Survey on Hate Speech Detection on Social Media

Shalini Kuswaha[1], Dr. Nitesh Dubey[2]
[1] Dept of CSE
[2]Professor, Dept of CSE
[1, 2] Global Nature Care Sangathan Group of Institution, Jabalpur, Madhya Pradesh, India.

Abstract- *Advances in Internet Technologies (ITs) and online social networks have made more benefits to humanity. At the same time, the dark side of this growth/benefit has led to increased hate speech and terrorism as most common and powerful threats globally. Hate speech is an offensive kind of communication mechanism that expresses an ideology of hate using stereotypes. Hate speech targets different protected characteristics such as gender, religion, race, and disability. Control of hate speech can be made using different national and international legal frameworks. Any intentional act directed against life or related entities causing a common danger is known as terrorism. There is a common practice of discussing or debating hate speech and terrorism separately.*

In recent years "Hate-Words" or "Offensive Words" have become increasingly sophisticated, making their detection more difficult. This paper presents a review of Hate-Words Detection and various approaches for detection of Hate-Words on social media, which find out open research questions and challenges in present methods.

Keywords- Machine learning, Hate-Words, Hate-Crimes, Online Social Network, Twitter and Facebook.

I. INTRODUCTION

Along with the rapid development of technology, the Internet became one of the means for the community, especially in Indonesia to communicate. Through the internet, communicating is done on social media. In communicating, especially in social media, people often issue opinions. Not only are good opinions, but many negative opinions are written on social media, including hate speech. Hate speech is a speech that intimidates people from certain social groups oriented towards differences, races, national origin, and gender [1].

Hate speech also has a complex connection with freedom of expression, individual rights, groups, minorities, and also related to the concepts of dignity, freedom, togetherness, and also the [2] context. In Indonesia a lot of social media is very popular with people like Instagram, Facebook, and Twitter. Twitter became one of the social media used by people in Indonesia because it can post various things, including opinions that contain hate speech to the internet easily anywhere, anytime, in real time [3]. As more and more people use social media, it is a big challenge to differentiate good opinions from bad ones. Hate speech is included in bad opinions when written on social media. In Indonesia there are laws that have regulated hate speech, namely Article 28 paragraph (2) of the ITE Act 2008. Although it has been regulated in the ITE Law, it has not been able to accommodate the violation as a whole because the traditional method is very limited because it is unable to handle a number of large existing data, so the need for data processing by text sentiment. To get hate speech information from the existing opinion data, in this final project will be done data processing with sentiment analysis using an Artificial Neural Network method optimized with back propagation algorithm. Artificial Neural Network is one of the artificial intelligence sciences that can solve problems in the fields of patterning and pattern recognition [4]. While the back propagation algorithm is one of the development of Artificial Neural Network methods with a single screen added one or more hidden screen [5]. From this analysis, it can separate opinions into negative classes which mean that opinions contain hate speech. With this system, can be known sentences that contain hate speech.

1.2 Hate Speech

Communication actions carried out by an individual or group in the form of provocation, incitement, or insult to other individuals or groups in various aspects such as race, color, gender, disability, sexual orientation, citizenship, religion, etc. are the meanings of hate speech. In the legal sense, hate speech is a prohibited speech, behavior, writing, or performance because it can lead to violence and prejudice from the offender or the victim of the action [6]. Almost all countries around the world have laws that govern Hate Speech.

Hate speech is a particular form of offensive language where the person using it is basing his opinion either on segregative, racist or extremist background or on stereotypes. Merriam- Webster1 defines hate speech as a ``*speech expressing hatred of a particular group of people."* From a legal perspective, it defines it as a ``*speech that is intended to insult, offend, or intimidate a person because of* *some trait (as race, religion, sexual orientation, national origin, or disability)."* This being the case, hate speech is considered a world-wide problem that many countries and organizations have been standing up against. With the spread of internet, and the growth of online social networks, this problem becomes even more serious, since the interactions between people became indirect, and people's speech tends to be more aggressive when they feel physically safer, not to mention that internet presents for many hate groups sees it as an *"unprecedented means of communication of recruiting"* [7].

In the context of internet and social networks, not only does hate speech create tension between groups of people, its impact can also influence businesses, or start serious real- life conflicts. For such reasons, websites such as Facebook, YouTube and Twitter prohibit the use of hate speech. However, it is always difficult to control and filter all the contents.

