14] to describe how multiple forms of oppression may intersect and create new forms of oppression that draw on the intersecting oppressions. One important note is that being on the intersection of several forms oppression is multiplicative of the separate forms of oppression, not additive. This is seen for instance in the near invisibility of the deaths of black women perpetrated by law enforcement contrasted with the deaths of black men at the hands of law enforcement [15].

2.1 Understandings of "Hate Speech"

In this work, we make use of existing definitions of hate speech and offensive language and do not introduce or modify the definitions of these concepts. Rather, we provide a discussion of the annotation methods and the definitions used in the previously published datasets.

The definition of hate speech proposed in [45] (and subsequently in [43]) is an 11-point test whereby a tweet is classified as hate speech if any one of the

test conditions (provided in Figure 1) are met. This test is based on work in the fields of Gender Studies and Critical Race Theory (CRT). Specifically, [45] draw on the work of [31] and [14] to create their test. While intersectionality is not explicitly considered in [45], it is specifically addressed in [43] through the selection of intersectional feminists annotators. In addition, in [43], annotators are asked to select between "racism", "sexism", "neither", and "both" while [45] do not annotate for "both".

Fig. 1 11-point test for hate speech provided by Waseem and Hovy.

A tweet is offensive if it

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

In the criteria for a tweet to be annotated as hate speech in [45] and [43], we observe there are three different groups of tests (see Table 1).

Table 1 Types of hate speech in annotation guidelines of [45,43].

Group description	**Test numbers**
Overt aggression	1, 2, 9, 11
Defense/Support of hate speech	5, 8, 10
Subversive aggression	3, 4, 6, 7

Considering the guidelines presented in Figure 1 and the categorization in Table 1 in further detail, it is apparent that the aim of these guidelines was to

capture a broad spectrum of the hostile experiences that oppressed groups in society face. We can visualize a quadrant describing the types of abuse these guidelines capture. Along one axis, abuse can range from explicit to implicit. And the second axis, abuse can range from directed to generalized [44]. These two datasets [43,45] attempt to capture both explicit and implicit hate speech that can be either directed or generalized.

In consideration of hate speech, offensive language, and more generally abusive language, it is important to note that the use of slurs and profanity may not be indicators of abuse. For instance [36] argues that while the *n-word* is considered an offensive term in many contexts, it is not an offensive term when used within the African American community, instead it can function as a way of communicating solidarity and framing oneself within the historical context of the oppression of African Americans in the United States, and as [36] writes:

> "Using nigga to address and refer can contribute to the construction of a speakers identity, but as in the segment above, it can also ascribe identity (Coupland 2007) to a referent or addressee as a coparticipant in the diaspora."

Waseem and Hovy's annotation method does not explicitly afford context dependent annotation, seen through test 1. As such, any use of the *n-word* may be annotated as hate speech.

In comparison, [16] employ a different definition, in which they move away from the categories of sexism and racism and employ the term "target group", which suggests a move away from the literature in Gender studies and CRT. Further, they more clearly move away from the literature by basing their definition of hate speech in the user guidelines of Facebook and Twitter. Thus, they reach a definition that erases the societal context within which hate speech occurs and those who are most frequently targets of it:

> "language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."

Using this definition, [16] ask their annotators to distinguish between "hate speech", "offensive", and "neither". They allow for distinguishing between these by instructing their annotators to take context in which the message was sent into account and explicitly state that the use of profanity or slurs does not necessarily indicate hate speech, it may simply be offensive depending on the context. Thus they seek to reintroduce *a context* after erasing a societal context from their definition.[4] Considering the case of the *n-word* and AAVE, this annotation method allows for it not to be tagged as hate speech:

> "Users were asked to think not just about the words appearing in a given tweet but about the context in which they were used. They were instructed that the presence of a particular word, however offensive, did not necessarily indicate a tweet is hate speech"

[4] Context is not defined more clearly in their paper.

Thus suggesting that while the use of a term may not be hate speech, it is still offensive and as such the *n-word* may still be flagged as offensive when used within the African-American community. Interestingly, [16] find that annotators tend to regard homophobic and racist language more likely to be hate speech whereas sexist language is more often flagged as offensive.

In this work, we do not distinguish between the two definitions and annotation methods as our aim is to investigate methods for domain adaptation from one datasets onto the other.

2.2 Commonalities & Differences

Here we give a brief overview of several of the commonalities and differences that are found amongst the three utilized datasets.

In [43] and [45], the same annotation guidelines are used. However, there are also key differences to be found between the two datatsets: in [45], two annotators label the datasets, whereas in [43] the datasets is annotated by a group of activist intersectional feminists and another set of annotations is obtained by crowdsourcing the annotation efforts on CrowdFlower. The annotations from [43] that we employ are the feminist annotations. In [16] the annotations are similarly crowdsourced on CrowdFlower.

All three datasets are collected from Twitter. However, while [16] collect tweets written within the United States of America, [45] and [43] do not limit by geographic location. To mitigate the different geographical (and thereby cultural) biases that arise from the different sampling of these datasets is one of the contributions we seek to make with our work

One of the key differences between the two definitions of hate speech is its positioning of within societal structures. By basing their test in Gender Studies and CRT, [45] implicitly place their work within the notion of structural inequality, By using charged terms such as "sexist and racial slur" and "attacks a minority", they explicitly frame their work within the context of abuse not being equally distributed amongst all groups.

On the other hand, [16] do not frame their work within this context nor do they base their definitions in the previous literature. Given that they base their definition in the guidelines of social media companies, it is based in law, as their guidelines are placed within the context of corporations that seek to react to a user base that highlights their discomfort on their platform while simultaneously navigating the legal realities of multiple nations. One such reality is that, within the U.S.A. anti-subordination is a complicated area to navigate, and for a corporation it is unnecessary to do so when it is possible to frame within an anti-discrimination context.

Beyond these differences, another difference occurs in the targets within the "hate speech" and "racism" classes. As previously noted, annotators were more likely to find "hate speech" to be racist or homophobic speech, while sexist speech was more likely to be "offensive" [16]. Thus considering the targets of racism between the datasets, the main targets of racism in [43] and [45] are Muslims, whereas the main targets of racism in [16] are African Americans.

Finally, the definitions, labels, and the annotation scheme differ slightly as covered in Section 2.1.

3 Model

We use a deep multi-task model to transfer knowledge between different tasks. While a number of different approaches to MTL have been explored in the past, the paradigm that has attracted most attention in deep learning and natural language processing (NLP) in particular is *hard parameter sharing*. As its name suggests, this MTL paradigm works by sharing a subset of a model's parameters between different tasks. From a different but equivalent perspective, this is building distinct models for each task, with these models sharing (and jointly optimizing) some of their parameters.

We will compare the performance of these MTL models against simple baseline models without hard parameter sharing. We first introduce and define the multi-layer perceptron feed forward neural network and then discuss modifications that allow hard sharing of parameters

3.1 Baseline Model Definition

We build a very simple feed-forward neural network without any parameter sharing. This model which takes as its input some fixed-size representation x of a tweet and computes a hidden latent representation h, which is a linear projection of x using a matrix of weights W_0 and a bias term b_0, followed by a non-linear transformation:

$$h_0 = \tanh(xW_0 + b_0) \tag{1}$$

For a deeper model, further hidden representations h_l are computed accordingly by stacking these layers. The respective previous hidden layer output h_{l-1} is provided as the inputs to following layer.

$$h_l = \tanh(h_{l-1}W_l + b_l) \tag{2}$$

The final hidden representation h_L is then used to compute the model output:

$$y = \sigma(h_L W_{out} + b_{out}) \tag{3}$$

Typically, σ is the softmax function which, for a k-dimensional input, normalizes the output to the range $[0, 1]$ such that its sum is 1, representing a categorical distribution over outputs.

$$\sigma(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{4}$$

3.2 Multi-task Model Definition

In Equation 3, the model uses the parameters W_{out} and b_{out} to predict outputs for a single classification task. This model can be extended to predict outputs for more than one task, with the use of hard parameter sharing through the introduction of additional parameters that are specific for each task t:

$$y_t = \sigma(h_L W_t + b_t) \tag{5}$$

In this setup, the weights W_l and bias terms b_l are thus the parameters that are shared between the tasks and learned jointly. In constrast, the weights W_t and bias terms b_t at the output layer are specific to only one task.[5]

3.3 Training

Our goal is to learn parameters for the model in a supervised learning scenario using labeled training data. We assume that training data is provided as set labeled pairs of instances x_i and labels y_i: $\{(x_i, y_i)\}_{i=1}^{N}$. Training the neural network is the process of optimizing the parameters with the objective of reducing the error rate of predicted outputs $\hat{y}_i$ with respect to the annotated labels y_i.

