Survey: Sensitivity and Counter Narratives

Gyana Prakash Beria
IIT Bombay
gyanaberia@cse.iitb.ac.in

Nihar Ranjan Sahoo
IIT Bombay
nihar@cse.iitb.ac.in

Pushpak Bhattacharyya
IIT Bombay
pb@cse.iitb.ac.in

Abstract

Sensitivity is a broad term that encompasses a range of negative behaviors in online communication that can cause harm to individuals and groups. The detection and moderation of sensitivity is a crucial task for natural language processing (NLP) research, given the growing concerns about the impact of online communication on social and political discourse. This paper provides an overview of the major approaches used in the literature to address sensitivity detection. In addition, we review the current research on counter narratives, which aim to counter hate speech and other forms of sensitivity by promoting positive and constructive discourse.

1 Introduction

Sensitivity, a term that encompasses hate speech, offensive language, insults, profanities and other forms of harmful content, has become a pervasive issue in our increasingly digital world. The internet provides a platform for individuals to express their thoughts and opinions, but it has also enabled the spread of harmful content that can have serious consequences, including inciting violence and promoting discrimination. As such, detecting and mitigating sensitivity has become a critical concern for policymakers, online platforms, and communities alike. However, traditional approaches to moderation such as censorship and content removal have been criticized for impeding freedom of speech and expression. To address this issue, counter narrative approaches have emerged as a promising alternative, where the solution to harmful speech is more speech.

In this survey paper, we will delve into the domain of sensitivity, particularly the research conducted in the areas of hate speech and offensive language. We will also discuss the concept of counter narratives, along with the various strategies such as positive tone and humor that are utilized within this domain. We will present various datasets that are popular within our topic. These datasets can be used to train language models to detect sensitivity and generate counter narratives.

2 Motivation

Content moderation is a complex issue, and there are many different approaches that have been taken to address it. Governments frequently implement policies and laws to limit the spread of hate speech and punish those who propagate extremist ideologies, like terrorist propaganda. In extreme cases, some governments have resorted to internet shutdowns or other measures to restrict access to online content. Social Media Platforms themselves may also employ a variety of moderation techniques, including account suspension or termination, the removal of specific posts or comments, and even complete censorship of certain topics or ideas. Although these methods have been widely used, they have not been very effective in combating sensitive texts. These methods result in selective free speech, which can have negative or harmful consequences in the future.((Mathew et al., 2019)) Thus implementing these methods can be a complex and delicate process that requires balancing the right to freedom of expression with the need to maintain social peace and security.

The limitations of traditional methods of sensitivity detection and moderation have created a need for alternative approaches that can address the complexities and nuances of sensitive content. Counter-narratives have emerged as a promising alternative, as they seek to address the root causes of sensitivity by promoting dialogue, empathy, and understanding. By providing alternative narratives that challenge and deconstruct harmful content, counter-narratives can prevent the spread of sensitivity while also promoting free expression and healing any harm caused by such content.

3 Sensitivity

Sensitive content can be defined as any content, be it text, audio, or visual that may offend a person, particularly in relation to religion, race, gender, politics, sexuality, disability, or vulgar language. In simple terms, sensitive content is any content that can create a negative surprise ((Tripathi et al., 2019)). Sensitivity is an umbrella term, and various works have utilized terms such as offensive, hate speech, aggression, abusive, toxicity, and others to annotate their datasets. Table 1 contains some of the definitions of popular terms used in this domain. These hostile dimensions often sound similar at an abstract level (e.g., hate and offensive, aggressive and abusive), and many researchers fail to fully understand what they are trying to annotate and recognize, leading to poor work.

From Table 1, one can observe the similarity between the definitions of some of these terms common in literature. The absence of a shared framework across diverse fields of study allows for personal interpretations, leading to instances where the same linguistic phenomenon can be labelled differently, or vice versa, different phenomena can be labelled under the same name [(Poletto et al., 2021)]. As noted by (Waseem et al., 2017a), this "lack of consensus has resulted in contradictory annotation guidelines, where some messages considered as hate speech by (Waseem and Hovy, 2016) are only considered derogatory and offensive by (Davidson et al., 2017). "

4 Counter Narratives

Counter-narratives can be defined as non-negative fact-based arguments against hate speech [(Chung et al., 2019)]. As the name suggests, it counters offensive and wrong information with credible evidence.

