Journal of Theoretical and Applied Information Technology
31st December 2023. Vol.101. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HATE SPEECH DETECTION USING LSTM AND NAÏVE BAYES ALGORITHM

ANANDA WONGWATANA NASUTION [1] , ARIEL ZAHRAN [2] ,MUHAMMAD FAZAR AKBAR [3] ,GHINAA ZAIN NABIILAH [4] ,ROJALI [5]

[1,2,3,4,5] School of Computer Science, Bina Nusantara University, Indonesia

E-mail: [1] ananda.nasution@binus.ac.id, [2] ariel.zahran@binus.ac.id,[3] Muhammad.akbar032@binus.ac.id, [4] ghinaa.nabiilah@binus.ac.id, [5]rojali@binus.edu

ABSTRACT

Hate speech detection requires effective strategies to ensure a safe and inclusive online environment. This research paper presents a comparative study of hate speech detection using Natural Language Processing (NLP) techniques, specifically Naïve Bayes and Long Short-Term Memory (LSTM) approaches. The objective is to develop models capable of automatically identifying and analyzing hate speech in written language. The prevalence and impact of hate speech are emphasized, as it can lead to psychological harm and incite criminal acts. NLP offers a valuable tool for automatically detecting potentially dangerous content and addressing this problem. The study utilizes a dynamically generated dataset containing diverse words and expressions to train and evaluate the Naïve Bayes and LSTM models. The results show that the LSTM and the Naïve Bayes model, achieving an accuracy of 74% and 64%.

Keywords: *Hate Speech, Naive Bayes, LSTM, Text Classification,NLP*

1. INTRODUCTION

Social media is used as a platform for users to share information across many networks. Social media has developed rapidly in the past years, with the emergence of technology that allows users to have freedom to express their ideas, opinion, and criticisms on social media.There are many social-media platform that are often used in todays time like Twitter and facebook.

According to Conover et al. (2013), Twitter has emerged as a widely used micro-blogging platform, allowing millions of users to express their thoughts and opinions through real-time status updates[1]. The platform boasts an impressive user base of 270 million active users, with a staggering 500 million tweets being posted daily (M.C. Wellons, 2015)[2]. Given its widespread popularity and extensive reach, social media websites like Twitter have also become instrumental for organizing and mobilizing events such as protests and public demonstrations (Muthiah et al., 2015)[3]. Additionally, Twitter serves as a prominent platform for sharing opinions and information surrounding live events, both prior to, during, and after their occurrence (Bollen et al., 2011)[4].

However, with the presence of social media, hate speech cannot be avoided. Online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. Online media prohibits users to post violent threats, harassment, and hateful contents. However, there are still tons of users who disobey the rules and spread hate speech and negative words.Hate speech has become a rampant problem on social media and can have an impact and influence on an individual's psychology. Hate speech can contribute and even lead to criminal acts, such as violence. With the use of natural language processing (NLP), a useful tool for identifying and analyzing hate speech in written or spoken language, this can be addressed. NLP is used to automatically flag potentially dangerous content on social media for review by humans or to inform policy decisions.

This paper presents an approach based on Naïve Bayes and Long short-term memory (LSTM) to detect hate speech. The collection of hate speech is Dynamically Generated so that it contains a variety of words, expressions, and emotional signals.

2. LITERATURE REVIEW

In recent years, hate speech on social media has become an increasingly prevalent issue, making the detection of such language a crucial area of research. Natural Language Processing (NLP) has emerged as a promising approach to automatically identifying hate speech in online content. A number of studies have compared the performance of various machine learning algorithms for hate speech detection, with some of the most successful approaches utilizing Support Vector Machines (SVM) and deep learning models, such as the attention-based model proposed by Azam et al. (2020).[5]

Other studies have explored the use of different feature sets and data preprocessing techniques, such as TF-IDF and stemming, to improve the accuracy of hate speech detection. The work of Santos et al. (2019), for instance, found that the bag-of-words feature set combined with Random Forest achieved high accuracy in detecting hate speech in a Brazilian Portuguese Twitter dataset[6]. Razzaque et al. (2021) similarly employed Random Forest for hate speech detection in a Twitter dataset, but also compared the performance of different text mining techniques, including Multinomial Naïve Bayes and SVM.[7] Overall, these studies highlight the potential of NLP and machine learning for automated hate speech detection, as well as the need for continued exploration and refinement of these methods.