The common way to optimize a deep learning model is via the so-called forward-backward algorithm, where we first compute some guess that the model produces for a given input example (following Eq. 5) and then compare this to the ground truth that is annotated in our training data. This comparison is a quantity that measures the error (or *loss*), which we then use to inform our model how strongly it should change its parameters during the backward pass in order to arrive at a better guess in the next iteration. This is typically done using some flavor of gradient descent.

A challenge that multi-task learning now poses in comparison to a classical single-task scenario is that we do not have a single loss that we can use to optimize our model, but one for every task. This raises the question of how to schedule the training across the different tasks. A common technique is to flip a coin at every training iteration (i.e. for every forward-backward pass) that decides for which task we are going to train. Depending on the outcome, we then sample a batch of training data for this task and optimize the parameters that are involved in predicting the respective output for this task. An obvious alternative to flipping a coin would be a strictly defined schedule that alternates between the tasks at every iteration, or a biased coin or schedule that gives preference to some tasks over others.

In this work, we select the coin-flipping strategy and sample training data for a task that we choose randomly with equal probability. We also follow standard practice in employing a dropout regularizer on each hidden layer during training, where we randomly set units in the hidden representations computed in Eq. 1 to zero with 0.2 probability.

[5] Note that in principle, hard parameter sharing also allows us to predict the different tasks at different depths of the model, e.g. to compute the output for task A from some hidden representation h_m and task B from h_n (with $m \neq n$). Yet another possible variation is to compute further hidden representations that are task-specific and not shared, but ultimately draw on some common lower-level representation.

3.4 Features

Our model utilizes fixed-size representations of the input (thereby differing from more complex deep learning architectures like recurrent or convolutional neural networks). We use two classes of features representation: (1) a Bag-of-Words representation of tweet words, bigrams and character n-grams, and (2) continuous word representations. We perform an evaluation of both in isolation as well as a combination of the two.

3.4.1 Bag-of-Words (BoW) Representation of Features

We construct a vocabulary of words occurring within our corpus of tweets and restrict our Bag-of-Words representation to the 5000 most frequently occurring words to prevent our model from overfitting the Zipf long tail. Each occurrence of a word is modeled as a one-hot vector that is summed for each tweet.[6]

In addition, we collect word pairs (bigrams) and concatenate these into a single word and also add the most frequent 5000 to our vocabulary. For example: "go away" would be added to the vocabulary as the token "go_away". The intuition behind this is that multi-word expressions that have a meaning that is distinct from their constituent words can be more accurately represented as its their own tokens rather than having the meaning diluted by other training examples.

As we are expecting to classify less formal and non-standard uses of English on Twitter, we must also account for word mis-spellings, alterations and colloquial style. For example: the words "yeah", "ye", "yep", "yea" and "yes" all convey similar meaning but would be represented as five distinct tokens using a Bag-of-Words model. We account for this through character segmentation as well as word segmentation. The segment "ye" appears 5 times and (in conjunction with other observed features) convey part of the meaning. We extract character bigrams (character pairs) and character trigrams (groups of three letters) and treat these as words in our vocabulary. Again, we only use the 5000 most common in our vocabulary.

3.4.2 Sub-word Embeddings

Rather than encoding the meaning of a word as a single one-hot vector that is the size of the vocabulary, word embeddings represent the meaning of tokens as low-dimensional real-valued vector. Typically, this vector may be between 100 to 300 dimensions. These dense meaning representations can be designed such that words which convey similar meanings or appear in similar contexts also have similar vector-based representations.

We choose to use embeddings because this allows us to capture from different, but related concepts that occur in our data that cannot so easily be represented with the symbolic Bag-of-Words representation. For example, the encoding of a tweet containing the word 'football' will be entirely different from the a tweet containing the word 'soccer' using a BoW representation - even though these are similar concepts. Using continuous representations enables some cross talk between

[6] A one-hot vector is a binary vector of indicator features that are 1 if that feature occurs in the document otherwise 0 in the feature does not occur in document.

different concepts encountered during training. We hypothesize this may yield a classifier that is less prone to over-fitting the distribution of data that it observed during training and more accurate for unseen out of domain data.

Our multi-layer perceptron models are designed with a fixed input representation size. However, the size of tweets is variable. To train our model, we must make a fixed size representation of a variable size input. While it is normal in text classification problems perform convolutions [26] over the input data or to train train a time-series model such as a Recurrent Neural Network, the limited size of the training data available for this task prevents use from using these techniques. Instead we a perform a pooling operation by averaging all the vectors in the tweet which is shown to yield an acceptable (yet suboptimal) performance on other text classification tasks [41].

Because language on Twitter is informal, we expect to encounter unseen words and variations of known words. Rather than using word representations, we use vector representations of sub-word units similar to morphemes [22] which will allow us to better capture common word units that are occurring in this informal language.

3.5 Pre-processing

We pre-process all tweets with the following steps: usernames and mentions are converted to a single type to aid anonymity and to also prevent bias in the training that may occur by learning associations between usernames rather than language. URLs and Hashtags are filtered out for the same reason. Furthermore, we convert all text to lower case and normalize numbers to a special digit symbol. Finally, all line breaks in tweets are replaced with spaces.

4 Experiments

To test our hypothesis, we construct three experimental configurations which we test out using our models. For our configurations we use the same two datasets described in Section 2, namely the composition of the datasets from [45] and [43], and the dataset from [16]. For all three configurations we test our models using each dataset as the training dataset in turn. Further, we conduct three different experiments with different features for each configuration, a lexical model using BoW, a model using only embeddings, and a model using both BoW and embeddings. In each case, we train our models on a total of 45 iterations over the available training data and finally test it using the parameters which yield the best performance on the held-out development set, preventing overfitting on the training data.

4.1 Baseline models

We construct our baseline models using by training a model on a single datasets and predicting on another as has been attempted in previous work [43]. We select this as our baseline as it has been been attempted in previous work with low

success and therefore will highlight the issues with attempting to predict on one dataset given that a model is trained on another. Consistent with previous work, we expect these models to have poor performance on out of domain data. By using each dataset in turn for training and predicting on the other, we show that it is not simply a questions of which dataset our models are trained on but rather that regardless of which dataset we train on, the capabilities of a model to predict on a dataset which is collected in a different culture, with different ways of using language, and with different targets and topics will be poor unless we specifically seek to address this.

To evaluate the performance of the classifier on the out of domain data, we defined a deterministic class mapping between the two datasets based on observations. We map the "Neither" class from [43, 45] to the "Not Offensive" class in [16]. We also observe that in [16], a large majority of the tweets annotated as offensive language are sexist so we map the "Offensive" class to "Sexist". A large majority of the tweets labeled as hate-speech contain racist slurs and remarks, so we map the "Hate Speech" in [16] class to the "Racism" class in [43, 45].

4.2 Composite data models

In this configuration, we build a composite of all three datasets into a single training set and test set. With this model we seek to build a strong baseline as we expect this will outperform the simple baseline models and simultaneously will allow for us to test the performance of a model where a composition of all known datasets is performed. Additionally, this method allows us to test whether the influence of using a multi-task learning configuration only shows benefits due to the model being exposed to all available datasets. Finally, using the composite datasets we test whether the distribution of documents from each dataset influences which evaluation dataset the model performs best on.

4.3 Multi-task learning models

Our third configuration uses a multi-task learning framework, in which we test for whether simultaneously learning to predict on two different datasets with a shared representation can outperform a strong baseline. Furthermore, by utilizing this approach, we test the potential of domain transfer for abusive language via multi-tasking. Finally, this setup allows us to test whether cultural influences and differences can be utilized such that prediction is improved on a dataset whose collection is based in a different cultural context. Given that it differentiates between the primary and auxiliary tasks while still learning from both datasets, we expect this configuration to outperform the other two model types. Specifically, we expect it to outperform our simple baseline models by a large margin and, to a lesser degree, our composite data models.