Example: *I hate Muslims. They should not exist.*

Counter-Narrative: *Muslims are human too. People can choose their own religion.* [1]

Counter-narrative is a technique where we counter hate speech with more speech [(Mathew et al., 2019)]. As such it doesn't affect anybody's freedom of expression. Counter-narratives can change the viewpoints of people who are blinded by stereotypes. This can lead to a peaceful exchange of opinion and mutual understanding.

[1] examples taken from (Chung et al., 2019).

4.1 Types Of Counter Narratives

There are many strategies that can be used to counter hateful messages in online media. (Benesch et al., 2016) identifies eight such strategies which are as follows:

1. **Tone:** Tone is the emotional quality that is conveyed by the language used in a sentence. (Benesch et al., 2016) considers the whole spectrum of tone from "hostile", which can make the original hate speaker delete their post, to "positive tone" which creates a gentle environment between people to continue the conversation and de-escalate the situation. Recent works like (Chung et al., 2019) and (Mathew et al., 2019) often use the "Positive tone" category because speech filled with empathy, and kindness is known to have a positive effect in decreasing hostility [(Hangartner et al., 2021)].

2. **Presenting facts to correct misstatements or misperceptions:** counter-narrative which provides factual evidence to correct any misperceptions and prevent the spread of misinformation. This can make the original speaker more informed about an issue.

   example: *Actually homosexuality is natural. Nearly all known species of animal have their gay communities.* [2]

3. **Pointing out hypocrisy or contradictions:** counter-narratives that point out any inconsistencies or hypocrisy in the hate-filled statement. Correcting the statements from hate speakers can prevent the spread misleading informations

   example: *The 'US Pastor' can't accept gays because the Bible says not to be gay. But...he ignores:The thing about eating shrimp or pork, The thing about touching the skin of a dead pig (Football). But when it comes to loving the wrong person (gays) this will not do! Christians only follow the parts ofthe bible that supports their bigotry. YOUR A HYPOCRITE.* [2]

4. **Warning of offline or online consequences:** counter-narratives that warn the user of the potential consequences of their actions. This can make the hate speaker retract their statements.

[2] Examples are taken from (Mathew et al., 2019) and modified

Terminology and definitions	Source
Definitions	
Language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group	(Davidson et al., 2017)
Act of offending, insulting or threatening a person or a group of similar people on the basis of religion, race, caste, sexual orientation, gender or belongingness to a specific stereotyped community	(Schmidt and Wiegand, 2017)
Offensive Language	
Any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct	(Zampieri et al., 2019)
Profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group	(Fortuna and Nunes, 2018)
Abusive/ Toxicity	
Hurtful language, including hate speech, derogatory language and also profanity	(Founta et al., 2018)
Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion	(Fortuna and Nunes, 2018)
Aggressiveness	
Intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target	(Sanguinetti et al., 2018)

Table 1: Definitions of terms used in Literature, source: (Poletto et al., 2021)

example: *You are beating up someone gay or straight, it is still an assault and by all means, this preacher should be arrested for sexual harassment and instigating!!!* [2]

5. **Affiliation:** counter-narratives which are relatable or can be affiliated with people.

   example: *Hey I'm Christian and I'm gay and this guy is so wrong. Stop the justification and start accepting.*[2]

6. **Denouncing hateful or dangerous speech:** counter-narratives where the target sentences are denounced as being hateful.

   example: *please take this down YouTube. this is hate speech.*[2]

7. **Humor and sarcasm:** counter-narratives that use satirical statements to mock or ridicule hate speech.

   example: *HAHAHAHAHAHAHAH...oh you were serious. That's even funnier :*[2]

8. **Visual Communication:** These counter arguments uses visual representation to counter fake and hurtful speech.