Yoon Kim's (2014) The paper proposes a method using Convolutional Neural Networks (CNNs) to build sentence classification models, with data from labeled categories in the IMDB dataset and different CNN architectures tested. The results show that the CNN model achieves better classification accuracy than traditional models such as Naïve Bayes and SVM. They demonstrated the superiority of CNNs in sentence classification tasks and made a significant contribution to the development of NLP based on Deep Learning.[8]

Xiaonan Li and Yijun Li's (2018) provides an overview of various text classification algorithms and their applications. They collect and analyze various text classification techniques, such as Naive Bayes, SVM, and Deep Learning, using different datasets, including 20 Newsgroups and Reuters-21578. Their research shows the advantages of each text classification technique under different data conditions and illustrates current trends in NLP development. They provided a broad perspective on the existing text classification techniques.[9]

Sebastian and Samuel's (2016) provides an overview of text classification, with a focus on machine learning techniques. They used different machine learning techniques for text classification and provided an overview of their strengths and weaknesses.[10]

M Shubhadeep and Kumar Bala (2017) proposed a method for detecting sarcasm over Twitter. They arranged hundreds of tweets to analyze keywords indicating sarcasm. To judge the content of the sentence as well as the authorial or writing style of the author. A training dataset of 10,000 tweets was preprocessed to correct spelling variation ,cleaned by removing the retweets and The remaining tweets were manually labeled as sarcastic or non-sarcastic . In order to classify tweets, Naïve Bayes classifier highlighted sarcastic and non-sarcastic tweets and prominent features were identified from those tweets. The classifier showed an accuracy of 65%.[11]

Watanabe (2018) proposes an approach to detect hate expressions on Twitter. Their approach is based on unigrams and patterns that are automatically collected from the training set. These training sets contained 21,000 tweets, distributed evenly among the three classes (i.e., “Clean,” “Offensive” and “Hateful”).Their results demonstrated an improvement on standard learning techniques. The classifier showed an accuracy of 87.4%.[12]

Campos et al(2019) have done the experiment of Text Mining in Hotel Reviews. The paper stated that Naïve Bayes is the algorithm is a very simple algorithm that spend less time on text classification, because the algorithm only uses the Bayes Theorem to find the class of the sentence.[13]

3. METHODOLOGY

3.1 Dataset

This research uses a dataset of Dynamically Generated Hate Speech Dataset of 4070 from kaggle. The data are divided into nothate and hate speech, it has multi labels of classification.
Table 1 contains the representation of the dataset used.

Table 1: representation of the dataset

Type	Example
none	I dont hate Muslims
derogation	I'm so happy about brexit, send them all home
Animosity	Their disdainful attitude makes me think they’re Chinese

Threatening	I would love to crucify a Romanian

●None, this label contains sentences that don't contain hateful words towards an individual or group.
●derogation, this label contains sentences that contain words that have the meaning of belittle, degrade, or demean individuals or groups based on their characteristics, against certain races and religions.
●Animosity, this label contains sentences that contain words that express hostility, anger, or enmity against certain races and religions.
●Threatening, this label contains sentences that contain words that explicitly or implicitly threaten harm, violence, or negative consequences.

3.2 Proposed Method

The approach consists of the following steps: Creating a dataset. First, we collect hate speech through kaggle. Then we pre-processed these data so that they can be fit for feature extraction. After preprocessing we pass this data in our classifier, which then classifies them into positive or negative classes based on trained results, which will enable in evaluation of how hate speech can be curbed. Fig. 1. contains the flow of the process proposed in this study.