We test two conditions for this set of experiments, alternating between which of the two datasets we use as a main task, with the other serving as the auxiliary training data. While our model internally treats both tasks equally, the difference between these two scenarios is that we only tune the model on the development set of the respective main task.

4.4 Dataset statistics

We construct a dataset for "racism"/"sexism" detection by merging the [43] and [45] datasets. In [45], only the classes "racism", "sexism", and "neither" are utilized, however due to the focus on intersectional abuse [43] also annotate for "both". In our experiments we augment the "racism" class with documents labeled as "both" as there are only 49 documents labeled as "both" and because [16] is not annotated for the intersections but rather "hate speech" and "offensive." Therefore, training to detect this composite class, regardless of which dataset is trained on or is used as the primary task, would hardly be successful. Furthermore, the issues with detecting the class would become greater as we create splits in our dataset for training and testing purposes. Finally, we augment the "racism" class with the documents from the "both" class, as "racism" is the class with the fewest documents, and as such increasing the number of documents is more likely to improve performance on the class rather than the "sexism" class which has more documents, even if the increase in number of documents is negligible.

To build our model, we create stratified splits of our dataset to ensure that class balance across different splits remains the same. We generate a split for training our model, a split for development evaluation, and a final evaluation (see Tables 2 and 3 for dataset statistics across the splits) dataset which is entirely unseen for our models at test time.

Table 2 Dataset statistics of the Waseem[43]/Waseem-Hovy[45] and splits produced for training, developing and evaluating the models.

Dataset Split	Racism	Sexism	Neither
Training	1697	3365	11688
Development	211	420	1461
Test	214	423	1461
Total	2122	4208	14610

Table 3 Dataset statistics of the Davidson[16] and splits produced for training, developing and evaluating the models.

Dataset Split	Hate Speech	Offensive	Not Offensive
Training	1144	15352	3330
Development	143	1919	416
Test	143	1919	417
Total	1430	19190	4163

4.5 Evaluation Metrics

Given the high imbalance between positive classes seen in Tables 2 and 3, that is the "racist", "sexist", "offensive", or "hate speech" classes, it is important that we evaluate using metrics that are not susceptible to class imbalances. For instance, a metric that would be susceptible to class imbalances is accuracy, which simply

calculates the fraction of all correct predictions over all documents in the evaluation set. Thus, if one class dominates the dataset, and a classifier performs well on that class but poorly on all other classes, the accuracy score would still show a quite high score. For this reason, we provide precision, recall, and weighted-average F1-scores for each class as well as their average. As such we can show the actual performance on our task, rather than a biased sample. Below we provide definitions and explanations of our metrics.

4.5.1 Precision, Recall, and F1-score

We compute the precision, recall, and F_1-score, and report F_1-scores, as these measures are robust against class imbalance while providing insight into the performance of our models. For all three, in the class-based representation the "positive" class refers to the class which we are predicting for.

Precision describes the fraction of how many of the examples which our model predicted to belong to one of the positive classes actually belonged to those positive classes. Thus, it provides us with a insight of how often other classes are misclassified as this class.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{6}$$

Recall, on the other hand, describes how often our models predicted the correct class as a proportion of all predictions; providing insight how often the classifier misclassifies the this class as another class.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{7}$$

The F_1-score is the harmonic mean between precision and recall which penalizes imbalance between precision and recall.

$$F_1\text{-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

5 Experimental Results

In this section, we present the results of our experiments. Results of all experiments are presented in Table 4. Each subsection will highlight one type of model and analyze and discuss the performance of that model. We will be comparing across datasets with respect to the class distributions. Please refer to Tables 2 and 3 for class distributions.

5.1 Single-task baseline models

Our single-task baseline models are built using the same method that has been used in previous work to predict on out-of-domain data, namely training on an in-domain training set and predicting on an out-of-domain test set [43]. In this we show findings consistent with previous work, namely that in-domain prediction

Table 4 Comparison of test-set performance of within-domain and out-of-domain datasets using models trained only on one dataset (first four rows), models trained by concatenating both datasets (middle two rows), and using both datasets in a multi-task learning environment (final four rows). For each training regime, we compare using Bag-of-Words (BoW), the Average of Subword Embeddings (Emb) and both (B+E) as features for each tweet. Key: (R)acism, (S)exism, (H)ate-Speech, (O)ffensive, (N)either. Datasets: Davidson [16], Waseem[43]/Waseem-Hovy[45] (W/W+H)

Training Objective		**Feats**	**F_1-Scores of Predictions on Test Sets**							
Primary	**Aux**		**W/W+H**				**Davidson**			
			R	S	N	Avg	H	O	N	Avg
W/W+H	-	BoW	0.70	0.65	0.88	0.82	0.00	0.64	0.42	0.57
W/W+H	-	Emb	0.30	0.42	0.85	0.71	0.01	0.04	0.29	0.08
W/W+H	-	B+E	0.00	0.00	0.82	0.57	0.00	0.00	0.29	0.05
Davidson	-	BoW	0.22	0.29	0.69	0.56	0.32	0.94	0.84	0.89
Davidson	-	Emb	0.00	0.32	0.60	0.48	0.19	0.92	0.69	0.84
Davidson	-	B+E	0.25	0.33	0.70	0.58	0.39	0.82	0.94	0.89
Both	-	BoW	0.21	0.54	0.81	0.70	0.20	0.92	0.77	0.86
Both	-	Emb	0.21	0.45	0.76	0.64	0.05	0.90	0.64	0.80
Both	-	B+E	0.17	0.53	0.81	0.69	0.31	0.92	0.77	0.86
W/W+H	Davidson	BoW	0.64	0.63	0.87	0.80	0.39	0.94	0.84	0.89
W/W+H	Davidson	Emb	0.32	0.50	0.84	0.72	0.10	0.91	0.64	0.82
W/W+H	Davidson	B+E	0.51	0.53	0.86	0.75	0.16	0.93	0.78	0.86
Davidson	W/W+H	BoW	0.66	0.62	0.86	0.79	0.37	0.94	0.83	0.89
Davidson	W/W+H	Emb	0.39	0.49	0.84	0.73	0.09	0.91	0.62	0.81
Davidson	W/W+H	B+E	0.60	0.57	0.85	0.77	0.14	0.93	0.78	0.86

performs reliably when using simple features and models such as our MLP with BoW features.

Considering the results of out of domain classification presented in the first six rows in Table 4, we observe that the performance is extremely poor. While the F_1-scores for minority classes performs below chance, the *Offensive* class out of domain the F_1-scores for the majority class are, in some instances, slightly more respectable.

Considering the average performance over all classes, we observe significant drop in F_1-score from the in-domain dataset to the out-of-domain dataset. This baseline shows that performance on out-of-domain datasets will be poor regardless of which single-domain dataset is used as the training set when the datasets have different underlying distributions and label schemata.

5.2 Composite dataset models

With our composite dataset models, we sought to build a strong baseline which used both datasets to allow comparison against our multi-task learning models. In our results, we observe that the performance on the minority classes is around the level of random chance while the performance of majority class is satisfactory. Considering the average F_1 score, these models perform well compared to our single-task baseline models. However, this performance is less than desirable.

We observe that inclusion of the second dataset in training reduces average classification performance of BoW models in comparison to our in-domain baselines which only use a single dataset. While in comparison the in-domain performance

the accuracy is reduced, in comparison to the out-of-domain performance we observe a marked rise. This provides evidence to suggest that while the model will be better at "generalizing" between the multiple datasets, it will do so at the cost of in-domain performance on the distinct datasets from which it is built.

In our single-task baseline, we observe that for the Waseem/Waseem-Hovy data, the Embedding-based features yield poor classification performance. This may be due to data scarcity for the minority classes (racism and sexism). In the composite dataset, we observe an improvement in classification performance for these values when using embedding-based features due to the inclusion of the additional data.

5.3 Multi-task learning models

In all cases, the application of multi-task learning (presented in the final six rows of Table 4) yields clear improvements in the average F_1 classification performance in comparison to the composite dataset as well as to the cross-domain scenarios, outperforming our strong and weak baselines. Notably, these improvements are achieved with minimal loss of performance compared to the in-domain performance of the single task model. We observe four instances where the score was reduced: the average reduction in these cases was 0.025.

The choice of a primary versus auxiliary task appears to have little effect on either test set, which is relatively unsurprising given that the main task choice solely impacts the final model selection criterion rather than training itself.

