2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 979-8-3503-3086-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAC3N60023.2023.10541516

Developing and Evaluating a Binary Hate Speech Dataset: A Comparitive Study of Machine Learning Models

1st Ayush Kumar
Delhi Technological University
India
ayushkm826@gmail.com

2nd Harsh Mittal
Delhi Technological University
India
24mithar@gmail.com

3rd Ajeet Kumar
Delhi Technological University
Delhi, India
ajeetdph@gmail.com, ajeetdph@dtu.ac.in

***Abstract*—This research article presents an investigation into a hate speech dataset that contains 2,478 tweets. The dataset was meticulously annotated by human annotators, and it includes two types of labels: hate (45.56%) and not-hate. We elaborate on the annotation process and the inter-annotator agreement attained. Additionally, we carry out a thorough analysis of the dataset, which encompasses exploratory data analysis, statistical analysis, and machine learning experiments. Our results demonstrate that recognizing hate speech is a difficult task due to its highly subjective nature among annotators. However, our dataset can be an advantageous resource for training and assessing hate speech detection models. We also discuss the ethical considerations that must be addressed while collecting and using hate speech data and provide recommendations for future research in this field.**

***Index Terms*—Hate Speech, Dataset, Machine Learning, Deep Learning**

I. Introduction

The proliferation of hate speech on digital platforms is a growing concern in modern times. It is defined as any form of expression or action that shows hostility, hatred, or prejudice towards an individual or group based on their race, ethnicity, religion, sex, sexual preference, or other identifying characteristic. This has led to an increasing need for automated detection systems to prevent the spread of such content. To achieve this, the development of accurate hate speech detection models is critical, which requires reliable and large datasets. However, developing such datasets is a complex and challenging task since hate speech is a multifaceted and subjective phenomenon [1].

To overcome these challenges, researchers have developed several hate speech datasets that are typically annotated by humans. Nevertheless, the quality and consistency of annotations can vary, depending on several factors, including the annotator's background, the hate speech definition, and the annotation guidelines. This paper presents a new hate speech dataset containing 2,478 tweets, manually annotated to indicate hate speech, offensive language, or neutral content [2] scribe the annotation process, and the inter-annotator agreement, and present a comprehensive analysis of the dataset through exploratory data analysis, statistical analysis, and machine learning experiments.

The subsequent sections provide an overview of related work in hate speech detection and datasets (Section 2), details about our hate speech dataset and various machine learning and deep learning models(Section 3), a comprehensive analysis of the dataset (Section 4), and finally, ethical considerations of collecting and using hate speech data and concluding remarks and recommendations for future work in this area (Section 5, 6, 7).

In summary, this paper contributes to the field of hate speech detection by introducing a large and reliable hate speech dataset, providing insights into hate speech characteristics, and the challenges of detecting it. The dataset and analysis are valuable resources for developing and evaluating hate speech detection models and enhancing the understanding of hate speech on digital platforms.

II. Related Work

In this section, we will discuss state-of-the-art work on online hate speech detection.

Capozzi et al. [3] have introduced a data visualization tool that facilitates the analysis and evaluation of diverse collections of text data related to hatred on online channels. The study put forth a corpus comprising tweets aimed at multiple ethnic minority groups that are controversial in Italian public discourse

Kapil and Ekbal's [4] research addresses the challenge of improving the performance of individual categorization tasks by leveraging relevant information from multiple related tasks. They have developed a "deep Multi-Task Learning (MTL) framework" that utilizes a shared-private architecture to extract both common and task-specific features from five categorization tasks.

Roy et al. [5] have suggested an automated approach that employs a Deep Convolutional Neural Network (DCNN) for hate speech detection. The proposed DCNN model utilizes tweet text along with GloVe embedding vector to capture tweet semantics via convolution operation, and it achieves high precision, recall, and F1-score values of 0.97, 0.88, and 0.92 for the optimal scenario, surpassing the performance of existing models.

