A Review on Hate Speech Recognition on Social Media

Chandan Kumar[1], Ritesh Kumar Yadav[2], Varsha Namdeo[3]

Department of Computer Science & Engineering[1,2,3]

SRK University, Bhopal, (M.P.), India[1,2,3]

ABSTRACT

Social media platforms provide an inexpensive communication medium that allows anyone to quickly reach millions of users. Consequently, in these platforms, anyone can publish content, and anyone interested in the content can obtain it, representing a transformative revolution in our society. However, this same potential of social media systems brings together an important challenge—these systems provide space for discourses that are harmful to certain groups of people. This challenge manifests itself with several variations, including bullying, offensive content, and hate speech. Specifically, authorities of many countries today are rapidly recognizing hate speech as a serious problem, especially because it is hard to create barriers on the Internet to prevent the dissemination of hate across countries or minorities. In this paper, we provide the first of a kind systematic large scale measurement and analysis study of hate speech in online social media. We aim to understand the abundance of hate speech in online social media, the most common hate expressions, the effect of anonymity on hate speech, and the most hated groups across regions. This survey organizes and describes the current state of the field, providing a structured overview of previous approaches, including core algorithms, methods, and main features used.

Keywords: Hate Speech, NLP, Machine Learning, Social media, patter Recognition.

Introduction

Online social media sites today allow users to freely communicate at nearly marginal costs. Increasingly users leverage these platforms not only to interact with each other but also to share the news. While the open platforms provided by these systems allow users to express themselves, there is also a dark side of these systems. Particularly, these social media sites have become a fertile ground for inflamed discussions, that usually polarize ‘us’ against ‘them’, resulting in many cases of insulting and offensive language usage.

Another important aspect that favors such behavior is the level of anonymity that some social media platforms grant to users. For example, “Secret” was created, in part, to promote free and anonymous speech but became a means for people to defame others while remaining anonymous. The secret was banned in Brazil for this very reason and shut down in 2015 1. There are reports of cases of hateful messages in many other social media independently of the level in which the online identity is bonded to an offline identity – e.g., in Whisper, Twitter, Instagram, and Facebook.

With this context, it is not surprising that most existing efforts are motivated by the impulse to detect and eliminate hateful messages or hate speech [1, 2]. These efforts mostly focus on specific manifestations of hate, like racism [3]. While these efforts are quite important, they do not attempt to provide a big picture of the problem of hate speech in the current popular social media systems. Specifically providing a broad understanding of the root causes of online hate speech was not the main focus of these prior works. Consequently, these prior works also refrain from suggesting broad techniques to deal with the generic offline hate underlying online hate speech. In this paper, we take the first step towards a better understanding of online hate speech. Our effort consists of characterizing how hate speech is spread in common social media, focusing on understanding how hate speech manifests itself under different dimensions such as its targets, the identity of the haters, geographic aspects of hate contexts. Particularly, we focus on the following research questions.

What is hate speech about? We want to understand not only which the most common hated groups of people are, but also what are the high-level categories of hate targets in online hate speech.
What role does anonymity play on hate speech? Is anonymity a feature that exacerbates hate speech or is social media users not worried about expressing their hate under their real names? What fraction of haters uses their names in social media?
How does hate speech vary across geography? Does hate speech targets vary across countries? And, within states of a country like the USA? Are there categories of hate speech that are uniformly hated and others that are hated only in specific regions?

Answering these questions is crucial to help authorities (including social media sites) for proposing interventions and effectively deal with hate speech. To find answers, we gathered one-year data from two social media sites: Whisper and Twitter. Then, we propose and validate a simple yet effective method to detect hate speech using sentence structure and using this method to construct our hate speech datasets. Using this data, we conduct the first of a kind characterization study of hate speech along multiple different dimensions: hate targets, the identity of haters, geographic aspects of hate, and hate context. Our results unveil a set of important patterns, providing not only a broader understanding of hate speech but also offering directions for detection and prevention approaches.