Therefore, in the research field, hate speech has been subject to some studies, trying to automatically detect it. Most of these works on hate speech detection have goals such as the construction of dictionaries of hate words and expressions [8] or the binary classification into *``hate"* and *``non-hate"* [9]. However, it is always difficult to clearly decide on a sentence whether it contains hate or not, in particular if the hate speech is hiding behind sarcasm or if no clear words showing hate, racism or stereotyping exist. Furthermore, OSN are full of ironic and joking content that might sound racist, segregative or offensive, which in reality is not. An example is given in the following two tweets: *``Hey dummy. It has been a while since we last read one of your useless comments.". ``If we want the opinion of a WOMAN, we'll ask you dear... For now keep quiet."*

The first tweet sounds offensive and demeaning the person target of the tweet. However, given the mutual follow of both users, the tweet is actually a joke between two friends. The second also presents the same problem, even though the user seems to be offending women, given the context of the message (i.e., a small discussion between a group of friends), the tweet in itself was not posted to offend women, or even the person targeted by the tweet. Such expression and others that include reference to a particular gender, race, ethnic group or religion are widely used in a joking context, and have to be clearly distinguished from hate speeches. Therefore, the use of dictionaries, and *n*-grams in general, might not be the optimal option to perform the distinction between expressions showing hate, and those that do not. It is arguable that sentiment analysis techniques can be used to perform hate speech detection. However, this is a different task, which requires more sophisticated techniques: In sentiment analysis, the main task is the detection of sentiment polarity of the tweet, which goes back to the idea of the detection of any existing positive/negative word or expression. This makes it easy to rely on the direct meaning of words: words have usually the same sentiment polarity regardless of the context or the actual meaning with very few exceptions (e.g. the word ``bad" cannot be interpreted, under any circumstance, in a positive way). However, in the case of hate speech, some words might be negative, might even have the meaning of hate, but the context makes them not hate speech-related. A typical example can be seen in the following two examples:

"I hate seeing them losing every time! It's just unfair!":

Even though the word ``hate" has been employed here, the given sentence does not fall under the category of hate speech, simply because the context is not a context of offending a person, let alone to be offending him for his gender, race, etc.*``I hate these neggers, they keep making life much painful"*:

This is obviously a hate speech towards a specific ethnic group. This makes the task of hate speech detection quite different and more challenging than sentiment analysis: not only is it context-dependent, but also, we should not rely on simple words or even n-grams to detect it. On a related context, writing patterns have proven to be effective in text classification tasks such as sarcasm detection [10], multi-class sentiment analysis [11] or sentiment quantification [12]. The types of patterns, and the way they are built and extracted depend on the application. Therefore, during this work, we try to extract patterns of hate speech and offensive texts using a pragmatic approach, and use these, along with other features to detect hate speech in short text messages on Twitter.

II. RELATED WORK

Automatic hate speech detection does not have a long history, but there has been a huge interest in the recent ten years. Since in most scenarios comments, posts were used for hate speech detection, the problem is classified as a natural language processing problem. Three main types of approaches were identified for hate speech detection.

A. Lexical Based Approaches

Lexical based approaches rest on the idea that most important part of a text classification task is being able to understand lexical phrases. Machine is fed with patterns of language, grammar, manually created rules describing certain type of texts or else domain base knowledge describing certain type of texts. N.D.Gitari et al [13] presents a classifier model for hate speech detection using a lexicon. The methodology proposed by them is comprised of three steps. A rule based and learning approach is used for subjectivity detection as first step. Then a lexicon for hate speech had been built in the 2nd step. Negative polarity words, hate verbs and theme based grammatical patterns were used as features to build the lexicon. Using those three types of features rules were generated to classify a sentence as hate or not as the 3rd step. An F-score of 70.83 was achieved for the combination of all three feature types.

B. Machine Learning Approaches

Machine learning approaches are the most commonly seen approach used then. A multi-class classifier to distinguish between hate speech, offensive language and none of them is presented by Davidson et al [14]. Logistic regression with L2 regularization has been used to build the final model. Their best performing model has an overall precision of 0.91, recall of 0.90 and F1-score of 0.90. Z. Waseem et al [15] have evaluated the influence of different features for prediction of hate. A logistic regression classifier with 10-fold cross validation had been used to test the influence of various features on prediction performance. They have found that character n-gram is better than word n-gram in accordance with their features. They have used gender, location and length of the tweet as additional features mainly. Best performance has been achieved with character n-grams of lengths up to 4 with the additional feature gender with an Fscore 73.93%. Usage of additional features location and length hasn't given improvements to F1-score.

