The Effect of Geography on Hate Speech: A Study in Computational Linguistics

Monoshi Kumar Roy
Department of Computer Science
Iowa State University
Ames, Iowa
monoshi@iastate.edu

Anwar Hossain Zahid
Department of Computer Science
Iowa State University
Ames, Iowa
ahzahid@iastate.edu

***Abstract*—Hate speech has become a pressing issue in the digital age, with social media platforms providing a global platform for disseminating such content. For example, in the last couple of years, social media like Facebook has been extensively used in subcontinent countries(India, Bangladesh, Pakistan, etc.) to spread blasphemous speech and communal riots. However, the definition of hate speech is not universal and varies based on cultural and geographical contexts. Firstly, our project tries to identify whether a comment is hate speech in the first place. After that, we also explore the effect of geography on hate speech by building a model that predicts the likelihood of a text being regarded as hate speech in different geographical locations.**

***Index Terms*—Hate Speech, Geographic Specific Hate Speech, Natural Language Processing, Machine Learning**

I. Introduction

Hate speech is a form of communication that aims to degrade, intimidate, or dehumanize individuals or groups based on their identity. It can be expressed through different means, such as speech, writing, or gestures, and can have severe consequences for its targets. Hate speech is a complex phenomenon dependent on several factors, such as cultural context, history, demography, and geographical location. In recent years, social media has become a dominant platform for communication and has made it easier for people to express their opinions. And with the rise of social media platforms, hate speech has become more prevalent and has increased hate crimes, discrimination, and violence. Therefore, developing tools to detect hate speech and prevent its spread is crucial.

In this project, we aim to analyze the comments of different individuals on various social media platforms (Facebook, YouTube, Twitter) in five different languages and use the natural language processing pipeline to train the machine learning models and predict whether a given comment is hateful.

Hate speech is not constant in context to the location, and a text that ignites hate in one location may not have the same effect in another. For example, in the Indian subcontinent, people tend to be very protective regarding their religion. They get offended if someone criticizes their belief or religion. But this may be different in many parts of the world. Therefore, geographical information can help improve the accuracy of hate speech detection. In this paper, we aim to study the effect of geography on hate speech and develop a model that can predict the likelihood of a text being regarded as hate speech in different geographical locations. Our model is based on machine learning algorithms that use textual features and geographical information to identify hate speech. We conducted experiments on a dataset consisting of text data from different locations and evaluated the performance of our model using various metrics. The results show that our model achieves a high level of accuracy in predicting hate speech in different geographical locations, indicating the potential of our approach to detect hate speech in a social context.

II. Related Work

Detecting hate speech in natural language processing is a relatively new field. Previous research has focused on developing machine learning algorithms that identify hate speech based on various linguistic features, such as profanity, slurs, and aggressive language. As the popularity of social media platforms is increasing, researchers are also trying to develop methods to prevent hate speech across these platforms. Now, numerous Machine Learning(ML) methods are available to classify hate speech in a comment. However, many researchers are also trying to understand the interpretability of the results given by ML models [1]. Another problem regarding hate speech detection is generalizing models and reducing the overfitting problems of ML models. Again, while detecting hate speech, ML models should also take into account the free speech right of every human being [2]. Deep learning models have also become popular for processing large amounts of posts and detecting whether those have any hate speech [3]. These studies have not considered the effect of geography on hate speech. However, only a few studies have focused on the effect of geography on hate speech. One study analyzed hate speech in tweets and found that hate speech was more prevalent in the United States compared to other countries [4]. Another study by Schmidt and Wiegand (2017) explored the cross-lingual detection of hate speech and found that the context of hate speech varies across languages [5]. Researchers are also trying to analyze the behavior of general people towards different social groups, such as Immigrants, Minorities, LGBTQ, etc., based on Twitter data and using only a lexicon-based approach [11].

