Detection of Social Biases in Hate Speech and Offensive Text

Himanshu Gupta, Pushpak Bhattacharyya
Indian Institute of Technology Bombay
Mumbai, India
{himnahsuhg786, pushpakbh}@gmail.com

Abstract

Warning: *This paper has contents which may be offensive, or upsetting however this cannot be avoided owing to the nature of the work.*

Hate speech and offensive texts are examples of damaging online content that target or promote hatred toward a group or individual member based on their actual or perceived features of identification, such as ethnicity, religion, or sexual orientation. Sharing violent and offensive content has had a significant negative impact on society. These hate speech and offensive content generally contains societal biases in them. With the rise of online hate speech, automatic detection of such biases as a natural language processing task is getting popular. However, not much research has been done to detect unintended social bias from these toxic language datasets. This report attempts to summarise what are existing hate speech detection and offensive text detection models are. Then it will reason why hate speech models struggle to generalise, which sums up existing attempts at addressing the main obstacles. Finally, this report introduces a new dataset from an existing toxic language dataset to detect social biases, their categories, and targeted groups in English. The dataset contains instances annotated for five different bias categories, viz., *gender, race/ethnicity, religion, political, and LGBTQ*. We then report baseline performances of both classification tasks on our curated dataset using transformer-based models. The input to the models is English texts which are probably hate speech or toxic texts. The models will then classify these texts into biased or neutral along with bias categories. Model biases and their mitigation are also discussed in detail. Our study motivates a systematic extraction of social bias data from toxic languages.

1 Problem Statement

The movies and television shows we watch, and the books and articles we read, as well as the social media and meetings in which we participate and the people we surround ourselves with, all influence us. We have different perspectives based on our race, gender, ethnicity, religion, sexual orientation, socioeconomic status, nationality, and a whole array of other factors. These perspectives sometimes lead to biases that influence how we see the world, even if we aren't conscious of them. Biases like this have the potential to lead us to make decisions that are neither intelligent nor just. And when these biases are expressed in the form of hate speech and offensive texts, it becomes painful for certain community. While some of these biases are implied, most of the explicit biases can be found in the form of Hate Speech and offensive texts. Use of hate speech not only incites violence but sometimes also leads to societal and political instability. BLM (Black Lives Matter) movement is the consequence of one such bias in America. So, in order to address these biases, we must first identify them.

1.1 Problem of Offensive Texts

Offensive text is something that upsets or embarrasses people because it is rude or insulting (Fortuna and Nunes, 2018). Offensive content/language is used as a broader concept or a type of socially unacceptable discourse. There are also several other related terms like trolling, toxic/obscene/insult + content/language/speech (Davidson et al., 2019a). Also some more specific racist content or misogyny can be treated as offensive (Davidson et al., 2019b).[1] Examples of offensive texts -

1. *IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE.*
2. *@USER Figures! What is wrong with these idiots? Thank God for @USER*

Offensive text detection will be discussed more in chapter 3.

[1] https://www.sciencedirect.com/science/article/pii/S0306457321001333

1.2 Hate Speech in NLP

Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity (Mathew et al., 2020). It poses serious threats to a democratic society, human rights protection, and the rule of law. If left neglected, it can evolve to larger-scale acts of violence and war. Its not the same as using certain profane words in text. A sentence can use profane words and still might not be hate speech (Warner and Hirschberg, 2012). E.g., Oh shit! I forgot to call him.

Examples of Hate Speech:

1. *The jews are again using holohoax as an excuse to spread their agenda . Hilter should have eradicated them.*

2. *We Muslims have no military honour whatsoever we are sub human savages that slaughter unarmed men women and children*

We will study about Hate Speech in more detail in chapter 2 and we will talk about hate speech detection in chapter 4.