Our results imply that the MTL approach can overcome the problems that arise from differing annotation schemes for hate speech detection stemming from cultural influences and differences. This poses the central contribution of our work and, extrapolated to a more general case, suggests that the improved generalization that comes with a multi-task learning approach can bridge gaps between different domains and annotation schemes in several other tasks. This is, to our knowledge, an application of multi-task learning that has previously received little attention and is worth exploring further.

5.4 Critiques of Datasets

Referring back to Section 2, we find one troubling aspect of the data released by [16]. While their work is interesting and profound there is a serious issue in their data which we discovered quite late in our process of writing this, and had we been aware of it at an earlier stage we would not have used their datasets. The issue that we found is that a large part of their positive classes consist of African American Vernacular English, and while we encourage research to work on abusive language and AAVE, the combination should be handled with care. As a large majority of the datasets is written in AAVE, we consider the use of the *n-word*. The *n-word* occurs with a 'ga' ending 2167 times. It is labeled as either "offensive" or "hate speech" a total of 2161 times. This includes examples such as:[7] "This Niggah Kevin Hart couldn't sit down lmaooooooooooo My niggah My

[7] Emoticons used in the text are removed, urls are replaced with "<url>" token, and usernames are replaced with "@user".

Niggah", "If I wanted my ex back believe me I'd fucking go get they ass. but I ain't bout to dig through the trash.", and "@user Police just tried to Rodney King a nigga... happen to my nig out here". Considering these examples within the frame of AAVE, it is clear that these are not offensive, nor do they appear to contain other signals of abuse or offensive language, yet they were all labeled as "offensive". We determine that these tweets are in fact AAVE using the references to African American celebrities,[8] the use of phonologically motivated spelling variations and contractions [25], and the reference to the police brutality, including the fact that not only is the user describing the threat of police brutality to themselves, but also referring to someone they know who has been a victim of police brutality from which we illicit is AAVE due to the over-policing of black communities [17]. Considering these factors, some of which common to large sets of the dataset it becomes clear that these examples, as so many other in the dataset, are AAVE.

By training models to detect offensive language and hate speech using this dataset, researchers are implicitly also passing judgment on what is deemed acceptable sociolects and dialects. To seek to control the dialect spoken by communities that are marginalized through over-policing [17], mass incarcerated [38] and under represented in academia [1], media [18], and leadership positions [12] is callous at best and malicious at worst. While it is our contention that this dataset in its current state should not be used in terms of abusive language detection research without re-annotation, we encourage a re-annotation of this resource as it can be a valuable resource into the nature of abusive and offensive language within African American communities. Furthermore, it goes to highlight the argument in [43], that the identities of annotators is important. We find it unlikely that people from marginalized African-American communities would annotate the examples above, or the many other instances in the dataset as offensive or hate speech. Therefore, we encourage for a re-annotation with members of marginalized African American communities as the primary annotators.

We acknowledge that through their instruction for annotators to consider context, the fact that AAVE is so frequently annotated as either "hate speech" or "offensive" directly goes against the intentions of [16] and the instructions they provided their annotators with. This further highlights the importance of selecting the correct annotators for tasks such as abusive language detection.

6 Related Work

6.1 Abusive Language

Abusive language research has seen a recent increase in attention from researchers in NLP [44,43,16,24,11,46,33,39] yet the focus of bridging across geographical context, cultural context, or dataset has to our knowledge only been addressed by [43], [11], and [33]. In [43], they collapse their annotations and the annotations from [45] into a "hate speech" and "not hate speech" classes, and train on [43] and predict on [45]. In [33], they build models using a mixture of lexical features and word embeddings, additionally, they split their dataset up by when the documents

[8] "My n*ggah my n*ggah" is a reference to Denzel Washington's character in the movie Training Day.

were posted and find that by adding data as the model learns improves the performance of the model. Finally, [11] take a very different approach by specifically aiming to make their models function on new datasets. Rather than assigning focus on individual documents, as with a BoW model, [11] choose to approach their task by considering multiple communities, some of which are known to be abusive. Given these abusive and non-abusive communities, they compute the distance of a comment to the communities using a BoW representation of the comment. Using this approach, [11] find that their model outperforms models that are trained within domain using lexical features. One important distinction between [11] and this work, is that [11] requires multiple distinct data sources to perform well. As shown through our use of two datasets, we do not require multiple data sources to obtain generalization.

Other work in the field has dealt with using neural networks for predicting hate speech [20,34,2]. In all three papers, they experiment with Convolutional Neural Networks (CNN). In [34] they model the task slightly differently from other previous works by first building a model to detect whether something is hate speech and then classifying it into the specific form of hate speech ([34] consider "racism", "sexism", and "neither").

6.2 Multi-task learning

As noted above, this is to our knowledge the first work that employs multi-task learning strategies to tackle hate speech detection, aiming at transferring knowledge between domains and differently annotated datasets.

An example of previous work that has used multi-task learning to build models to work well across domains is [47], where sentence representations are learned through auxiliary tasks in order to improve cross-domain sentiment classification. However, this approach critically differs from ours in that the authors do not perceive different annotations as different tasks, but create synthetic data for their auxiliary tasks that are exclusively used to learn better representations of the input.

Another interesting case is that of [29], which proposes a multi-input and multi-output sequence-to-sequence model for neural machine translation that can handle different source and target languages, encoding input from any language into the same language-independent intermediate representation, from which they decode into any available target language. While fundamentally different in model architecture and learning problem, this work shares our idea of perceiving heterogeneous datasets from different 'domains' as separate tasks to build a robust cross-domain model.

7 Conclusion

In this work, we applied the use of multi-task learning to develop classifiers for hate speech and abusive language. We find that utilizing an MTL framework for detecting hate speech allows for vastly improving the ability of a hate speech detection model to generalize to new datasets and distributions. In this work, we specifically chose datasets which were collected with distinct cultural groundings

and bias to examine the utility of MTL to overcome such biases. With this in mind, we show that MTL does in fact allow for generalization onto a different cultural context. A particular strength of our MTL approach is that its better generalization allows for a more robust application to completely novel data. In such a scenario, the outputs from the MTL model could act as a mixture of experts that jointly vote on new data. Prior knowledge could be easily integrated here in giving more weight to the sub-model whose training data we believe is closest to our new data.

Our results further show that MTL allows for comparable results to using single task models that predict in-domain, while also allowing for prediction on other datasets. Additionally, we find that a high performance model can be built using composite datasets however, MTL allows for overall improvements over it. Furthermore, we find that in our experiments the choice of primary and auxiliary task had little influence on the performance of the model. We show that applying MTL to classify hate speech on out of domain data is a vast improvement over single-task models and has a slight average improvement over the composite dataset models.

In a more practical sense, our approach simplifies the construction of broad-domain filters for moderation of content by a classifier to learn from examples from multiple different domains and tasks. This minimizes the barrier of entry for detecting hate speech from different domains and communities and thus mitigates the risk of exposing users to previously unseen forms of online hate speech and abuse. By using the confidence from scores from this approach to only expose moderators to borderline content where absolutely necessary, we can reduce the volume of explicit materials that staff members are exposed to which has the potential to reduce harm.

In conclusion, while our method does not guarantee improvements on in-domain prediction of single-task models, we introduce the use of a method that can allow for lower barriers to training and detecting new forms of hate speech and abusive language. Considering the correlation between online hate speech and hate crime, lowering entry barriers for hate speech and abusive language detection may allow for platforms to more easily protect their users from undue harm online and offline.

8 Future work

Our work raises a number of questions on how to deal with domain adaptation and abusive language. First and foremost, future work should seek to address making improvements on the minority classes. Second, this paper explores multi-task learning for domain adaptation, it could therefore be beneficial to consider other methods for domain adaptation. Additionally, future work could seek to address the use of user information and the use of demographic variables such as age, gender, and income as additional signals for detection of abusive language and hate speech across datasets. As far as our multi-task approach is concerned, future work may investigate relationships between the datasets and how they reflect in optimal hyper-parameters for the network architecture and training. For example, specific task combinations could benefit from a more fine-tuned training schedule or learning rate ratio, or the integration of further task-specific hidden layers.