978-1-6654-0906-3/21/$31.00 ©2021 IEEE

III. Methodology

Our study aims to build a model that predicts the likelihood of a text being regarded as hate speech in different geographical locations. To achieve our goal, we accomplished a set of works. They are given below:

A. Data Collection and Preprocessing

As we wanted to detect hate speech on the basis of different geographical locations, and so building a model, we needed a dataset consisting of text data labeled with geographical locations. But there was no available dataset for our purpose. To make such dataset, we decided to develop a dataset using available hate speech datasets in various languages, such as Arabic, Bangla, Chinese, Hindi, and Russian [6]- [9]. Most of the data are collected from social media platform, such as twitter, and already labeled as hate speech or non-hate speech based on a set of predefined criteria. We then labeled the data with geographical locations. Our dataset has five features, Comments, Translated_Comments, Category, Hate Speech, and Geography. We collected a total of 9988 samples across mentioned five languages, and our target attributes were both Hate Speech(binary) and Geography. We cleaned the data to remove noise from the text data including HTML tags, punctuation, stop words, and non-alphabetic characters, and made it ready for feature Extraction.

B. Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) [10] is a popular algorithm for extracting features from text data. It works by tokenizing the text data into individual words or n-grams and then calculating the frequency of each term. Once the TF-IDF scores are calculated for each term, they can be used as features for machine learning algorithms such as binary classifiers. So, we used TF-IDF algorithm to convert the collection of raw text data to a matrix of TF-IDF features. These features can help the classifier to distinguish between different classes of documents. However, not all features generated from word embeddings may be useful for predicting the target attribute, and feature selection techniques such as Chi-square, mutual information, and feature importance can be used to select the most relevant features. Therefore, we drop the Category attribute and decided to use it for future research. After performing all these things, the final dataset is now ready to be used for training and testing our model. But before that we have to select appropriate model to serve our purpose.

C. Model Selection

Model selection is a little bit trickier part, as there are many models that can be used for binary classification in NLP, including logistic regression, support vector machines, decision trees, and neural networks. In our experiment, we used Logistic Regression Classifier. We know that Logistic Regression Classifier is less prone to over-fitting than other models like decision trees and neural networks, so, using this classifier is a good choice since we are dealing with limited training data.

D. Model Training

After choosing the classifier, we are ready to train the model with our dataset. As we're trying to develop generalized Hate Speech Detector and demonstrate the impact of geography on Hate Speech Detection, hence, we're required to train two separate model. Our first model can be called generalized hate speech detection model. And our second model will evaluate the possibility of a comment being regarded as a hate speech in a particular geographical location, given that the comment is a hate speech. Ensembling these two models, we can achieve our generalized hate speech detector. Before starting the training, we split the dataset into training and testing sets. We used 80% of data for training and 20% of data for testing purpose. We also used k-fold cross-validation technique to ensure the model's generalization performance. In our experiment, we used 5-fold validation to get generalize performance of our models.

IV. Experimental Result

A. Model Testing

We trained our models with Logistic Regression classifier on the transformed features, with L2 regularization and default hyperparameters. Our first model achieved 73% accuracy on the test set. Under the ethical consideration, we have to achieve minimum false positive result from our model, otherwise, it might violates the freedom of speech. Considering this phenomena, we got the 83.21% precession, which is quite good enough to deal with and train the model with limited amount of data. Our Second model achieved 93% accuracy on the test set. The confusion matrix showed that both of the models had a high precision and recall for the target classes, indicating a balanced performance.