II Related Work

We review existing work on hate speech along three dimensions.

a. Understanding hate speech

Hate speech has been an active research area in the sociology community [4]. Particularly, claims that some forms of hate speech are far from being solved in our society, especially those against black people and women. Hate speech originating from such prejudices is quite abundant and authorities have created standard policies to counter it. However, there has been a multitude of undesirable social consequences of these very policies (e.g., incivility, tension, censorship, and reverse discrimination) due to the suppression of haters and the protection of hate targets. Over time, this tension has driven the evolution of standard policies to regulate hate speech.

A very recent study [5], supported by UNESCO, reviews the growing problem of online hate speech with the advent of the internet from a legal and social standpoint. They pointed out that platforms like Facebook and Twitter have primarily adopted only a reactive approach to deal with hate speech reported by their users, but they could do much more. More specifically, their study reports “These platforms have access to a tremendous amount of data that can be correlated, analyzed, and combined with real-life events that would allow a more nuanced understanding of the dynamics characterizing hate speech online". Our work is motivated by this vision.

Even before the popularity of social networks, the problem of racism and hate detection was already a research theme in computer science. Back in 2004, there have been efforts that attempt to identify hateful web pages, containing racism or extremism [6]. Nowadays, there has been a multitude of related problems under investigation in social media systems [7]. However, these approaches do not give a data-driven global view of hate speech in online media today; we aim to bridge this gap.

b. Detecting hate speech in online media

In recent years, there have been several studies that focus on computational methods to find hate speech on social media.

[3] reviews three different recent studies that aim to detect the presence of racism or offensive words on Twitter. They point out that, while simple text searches for hate words in tweets represent a good strategy to collect hate speech data, it also creates a problem: the context of the tweets is lost. For instance, the word “Crow” or “Squinty” is a racial slur in the United Kingdom, but it can also be used in multiple different non-hate related contexts. Multiple researchers try to solve this problem using manual inspection or a mix of crowdsourcing labeling and machine learning techniques [1, 2]. Their basic framework consists of the creation of a corpus that contains a set of known hate keywords. This corpus is then manually annotated to construct a training dataset that contains positive and negative hate posts. Finally, they learn from this training dataset to build automated systems (via machine learning approaches) for detecting hate speech. Overall, these types of approaches have two shortcomings. Firstly, it is hard to detect new hate targets using hate keywords. Secondly, manual labeling, although useful, but is not scalable if we want to understand and detect hate speech at large scale. Aside from leveraging text-based features researchers also explored other features like leveraging user history [8] or even community detection [9]; these techniques can be used in addition to the text-based features. Although all these efforts offer advances in this field, it is safe to say that computational methods to detect hate speech currently are in a nascent stage.

Most of these prior efforts focus on detecting online hate speech. Differently, our research goal is to use computational techniques to understand the social phenomena of online hate speech. Our approach, based on sentence structure, provides a reasonably accurate data set to answer our research questions. Our strategy also allows us to identify several explicit hate speech targets (or communities), which directly complements (and benefits) the existing keywords search based semi-automated approaches.

c. Hate speech and anonymity

The problem of hate speech inspired a growing body of work in effectively detecting such speeches on various social media platforms. However, so far these efforts focused on either non-anonymous social media platforms, like Twitter or Facebook [10], or on radical forums and known hate groups. However, there is an interesting and unexplored middle ground between—Anonymous social media like Whisper or Secret. These media sites are recently becoming quite popular within normal users. These platforms do not require any account or persistent identity to post on their sites. Recent efforts [7, 30] reviewed content posted on such forums in depth. They found that users post more sensitive content on such forums and a significant fraction of such posts are confessions about their personal lives. Existing efforts in sociology [11] already pointed out that in the presence of anonymity, humans show a disinhibition complex. In other words, the posters might be much less inhibited and express their otherwise suppressed feeling or ideas on anonymous social media sites. Thus, intuitively, in the presence of anonymity one will expect to find the presence of hate speech from a diverse set of users who are not radicalized, but they have certain prejudices which otherwise they will not express in their posts. Based on this intuition we made an effort to investigate Whisper, an anonymous social media system, in our analysis. We hope to provide a more inclusive picture of hate speech in social media in that way.