References

1. Allen, W.R., Epps, E.G., Guillory, E.A., Suh, S.A., Bonous-Hammarth, M.: The black academic: Faculty status among african americans in u.s. higher education. The Journal of Negro Education **69**(1/2), 112–127 (2000). URL http://www.jstor.org/stable/2696268
2. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, pp. 759–760. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017). DOI 10.1145/3041021.3054223. URL https://doi.org/10.1145/3041021.3054223
3. Bingel, J., Søgaard, A.: Identifying beneficial task relations for multi-task learning in deep neural networks. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 164–169. Association for Computational Linguistics, Valencia, Spain (2017). URL http://www.aclweb.org/anthology/E17-2026
4. Bjerva, J.: One model to rule them all: Multitask and multilingual modelling for lexical analysis. arXiv preprint arXiv:1711.01100 (2017)
5. Bjerva, J.: Will my auxiliary tagging task help? estimating auxiliary tasks effectivity in multi-task learning. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden, 131, pp. 216–220. Linköping University Electronic Press (2017)
6. Boeckmann, R.J., Liew, J.: Hate speech: Asian american students justice judgments and psychological responses. Journal of Social Issues **58**(2), 363–381 (2002). DOI 10.1111/1540-4560.00265. URL http://dx.doi.org/10.1111/1540-4560.00265
7. Bollmann, M., Bingel, J., Søgaard, A.: Learning attention for historical text normalization by learning to pronounce. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 332–344 (2017)
8. Boyle, K.: Hate speech–the united states versus the rest of the world. Maine Law Review **53**(2), 487–502 (2001)
9. Caruana, R.: Multitask learning. In: Learning to learn, pp. 95–133. Springer (1998)
10. Caruana, R.A.: Multitask connectionist learning. In: In Proceedings of the 1993 Connectionist Models Summer School. Citeseer (1993)
11. Chandrasekharan, E., Samory, M., Srinivasan, A., Gilbert, E.: The bag of communities: Identifying abusive behavior online with preexisting internet data. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pp. 3175–3187. ACM, New York, NY, USA (2017). DOI 10.1145/3025453.3026018. URL http://doi.acm.org/10.1145/3025453.3026018
12. Cohen, P.N., Huffman, M.L.: Black under-representation in management across u.s. labor markets. The ANNALS of the American Academy of Political and Social Science **609**(1), 181–199 (2007). DOI 10.1177/0002716206296734
13. Crawford, K., Gillespie, T.: What is a flag for? social media reporting tools and the vocabulary of complaint. New Media & Society **18**(3), 410–428 (2014). DOI 10.1177/1461444814543163. URL https://doi.org/10.1177/1461444814543163
14. Crenshaw, K.: Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist eory and antiracist politics. University of Chicago Legal Forum **1989**(1) (1989)
15. Crenshaw, K.: The urgency of intersectionality (2016). URL https://www.ted.com/talks/kimberle_crenshaw_the_urgency_of_intersectionality
16. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of ICWSM (2017)
17. Desmond-Harris, J.: Are black communities overpoliced or underpoliced? both. (2015). URL https://www.vox.com/2015/4/14/8411733/black-community-policing-crime
18. Dixon, T., Linz, D.: Overrepresentation and underrepresentation of african americans and latinos as lawbreakers on television news. Journal of Communication **50**(2), 131–154 (2000). DOI 10.1111/j.1460-2466.2000.tb02845.x. URL http://dx.doi.org/10.1111/j.1460-2466.2000.tb02845.x
19. European Commission: Code of conduct on countering illegal hate speech online. Tech. rep. (2016)
20. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, pp. 85–90. Association for Computational Linguistics (2017). URL http://aclweb.org/anthology/W17-3013

21. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448 (2015)
22. Heinzerling, B., Strube, M.: Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. CoRR **abs/1710.02187** (2017). URL http://arxiv.org/abs/1710.02187
23. Home Office: Action against hate the uk governments plan for tackling hate crime. Tech. rep. (2016)
24. Jha, A., Mamidi, R.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In: Proceedings of the Second Workshop on NLP and Computational Social Science, pp. 7–16. Association for Computational Linguistics (2017). URL http://aclweb.org/anthology/W17-2902
25. Jørgensen, A., Hovy, D., Søgaard, A.: Learning a pos tagger for aave-like language. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1115–1120. Association for Computational Linguistics, San Diego, California (2016). URL http://www.aclweb.org/anthology/N16-1130
26. Kim, Y.: Convolutional neural networks for sentence classification. CoRR **abs/1408.5882** (2014). URL http://arxiv.org/abs/1408.5882
27. Klerke, S., Goldberg, Y., Søgaard, A.: Improving sentence compression by learning to predict gaze. In: Proceedings of NAACL-HLT, pp. 1528–1533 (2016)
28. Levin, S.: Moderators who had to view child abuse content sue microsoft, claiming ptsd (2017)
29. Luong, M.T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114 (2015)
30. Martínez Alonso, H., Plank, B.: When is multitask learning effective? semantic sequence prediction under varying data conditions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 44–53. Association for Computational Linguistics, Valencia, Spain (2017). URL http://www.aclweb.org/anthology/E17-1005
31. McIntosh, P.: White privilege and male privilege: A personal account of coming to see correpondences through work in women's studies (1988)
32. Müller, K., Schwarz, C.: Fanning the flames of hate: Social media and hate crime (2017)
33. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web, WWW '16, pp. 145–153. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016). DOI 10.1145/2872427.2883062. URL http://dx.doi.org/10.1145/2872427.2883062
34. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter. In: Proceedings of the First Workshop on Abusive Language Online, pp. 41–45. Association for Computational Linguistics (2017). URL http://aclweb.org/anthology/W17-3006
35. Pew Research Center: Online harassment (2017). URL http://www.pewinternet.org/2014/10/22/online-harassment/
36. Rahman, J.: The n word: Its history and use in the african american community. Journal of English Linguistics **40**(2), 137–171 (2012). DOI 10.1177/0075424211414807. URL https://doi.org/10.1177/0075424211414807
37. Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., Pande, V.: Massively multitask networks for drug discovery. arXiv preprint arXiv:1502.02072 (2015)
38. Roberts, D.E.: The social and moral cost of mass incarceration in african american communities. Stanford Law Review **56**(5), 1271–1306 (2004)
39. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: M. Beißwenger, M. Wojatzki, T. Zesch (eds.) Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, *Bochumer Linguistische Arbeitsberichte*, vol. 17, pp. 6–9. Bochum (2016)
40. Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., Tippett, N.: Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry **49**(4), 376–385 (2008). DOI 10.1111/j.1469-7610.2007.01846.x. URL http://dx.doi.org/10.1111/j.1469-7610.2007.01846.x
41. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642. Association for Computational Linguistics, Stroudsburg, PA (2013)

42. The Guardian: Germany approves plans to fine social media firms up to €50m (2017)
43. Waseem, Z.: Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 138–142. Association for Computational Linguistics, Austin, Texas (2016). URL http://aclweb.org/anthology/W16-5618
44. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: A typology of abusive language detection subtasks. In: Proceedings of the First Workshop on Abusive Language Online. Association for Computational Linguistics (2017)
45. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop. Association for Computational Linguistics, San Diego, California (2016)
46. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, WWW '17, pp. 1391–1399. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017). DOI 10.1145/3038912.3052591. URL https://doi.org/10.1145/3038912.3052591
47. Yu, J., Jiang, J.: Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. Association for Computational Linguistics (2016)

Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling

Notebook for PAN at CLEF 2017

Adam Poulston, Zeerak Waseem, and Mark Stevenson

Department of Computer Science
University of Sheffield, UK
{arspoulston1, z.w.butt, mark.stevenson}@sheffield.ac.uk

Abstract This paper presents our approach and results for the 2017 PAN Author Profiling Shared Task. Language-specific corpora were provided for four langauges: Spanish, English, Portuguese, and Arabic. Each corpus consisted of tweets authored by a number of Twitter users labeled with their gender and the specific variant of their language which was used in the documents (e.g. Brazilian or European Portuguese). The task was to develop a system to infer the same attributes for unseen Twitter users. Our system employs an ensemble of two probabilistic classifiers: a Logistic regression classifier trained on TF-IDF transformed n–grams and a Gaussian Process classifier trained on word embedding clusters derived for an additional, external corpus of tweets.

1 Introduction

Author profiling is the task of determining the characteristics of the individual who wrote a document. Many different characteristics can be determined (e.g. personal characteristics such as gender, age, personality [19] and socioeconomic indicators [5,13,14,15]) across a variety of media (e.g. written essays, books, blogs and other social media). Despite their potential ethical concerns, author profiling techniques can be a valuable component in various applications, such as bias reduction in predictive models [2] and language-variant adaption in part-of-speech taggers [1].