B. Evaluation Report

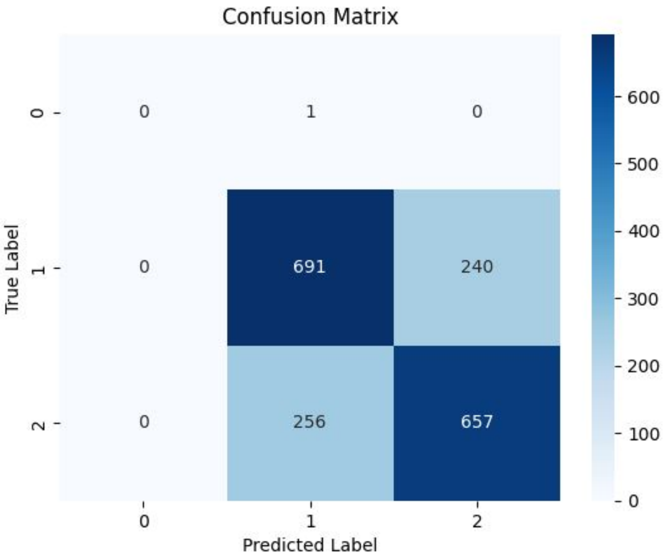

Fig. 1: Confusion Matrix for Generalized Hate Detection

According to the confusion matrix, our binary classifier successfully detected 1,384 out of 1,8800 hate speech comments, whereas 240 out of 1,8800 were false positives, yielding an accuracy of 73%. We obtained this result using Logistic Regression, SVM with Linear and RBF kernel respectively. We expected SVM to provide better results than Logistic Regression but the mentioned algorithms performed uniformly on our dataset.

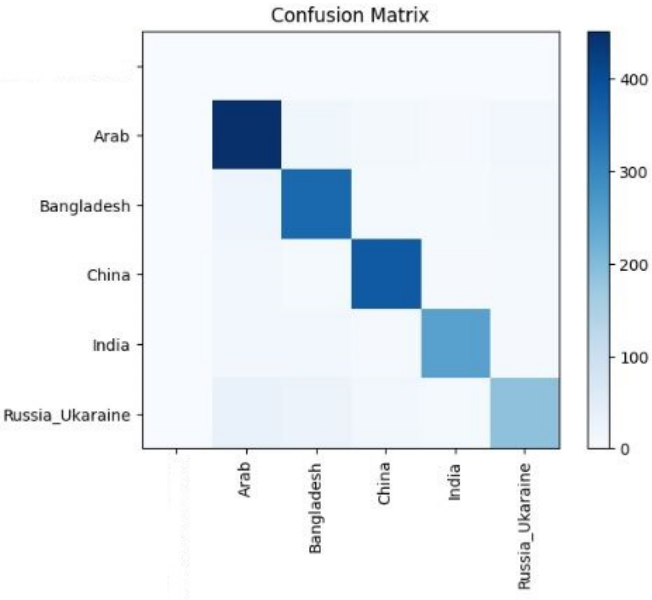

Fig. 2: Confusion Matrix for Location Detection

The confusion matrix for the second model demonstrates that our is highly effective in identifying hate speech comments on different location, given that a comment is detected as hate speech, with a high degree of accuracy and discrimination.

TABLE I: Performance of different classifiers on Hate Speech detection model

Evaluation Matrics	Logistic Regression	SVM (Linear)	SVM(RBF)
F1 Score	0.730	0.728	0.731
Precision	0.731	0.728	0.736
Recall	0.487	0.483	0.491
Accuracy	0.731	0.728	0.735

Table 1 demonstrates the comparison of the performance of three different classifiers. The results suggest that the SVM with an RBF kernel outperforms the other two classifiers in terms of precision, recall, and accuracy, while the Linear Regression and SVM with a linear kernel have similar performance across all metrics. However, it's important to note that the performance of the classifiers can vary depending on the dataset and the specific problem being addressed, so, as we're dealing with limited text data, and want to avoid overfitting, hence Logistics Regression was used to train our first model.

TABLE II: Performance of Logistic Regression on Geography Model

Location	Precision	Recall	F1 Score
Arab	0.85	0.91	0.88
Bangladesh	0.85	0.89	0..87
China	0.92	0.92	0.92
India	0.93	0.88	0.90
Russia_Ukaraine	0.84	0.72	0.78

TABLE III: Performance of SVM (Linear) on Geography Model

Location	Precision	Recall	F1 Score
Arab	0.84	0.92	0.87
Bangladesh	0.88	0.90	0..89
China	0.94	0.93	0.93
India	0.94	0.88	0.91
Russia_Ukaraine	0.83	0.73	0.78

TABLE IV: Performance of SVM (RBF) on Geography

Location	Precision	Recall	F1 Score
Arab	0.84	0.93	0.88
Bangladesh	0.84	0.91	0..88
China	0.93	0.86	0.92
India	0.93	0.86	0.89
Russia_Ukaraine	0.86	0.69	0.76

Tables 2, 3, and 4 compare the performance of three alternative classifiers on the Geography detection model. It is worth noting that the SVM with RBF kernel classifier fared the worst of all. However, the results suggest that the choice of classifier can have an impact on the performance of the geography model , and we used the logistic regression classifier for exactly the same reason as the first model.