1.3 Bias in NLP

Bias is a prejudice in favor or against a person, group or thing that is considered to be unfair (Singh et al., 2022a). There are many kinds of bias like gender bias, racial bias, age bias, etc.. They are seen to have negative consequences on people. They have negative implications for both the group biasing and the recipient. The offender, on the other hand, may suffer less than the recipient. The recipient is subjected to severe consequences. This is due to the fact that they are the ones who are at the receiving end of the emotional distress. Furthermore, social biases have negative consequences since they influence the entire group, not just the individuals involved. Examples of bias in NLP:

1. *We need to be vigilant and look out for radical islamic terrorism, the forced marriages, the forced genital mutilation, the misogamy, and if you are gay watch out as you will be thrown off a building.*

2. *There are lot more blacks killing blacks than cops killing blacks*

2 Motivation

An enormous amount of user-generated material has been produced as a result of the quick adoption of social media platforms by billions of people worldwide. Sadly, these sites are frequently used for online harassment and abuse, which can have major consequences for the victims. As a result, online communities and social media platforms start to have serious concerns about the detection of inappropriate information. Children who were both victims and perpetrators of cyberbullying (Schmidt and Wiegand, 2017), according to a study, were twice as likely to try suicide. This was compared to youth who had not encountered this type of peer harassment. Finding ways to early detect and monitor hate speech in cyberspace is necessary for mitigating such negative impacts before there are significant escalation and spread of negative ideas outside of the boundaries of the internet (Badjatiya et al., 2019). Bias and toxicity have become a grave problem for many communities and have been growing across many languages. Hate speech creates an environment of intimidation, discrimination, and may even incite some real-world violence (Halder et al., 2020). Both researchers and social media platforms have been focused on developing models to detect bias and hate speech in online communication for a while now. These biases can be expressed in different ways. In recent years social media has emerged as a go to platform for Hate speech and offensive texts containing social biases and prejudice. Before Social media, movies and TV shows were the major sources of bias and hate speech. To mitigate the bad impacts of biases and hate speech, it is necessary to discover and monitor them early on, before large escalations and the spread of unfavorable ideas outside the internet happens. Manual identification of hate speech is judged ineffective due to the vast number of internet users and the enormous volume of online content. Consequently, it is essential that objectionable and profane language is automatically detected and removed in online situations. In all areas of text processing, including text translation, natural language modelling, and sentiment analysis, advances in machine learning and natural language processing, specifically speaking transformer-based models, demonstrated exceptional results. But what's tough is identifying such biases from profane languages. Several models have been proposed to detect hate speech from toxicity automatically but they all suffer from

their own biases referred as model biases. These model biases leads to low precision and low recall values and thus hamper overall performance of a model. Even if we are somehow able to detect hate speech and toxicity accurately, the next challenging task would be to filter out biases from them. Not all hate speech and offensive texts contain bias. Some of them are personal attacks too. We can't just rely on humans to identify hate speech and bias in a sentence, as it will be too expensive and time consuming. Therefore we believe contributing to improving and comparing different machine learning models to fight such harmful contents is an important and challenging goal.

3 Literature Survey

In this chapter we will discuss about various definitions of *offensive language*, its type and target. A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. We will discuss about some existing approaches to handle it and then finally we will talk about presence of biases in hate speech and offensive text.

3.1 Offensive Text

In this section we will discuss about various definitions of "offensive language", its type and target.

3.1.1 Definition of Offensive Language

Unfortunately, offensive content poses some unique challenges to researchers and practitioners. First and foremost, identifying what constitutes abuse/offensive text is challenging, making it impossible to derive ground truth on which to base further investigation of offensive content. Unlike other sorts of destructive activities, such as spam or malware, this type of conduct is usually controlled by humans rather than bots (Founta et al., 2018). The term "offensive language" describes a broad category of content that includes hate speech, profanity, threats, cyberbully and various ethnic and racial slurs (Kaur et al., 2021). Each of these categories has the potential to be abusive, and they aren't mutually exclusive. There is no universally accepted definition of abuse, and phrases like "harassment", "abusive language", and "damaging speech" are frequently used interchangeably. Because of its frequency and serious effects publicised in the media, online abusive behaviour has gotten a lot of attention in the last few years. Online abuse is linked to low self-esteem, poor academic performance, anxiety, despair, and suicide ideation among teenagers, according to research (Sap et al., 2019). There have been countless examples of youngsters committing suicide around the world as a result of online harassment. All of these facts and allegations have generated concerns about the lack of appropriate alternatives for dealing with occurrences of internet abuse. As a result, social media platforms must be made secure enough for users to avoid being exposed to objectionable content on a frequent basis while also being accessible enough to discuss complicated and controversial themes. It appears that defining objectionable content terms is as difficult as determining what might be offensive to a single person. In traditional annotation systems, a precise definition is also vital from the standpoint of annotator agreement. During the content annotation procedures, it is critical for the dataset suppliers that the concepts are understood in the same way. Table 2.1 contains some of the examples of offensive and not offensive texts from OLID dataset[2]:

3.1.2 Categorization of Offensive Language

(Zampieri et al., 2019) categorized Offensive texts into two types:

- **Targeted Insult (TIN):** Posts containing insult/threat to an individual, a group, or others;
- **Untargeted (UNT):** Posts containing non-targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

Table 2.2 contains some of the examples:

3.1.3 Targets of Offensive Language

(Zampieri et al., 2019) categorized targets of Offensive texts into following types:

- **Individual (IND):** Posts targeting an individual. This can be a famous person, a named individual or an unnamed participant in the conversation. Insults and threats targeted at individuals are often defined as cyberbulling.
- **Group (GRP):** Posts targeting a group of people considered as a unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or other

[2] https://scholar.harvard.edu/malmasi/olid

Text	label
Someone should'veTaken"" this piece of shit to a volcano.	offensive
Liberals are all Kookoo	offensive
I feel like he is better chasing the title	not offensive
Grateful Trump doesn't have a dog in the White House. He is a cruel man.	offensive
Yes you are but I was asking what is it about the movie ?	not offensive

Table 1: Examples of offensive and not offensive tweets

.

Offensive Text	Label
@thecomeback @JABItalia Fuck @APrecourt	UNT
I mean I'm dating to get fucking attention	UNT
Hey @LIRR , you are disgusting.	TIN
@BreFields1 @jonesebonee18 fuck you lol	TIN
@karlsantix You are a complete knob! It's ppl like you who are messing up this country	TIN
If I pull up to yo crib and you offer me cockroach milk you getting yo ass beaten	TIN
@TopSergeant Assuming liberals are unarmed would be a grave mistake by the deplorables.	TIN

Table 2: Types of offensive text.Here OFF-offensive, UNT-untargeted insult, TIN-targeted insult

.

common characteristic. Many of the insults and threats targeted at a group correspond to what is commonly understood as hate speech.

- **Other (OTH):** The target of these offensive posts does not belong to any of the previous two categories (e.g., an organization, a situation, an event, or an issue.)

Table 2.3 contains some of the examples.

3.2 Hate Speech

A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language.

3.2.1 Offensive Text Vs Hate Speech

What constitutes hate speech and when does it differ from offensive language? No formal definition exists. The presence of profane content does not in itself signify hate speech. General profanity is not necessarily targeted towards an individual and may be used for stylistic purposes or emphasis. On the other hand, hate speech may denigrate or threaten an individual or a group of people without the use of any profanities (Malmasi and Zampieri, 2017). A recurrent issue with the majority of previous research is that many of them tend to conflate hate speech and abusive/offensive language. (Davidson et al., 2017) define hate speech as 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. In extreme cases this may also be language that threatens or incites violence, but limiting this definition only to such cases would exclude a large proportion of hate speech. This definition, however, also does not include all instances of offensive language because people frequently use expressions that are very offensive to specific groups but are used in a qualitatively different way. For example, some African Americans use the term n*gga in everyday language online, people quote rap lyrics using terms like h*e and b*tch, and teens use homophobic slurs like f*g while playing video games. Due to the prevalence of such language on social media, this boundary condition is critical for any practical hate speech detection system.

3.2.2 Definition of Hate speech

Hate Speech is a speech that targets disadvantaged social groups in a manner that is potentially harmful to them (Davidson et al., 2017). According to (Fortuna and Nunes, 2018) hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific

Offensive Text	Type
Hey @LIRR , you are disgusting.	OTH
@BreFields1 @jonesebonee18 fuck you lol	IND
@karlsantix You are a complete knob! It's ppl like you who are messing up this country	IND
If I pull up to yo crib and you offer me cockroach milk you getting yo ass beaten	IND
@TopSergeant Assuming liberals are unarmed would be a grave mistake by the deplorables.	GRP

Table 3: Targets of offensive text. IND-Individual, GRP-Group, OTH-Others

.

Source	Definition
Facebook	Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.)[3].
Youtube	Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity.
Twitter	Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease.