ISBN: 979-8-3503-3086-1/23/$31.00©2023 IEEE

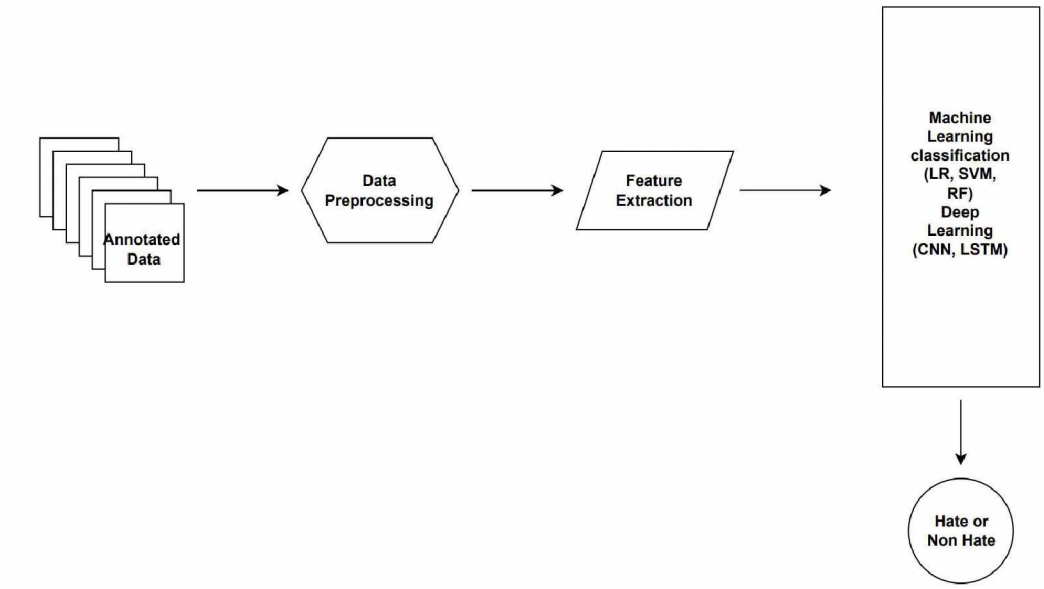


Fig. 1: Flowchart for Labelling a Tweet Hate or Not-Hate

Corazza et al. [6] have presented a neural network that has been validated to perform accurately in three different languages, namely English, Italian, and German.

Raufi and Xhaferri [7] have tackled the challenge of detecting hate speech in Albanian by creating a classification system that uses machine learning and has a lightweight design. Their system is well-suited for mobile applications, which is becoming increasingly popular as more people use mobile devices to access the internet. The fact that their system is lightweight is particularly beneficial since mobile devices typically have limited resources. Overall, their focus on mobile applications and lightweight design make their system a promising solution for detecting hate speech.

Koushik et al. [8] have proposed a machine learning-based approach to detect hate speech in tweets automatically. The method involves using a bag of terms and the TFIDF (Term Frequency-Inverse Document Frequency) approach for tweet classification into two categories, which was achieved by employing a logistic regression classifier. The researchers have separated 30 per cent of the dataset for testing purposes and used the remaining 70 per cent as training data. The model attained high accuracy, with 94.11 per cent accuracy using the bag of words feature and 94.62 per cent accuracy using the TFIDF feature.

Al-Makhadmeh and Tolba [9] have presented a methodology for predicting hate speech from social media platforms that combines natural language processing and machine learning techniques. The proposed method is a hybrid approach that leverages both methods to achieve accurate predictions of hate speech.

Kovacs [10] and his team have put forth a deep natural language processing (NLP) model that fuses CNN and RNN to automatically flag hateful text on online platforms. They assessed the model's effectiveness on HASOC2019 dataset. It resulted in 63% in macro F1 score.

In this research, Plaza-del-Arco [11] and colleagues tackle the problem of identifying hate speech in Spanish on social media platforms, aiming to enhance the comprehension of the potential of advanced machine learning techniques. The researchers compare Deep Learning techniques with traditional machine learning models and pre-trained language models built on Transfer Learning. This work makes a substantial contribution by achieving positive results in Spanish while utilising multilingual and monolingual pre-trained language models like BETO, BERT and XML.

Alsafari et al. [12] develop and train single and ensemble classifiers using non-contextual and contextual word-embedding models, including Fast text-SkipGram, Multilingual Bert, and AraBert, on Arabic-Twitter datasets with binary, tertiary, and six-class hate/offensive classification problems. For each classification assignment, the researchers ran many tests to assess how well single and ensemble classifiers performed on test datasets. The ensemble strategy based on the average-based technique yielded the best results, achieving F-scores of 91 per cent, 84 per cent and 80 per cent for binary, tertiary, and six-class prediction tasks, apiece.

Vidhen and Yasseri [13] conduct comprehensive conceptual research and create an automated software tool that can differentiate between non-Islamophobic, mildly Islamophobic, and strongly Islamophobic content. The method achieves an accuracy of 77.6 per cent and a balanced accuracy of 83 per cent. This tool will facilitate further quantitative investigations on the sources, distribution, incidence, and consequences of Islamophobic hate speech on social media.