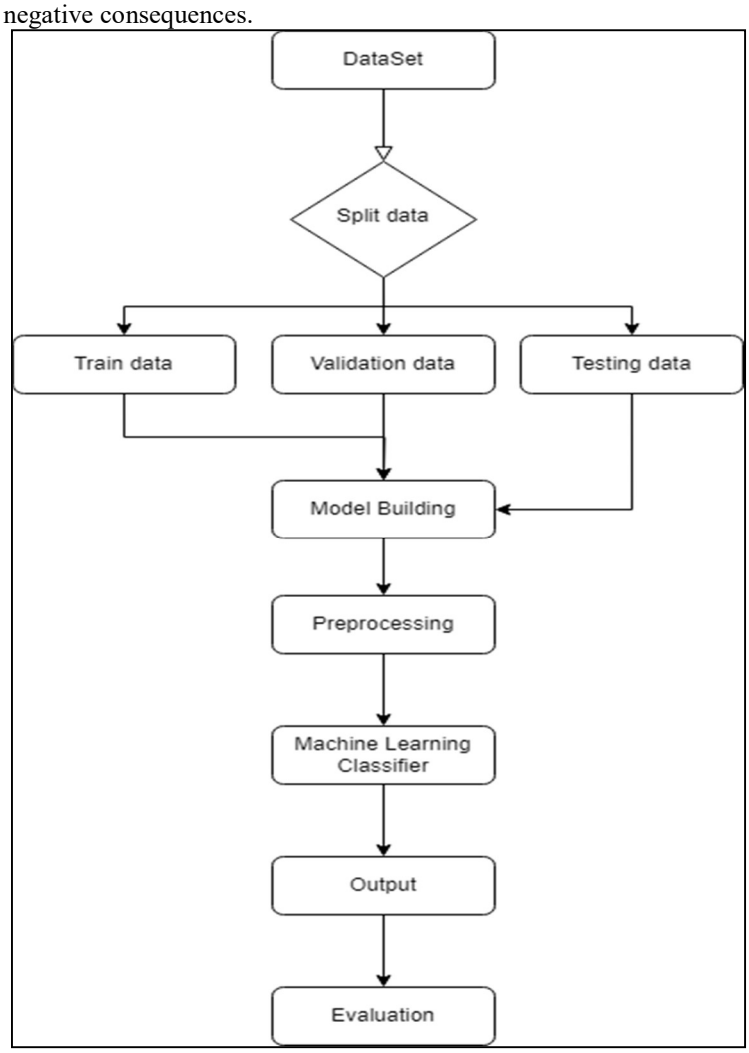

Figure 1: flow of proposed method

The pre-trained model used is the Naïve Bayes and LSTM models. Naïve Bayes is a simple yet powerful probabilistic model that excels in handling text classification tasks. It is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

computationally efficient and works well with high-dimensional feature spaces, making it particularly useful when dealing with large-scale datasets. Naïve Bayes assumes independence between features, which allows for fast training and inference. It can provide quick insights into the data and perform well in scenarios where the data exhibits strong feature independence. LSTM (Long Short-Term Memory) is specifically designed for sequential data processing. It is particularly effective in capturing long-term dependencies and contextual information in textual data. LSTM models excel in tasks that require understanding the temporal dynamics and semantic relationships within the text. They can capture complex patterns and handle variable-length sequences.

Using both of the models allows us to benefit from the efficiency and simplicity of Naïve Bayes and use the power of LSTM to capture complex patterns. This approach can provide a comprehensive framework to address the problem ahead.

3.3 Preprocessing

Preprocessing is a stage for processing and cleaning raw data before it is used for analysis. Preprocessing involves several stages to eliminate unnecessary words or characters that are not needed in the classification process. The stages in preprocessing generally involves deleting characters and numbers, changing letters to the same capitalization, stopword removal,Tokenization and stemming[14]. These preprocessing stages clean and transform the raw text data into a format that can be used as input for training the model. Figure 2. contains the steps taken to clean the data.

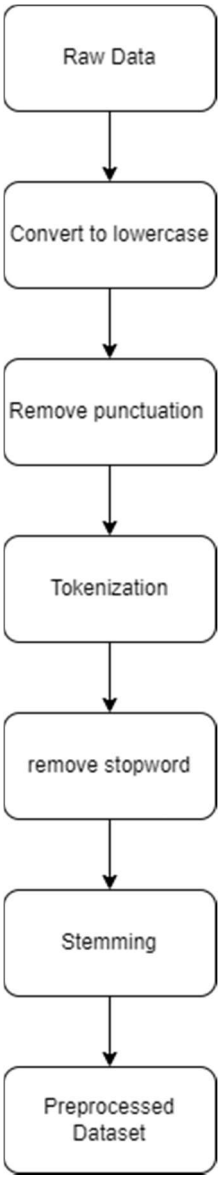

Figure 2: Preprocessing Stage.

- Convert the text to lowercase
- Remove punctuation from text
- Tokenizing is used to break sentences into words
- Stemming removes prefixes,suffixes and affixes
- Removing stopword removes words that do not have a significant meaning, such as prepositions or conjunctions

These preprocessing steps clean and transform the raw text data into a format that can be used as input for training the model.