Table 4: Hate Speech Definitions

.

characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used. (Schmidt and Wiegand, 2017), (Singh et al., 2022b) provides a slightly different definition in her review, where hate speech is defined as "Commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics." Other definitions of hate speech are provided in the Table 4

(Warner and Hirschberg, 2012) discuss about numerous issues involved in defining what constitutes hate speech, which need to be resolved in order to annotate a corpus and develop a consistent language model. To begin with, merely mentioning or even applauding a group linked to hate crimes does not constitute hate speech. The name "Ku Klux Klan" by itself is not hateful, as it may appear in historical articles, legal documents, or other legitimate communication. Even endorsing the organisation does not imply a verbal attack on another organisation. Similarly, an author's overwhelming pride in his or her own race or group can not qualify as hate speech. While boasting in this manner may appear rude and is likely to be accompanied by hateful language, the definition requires a disparagement of others. For example, the following sentence does not constitute hate speech, even though it uses the word "Aryan".

And then Aryan pride will be true because humility will come easily to Aryans who will all by then have tasted death.

On the other hand, it is believe that arbitrarily labelling someone as a member of a group is frequently hate speech. The author unnecessarily changes bankers and workers with "jew" and "white" in the following example, which conveys hatred.

The next new item is a bumper sticker that reads: "Jew Bankers Get Bailouts, White Workers Get Jewed!" These are only 10 cents each and require a minimum of a $5.00 order

Unnecessarily bringing up a person's race or ethnicity appears to be a means for an author to invoke a well-known, negative stereotype. While derogatory phrases and racial epithets are always hateful language when used with the purpose to damage, there are some circumstances in which they are permissible. Such phrases, for example, may be allowed in a discussion of the words themselves. For example:

Kike is a word often used when trying to offend a jew.

When a speaker from the targeted group uses such words, it can be difficult to classify them without that knowledge. For example:

Shit still happenin and no one is hearin about it, but niggas livin it everyday.

To express communal solidarity, African American authors appear to utilise the "N" word with a specific variant spelling, replacing "er" with "a". Hate speech mentions must be differentiated from such usage.

3.2.3 Hate Speech categories and targets

Hate speech, according to definitions, is directed at groups or individuals based on specific characteristics such as ethnicity, religion, disability, gender identity, age, veteran status, sexual orientation, or other factors. Studies have been carried out with the purpose of characterising online hate speech and determining which groups are more vulnerable. This subsection summarises the main findings from papers categorised as taking a more descriptive approach to the problem of hate speech identification. Racism, sexism, prejudice against refugees, homophobia, and general hate speech all have descriptive articles. Some other categories and targets are mentioned in the Table 6.

- ***Racism:*** The authors of one study wanted to know when hate speech happens and why statements on social media are labelled as racist. They came to the conclusion that the most of the time (86%) it was due to the "presence of derogatory language." "References to traumatic historical circumstances" and "presence of stereotypes or threats" are two more motives. Another study attempted to explain the regional distribution of racist tweets by describing racism across the United States. They used data obtained from Twitter

Text	Label
Migrants are filthy cockroaches that will infect our country	Hate Speech
Don't try to explain-Irish Catholics are just idiots	Hate Speech
People should stop to use the word nigger.	Normal
Refugees! More like rape-fugees!	Hate Speech

Table 5: Hate Speech Examples

.

to describe the frequency of tweets in various states, based on the messages' physical location.

- ***Sexism:*** A fairly rudimentary technique was used in a study on sexism. The Twitter search API was used to collect tweets that contained derogatory terms directed at women. A single researcher retrieved and coded approximately 5,500 tweets using a basic binary model. Despite the study's limitations (many of the tweets were repeating the title or lyrics of popular songs that contained the searched harmful terms), it was nevertheless useful in learning that offensive speech toward women occurs on Twitter. Misogynistic language on Twitter is also described in a second study. The major findings were that 100,000 instances of the term rape were detected in UK-based Twitter accounts, with approximately 12% of them appearing to be threatening. Furthermore, almost 29% of the rape tweets appeared to use the term in a casual or metaphorical manner. However, women are almost as likely as men to use insulting phrases against women on Twitter, according to one survey.

- ***Prejudice Toward Refugees:*** Another study focused on the annotation of a German dataset for anti-refugee hate speech. The study's major purpose was to highlight the difficulties and obstacles that come with annotating a dataset.

- ***Homophobia:*** In Africa, another study was carried out utilising ethnographic technique. Data was gathered from a variety of sources (e.g., newspapers, websites) to determine that homophobic discourses used arguments relating to abnormality, xenophobia, racism, barbarism, immorality, unpatriotism, heterosexism, anti-Christianity, un-African, animalistic behaviour, inhumane, criminality, pathology, and Satanism.