C. Hybrid Approaches

Hybrid approaches are used by many researchers. Combination of learning-based approaches with lexical based approaches is done in here. In some scenarios first, the lexical based approach is used, and data is filtered and then those filtered data is fed in to a machine learning model. Meantime in some scenarios lexical resources are used to extract features from text data and those features are fed to the machine learning model.

Results of the research conducted by A. Wester et al. [16] shows that combination of lexical features outperforms the use of more complex syntactic and semantic features for the task of detecting online hate. Maximum Entropy, SVM and Random Forest are the three different classification frameworks used. Basic lexical features, word forms, lemmas and n-grams were used as initial set of features. Then in the second-round different combinations of mentioned features were used. According to their analysis Bag-of-Word model and lexical n-gram model with both Maximum Entropy and SVM classifiers were selected to build the final model. From them n-gram model has outperformed BoW model with both SVM and MaxEnt with F-scores of 0.6885 and 0.6860 respectively.

Usage of paragraph level features for the first stage of classification was proposed by Warner et al [17]. They have used the hypothesis that hate speech resembles a word sense disambiguation task.

2.1 Some Existing Mechanism

Nabiila Adani Setyadi et. al. [18] proposed Hate Speech Detection Using Back propagation Neural Network. The author hopes after this application the computer can know and classify the existence of hate speech on a text from social media twitter.

Hajime Watanabe et. al. [19] proposed approach to detect hate expressions on Twitter. Approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm.

Ricardo Martins et. al. [20] proposed to examine methods to classify hate speech in social media. Aim to establish lexical baselines for this task by applying classification methods using a dataset annotated for this purpose. As features, our system uses Natural Language Processing (NLP) techniques in order to expand the original dataset with emotional information and provide it for machine learning classification.

Nevertheless, our analysis still has limitations that lead to exciting future research directions. Firstly, it is reasonable to question the definition of hateful content, in the sense that it is not clear what is the threshold a published text shared in social media has to violate to be considered hateful due to the subjectivity of the definition of hate-speech. Secondly, this work does not address the issue of user's characterisation and their potential use of code to overcome anti-hate speech policies and automatic detection systems. Thirdly, since the words used in hate speech change rapidly - new words creation, expressions used locally or within a given context - it is a somewhat arduous task to be up to date with the new expressions used.

As future work, it is planned the creation of a classification module for new emotional words, to increase the ability to analyse new words without the dependence of specialised and updated lexicons and consequently increase the prediction of hate speech. Another line of future work is to explore computational strategies and approaches to characterise and monitor user-centric content in social media.

Axel Rodriguez et. al. explores a novel framework to effectively detect highly discussed topics that generate hate speech on Facebook. With the use of graph, sentiment, and emotion analysis techniques, we cluster and analyze posts on prominent Facebook pages. Consequently, the proposed framework is able to identify the pages that promote hate speech in the comment sections regarding sensitive topics automatically. In this paper, we proposed a new approach to identify hate speech on Facebook, which is a challenging task since many Facebook users have been trying hard to cover their real intentions. To tackle this problem, we used graph analysis to identify pages that potentially promote hate speech. By applying sentiment and emotion analysis, the most negative posts and comments were obtained. *K-means* clustering was then applied to determine the most discussed topics. By analyzing the topics generated by the best combination of parameters, it can be observed that this new approach yields promising results. As for the future work, we would like to address how to incorporate the latest comments posted on Facebook, since the social media is very dynamic. We would also like to include replies to the comments to be considered in the process, in order to capture the full conversation and also to identify ironical sentences. More parameter combinations will be tested, with different seed pages known for discussing various topics, to evaluate the effectiveness of the proposed approach under different types of contents.

In this paper [23] the final objective was to achieve hate speech and offensive language detection and prevention. The results indicated high level of classifier accuracy as well as applicability of the machine learning algorithms in mobile environments. Likewise, the use of machine learning in this kind of text analytics is highly suggested because the need to quickly handle requests/responses is high in the mobile ‘world’. The neural network algorithms have been proven to produce very good results even though we have tested it in a simple feed-forward network library.