5 Works on Sensitivity

Sensitivity is difficult to detect because we have to look at the intent and context behind the conversation [(Tripathi et al., 2019)]. Keyword or phrase-based rules that look at the presence of certain words are not enough as some sentences may be implicitly offensive. For ML-based models, another issue that comes up is the availability of good-quality datasets. Often the classification of data tends to reflect the annotator's subjective biases [(Davidson et al., 2017)]. For example, people identify racist and homophobic statements as hateful but tend to see sexist jokes as merely offensive. The model should also be able to differentiate between whether the sensitive statement is directed toward a specific individual or community and whether it is explicit or implicit [(Waseem et al., 2017b)]. We need datasets annotated with extensive labeling in order to train the models in these tasks. Sensitivity also depends on time, thus it becomes necessary that the dataset contains data relevant to current society. Sensitivity can also become culture-specific, which is sometimes only captured by its local languages. This calls for the creation of datasets in various languages.

As we try to predict more labels, the model be-

comes more complex, which makes it difficult to explain their predictions. Thus a shift from simple predictive models to interpretable models is needed. [(Mathew et al., 2020)] observed that a good performance model doesn't always perform well in terms of explainability. Models that use rationales help reduce unintended bias towards the target.

In the following subsections, we will discuss some of the works done in the detection of offensive language and hate speech. We will also delve into the works that have been done specifically in the domain of Indian languages.

5.1 Detecting Offensive language

A lot of work has been done on the detection of offensive text in online communities. [(Cheng et al., 2015)] studied antisocial behavior in online communities by investigating the behavior of users who are eventually banned from an online site. He observed how their posts worsen over time and how other members of the community react to them. [(Yenala et al., 2018)] dealt with the task of detecting inappropriate content in query completion systems and user conversations in messaging systems.

A large number of datasets have also been created to tackle this issue and detect the subcategories of offensive language. Online sites like Twitter ((Davidson et al., 2017)) and Facebook ((Bhardwaj et al., 2020)) are good sources for creating datasets. A good quality dataset should ideally have good annotator agreement. A poor agreement can occur due to the annotator's bias. Sometimes it is not possible due to genuine difficulty in interpreting posts, which can lead to differences in opinion.[(Leonardelli et al., 2021)] found out that many popular datasets have a very less quantity of such challenging data. It suggested increasing the number of hard cases in the benchmark datasets which can lead to an increase in the robustness of the model. To create large-sized datasets, [(Tripathi et al., 2019)] combined manual annotation with a template based approach and semi supervised learning. [(Zampieri et al., 2019)] provided a fine-grained three-layer annotation procedure. It presented the OLID dataset, which has high-quality annotation of types and targets of the offenses. Various work has also been done to look into the other aspects of offensive languages such as abusive language, cyberbullying and cyber aggression ((Founta et al., 2018)), hate speech((Davidson et al., 2017), (Mathew et al., 2020)), etc.

5.2 Detecting Hate Speech

A common issue with the majority of research done in this sphere is that many of them combine hate speech and offensive language [(Davidson et al., 2017). Although it is not wrong to classify hate speech as offensive, one should note that hate speech often has a grave impact on society. Hate speech is an extreme case of offensive language which can spread discriminatory hatred and violence [(Assimakopoulos et al., 2020)]. Frequent exposure to hate speech could increase a person's prejudice against other groups, and on a large scale can degrade a nation's security and integrity. Due to this various countries have laws that penalize any citizen that spread hate speech. We shouldn't consider people as hate speakers because we failed to detect the difference between usual offensive language and serious hate speech [(Davidson et al., 2017)]. Thus it becomes necessary to detect hate speech separately. Models have evolved from using lexicons to using deep learning techniques like LSTM and BERT [(Mathew et al., 2020)].

To train such models we need datasets of good quality and large quantity. Annotators often fail to develop an understanding of what constitutes hate speech, which affects the quality of data. Although hate speech is defined in legal discourse as a statement(s) that incite discriminatory hatred, it is mistakenly used as an umbrella term for abusive or insulting statements. [(Assimakopoulos et al., 2020)] provides a 3 step annotation scheme that decreases the confusion between annotators that occurs due to them having different backgrounds and opinions. HateXplain is another popular dataset that contains more than 25000 English posts from Twitter and Gab that are labeled as hateful, offensive, and normal. Dataset mentioned in [(Davidson et al., 2017)] has over 20000 English tweets labeled as non-offensive, hate speech, and profanity(offensive).

5.3 Work on Indian Languages

The online presence of Indian people is ever-increasing and we also see a trend of using native languages on these social sites. This calls for measures to detect and mitigate the spread of hate speech in these languages.