- ***General Hate Speech:*** Finally, other research takes into account multiple sorts of hate speech at the same time. In one particular case, two social networks (Twitter and Whisper) were crawled with expressions that follow a rigid pattern:

  $$I < intensity >< userintent >< hatetarget > .$$

  One message following this pattern would be "I really hate people." After collecting the messages, the researchers tried to infer the target of hate in the tweets. With this method, they concluded that "race," "behavior," and "physical" were the most hated categories. Finally, a review of FBI data from victims of single-bias hate crime occurrences in the United States in 2015 revealed that the offender's bias was toward different targets in varying amounts.

3.3 Bias Detection in Hate Speech and Offensive Text

While a lot of the research has been done to detect and mitigate model biases, very little research has been done to detect social biases and stereotypes in Hate Speech and offensive text. In (Davidson et al., 2019b) only racial bias in five different sets of Twitter data annotated for hate speech and abusive language are studied. While similar studies have been done in (Sap et al., 2020), but here again bias category detection was framed as generation task as opposed to our classification task. In this section we will define social biases and stereotypes, their types and about communities who are victims of these biases in detail.

3.3.1 Social Bias

People frequently hold prejudices, stereotypes, and discrimination against those outside their own social group. Positive and negative social bias refers to a preference for or against persons or groups

Categories	Targets
Race	nigga, black people, white people
Behavior	insecure people, sensitive people
Physical	obese people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people

Table 6: Types and Targets of Hate Speech (Fortuna and Nunes, 2018)

.

based on their social identities (e.g., race, gender, etc.) (Caliskan et al., 2017). When individuals act on their biases, this is considered discrimination. Members of certain social groups (e.g., women, Muslims, transgender people) are more likely to experience discrimination due to living in a society that does not adequately promote equity (Chandra et al., 1981).

- It can be positive or negative
- It can be based on stereotypes
- Bias is an individual preference
  e.g. if you hire an Asian for a job that also has an equally qualified black applicant because you think blacks are not as smart as Asians, this is bias.

3.3.2 Stereotypes

A stereotype is a widely held, simplified, and essentialist belief about a specific group. Groups are often stereotyped on the basis of sex, gender identity, race and ethnicity, nationality, age, socioeconomic status, language, and so forth (Muralidhar, 2021). Stereotypes are deeply embedded within social institutions and wider culture. They are often evident even during the early stages of childhood, influencing and shaping how people interact with each other (Dixon et al., 2018). For example, video game designers designed a game platform for girls in pink because that is what the parents (who purchase the game) perceived their girls wanted[4]. Some characteristic features of a stereotype:

- It happens at group level.
- It is based on ideas and experiences with certain groups and then extended to apply to an entire group.
- It is not fixed and can change over time
- It is an expectation that people might have about every person of a particular group in context.
  e.g. Jew moneylender (this example demonstrates how JEWS, a specific religious group are expected and assumed to be rich and moneylenders or bankers).
  e.g. The sexy mommy was well liked by boys.
- It is used to reduce the processing time while judging people.
- It is grounded in the observations of everyday life and has some degree of truth.
- It may be statistically accurate but not universally valid.
  e.g. “Asians are good in maths” but other people are also good in maths,
  e.g. “African-Americans have greater athletic ability” but there are good athletes from other races also,
  e.g. “English loves their tea” but besides English many other people also love their tea.

3.3.3 Categorization of Social Biases

Social Biases can be categorized (Sap et al., 2020) into many types but for our research purpose we are considering only following 7 bias category:

- **Gender:** Favoritism towards one gender over other. It can be of the following types: Alpha,

[4] https://genderedinnovations.stanford.edu/terms/stereotypes.html

Beta or Sexism. The target term is towards which gender the bias is directed.