In this paper, we present our approach to the 2017 edition of the PAN Author Profiling shared task [10,11,16]. A dataset was provided consisting of Twitter users across four languages and their variants. Each user was labeled with a binary gender label (male/female) and the particular variant of their language (e.g. Brazilian vs European Portuguese). The dataset was balanced by both gender and language variant. Given an unseen user (and their native language), the task is to determine their gender and language variant being used.

To predict gender and language variant, we applied an ensemble of probabilistic machine learning classifiers (described in detail in Section 2). First, an external Twitter corpus was acquired and Tweets geo-located within the countries covered in the tasks languages were extracted (except for the Arabic language variants). This corpus was divided into individual languages (Portuguese, English and Spanish) and used to derive Word2Vec word embeddings [7,8] for each language. Then, each set of language

specific word embeddings were clustered using K-Means to derive a set of word to cluster mappings, which can be thought of as roughly analogous to topics in a topic model. The normalised frequency of each word cluster across a user's tweets was used to train a Gaussian Process classifier. Second, a Logistic Regression classifier was then trained using TF-IDF transformed unigram and bigram frequencies. Both classifiers were employed in an ensemble approach by averaging the predicted probabilities for each sample to determine the label.

2 Approach

Our approach combines two probabilistic classifiers trained on distinct feature sets in an ensemble to predict gender and language variant. Two classifiers were applied: a Logistic Regression classifier trained on TF-IDF n–grams (Section 2.1) and a Gaussian Process classifier trained on word cluster frequencies (Section 2.2). For each unseen document, probabilities from both classifiers are taken and averaged, and the highest average probability class is taken as the prediction. Models were trained using the implementations found in scikit-learn [9] unless stated otherwise.

For Arabic data, only the Logistic Regression classifier is applied, as the volume of geo-located Arabic tweets collected was too low to allow for training of robust Word2Vec models for use with the Gaussian Process classifier.

2.1 Logistic regression classifier with TF-IDF n–grams

Word unigram and bigram features were extracted for each training document. The text was tokenised using a Twitter-aware tokeniser [4]; no additional steps were taken to deal with the extra complexities of Arabic text. A list of stop words was not used while deriving n–gram features, instead tokens that appeared in more than 90% of the documents were removed, as this allows for the removal of n–grams common across a language's variants while also removing stop words.

TF-IDF weighting was applied to down-weight n–grams common across the documents and assign a higher weight to n–grams which are rare.

A Logistic Regression classifier was trained for each language using the n–gram features. Logistic Regression was chosen for use with the n–gram features because it has been shown to perform well on similar high-dimensional classification tasks, and produces probabilistic predictions [3].

2.2 Gaussian process classifier with word embedding clusters

We obtained the data for our word embedding clusters from a Twitter Firehose[1] sample collected throughout 2015. We only used tweets that were geo-located in the specific language regions determined by the shared task (see Table 1).

Some language variants were less frequent in the resulting datasets than others, for instance we collected very few tweets from Ireland compared to the U.S.A. Downsampling was used to avoid over representation of the more prevent language variants.

[1] Twitter Firehose has since been discontinued and can no longer be accessed.

Table 1. Countries scraped for each language.

English (F_{en})	**Spanish** (F_{sp})	**Portuguese** (F_{pt})
Australia	Argentina	Brazil
Canada	Chile	Portugal
Great Britain	Colombia	
Ireland	Mexico	
New Zealand	Peru	
United States	Spain	
	Venezuela	

Data for the language variant with the largest volume of documents was reduced so that it contained no more than 10 times number of tweets of the smallest language variant.

Word embeddings For each language dataset (F_{en}, F_{es}, and F_{pt}) were trained using the Word2Vec [7,8] implementation in gensim [18] with Continuous Bag of Words (CBOW), negative sampling, 200 dimensions, and a window size of 10.

We applied K-Means clustering [6] to the word embeddings to derive a set of 100 clusters for each language, in which each word is assigned a cluster based on its nearest cluster in the embedding space. We then computed the frequency distribution of the clusters for every training document, and used them as features to train a Gaussian Process classifier with an RBF kernel [17].

Similar word embedding clusters have been applied with Gaussian Processes to perform other author profiling tasks such as socio-economic status detection [5]; furthermore, the derived clusters are similar to topics derived in a topic model, in that they identify semantically similar groups of words in documents, which we found to perform well in a similar task [12].

3 Results

Table 2 shows the accuracy scores achieved by a Support Vector Machine (SVM) classifier with a linear kernel, trained on the same TF-IDF n–grams described in Section 2.1. We chose this approach as our baseline, as it has been shown to perform well on similar tasks and represent a strong baseline.

Table 2. Baseline accuracy scores for gender and language variant prediction for each language derived from a SVM classifier trained on TF-IDF n–grams.

Target	**Spanish**	**English**	**Portuguese**	**Arabic**
Gender	0.7361	0.7896	0.8263	0.7450
Language variant	0.9532	0.8617	0.9800	0.8150
Joint	0.7007	0.6838	0.8113	0.6275

Table 3. Accuracy scores for gender and language variant prediction for each language as submitted for the PAN: Author Profiling task 2017.

Target	Spanish	English	Portuguese	Arabic
Gender	0.7939	0.7829	0.8388	0.7738
Language variant	0.9368	0.8038	0.9763	0.7975
Joint	0.7471	0.6254	0.8188	0.6356

Table 3 shows the results of our final submitted run for the PAN: Author Profiling task 2017. For Spanish, English and Portuguese the results were attained by applying the ensemble of Logistic Regression and Gaussian Process classifiers described in Section 2; for Arabic only the Logistic regression classifier was applied (Section 2.1). In the rankings for the PAN Author Profiling shared task [16], our approach achieved 7th place out of 22 entries for joint prediction and 6th for gender, exceeding reported baselines. We achieved poorer results for language variant prediction at 9th place, and did not exceed the baseline approach.

3.1 Discussion

In Table 3, we see that the our ensemble performs quite well for identifying language variant or gender individually. For joint prediction our ensemble performs less well, likely due to errors in either gender or language variant prediction propagating through to incorrect joint predictions. Of the three languages the ensemble was applied to, the best performance was observed for Portuguese and the worst for English. Broad topics of interest appear to be effective for the gender prediction problem while individual terms that are unique to specific language variants are more discriminating for language variant prediction.

Similar to our results in a previous PAN: Author Profiling Profiling shared task entry [12], in which LDA topic models were able to improve predictive performance over word n–grams, word embedding clusters improved predictive accuracy for gender classification. For the language variant differentiation task, introducing the word embedding clusters in fact reduced accuracy scores over earlier runs.

Under our current clustering scheme, each term was assumed to be equally as representative of its cluster as each other term; in practise though, certain terms were closer to the centroid in embedding space than others. Prior to submission we had begun experimenting with weighting terms based on their proximity to their closest centroid, and our initial findings were promising. In future work we would like to investigate the effect of weighting terms in more detail.

4 Conclusion

In this notebook, we have shown that by employing an ensemble of classifiers and utilising clusters of word embeddings reasonable results can be achieved. We propose,

that our approach can be improved by weighting the word embedding clusters by the distance to the cluster centroid.