In a preliminary short paper [12], we attempted to correlate hate crimes incidents with hate speech in Whisper and Twitter. In this paper, we used the same methodology to gather data from Twitter and Whisper, but we provide a much wider and deeper understanding of hateful messages in these systems.

III Approaches in Hate Speech Detection

In this section, we analyze features described in the papers focusing on algorithms for hate speech detection, and also other studies focusing on related concepts (e.g., Cyber-bullying). Finding the right features for a classification problem can be one of the more demanding tasks when using machine learning. Therefore, we allocate this specific section to describe the features already used by other authors. We divide the features into two categories: general features used in text mining, which are common in other text mining fields; and the specific hate speech detection features, which we found in hate speech detection documents and are intrinsically related to the characteristics of this problem. We present our analysis in this section.

General Features Used in Text Mining: The majority of the papers we found to try to adapt strategies already known in text mining to the specific problem of automatic detection of hate speech. We define general features as the features commonly used in text mining. We start by the most simplistic approaches that use dictionaries and lexicons.

Dictionaries: One strategy in text mining is the use of dictionaries. This approach consists of making a list of words (the dictionary) that are searched and counted in the text. These frequencies can be used directly as features or to compute scores. In the case of hate speech detection, this has been conducted using: Content words (such as insults and swear words, reaction words, personal pronouns) collected from www.noswearing.com [13]. The number of profane words in the text, with a dictionary that consists of 414 words, including acronyms and abbreviations, where the majority is adjectives and nouns. Label Specific Features that consisted of using frequently used forms of verbal abuse as well as widely used stereotypical utterances.

Ortony Lexicon was also used for negative affect detection; the Ortony lexicon contains a list of words denoting a negative connotation and can be useful, because not every rude comment necessarily contains profanity and can be equally harmful.

This methodology can be used with an additional step of normalization, by considering the total number of words in each comment. Besides, it is also possible to use this kind of approach with regular expressions [14].

Distance Metric. Some studies have pointed out that in text messages the offensive words may be obscured with an intentional misspelling, often a single character substitution. Examples of these terms are “@ss,” “sh1t”, “nagger,” or homophones, such as “Joo”. The Levenshtein distance, i.e., the minimum number of edits necessary to transform one string into another, can be used for this purpose. The distance metric can be used to complement dictionary-based approaches.

Bag-of-words (BOW). Another model similar to dictionaries is bag-of-words. In this case, a corpus is created based on the words that are in the training data, instead of a pre-defined set of words, as in the dictionaries. After collecting all the words, the frequency of each one is used as a feature for training a classifier. The disadvantages of this kind of approach are that the word sequence is ignored, and also its syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation N-grams can be adopted.

N-grams. N-grams are one of the most used techniques in hate speech automatic detection and related tasks. The most common N-grams approach consists in combining sequential words into lists with size N. In this case, the goal is to enumerate all the expressions of size N and count all occurrences. This allows improving classifiers’ performance because it incorporates at some degree the context of each word. Instead of using words, it is also possible to use N-grams with characters or syllables. This approach is not so susceptible to spelling variations as for when words are used. Character N-gram features proved to be more predictive than token N-gram features, for the specific problem of abusive language detection.

However, using N-grams also have disadvantages. One disadvantage is that related words can have a high distance in a sentence and a solution for this problem, such as increasing the N value, slows down the processing speed. Also, studies point out that higher N values (5) perform better than lower values (unigrams and trigrams). In a survey researchers report that N-grams features are often reported to be highly predictive in the problem of hate speech automatic detection, but perform better when combined with others.

Profanity Windows: Profanity windows are a mixture of a dictionary approach and N-grams. The goal is to check if a second person pronoun is followed by a profane word within the size of a window and then create a boolean feature with this information.

TF-IDF: The TF-IDF (term frequency-inverse document frequency) was also used in this kind of classification problems. TF-IDF is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that a word appears in the document. However, it is distinct from a bag of words, or N-grams, because the frequency of the term is off-setted by the frequency of the word in the corpus, which compensates the fact that some words appear more frequently in general (e.g., stop words).