One quick solution to this problem is to translate the text from the regional languages to English and check if the translated statement can be labeled

as hate speech. This has many downsides. Different cultures can have different notions of hate speech captured in their local languages whose proper translation may not be found in English [(Malik et al., 2022)]. Translating also makes it difficult to identify which words make the statement hateful. It is also possible that some words may be offensive in one language while being identified as normal in another. For example, in Hindi, the term (ku**a) is used as a swear word, while its English term "dog" is not often used in that way [(Bhardwaj et al., 2020)].

5.3.1 Pure Indian Languages

The Indian constitution recognizes 22 major languages. Amongst the major languages, Hindi has the largest number of speakers. Despite being the third most spoken language in the world, there is a lack of significant datasets in the language [(Bhardwaj et al., 2020)]. Amongst the few available datasets, it is observed that either the dataset is small or they cater to a specific dimension. HASOC presented by [(Mandl et al., 2019)] is a dataset in 3 languages, namely Hindi, English, and German. Along with binary classification of hate speech, The dataset also has labels for types and targets of hate speech. [(Mathur et al., 2018)] presents the development process of a multi-dimensional hostility detection dataset in Hindi.

5.3.2 Code-Mixed Indian Languages

The mixing of two or more languages in speech is called code-mixing. For example, Hinglish is a code-mixed language derived from combining Hindi and English languages. Code mixing provides the ease of using characters of one language for another. It also disregards the grammatical rules of the parent languages. Both these properties help in ease of informal communication.

Code mixing provides new challenges for the detection of hate speech. The words in code mixed language don't have any proper spelling, which can increase the ambiguity in the language. For example, the term *me* and *mai* can both refer to "I" of the English language. Code mixing doesn't follow any fixed grammar rules. Due to all these leeways, hate speech written in code-mixed languages can easily bypass models that are trained to detect the same for the parent languages. This calls for the creation of datasets in code-mixed language and train models.

[(Mathur et al., 2018)] has presented the HOTS dataset, which has more than 3000 tweets annotated as non-offensive, abusive, and hateful. Aggression annotated corpus, described in [(Kumar et al., 2018)] has posts from Twitter and Facebook annotated using 3 top-level tags and 10 level 2 tags.

6 Works on Counter narratives

A counter narrative is a response to hate speech that utilizes fact-based arguments, counters stereotypes and false information, and alters the viewpoints of people, especially individuals who spread hate speech [(Chung et al., 2019)]. Many studies have found it to be an effective means to not only combat hate, but also address the harm that it causes [(Yadav, 2018)]. A study conducted by (Hangartner et al., 2021) found that counter-narratives imbued with empathy can effectively reduce instances of xenophobic hate speech. (Mathew et al., 2019) observed that counter narrative comments on youtube videos received more likes than non-counter narrative comments, indicating a positive audience response to counter narratives. According to (Yadav, 2018), the most popular counter narrative pages on Facebook in India are related to 'satirical or religious criticism'.

The increasing popularity of counter-narratives has led to the creation of Counter Narrative datasets using various methods, including social media scraping, crowd or niche sourcing, and hybrid approaches [(Tekiroğlu et al., 2020), (Fanton et al., 2021)]. (Chung et al., 2019) developed a multilingual hate speech/counter-narrative dataset on islamophobia. To expedite the data collection process, (Fanton et al., 2021) proposed a hybrid methodology that involved iteratively training a language model to generate pairs of hate speech and counter-narrative, which were then validated by human annotators. Expanding upon this methodology, (Bonaldi et al., 2022) introduced a dialog-based data collection approach, which simulates real-life conversations involving multiple exchanges between people. Counter-narratives that effectively refute hate speech through factual information, statistics, and relevant examples are more likely to be accepted. To streamline the process of creating such informative counter-narratives, (Chung et al., 2021) developed a generative pipeline that leverages external knowledge acquired through key phrases.

7 Datasets

Here we will give a comprehensive overview of popular datasets in the field of sensitivity and counter narratives

7.1 Sensitivity Datasets

1. **HateXplain:** It is a benchmark dataset published with (Mathew et al., 2020) that covers the bias and interpretability aspects of hate speech. It contains text classified exclusively as hateful, offensive, or normal. It also contains target group labels as well as word and phrase level span annotations that capture human rationales.