- **Occupation:** Unequal treatment at workplace based on gender, race, sex. It can be an economic bias, administrative bias or societal perception type of bias. The target term is identified from the broad category as:
  - Healthcare: doctor, nurse, counsellor
  - Hospitality: hotel manager, chef, bartender, cook, wedding planner etc.
  - Anti-social (Criminals): Gangsters, thieves, con-artist
  - Defence: Military officers
  - Entertainment: actors, dancers, musician , painter
  - Politicians: government office bearers
  - Financial and management services: bankers, traders, managers
  - Entrepreneurs: self-employed, businessman
  - Security: police, watchman, guards
  - Academia and Research: scientist, professor, teacher
  - Sports: players, sports manager, referees, coaches
  - Services: delivery boy, driver
  - worker: daily wage worker
- **Religion:** Bias against individuals on the basis of religion or belief. e.g. Christianity, Islam, Scientology etc. The target term is the religion towards which the bias is directed.
- **Race:** Favouritism for a group of people having common visible physical traits, common origins, language etc. It is related to dialect, color, appearance, regional or societal perception. The target term is towards which race the bias is targeted.
- **Ageism:** Discrimination based on age. It can be seen as societal perception, physical features or behaviour. The target term is towards which age group the bias is targeted.
- **LGBTQ+:** Bias towards LGBTQ community people. It can be due to societal perception or physical appearance. The target term is towards which LGBTQ+ group the bias is targeted.
- **Other:** This is a miscellaneous category marked for all kinds of bias in the above 6 category e.g body shaming, Physically Challenged, Mentally Challenged. The target term is the group towards whom the bias is targeted.

3.3.4 Social Bias Vs Hate Speech

While it may appear that Social Bias and Hate Speech are the same thing at first glance, they are not. The differences between them are quite subtle (Bolukbasi et al., 2016). While hate speech is always associated with negative sentiment, social bias can also be associated with positive sentiments. Bias is favouritism or against, whereas hate speech is a strong negative feeling (Borkan et al., 2019). Not all hate speech is biased, and not all biased speech is hate speech. We'll use the following examples to demonstrate the differences:

- Some comments are merely toxic without containing any social biases in them, e.g.
  *IM FREEEEE!!!! WORST EXPERIENCE OF MY F**K-ING LIFE*
- Toxic comments can be hate speech but not necessarily biased, e.g.
  who gives a shit what this arrogant jew thinks about anything.What good did he do for NYC?
- Some comments are just biased without containing any toxicity or hate speech in them, e.g.
  All Asian people are bad drivers

3.3.5 Counter-Narratives

Despite an extraordinary effort to establish sufficient legal and policy responses to hate content on social media platforms, dealing with hatred online remains a difficult problem (Blodgett et al., 2021), (Bosselut et al., 2019). Censorship and overblocking may be charged if hate speech is dealt with in the traditional fashion of content deletion or user suspension. One alternative option, which has gotten little attention in the research community so far, is to actively fight hate content using counter-narratives (i.e. informed textual responses) (Chung et al., 2019), (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco, 2021). A counter narrative (sometimes called counter-comment or counter-speech) is a response that provides non-negative feedback through fact-bound arguments and is considered as the most effective approach to withstand

hate speech. In fact, it protects the right to free speech and uses credible data to refute stereotypes and false information. It can also help de-escalate the argument and change the minds of haters and onlookers by encouraging the exchange of ideas and mutual understanding. A counter-narrative such as the one in Sentence 2 is a non-negative, appropriate response to Sentence 1, while the one in 3 is not, since it escalates the conversation.

1. I hate Muslims. They should not exist.

2. Muslims are human too. People can choose their own religion.

3. You are truly one stupid backwards thinking idiot to believe negativity about Islam.

4 Summary and Conclusion

In this report, we have discussed about problems of offensive languages and hate speech, and how they lead to social biases towards certain communities. Our main goal was to filter out biases from profane languages. We handled it systematically. We started with offensive text and tried to narrow it down to hate speech and identity attacks to detect biases.

In chapter 2 we did a background study of offensive text, hate speech and bias in detail. We started with various definitions of offensive texts, and their types and targets defined in the literature. We found that most of the offensive languages are prevalent on social media and World Wide Web such as Twitter, Yahoo!, etc. On social media people generally use two kinds of offensive languages, one that is targeted to someone or something and the other without any target i.e. languages used by people out of sheer frustration. Targeted insult was further classified into three categories based on whether it is targeted to an individual, group, or other. We also discussed about a hierarchical method of offensive text annotation, where in step 1 data was annotated at much finer levels of granularity. It was annotated into seven labels in Round 1, they are i) Offensive Languages ii) Abusive Languages iii) Hate Speech iv) Aggressive Behaviour v) Cyberbullying vi) Spam and vii) Normal. Later, in further rounds it was found that most of these labels are correlated to each other, hence they are merged into a single label. Final labels constitute of only 4 classes i) Abusive ii) Hateful iii) Normal and iv) Spam. In experimentation part in chapter 3 we talked about various features used in detecting offensive texts. While textual, semantic and sentiment features are the most popular ones, user and activity based features were also used in some of the studies. After feature extraction, two types of experiments were performed based on publicly available dataset. One was binary class classification in which offensive texts were classified from normal texts using SVM and logistic regression. Precision, recall and F1 score close to 0.9 were reported in both the models. Other experiment was performed to detect types and targets in OLID dataset, where models couldn't perform that well. This happened because of class imbalance and increased number of classes in downstream tasks.