References

1. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: A case study of african-american english pp. 1119–1130 (November 2016)
2. Culotta, A.: Reducing sampling bias in social media data for county health inference. In: Joint Statistical Meetings Proceedings (2014)
3. Freedman, D.A.: Statistical models: theory and practice. cambridge university press (2009)
4. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.a.: Part-of-speech tagging for Twitter: annotation, features, and experiments. Human Language Technologies 2(2), 42–47 (2011)
5. Lampos, V., Aletras, N., Geyti, J.K., Zou, B., Cox, I.J.: Inferring the socioeconomic status of social media users based on behaviour and language (2016)
6. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA. (1967)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
10. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
11. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017)
12. Poulston, A., Stevenson, M., Bontcheva, K.: Topic models and n-gram language models for author profiling-notebook for pan at clef 2015. (2015)
13. Poulston, A., Stevenson, M., Bontcheva, K.: User profiling with geo-located posts and demographic data pp. 43–48 (November 2016)
14. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) pp. 1754–1764 (2015)
15. Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N.: Studying user income through language, behaviour and affect in social media. PloS one 10(9), e0138717 (2015)

16. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (sep 2017)
17. Rasmussen, C.E., Williams, C.K.: Gaussian processes for machine learning, vol. 1. MIT press Cambridge (2006)
18. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010)
19. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE 8(9), e73791 (09 2013)

Appendix B

Applications for Ethical Approval

Zeerak Butt
Registration number: 160260775
Computer Science
Programme: Computer Science (PhD/Computer Sci E FT) - COMR33

Dear Zeerak

PROJECT TITLE: Domain adaptation for abusive language
APPLICATION: Reference Number 017099

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 30/01/2018 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 017099 (dated 03/12/2017).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely

Alice Tucker
Ethics Administrator
Computer Science

Application 017099

Section A: Applicant details

Date application started:
Wed 29 November 2017 at 11:07

First name:
Zeerak

Last name:
Butt

Email:
z.w.butt@sheffield.ac.uk

Programme name:
Computer Science (PhD/Computer Sci E FT) - COMR33

Module name:
COMR33
Last updated:
30/01/2018

Department:
Computer Science

Applying as:
Postgraduate research

Research project title:
Domain adaptation for abusive language

Similar applications:
Multi-task learning for hate speech detection;

Section B: Basic information

Supervisor

Name	Email
Kalina Bontcheva	k.bontcheva@sheffield.ac.uk

Proposed project duration

Start date (of data collection):
Wed 20 December 2017

Anticipated end date (of project)
Wed 5 February 2020

3: URMS number (where applicable)

URMS number
- not entered -

Suitability

Takes place outside UK?
No

Involves NHS?
No

Human-interventional study?
No

ESRC funded?
No

Likely to lead to publication in a peer-reviewed journal?
Yes

Led by another UK institution?
No

Involves human tissue?
No

Clinical trial?
No

Social care research?
No

Involves adults who lack the capacity to consent?
No

Involves research on groups that are on the Home Office list of 'Proscribed terrorist groups or organisations?
- not entered -

Vulnerabilities

Involves potentially vulnerable participants?
No
Involves potentially highly sensitive topics?
Yes

Section C: Summary of research

1. Aims & Objectives

Reports of hate crime on and offline have increased in the U.K. and U.S. following the Brexit referendum with a 29% percent increase in recorded hate crimes from 2015/2016 to 2016/2017 (O'Neill, 2017). In addition, online abuse has also become a focal point for high profile initiatives such as Prince William's initiative to counter online bullying (Furness, 2017). Finally, online hate crimes are a part of the Government's Hate Crime Action Plan (Home Office, 2016).
Considering an international scope, it has been made clear by the European Union as well as individual member states that online hate speech must be removed and addressed. Specifically, in 2016 the European Commission and a number of technology companies agreed to a code of conduct for the treatment of illegal hate speech online (European Commission, 2016) and Germany imposed 50M Euro fines on social media companies for systematically failing to remove illegal hate speech online (The Guardian, 2017).

Some of the main issues with abusive language research are related to the question of data set construction and usability to closely related tasks. In this project will seek to improve upon methods for abusive language detection. Specifically we will be aiming to consider methods for abusive language research that can allow for the reuse of data sets created for disparate tasks (i.e. cyberbullying, hate speech detection, toxicity detection) to allow for the reuse of data from semantically similar tasks.

This project aims to expand upon existing methods to incorporate an intersectional feminist methodology (McIntosh, 1988, Crenshaw, 1989) to the area of abusive langauge research. We will aim to apply domain transfer methods to overcome annotation gaps in abusive langauge research, as data sets are often annotated either for bullying, hate speech, or toxicity and models to detect hate speech perform poorly even when shifting shifting domain between different forms of hate speech (Waseem, 2016). By considering methods for models that can shift domains, we allow for building models that seek to take advantage of the commonalities of distinct forms of abuse (Waseem et al., 2017).

The main research areas and questions we will seek to explore are:

- How can machine learning methods for domain transfer be applied to various forms of abusive language detection tasks?
- Design and develop methodology for applying intersectional feminist methodology to computational abusive language research.

2. Methodology

The abusive language data will be obtained by using previously released data sets by Waseem (2016), Waseem and Hovy, (2016), Davidson et al. (2017), and Wulczyn et al. (2017). In addition, publicly available data sets from Reddit communities that are known to be toxic will be utilised to train a model to recognise potentially offensive tweets.

We will address the issue of domain adaptation by treating each data set as being distinct and applying machine learning methods such as self-training, co-training, ensemble methods, multi-task learning and joint-learning. We will work with both neural networks and linear models.

3. Personal Safety

Raises personal safety issues? No

There are no personal safety issues as there are no human participants involved in this project.

Section D: About the participants

1. Potential Participants

As we will use previously published data sets, we will not be identifying any new participants.

2. Recruiting Potential Participants

Recruitment will not be necessary

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by CiCS? No

- not entered -

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) No

Given that we will not be recruiting human participants, no informed consent needs to be collected.

4. Payment

Will financial/in kind payments be offered to participants? No

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

There are no harms to the participants.

How will this be managed to ensure appropriate protection and well-being of the participants?

To ensure that there will be no harm to participants, we will ensure that all data is stored on encrypted devices in efforts to further minimise any risk to the participants.

Section E: About the data

1. Data Confidentiality Measures

Data access will be restricted to the researchers (Zeerak Waseem) and the supervisors (Kalina Bontcheva, Andreas Vlachos). Further, all data will be stored on encrypted and password protected devices.

2. Data Storage

Only the researchers involved with the study will have access to the data, which will be stored in password protected encrypted folders. The data will be analysed by the research team all of whom are associated with the university of Sheffield. The project will take place at the university of Sheffield.
The data generated by the project will not be stored after the end of the project and will not be made available for future research projects. The data will be deleted at the end of the project.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
No

Consent forms relevant to project?
No

Additional Documentation

External Documentation

References:

(Crenshaw, 1989) Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 1989(1).
(Davidson et al., 2017) Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of ICWSM.
(European Commission, 2016) European Commission (2016). Code of conduct on countering illegal hate speech online. Technical report.
(Furness, 2017) Furness, H. (2017). Prince william launches anti-bullying plan to combat 'banter escalation scenarios'. Last accessed Nov. 20, 2017. The Telegraph.
(Home Office, 2016) Home Office (2016). Action against hate the uk government's plan for tackling hate crime. Technical report.
(Levin, 2017) Levin, S. (2017). Moderators who had to view child abuse content sue microsoft, claiming ptsd.The Guardian
(McIntosh, 1988) McIntosh, P. (1988). White privilege and male privilege: A personal account of coming to see correpondences through work in women,Äôs studies.
(O'Neill, 2017) O'Neill, A. (2017). Hate crime, england and wales, 2016/17. Report.
(The Guardian, 2017) The Guardian (2017). Germany approves plans to fine social media firms up to €50m.
(Waseem, 2016) Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, pages 138-142, Austin, Texas. Association for Computational Linguistics.
(Waseem and Hovy, 2016) Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, California. Association for Computational Linguistics.
(Wulczyn et al., 2017) Wulczyn, E., Thain, N., Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale (to appear in Proceedings of the 26th International Conference on World Wide Web – WWW 2017).

Section G: Declaration

Signed by:
Zeerak Butt
Date signed:
Sun 3 December 2017 at 23:38

Offical notes

- not entered -

Approved: 10/04/2018

Zeerak Butt
Registration number: 160260775
Computer Science
Programme: Computer Science (PhD/Computer Sci E FT) - COMR33

Dear Zeerak

PROJECT TITLE: Automatic Annotation of Abusive Language
APPLICATION: Reference Number 017306

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 10/04/2018 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 017306 (dated 27/12/2017).

