Hate Speech Detection

Using Machine Learning Algorithm

SAMREEN SOHAIL[1] , ,Mohd.Istiaq Hossian Junaid[2], Nishat Salsabil Rainy[3]

North South University

Bashundhara, Dhaka

Samreen.sohail@northsouth.edu((1711648642)

Mohd. istiaq@northsouth,edu(1821577642)

nishat.rainy@northsouth.edu(1812620642)

Abstract:

The increasing use of social media and information sharing has given major benefits to humanity. However, this has also given rise to a variety of challenges including the spreading and sharing of hate speech messages. Thus, to solve this emerging issue in social media sites, recent studies employed a variety of feature engineering techniques and machine learning algorithms to automatically detect the hate speech messages on different datasets. However, to the best of our knowledge, there is no study to compare the variety of feature engineering techniques and machine learning algorithms to evaluate which feature engineering technique and machine learning algorithm outperform on a standard publicly available dataset. Our study holds practical implication and can be used as a baseline study in the area of detecting automatic hate speech messages.

Moreover, the output of different comparisons will be used as state-of-art techniques to

compare future research for existing automated text classification techniques.

Introduction:

In recent years, hate speech has been increasing in-person and online communication. The social media as well as

other online platforms are playing an extensive role in the breeding and spread of

hateful content – eventually which leads to hate crime. For example, according to recent surveys, the rise in online hate speech content has resulted in hate crimes

including Trump's election in the US , the Manchester and London attacks in the UK , and terror attacks in New Zealand . To tackle these harmful consequences of hate speech, different steps including legislation have been taken by the European Union Commission. Recently, the European Union Commission also enforced social media networks to sign an EU hate speech code to remove hate speech content within 24 hours . However, the manual process to identify and remove hate speech content is labor-intensive and time-consuming. Due to these concerns and widespread hate speech content on the internet, there is a strong motivation for automatic hate speech detection.

The automatic detection of hate speech is a challenging task due to disagreements on different hate speech definitions.Therefore, some content might be hateful to some individuals and not to others, based on their concerned employed the different feature engineering techniques and ML algorithms to classify content as hate speech. Regardless of this extensive amount of work, it remains difficult to compare the performance of these approaches to classify hate speech content. To the best of our knowledge, the existing studies lack the comparative analysis of different feature engineering techniques and ML algorithms.

Therefore, this study contributes to solving this problem by comparing three feature engineering and eight ML classifiers on standard hate speech datasets. Table I shows major concepts related to automatic text classification along with their explanations and references. This study holds practical importance and served as a reference for new researchers in the domain of automatic hate speech detection.

Related works:

automatically detect racism on Dutch social media. In this study, the author uses the distribution of words in three dictionaries as characteristics. They provide the generated functions to the SVM classifier. As a result of their experiment, they scored 0.46 F.

features, and have a tendency to target hate speech. Subsequently, the author provides the vector of main features to the rule-based classifier. In the

definitions.According to,hate speech is:“the content that promotes violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, veteran status, and sexual orientation/gender identity”.Despite these different definitions, some recent studies claimed favorable results to detect automatic hate speech in the text .The proposed solutions

Today, hate speech is very common on social media.Therefore, in recent years, some researchers have applied ML-based supervised text classification methods to classify hate speech content. Different researchers have adopted different feature representation techniques, namely dictionary-based, bag-of-words-based, N-gram-based, TFIDF-based, and deep learning-based.

Peter Burnap et al. (Burnap & Williams, 2016) uses a dictionary-based method to identify online hate on Twitter. In this research, they used N-gram feature engineering technology to generate digital vectors from a predefined dictionary of hateful words. The author provides the generated numeric vector to the ML SVM classifier and gets an F score of up to 67%.

Stéphan Tulkens et al. (Tulkens et al., 2016) also used a dictionary-based method to

Njagi Dennis et al. (Gitari et al., 2015) used ML-based classifiers to classify hate speech on online forums and blogs. The author uses a dictionary-based method to generate the main feature vector. These features are based on emotional expressions using semantic and subjective

experimental setting, the author evaluated the classifier using precision performance indicators and obtained an precision of 73%.

However, the combination of dictionary-based methods and machine learning methods showed good results. However, the main disadvantage of this method is that it requires a dictionary based on a large corpus to search for domain words. In order to overcome this shortcoming, many researchers have used a BOW-based method, which is similar to a dictionary-based method, but the characteristics of the words are obtained from training data, rather than from a predefined dictionary.

Edel Greevy et al. (Greevy & Smeaton, 2004) uses supervised machine learning methods to classify racist texts. To convert the original text into a digital vector, the author uses two-letter group feature extraction technology. The author uses the bigram function and the BOW function rendering technology. They use the SVM classifier to perform the experimental results. The results showed that they obtained 87% certainty.

In recent years, the author has adopted deep learning-based NLP technology to classify hateful voice messages. Sebastian Köffer et al. (*Discussing the Value of Automatic Hate Speech Detection in Online Debates | Institut Für Wirtschaftsinformatik*, n.d.) uses the word2vec function and the SVM classifier to classify German text that hates voice messages and gets an F score of 67%. Word2Vec has the lowest results because such methods require a large amount of data to learn complex word semantics.

Schmidt et al. (Schmidt & Wiegand, 2017) conducted a survey of using natural language processing to detect hate speech in 2017. The author discussed in detail various functional engineering techniques that will be used to supervise the classification of hate voice messages. The biggest disadvantage of this survey is that the mentioned technology has no experimental results.

Previous research has shown that many researchers around the world are working on identifying hate speech written in different languages such as German, Dutch, and English. However, according to our information, there are no studies that can compare various ML algorithms and features in the standard data set, and these studies can be used as baseline studies for future researchers in the field of hate speech recognition.

Methodology:

This section introduces the proposed system that we used to classify tweets into a category, namely hate speech. Figure 1 shows the complete research methodology. As shown in the figure, the research method includes six key steps, namely data collection, data preprocessing, feature engineering, data division, classification , model construction and classification model evaluation.

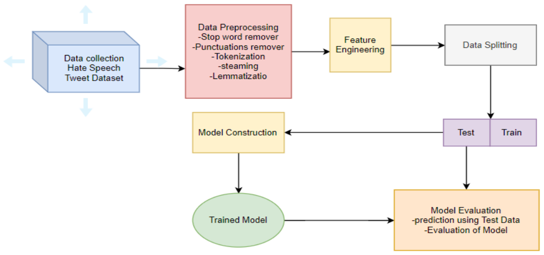

Fig 1: System Overview

Result:

We have train the model with linear SVC and Multinomial Naive Bayes. We get 96% accuracy in linear SVC and Multinomial Naive Bayes. 97%. The accuracy reflect that our model is perfect for the problem and we get best accuracy .

Conclusion and future works:

In this modern age social media plays a very important role. We have to be concerned about the people . So hate speech detection is very important for this morder age to detect and warn or disable hate speeches . It will save us from a huge clash.

In these model we have add only english bad speech detectors . In the future we will bangla language hate speech detector. It will help to control the people in our country.

References:

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 11. https://doi.org/10.1140/epjds/s13688-016-0072-6

Discussing the Value of Automatic Hate Speech Detection in Online Debates | Institut für Wirtschaftsinformatik. (n.d.). Retrieved May 25, 2021, from https://www.wi.uni-muenster.de/de/forschung/publikationen/131445

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 468–469. https://doi.org/10.1145/1008992.1009074

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing*. 1–10. https://doi.org/10.18653/v1/w17-1101

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). *A Dictionary-based Approach to Racism Detection in Dutch Social Media*. http://arxiv.org/abs/1608.08738