III. Methodology

A. Data Collection

The first step in creating a hate speech dataset is data collection. In this study, tweets from Twitter were collected using specific keywords related to hate speech and offensive language. The keywords were chosen based on earlier research on hate speech detection and input from experts in the field. The dataset included 2,478 tweets that were collected over two weeks, using diverse search queries. The number of Not-Hate Tweets were 1349 and Hate Tweets were 1129, which is 45.56% Hate label Tweets.

B. Annotation

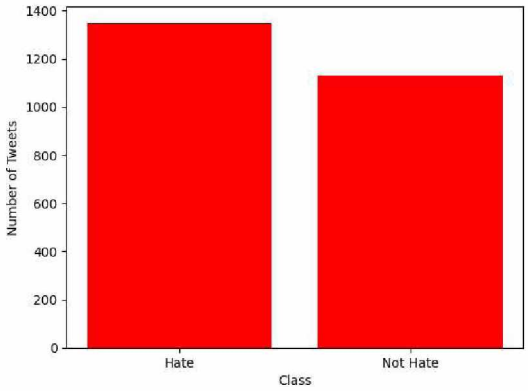


Fig. 2: Labels for the Proposed Dataset

In the annotation stage, two annotators were employed and trained to identify hate speech in the dataset using a set of guidelines. The annotators marked a tweet as hate speech if it targeted an individual or group based on characteristics such as race, ethnicity, religion, gender, or sexual orientation, or contained offensive language that was intended to offend or harm.

The annotation guidelines were developed based on prior research in the field and input from experts. The annotators were trained on the guidelines and provided feedback to ensure consistency and quality in annotations. The annotators were required to achieve high inter-annotator agreement to ensure the transcendence of the dataset.

Figure 2 displays the balance for our dataset for different classes of tweets.

C. Preprocessing

Preprocessing of the dataset involved the removal of usernames, URLs, and stop words, to protect privacy and prevent bias in model training. The preprocessed dataset was then split into training and testing sets. The training set was used for model training, while the testing set was used for evaluation.

Different machine learning models, such as logistic regression, support vector machines, random forests, and convolutional neural networks were trained using the annotated dataset. A combination of supervised and unsupervised learning techniques was utilized for model training to detect patterns in the data and make predictions.

The models were trained using various feature extraction and vectorization techniques, including bag-of-words, word embeddings, and character-level embeddings. The models were also trained using different hyperparameters to optimize their performance.

Figure 1 shows the end to end pipeline for labelling of tweets as hate or non hate using different machine and deep learning models.

D. Evaluation

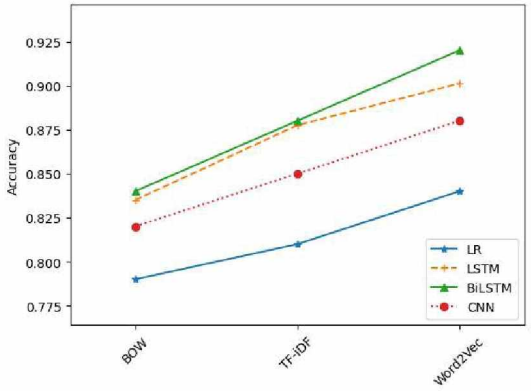


Fig. 3: Comparison of different Feature Extraction Techniques on Standard Models

The performance of the models was evaluated using metrics such as accuracy, precision, recall, and F1-score.

Standard comparision metrics like F1 Score, Accuracy etc. have been used to evaluate the models. The hold-out set of tweets, which were not used for training, was used to evaluate the models' ability to recognize hate speech and offensive language, and performance was compared across models.