4. RESULT AND DISCUSSION

The study was conducted utilizing the Python programming language and its associated libraries, which are commonly employed for training and building neural network models. Furthermore, experiments were conducted using the Google Colab platform. Google Colab is particularly recommended for research involving

pre-trained models due to its ability to allocate a significant amount of memory for the experimental procedures.

Table 2 Experiment Result

Algorithm / Model	Accuracy-Training	Accuracy-Validation
LSTM Classifier	0.9057	0.7457
Naïve Bayes Classifier	0.64	0.60

From the experiments conducted, Table 2 shows the The training data for the LSTM model achieved an accuracy of 90%, the model performed well in learning patterns and features from the provided training dataset.The Naïve Bayes model demonstrated a satisfactory performance in detecting hate speech. It achieved an accuracy of 64% on the test set. On the other hand, the LSTM model achieved an accuracy of 74%. The LSTM model's ability to capture sequential patterns in the text data proved advantageous for hate speech detection. It showcased higher precision, suggesting a better overall performance in identifying hate speech instances.

The higher accuracy of the LSTM model can be attributed to its ability to capture sequential patterns in the text data. Hate speech often relies on subtle linguistic cues and context, which can be effectively captured by LSTM's ability to analyze the sequential nature of language.

5. CONCLUSIONS

Based on the experiments conducted to create a classification model for hate speech detection, the results indicate that both the Naïve Bayes and LSTM models have shown promising performance. The Naïve Bayes model achieved an accuracy of 64%, while the LSTM model achieved a higher accuracy of 74%. These findings demonstrate the usefulness of utilizing machine learning algorithms in identifying and categorizing hate speech instances across multiple labels. The ability of the models to analyze textual data and capture complex patterns contributes to their success in hate speech detection. Hate speech detection is a challenging task due to the dynamic and evolving nature of language on social media. Therefore, continuous research and development efforts are required to enhance the models' performance and adaptability to new forms of hate speech.

REFERENCES:

[1] Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. "Political polarization on Twitter". *ICWSM*,2013, 133, 89-96.

[2] M.C. Wellons. "The politics of protest: Social media and black youth". *Journal of Social Media in Society*,2015, 4(2), 53-78.

[3] Muthiah, S., Lampos, V., & Williamson, G. "Twitter Floods, Facebook Bursts and the Power of Sharing: The Role of Social Media in London 2011 Riots". *IEEE Transactions on Computational Social Systems*, 2015,2(1), 7-13.

[4] Bollen, J., Mao, H., & Zeng, X. "Twitter mood predicts the stock market". *Journal of Computational Science*,2011, 2(1), 1-8.

[5] Azam, M. A., Islam, S. S., Moniruzzaman, M., & Chowdhury, S. U. "A novel approach to hate speech detection using an attention-based deep learning model". *IEEE Access*,2020, 8, 203429-203440.

[6] Santos, J. C. S., de Souza, L. F. T., & Braga, R. H. G. . "Towards automated hate speech detection: A comparison of feature sets and machine learning models".3*4th ACM/SIGAPP Symposium on Applied Computing* ,2019,(pp. 10-15).

[7] Razzaque, G. M. A., Sajib, S. A. M., Rony, M. S. I., Islam, M. A., & Hossain, K. M. A. "Detecting hate speech in social media using text mining and machine learning techniques". *International Conference on Informatics, Electronics & Vision (ICIEV)*,2021, (pp. 1258-1262).

[8] Yoon Kim. "Convolutional Neural Networks for Sentence Classification".*Empirical Methods in Natural Language Processing (EMNLP)*,2014, 1746-1751.

[9] Xiaonan Li and Yijun Li. . "A survey on text classification algorithms and applications". *Journal of Advances in Computer Networks*,2018, 6(1), 10-16.

[10] Sebastian, S., & Samuel, S. "Text classification using machine learning: An overview". *International Journal of Computer Applications Technology and Research*, 2016,5(2), 143-150.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] Shubhadeep Mukherjee, Pradip Kumar Bala,"Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering",*Technology in Society*,Volume 48,2017,Pages 19-27,ISSN 0160-791X.

[12] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*,2018 ,vol. 6, pp. 13825-13835.

[13] Campos D, Silva RR, Bernardino J. "Text Mining in Hotel Reviews: Impact of Words Restriction in Text Classification." *InKDIR* ,2019, (pp. 442-449).

[14] J. Mothe et al., Proceedings of 2019 11th International Conference On Knowledge And Systems Engineering : KSE 2019 : October 24-26, 2019, Da Nang, Vietnam.