The following optional amendments were suggested:

Please, address the following: Some of the jargon in the "Aims & objectives" is too specific and difficult to grasp for non experts. More specifically, please, amend the research objectives and fix some of the minor typos present in the application.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely

Alice Tucker
Ethics Administrator
Computer Science

Application 017306

Section A: Applicant details

Date application started:
Tue 12 December 2017 at 16:34

First name:
Zeerak

Last name:
Butt

Email:
z.w.butt@sheffield.ac.uk

Programme name:
Computer Science (PhD/Computer Sci E FT) - COMR33

Module name:
COMR33
Last updated:
10/04/2018

Department:
Computer Science

Applying as:
Postgraduate research

Research project title:
Automatic Annotation of Abusive Language

Similar applications:
Domain Adaptation for Abusive Language, Multi-task Learning for Hate Speech Detection

Section B: Basic information

Supervisor

Name	Email
Kalina Bontcheva	k.bontcheva@sheffield.ac.uk

Proposed project duration

Start date (of data collection):
Sat 7 April 2018

Anticipated end date (of project)
Wed 12 February 2020

3: URMS number (where applicable)

URMS number
- not entered -

Suitability

Takes place outside UK?
No

Involves NHS?
No

Human-interventional study?
No

ESRC funded?
No

Likely to lead to publication in a peer-reviewed journal?
Yes

Led by another UK institution?
No

Involves human tissue?
No

Clinical trial?
No

Social care research?
No

Involves adults who lack the capacity to consent?
No

Involves research on groups that are on the Home Office list of 'Proscribed terrorist groups or organisations?
- not entered -

Vulnerabilities

Involves potentially vulnerable participants?
No
Involves potentially highly sensitive topics?
Yes

Section C: Summary of research

1. Aims & Objectives

Content moderation, and in particular abusive language and hate speech on social media platforms has recently received a growth in media attention [Flynn, 2017], political attention [Furness, 2017, Home Office, 2016], and research attention in NLP [Waseem, 2016, Waseem and Hovy, 2016, Davidson et al., 2017, Wulczyn et al., 2017]. To combat inappropriate content, social media platform employ either manual moderation [Chen, 2012] or a mixture of manual and automated moderation [Bogle, 2016]. Given the volume of data published on a daily basis, with over 300M tweets per day [Twitter Inc, 2012], it is necessary to find methods for dealing with content moderation at scale.

Considering an international scope, it has been made clear by the European Union as well as individual member states that online khate speech must be removed and addressed. Specifically, in 2016 the European Commission and a number of technology companies agreed to a code of conduct for the treatment of illegal hate speech online [European Commission, 2016] and Germany imposed e50M fines on social media companies for systematically failing to remove illegal hate speech online [The Guardian, 2017].

Some of the main issues with abusive language research are related to the question of data set construction and usability to closely related tasks. In this project will seek to improve upon methods for abusive language detection. Specifically, we will be considering methods for abusive language research that can allow for computationally gathering data sets with a minimised need for human annotation, thus limiting adverse mental health risks associated with annotating large corpora of abusive language.

This project aims to expand upon existing methods and data sets to incorpo- rate an intersectional feminist methodology [McIntosh, 1988, Crenshaw, 1989] to the area of abusive language research to show that considering societal privilege and oppression can aid in the collection of data sets can have a profound analysis and prediction of abuse and allow for a richer analysis and understanding of the abuse which can allow for policy teams in industry and government to make decisions on a more informed platform.

Research Questions
The research questions we will seek to address in this project pertain the interaction of privilege and different forms of oppression.
1. Design and develop methodology for applying intersectional feminist methodology to automatic gathering of abusive language data.
2. Explore the application of natural language processing and machine learning methods to collect data sets of abusive comments.

2. Methodology

Initially, we will apply the previous released data sets [Waseem, 2016, Waseem and Hovy, 2016, Davidson et al., 2017, Wulczyn et al., 2017] as well as public data sets from Reddit. In addition, we will collect tweets around hashtags that see a great deal of abusive comments such as #whitepride, #blacklivesmatter, #religionofpeace, #feminazi, #feminism, outright slurs such as “nigger”, “kike”, and “paki”, and “raghead”, and finally code words employed by the extreme right, i.e. “googles”, “skypes”, and “skittles” referring to black, jewish, and muslim people respectively. Using such seed words to collect tweets, we will gather all tweets by all of the users. Henceforth, user tweets will refer to all tweets from a given user. Using the previously previously published and public data sets, we will use the positive classes to measure distances to each tweet written by each user. Initially focusing on high recall, if a single tweet by a user has a small distance to the labelled positive classes, all tweets are labeled as abusive. From this point, a triage system using multiple natural langauge processing tools to increase precision, selecting only the tweets that are hate speech. This will, amongst other methods, include using the same distance metric as previously used, sentiment analysis, sentiment towards individuals or demographic targets (cross-referenced with legally protected classes), examine for co-occurrences of terms in our positive classes and each tweet.

3. Personal Safety

Raises personal safety issues? No

- not entered -

Section D: About the participants

1. Potential Participants

Participants will be identified using keywords that generate a great deal of hate speech (e.g. #whitepride, #blacklivesmatter, #religionofpeace, #feminazi, #feminism), outright slurs (e.g. “nigger”, “kike”, and “paki”, and “raghead”), and finally code words employed by the extreme right, (e.g. “googles”, “skypes”, and “skittles” referring to black, jewish, and muslim people respectively).

2. Recruiting Potential Participants

Participants will be recruited using Twitter's API to collect tweets of users that use the aforementioned keywords.

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by CiCS? Yes

We will be recruiting using the university mailing lists to ensure that volunteers are physically located in Sheffield. We choose to do this as being exposed to abusive comments can have strains on mental health and by ensuring physical proximity, we can ensure that our volunteers can be given mental health counselling should they require it.

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) No

Informed consent will not be sought as collecting data from social media does not require informed consent. In addition, our aim is to observe behaviour and contacting users may alter their communication patterns.
Further, no tweets from protected accounts will be collected, as these explicitly re- strict access and communicate that they do not give consent for their use.

4. Payment

Will financial/in kind payments be offered to participants? No

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

There are no harms to the participants.

How will this be managed to ensure appropriate protection and well-being of the participants?

To ensure that there will be no harm to participants, we will ensure that all data is stored on encrypted devices in efforts to further

minimise any risk to the participants.

Section E: About the data

1. Data Confidentiality Measures

We will comply to the Data Protection Act (DPA) to further ensure safety and privacy of participants. Furthermore, data access will be restricted to the researchers (Zeerak Waseem) and the supervisors (Kalina Bontcheva, Andreas Vlachos). Further, all data will be stored on encrypted and password protected devices.

2. Data Storage

Only the researchers involved with the study will have access to the data, which will be stored in password protected encrypted folders. The data will be analysed by the research team all of whom are associated with the university of Sheffield. The project will take place at the university of Sheffield.
The data generated by the project will not be stored after the end of the project and will not be made available for future research projects. The data will be deleted at the end of the project.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
No

Consent forms relevant to project?
No

Additional Documentation

External Documentation

References

[Bogle, 2016] Bogle, A. (2016). Instagram is rolling out its tool to filger offensive comments to all users. Last accessed, Dec. 12.
[Chen, 2012] Chen, A. (2012). Inside facebook's outsourced anti-porn and gore brigade, where 'camel toes' are more offensive than 'crushed heads'.
[Crenshaw, 1989] Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist eory and antiracist politics. University of Chicago Legal Forum, 1989(1).
[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive lan- guage. In Proceedings of ICWSM.
[European Commission, 2016] European Commission (2016). Code of conduct on countering illegal hate speech online. Technical report.
[Flynn, 2017] Flynn, K. (2017). Why celebrities leave twitter. Last accessed, Dec. 12.
[Furness, 2017] Furness, H. (2017). Prince william launches anti-bullying plan to combat 'banter escalation scenarios'. Last accessed Nov. 20, 2017.
[Home Office, 2016] Home Office (2016). Action against hate the uk govern- ment's plan for tackling hate crime. Technical report.
[McIntosh, 1988] McIntosh, P. (1988). White privilege and male privilege: A personal account of coming to see correpondences through work in women's studies.
[The Guardian, 2017] The Guardian (2017). Germany approves plans to fine social media firms up to e50m.
[Twitter Inc, 2012] Twitter Inc (2012). Twitter turns six. Last Accessed Dec. 12.
[Waseem, 2016] Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, pages 138–142, Austin, Texas. Association for Computational Linguistics.
[Waseem and Hovy, 2016] Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, Califor- nia. Association for Computational Linguistics.
[Wulczyn et al., 2017] Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In Proceedings of the 26th Inter- national Conference on World Wide Web, WWW '17, pages 1391–1399, Re- public and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Section G: Declaration

Signed by:
Zeerak Waseem
Date signed:
Wed 27 December 2017 at 22:44

Offical notes

- not entered -