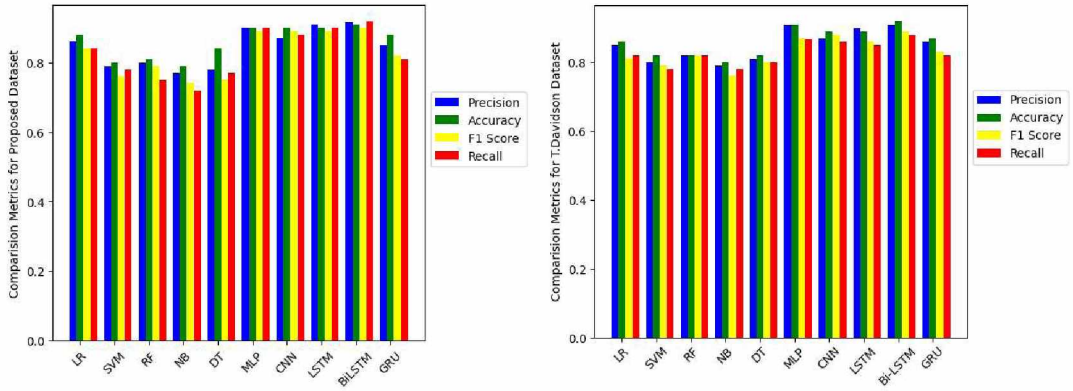
IV. Results

A. Model Training

In this study, we developed a hate speech dataset and trained various machine learning and deep learning models to identify hate speech in tweets. The dataset comprised of 2,478 tweets, with 45.56 % hate tweets on a set of guidelines by two annotators. After labelling, the dataset underwent preprocessing where usernames, URLs, and stop words were removed, and the data was split into training and testing sets. The machine learning models were trained using different feature extraction techniques and hyperparameters.

B. Comparison of Different Feature Extraction Techniques

Different extraction and vectorization techniques were used to train the machine learning models, including bag-of-words, word embeddings, and character-level embeddings. The performance of the models varied depending on the vectorization technique used. Figure 3 shows comparison of different vectorization techniques for standard models.



(a) (b)

Fig. 4: Comparison of Different Models using word embedding on (a) Proposed Dataset (b) T. Davidson Dataset

1) Bag-of-Words: Bag-of-words achieved an accuracy of 83.5% for LSTM model on the propsed datset and an accuracy of 85.73% for LSTM model on the T.Davidson dataset, after applying SMOTE.

2) TF-iDF: TF-iDf achieved an accuracy of 87.77% for LSTM model on the propsed datset and an accuracy of 88.21% for LSTM model on the T.Davidson dataset, after applying SMOTE.

3) Word Embeddings: Word embeddings used for this experimentation was Word2Vec, which achieved an accuracy of 90.12% for LSTM model on the propsed datset and an accuracy of 89.38% for LSTM model on the T.Davidson dataset, after applying SMOTE.

C. Performance of Different Machine Learning and Deep Learning Models:

The functioning of the learning models is a function of the feature extraction technique and model type used. The best results are obtained for word Embeddings and for Bi-LSTM architecture. The results for different models are as follows:

1) Logistic Regression (LR): achieved accuracies of 79%, 81%, 84% for BOW, TF-iDF and Word Embeddings correspondingly for the proposed dataset and 81%, 83%, 85% for T.Davidson dataset on BOW, TF-iDF and Word Embeddings correspondingly.

2) Support Vector Machines (SVM): achieved accuracy of 80% on word embeddings for the proposed dataset and 81% for T.Davidson dataset.

3) Random Forests (RF): achieved accuracy of 81% on word embeddings for the proposed dataset and 82% for T.Davidson dataset.

4) Naive Bayes (NB): achieved accuracy of 79% on word embeddings for the proposed dataset and 80% for T.Davidson dataset.

5) Decision Tree (DT): achieved accuracy of 84% on word embeddings for the proposed dataset and 82% for T.Davidson dataset.

6) Multi Layer Perceptron (MLP): achieved best accuracy of 90% on word embeddings for the proposed dataset and 91% for T.Davidson dataset.

7) Convolutional Neural Networks (CNN): achieved accuracy of 90% on word embeddings for the proposed dataset and 88% for T.Davidson dataset.

8) Long Short Term Memory (LSTM): achieved accuracy of 90% on word embeddings for the proposed dataset and 89% for T.Davidson dataset.

9) Bidirectional Long Short Term Memory (Bi-LSTM): achieved the prime accuracy, of 91% on word embeddings for the proposed dataset and 92% for T.Davidson dataset.

10) Gated Recurrent Unit (GRU): achieved accuracy of 88% on word embeddings for the proposed dataset and 87% for T.Davidson dataset.

Figure 4 shows comparison of various learning models using Word2Vec on a) Proposed Dataset b) T.Davidson Dataset. on various evaluation metrics.

V. Conclusion

This study employed a robust methodology to generate a hate speech dataset and analyze different models' performance in identifying hate speech. The results show that the models' performance varies depending on the feature extraction technique and model type used. The best results are obtained by Bi-LSTM model when trained for around 50 epochs using word embedings. And the results are consistent with that obtained with T.Davidson dataset. This study's findings are valuable for researchers and policymakers interested in developing hate speech detection models and can help guide the development of effective hate speech detection tools.