V. Limitations

While the model showed promising results in detecting hate speech in different location, there are limitations due to the dataset and methodology used. The dataset used for training the model was limited and only contained a small number of comments. This may have limited the model's ability to generalize to new and unseen data. Our model is not capturing important linguistic features, such as, part-of-speech tagging, sentiment analysis, or named entity recognition that could have improved its performance. The dataset only contained comments in a few languages, namely Arabic, Bangla, Chinese, Hindi, and Russian. The model's performance may have been limited when applied to comments in other languages. Due to the limited size of the dataset, there is a risk of overfitting the model to the training data. This means that the model may perform well on the training data but poorly on new, unseen data.

VI. Conclusion and Future Work

Our study highlights the importance of considering the effect of geography on hate speech detection. The linguistic features that indicate hate speech are not universal and vary based on cultural and geographical contexts. By building a model that considers the linguistic features that indicate hate speech in different geographical locations, we can better understand the cultural and contextual factors that contribute to hate speech. In this paper, we have demonstrated the potential of using geographical information to improve hate speech detection. We recommend that hate speech detection models should be trained on data from different geographical locations to improve their performance. Our model provides a framework for identifying hate speech in different geographical locations and can be used to develop targeted interventions to combat hate speech. Our model achieves a high level of accuracy in predicting hate speech in different geographical locations, indicating the effectiveness of our approach. Our results also suggest that hate speech is not constant in context to the location and that cultural context and historical background play a crucial role in the spread of hate speech. Our approach can help prevent the spread of hate speech by identifying it in its early stages and taking appropriate actions to prevent it.

In our future work, we will try to overcome all the limitations by using larger and more diverse datasets, performing linguistic analysis, and considering ethical considerations when developing the model. Our dataset and codes are available here.

References

[1] MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. “Hate Speech Detection: Challenges and Solutions.” PLOS ONE 14, no. 8 (2019). https://doi.org/10.1371/journal.pone.0221152.

[2] Yin, Wenjie, and Arkaitz Zubiaga. “Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions.” PeerJ Computer Science 7 (2021). https://doi.org/10.7717/peerj-cs.598.

[3] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, ”A Framework for Hate Speech Detection Using Deep Convolutional Neural Network,” in IEEE Access, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073.

[4] Waseem, Zeerak, and Dirk Hovy. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” Proceedings of the NAACL Student Research Workshop, 2016. https://doi.org/10.18653/v1/n16-2013.

[5] Schmidt, Anna, and Michael Wiegand. “A Survey on Hate Speech Detection Using Natural Language Processing.” Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017. https://doi.org/10.18653/v1/w17-1101.

[6] Lingiardi, Vittorio, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. “Mapping Twitter Hate Speech towards Social and Sexual Minorities: A Lexicon-Based Approach to Semantic Content Analysis.” Behaviour amp; Information Technology 39, no. 7 (2019): 711–21. https://doi.org/10.1080/0144929x.2019.1607903.

[7] Romim, Nauros, Mosahed Ahmed, Hriteshwar Talukder, and Md. Saiful Islam. “Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation.” Algorithms for Intelligent Systems, 2021, 457–68. https://doi.org/10.1007/978-981-16-0586-4_37.

[8] Jiang, Aiqi, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. “SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection.” Online Social Networks and Media 27 (2022): 100182. https://doi.org/10.1016/j.osnem.2021.100182.

[9] Alakrot, Azalden, Liam Murray, and Nikola S. Nikolov. “Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic.” Procedia Computer Science 142 (2018): 174–81. https://doi.org/10.1016/j.procs.2018.10.473.

[10] Andrusyak, B., Rimel, M., & Kern, R. Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages. In 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT) (pp. 236-241).

[11] Ramos, Juan. ”Using tf-idf to determine word relevance in document queries.” Proceedings of the first instructional conference on machine learning. Vol. 242. No. 1. 2003.