In chapter 4, we discussed about hate speech detection and how it is different from offensive languages. Though the difference is quite subtle, it is worth noting that not all offensive texts constitute hate speech. Its definition varies from one social media to another social media and from one demography to another demography. For example, certain 'N' words are used frequently by African Americans in their daily lives to express solidarity but if the same 'N' words are used by some other community then it will be hate speech. Hate speeches are classified into several categories based on race, religion, gender, etc. Each of these categories has its own targets. For example, blacks and whites could be the targets in the race category. Since most of the authors believe that hate speeches are a subset of offensive texts, most of the publicly available hate speech datasets are labeled along with offensive texts. Certain hate dictionaries and templates are used to collect these datasets from social media and websites. To classify hate speech from normal texts we have used 4 model settings, they are i) Binary Class Classification ii) Multi-Class Classification iii) Hierarchical Classification iv) Multitask Learning. From the error analysis of the results, we concluded that separating hate speeches from offensive texts isn't trivial. Most of the slurs which are used in offensive texts are also used frequently in hate speeches. This induces confusion in Machine Learning models and models fail to classify them correctly.

We discussed social biases detection in Chapter 5. Then we came up with five bias categories and their corresponding targets that were relevant to our research. Some of the Hate Speech datasets such as HateXplain, SBIC, CONAN and Identity Attack dataset are also explored as potential causes of bias. Manual annotation, as well as machine annotation, are being used to enrich Identity Attack data. Several annotation challenges occurred over the course of annotation were also discussed in detail. Following the data preparation and annotation, many machine learning methods for detecting biases are explored. While few-shot learning did not perform well, hierarchical and multi-task learning models did. The Multi-task learning model trained on our annotated Identity Attack dataset was our best model. On error analysis, it was found that bias detection also suffers from model bias just like hate speech detection. Models were latching onto certain identity words (Muslims, Blacks, Whites, etc.) for bias and its category predictions. To reduce model biases, we had to augment our Identity Attack Dataset with CONAN and Multi-target CONAN datasets. We then talked about Hollywood Identity Bias Dataset collated from movie scripts. It has seven bias categories instead of five, and the dataset also has pre-context and post-context unlike Identity Attack dataset which has no-context. We did not observe much improvement in model performances by augmenting this dataset with our Identity Attack dataset, probably due to the fact that both the datasets are collated from two completely different domains. While the biases in hate speech domains are quite explicit, biases present in movie scripts are implicit and generally depends on contexts.

To help machines predict societal biases in profane text, we introduce a hierarchical and systematic study of offensive language and its subset hate speech. Our study combines knowledge about offensiveness, hate speech, and their types and targets. We show that while classifying the offensiveness of statements is easier, current models struggle to separate hate speech from offensive languages because of overlapping slurs/derogatory terms. We also conclude that most of the hate speeches represent societal bias and stereotypes. While this assumption works well in most cases, manual data annotation is needed to make this task more robust and accurate. Machine annotation along with human annotators speeds up the overall annotation process. At the time of hate speech and bias detection, the problem of model bias was also encountered. Merely the presence of certain community words (Muslim, blacks, whites, etc.) makes model to label a comment as hate speech and hence societal bias. This indicates that more sophisticated models are required to detect biases in profane languages. To encounter model biases we took help of counter-narratives and showed that it can reduce model biases to great extent. We also conclude that biases present in hate speech/offensive texts are quite explicit as compared to biases present in Movie scripts, where implicit biases depends on contexts.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. *The World Wide Web Conference on - WWW '19*.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN -

COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019a. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019b. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51:1–30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task aware representation of sentences for generic text classification. In *COLING 2020, 28th International Conference on Computational Linguistics*.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281. AI in Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection.

Deepa Muralidhar. 2021. *Examining Religion Bias in AI Text Generators*, page 273–274. Association for Computing Machinery, New York, NY, USA.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language.

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*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022a. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022b. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues.

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*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.