Part-of-speech: Part-of-speech (POS) approaches to make it possible to improve the importance of the context and detect the role of the word in the context of a sentence. These approaches consist in detecting the category of the word, for instance, personal pronoun (PRP), Verb non-3rd person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part- of-speech has also been used in hate speech detection problems. With these features, it was possible to identify frequent bigram pairs, namely PRP_VBP, JJ_DT, and VB_PRP, which would map as “you are”. It was also used to detect sentences such as “send them home,” “get them out,” or “should be hung”. However, POS proved to confuse the class identification, when used as features.

Lexical Syntactic Feature-based (LSF): In a study, the natural language processing parser, proposed by Stanford Natural Language Processing Group was used to capture the grammatical dependencies within a sentence. The features obtained are pairs of words in the form “(governor, dependent)”, where the dependent is appositional of the governor (e.g., “You, by any means, an idiot.” means that “idiot,” the dependent, is a modifier of the pronoun “you,” the governor). These features are also being used in hate speech detection.

Rule-Based Approaches: Some rule-based approaches have been used in the context of text mining. A class association rule-based approach, more than frequencies, is enriched by linguistic knowledge. Rule-based methods do not involve learning and typically rely on a pre-compiled list or dictionary of subjectivity clues. For instance, rule-based approaches were used to classify antagonistic and tense content on Twitter using associational terms as features. They also included accusational and attributional terms targeted at only one or several persons following a socially disruptive event as features, to capture the context of the terms used.

Participant-vocabulary Consistency (PVC). In a study about cyberbullying, this method is used to characterize the tendency of each user to harass or to be harassed, and the tendency of a key phrase to be indicative of harassment. For applying this method it is necessary a set of messages from the same user. In this problem, for each user, it is assigned a bully score (b) and a victim score (v). For each feature (e.g., N-grams) a feature-indicator score (w) is used. It represents how much the feature is an indicator of a bullying interaction. Learning is then an optimization problem over parameters b, v, and w.

Template Based Strategy: The basic idea of this strategy is to build a corpus of words, and for each word in the corpus, collect K words that occur around. This information can be used as a context. This strategy has been used for feature extraction in the problem of hate speech detection as well. In this case, a corpus of words and a template for each word was listed, as in “W-1: go W+0: back W+1: to.” This is an example of a template for a two-word window on the word “back.”

Word Sense Disambiguation Techniques: This problem consists of identifying the sense of a word in the context of a sentence when it can have multiple meanings. In a study, the stereotyped sense of the words was considered, to understand if the text is anti-Semitic or not.

Typed Dependencies: Typed dependencies were also used in hate speech related studies. First, to understand the type of features that we can obtain with this, the Stanford typed dependencies representation describes the grammatical relationships in a sentence that can be used by people without linguistic expertise. These were used for extracting Theme-based Grammatical Patterns and also for detecting hate speech specific other language that we will present within the specific hate speech detection features. Some studies report significant performance improvements in hate speech automatic detection based on this feature.

Topic Classification: With these features, the aim is to discover the abstract topic that occurs in a document. In a particular study, topic modeling linguistic features was used to identify posts belonging to a defined topic (Race or Religion).

Sentiment: Bearing in mind that hate speech has a negative polarity, authors have been computing the sentiment as a feature for hate speech detection Different approaches have been considered (e.g., multi-step, single-step) Authors usually use this feature in combination with others that proved to improve results.

Word Embeddings: Some authors use a paragraph2vec approach to classify language on user comments as abusive or clean and also to predict the central word in the message. Fast Text is also being used. A problem that is referred to in hate speech detection is that sentences must be classified and not words. Averaging the vectors of all words in a sentence can be a solution; however, this method has limited effectiveness. Alternatively, other authors propose comment embeddings to solve this problem.

Deep Learning: Deep learning techniques are also recently being used in text classification and sentiment analysis, with high accuracy.

Other Features: Other features used in this classification task were based in techniques such as Named Entity Recognition (NER), Topic Extraction, Word Sense Disambiguation Techniques to check Polarity, frequencies of personal pronouns in the first and second person, the presence of emoticons and capital letters. Before the feature extraction process, some studies have also used stemming and removed stop-words. Characteristics of the message were also considered such as hash tags, mentions, retweets, URLs, number of tags, terms used in the tags, number of notes (re-blog and like count), and link to multimedia content, such as image, video, or audio attached to the post.