One limitation of this study is that the hate speech dataset was limited to English tweets with a limited number of tweets. Future research could expand the dataset to include tweets in other languages, as hate speech is a global issue. Additionally, the dataset could be expanded to include more examples of hate speech to improve the performance of the machine learning models.

Overall, this study contributes to the ongoing efforts to combat hate speech and promote online inclusivity. The development of a hate speech dataset and the training of machine learning models to identify hate speech in tweets provide valuable tools for social media platforms and policymakers to combat hate speech. The results of this study can be used to inform the development of automated hate speech detection systems that can help to make social media a safer and more inclusive space for all users.

VI. Ethical Considerations

Ethical considerations were taken into account by obtaining informed consent from annotators, ensuring anonymity, and taking measures to prevent misuse of the dataset. The hate speech dataset is potentially harmful if misused or leaked to unauthorized parties, and therefore it was necessary to take appropriate measures to protect the privacy and security of the dataset.

VII. Future Scope

Based on the findings of this study, there are various opportunities for future research. One important area is to tweak hate content flagging systems since different models have varying degrees of accuracy. Researchers can investigate more advanced machine learning models or hybrid approaches that integrate machine learning with other techniques such as rule-based systems or deep learning.

Another area for future research is to address the biases that exist in current models. The study demonstrated that particular demographic factors are linked to hate speech, and it is necessary to ensure that such biases are not perpetuated in detection systems. Researchers can explore the use of fairness metrics and techniques to reduce the effects of biases in hate speech detection.

Lastly, future research can investigate the impact of hate speech and offensive language on individuals and society. Studies can explore the psychological and social consequences of exposure to hate speech, as well as the broader societal implications of its prevalence on social media. Researchers can also study the effectiveness of interventions, such as counter-speech or education programs, in reducing hate speech and promoting tolerance.

Overall, there are significant opportunities for further research in hate speech detection and its effects on society. The hate speech dataset developed in this study can serve as a crucial resource for such research and can aid in advancing the development of effective and impartial hate speech detection systems.

References

[1] d'Sa, Ashwin Geet, Irina Illina, Dominique Fohr, and Awais Akbar. "Exploration of Multi-corpus Learning for Hate Speech Classification in Low Resource Scenarios." In Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings, pp. 238-250. Cham: Springer International Publishing, 2022.

[2] Alkomah, Fatimah, and Xiaogang Ma. "A Literature Review of Textual Hate Speech Detection Methods and Datasets." Information 13, no. 6 (2022): 273.

[3] Capozzi, Arthur TE, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. "A data viz platform as a support to study, analyze and understand the hate speech phenomenon." In Proceedings of the 2nd International Conference on Web Studies, pp. 28-35. 2018.

[4] Kapil, Prashant, and Asif Ekbal. "A deep neural network based multi-task learning approach to hate speech detection." Knowledge-Based Systems 210 (2020): 106458.

[5] Roy, Pradeep Kumar, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 (2020): 204951-204962.

[6] Corazza, Michele, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. "A multilingual evaluation for online hate speech detection." ACM Transactions on Internet Technology (TOIT) 20, no. 2 (2020): 1-22.

[7] Raufi, Bujar, and Ildi Xhaferri. "Application of machine learning techniques for hate speech detection in mobile applications." In 2018 International Conference on Information Technologies (InfoTech), pp. 1-4. IEEE, 2018.

[8] Koushik, Garima, K. Rajeswari, and Suresh Kannan Muthusamy. "Automated hate speech detection on Twitter." In 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), pp. 1-4. IEEE, 2019.

[9] Al-Makhadmeh, Zafer, and Amr Tolba. "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach." Computing 102 (2020): 501-522.

[10] Kovács, György, Pedro Alonso, and Rajkumar Saini. "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources." SN Computer Science 2 (2021): 1-15.

[11] Plaza-del-Arco, Flor Miriam, M. Dolores Molina-González, L. Alfonso Urena-López, and M. Teresa Martín-Valdivia. "Comparing pre-trained language models for Spanish hate speech detection." Expert Systems with Applications 166 (2021): 114120.

[12] Alsafari, Safa, Samira Sadaoui, and Malek Mouhoub. "Deep learning ensembles for hate speech detection." In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 526-531. IEEE, 2020.

[13] Vidgen, Bertie, and Taha Yasseri. "Detecting weak and strong Islamophobic hate speech on social media." Journal of Information Technology Politics 17, no. 1 (2020): 66-78.