2. **OLID Dataset:** The Offensive Language Identification Dataset ((Zampieri et al., 2019)) contains English tweets annotated using a three-layer annotation scheme. In the first layer, a tweet is labeled as either **NOT**(Not offensive) or **OFF** (Offensive). In the second layer, the offensive tweets are categorized as **TIN** (Targeted insults) or **UNT** (untargeted insults). In the third layer, the targets are categorized as **IND** (individual), **GRP** (group), or **OTH** (other).

3. **The HASOC Fire 2019 Dataset**: published with (Mandl et al., 2019), this dataset consists of Twitter data labelled as either HOF (Hate and Offensive) or NOT (Not Hate Offensive). The HOF data is further labelled as HATE, OFFN (Offensive), or PRFN (Profanity).

4. **Hostility Detection Dataset:** This dataset was published with (Bhardwaj et al., 2020) has around 8200 Hindi posts taken from various social media platforms like Twitter, Facebook, WhatsApp, etc. The posts have been manually annotated as hostile and non-hostile. Furthermore, the hostile label has four dimensions which are *fake, defamation, hate,* and *offensive*. This second layer annotation is multi-label instead of multi-class.

5. **HOT Dataset:** The Hinglish offensive Tweet dataset was created by [(Mathur et al., 2018)]. It contains more than 3000 Hinglish tweets, out of which 65 per cent of posts were abusive. The tweets are labelled for hate speech and abusive speech.

7.2 Counter Narrative Datasets

1. **CounterSpeech Dataset:** The CounterSpeech Dataset was introduced by (Mathew et al., 2019) and it is the first-ever dataset on Counterspeech. They define counterspeech as a "direct response or comment (not a reply to a comment) that *counters* the hateful or harmful speech". To create this dataset, user comments from YouTube videos were collected that targeted three communities: Jews, African-Americans, and LGBT. The counterspeech comments were further annotated for different types of counterspeech present. The following types of counterspeech were labeled: presenting facts, pointing out hypocrisy or contradictions, warning of offline or online consequences, affiliation, denouncing hateful or dangerous speech, humor, positive tone, and hostile.

2. **CONAN Dataset:** The "COunter NArratives through Nichesourcing" or CONAN dataset was introduced in (Chung et al., 2019). It comprises hate speech-counter narrative pairs, consisting of 6654 pairs for English, 5157 pairs for French, and 3213 pairs for Italian. It mainly consists of hate speech targeting Islam. The counter-narratives were further classified as: Presentation of facts, Pointing out hypocrisy or contradiction, Warning of consequences, Affiliation, Positive tone, Negative tone, Humour, Counter questions, Other.

8 Summary

In this paper, we have examined the concept of sensitivity, including its definitions and various terminologies used in this domain. We have reviewed current methods for mitigating hate speech and sensitive posts, along with their limitations. We have discussed counter narratives as a promising emerging approach for countering hate speech with more speech, exploring different types of counter narratives and challenges in detecting sensitive content. Additionally, we have highlighted the lack of work in Indian languages and provided examples of studies done in both pure and code-mixed Indian languages. We have observed a steady improvement in counter-narrative generation methods and provided a list of popular datasets that can be used for training frameworks in sensitivity detection and counter-narrative generation tasks.

References

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France. European Language Resources Association.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on twitter: A field study. dangerous speech project.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. *arXiv preprint arXiv:2211.03433*.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *International Conference on Web and Social Media*.

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.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

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.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. 2019. Thou shalt not hate: Countering online hate speech.

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

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a

systematic review. *Language Resources and Evaluation*, 55:1–47.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Rahul Tripathi, Balaji Dhamodharaswamy, Srinivasan Jagannathan, and Abhishek Nandi. 2019. Detecting sensitive content in spoken language. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 374–381.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017a. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017b. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

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*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Anandita Yadav. 2018. Counterspeech: An alternative policy to combat hate speech in india. *Indian Journal of Law and Human Behaviour*, 4(2):169–78.

Harish Yenala, Ashish Jhanwar, Manoj Chinnakotla, and Jay Goyal. 2018. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6.

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.