IV Research Challenges and Opportunities

Hate speech is a complex phenomenon and its detection problematic. Some challenges and difficulties were highlighted by the authors of the surveyed papers:

- Low agreement in hate speech classification by humans, indicating that this classification would be harder for machines.
- The task requires expertise in culture and social structure.
- The evolution of social phenomena and language makes it difficult to track all racial and minority insults.
- Language evolves quickly, in particular among young populations that communicate frequently in social networks.
- Despite the offensive nature of hate speech, an abusive language may be very fluent and grammatically correct, can cross sentence boundaries, and the use of sarcasm in it is also common.
- Finally, hate speech detection is more than simple keyword spotting.

We find it relevant to present those difficulties so that we bear in mind the kind of challenges that researchers face in their work.

V Conclusion

In this survey, we presented a critical overview on how the automatic detection of hate speech in text has evolved over the past years. First, we analyzed the concept of hate speech in different contexts, from social networks platforms to other organizations. Based on our analysis, we proposed a unified and clearer definition of this concept that can help to build a model for automatic detection of hate speech. Additionally, we presented examples and rules for classification found in the literature, together with the arguments in favor or against those rules. Our critical view pointed out that we have a more inclusive and general definition about hate speech than other perspectives found in the literature. This is the case, because we propose that subtle forms of discrimination on the internet and online social networks should also be spotted. With our analysis, we also concluded that it would be important to compare hate speech with cyberbullying, abusive language, discrimination, toxicity, flaming, extremism and radicalization. Our comparison showed how hate speech is distinct from these related concepts and helped us to understand the limits and nuances of its definition.

Through a systematic literature review, we concluded that there are not many studies and papers published in automatic hate speech detection from a computer science and informatics perspective. In general, the existing works regard the problem as a machine learning classification task. In this field, researchers tend to start by collecting and annotating new messages, and often these datasets remain private. This slows down the progress of the research, because less data is available, making it more difficult to compare results from different studies. Nevertheless, we found three available datasets, in English and German. Additionally, we compared the diverse studies using algorithms for hate speech detection, and we rank them in terms of performance. Our goal was to reach conclusions about which approaches are being more successful. However, and in part due to the lack of standard datasets, we find that there is no particular approach proving to reach better results among the several articles.

References

[1] Swati Agarwal and Ashish Sureka. 2015. Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter. In Proceedings of The 11th International Conference on Distributed Computing and Internet Technology (ICDCIT’15).

[2] J. Bartlett, J. Reffin, N. Rumball, and S. Williamson. 2014. Anti-social media. DEMOS

[3] Irfan Chaudhry. 2015. #Hashtagging hate: Using Twitter to track racism online. First Monday 20, 2 (2015).

[4] Richard Delgado and Jean Stefancic. 2004. Understanding words that wound. Westview Press.

[5] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. Coun- tering online Hate Speech. UNESCO.

[6] Edel Greevy and Alan F. Smeaton. 2004. Classifying Racist Texts Using a Support Vector Machine. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[7] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.

[8] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In Proceedings of 35th European Conference on IR Research.

[9] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Anti- social Behavior in Online Discussion Communities. In

International Conference on Web and Social Media (ICWSM).

[10] I. Kwok and Y. Wang. 2013. Locate the hate: Detecting tweets against blacks. In Proceedings of The AAAI Conference on Artificial Intelligence (AAAI'13).

[11] John Suler. 2004. The online disinhibition effect. Cyberpsychology & behavior 7, 3 (2004), 321–326.

[12] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In International Conference on Web and Social Media (ICWSM).

[13] Shuhua Liu and Thomas Forss. 2015. New classification models for detecting hate and violence web content. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K'15), Vol. 1. IEEE, 487–495.

[14] Wilson Jeffrey Maloba. 2014. Use of Regular Expressions for Multi-lingual Detection of Hate Speech in Kenya. Ph.D. Dissertation. iLabAfrica.