III. CONCLUSION

The aim of this paper was to introduce a development approach and an integration of Hate Speech Detection with machine learning algorithms in real time environments. The final objective was to achieve development work done in past in hate speech and offensive language detection and prevention.

Likewise, the use of machine learning in this kind of text analytics is highly suggested because the need to quickly handle requests/responses is high in the Social Media World.

REFERENCES

[1] Fasold, Ralph W. 2006. An Introduction to Language and Linguistics. Cambridge: Cambridge University Press.

[2] Gagliardone, Iginio, Alisha Patel, and Matti Pohjonen. 2014. Mapping and Analyzing Hate Speech Online: Opportunities and Challenges for Ethiopia. Oxford: University of Oxford.

[3] Habibi, Robet, Djoko Budiyanto Setyohadi, Dan Ernawati. 2016. “Sentiment Analysis on student twitter using the Back-Propagation Method”, Indonesia: INFORMATIKA Vol. 12, No. 1, April 2016.

[4] G.Y.P, Cynthia. 2006. Enterprise Bankruptcy prediction using Artificial Neural Network, Surabaya: Ten Institute of Technology, November.

[5] Rachmatullah, M. Naufal, dan Anggina Primanita. 2015,” Implementation of Artificial Neural Networks in Automatic Text Summification System Using Feature Extraction. Palembang: 2015 National Seminar on Information and Communication Technology.

[6] Labhukum-2017. Review of Hate Speech, [Online] Available at: Available at: labhukum.com/2017/07/18/review-without-ujarankebencian-hate-speech.

[7] J. P. Breckheimer, ``A haven for hate: The foreign and domestic implications of protecting Internet hate speech under the first amendment," South California Law Rev., vol. 75, no. 6, p. 1493, Sep. 2002.

[8] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, “Offensive language detection using multi-level classification”, Advances in Artificial Intelligence, vol. 6085. Ottawa, Canada: Springer, Jun. 2010, pp. 16_27.

[9] W. Warner and J. Hirschberg, “Detecting hate speech on the World Wide Web,”, in Proc. 2nd Workshop Lang. Social Media, Jun. 2012, pp. 19_26.

[10] D. Davidov, O. Tsur, and A. Rappoport, “Semi-supervised recognition of sarcastic sentences in Twitter and Amazon," in Proc. 14th Conf. Comput. Natural Lang. Learn., Jul. 2010, pp. 107_116.

[11] M. Bouazizi and T. Ohtsuki, “Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter," in Proc. IEEE ICC, May 2016, pp. 1_6.

[12] M. Bouazizi and T. Ohtsuki, ``Sentiment analysis in Twitter: From classification to quantification of sentiments within tweets," in Proc. IEEE, GLOBECOM, Dec. 2016, pp. 1_6.

[13] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A Lexicon-based Approach for Hate Speech Detection," Int. J. Multimed. Ubiquitous Eng., vol. 10, no. 4, pp. 215–230, 2015.

[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," arXiv Prepr. arXiv1703.04009, 2017.

[15] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Res. Work., pp. 88–93, 2016.

[16] A. Wester, L. Ovrelid, E. Velldal, and H. L. Hammer, "Threat detection in online discussions," WASSA@ NAACL-HLT, pp. 66–71, 2016.

[17] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," Proceeding LSM '12 Proc. Second Work. Lang. Soc. Media, no. Lsm, pp. 19–26, 2012.

[18] Nabiila Adani Setyadi, Muhammad Nasrun, Casi Setianingsih, "Text Analysis For Hate Speech Detection Using Back propagation Neural Network", The 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), IEEE-2018.

[19] Hajime Watanabe, Mondher, Tomoaki Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection", IEEE Access 2018.

[20] Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, Pedro Henriques "Hate speech classification in social media using emotional analysis", 2018 7th Brazilian Conference on Intelligent Systems, IEEE-2018.

[21] Axel Rodriguez, Carlos Argueta, Yi-Ling Chen, "Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis", IEEE-2019.

[22] Bujar Raufi, Ildi Xhaferri," Application of Machine Learning Techniques for Hate Speech Detection in Mobile Applications", IEEE-2018.