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

Presented By

Monoshi Kumar Roy
Department of Computer Science
Iowa State University

Anwar Hossain Zahid
Department of Computer Science
Iowa State University

Presentation Outline

- Introduction
- Motivation
- Objective
- Related work
- Methodology
- Experimental Result
- Limitations
- Conclusions
- Future Work

Introduction

What is Hate Speech?

- Comment that aims to degrade, intimidate or dehumanize an individual group of people based on their identity, race, gender, disability or sexual orientation.
- Can be conveyed through different means like gesture, written comments or oral speech.
- Social media platforms are being used extensively to propagate hate speech and it has become more prevalent.

Motivation

- Social media platforms are extensively used in subcontinent to spread communal riots.
- Communal riots cause the violation of human rights, economic loss and even death.
- In India, from 2004-2017, around 1600 people were killed in communal riots. (NDTV news India)
- In Bangladesh, 7 people were killed and more than 150 were injured during the Durga Puja communal riot (The Guardian).

Objective

- Our main objective is to detect whether a comment is hate speech.
- We also want to explore the relationship between hate speech and specific geography.
- Automated detection of hate speech and the geographic information of that hate speech.

Related Work

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.

- Introduced a novel tweet annotated dataset for hate speech.
- Experimented with different machine learning models such as Linear Regression, Support Vector Machine.
- Considered only twitter data and showed twitter hate speech is more prevalent in the US.

Related Work

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.

- Explored cross lingual detection of hate speech.
- Reviewed different types of feature representation like lexical, syntactic and semantic.
- Found that the context of hate speech varies across languages

Methodology

Dataset Preparation

- There was no available dataset for our purpose
- We developed a dataset using available hate speech datasets in various languages
- We then labeled the data with geographical locations
- Our dataset has five features, Comments, Translated Comments, Category, Hate Speech, and Geography

Methodology

Feature Extraction

- Cleaned the data to remove noise from the text data including HTML tags, punctuation, stop words
- Used TF-IDF (Term Frequency-Inverse Document Frequency) algorithm to convert the collection of raw text data to a matrix of TF-IDF features
- Dropped the Category attribute and decided to use it for future research.

Methodology

Model Selection

- As we are dealing with limited training data, we decided to use Linear Regression, Support Vector Machine, Classifier which is less prone to overfitting than other models like decision trees and neural networks.
- We are considering two Machine Learning Models here, One is for hate speech detection, and the other one is for Location given the speech is hateful.

Model 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.

Experimental Result

Model Testing

- We trained a Logistic Regression model on the transformed features, with L2 regularization and default hyperparameters.
- Our generalized hate detection model achieved an accuracy of 73% on the test set.
- Our location detection model achieved an accuracy of 93% on the test set.

Experimental Result

Table 1: Evaluation Metrics on ML Models

Evaluation Metric	**ML Models**		
	Linear Regression	SVM (Linear Kernel)	SVM (RBF Kernel)
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

Experimental Result

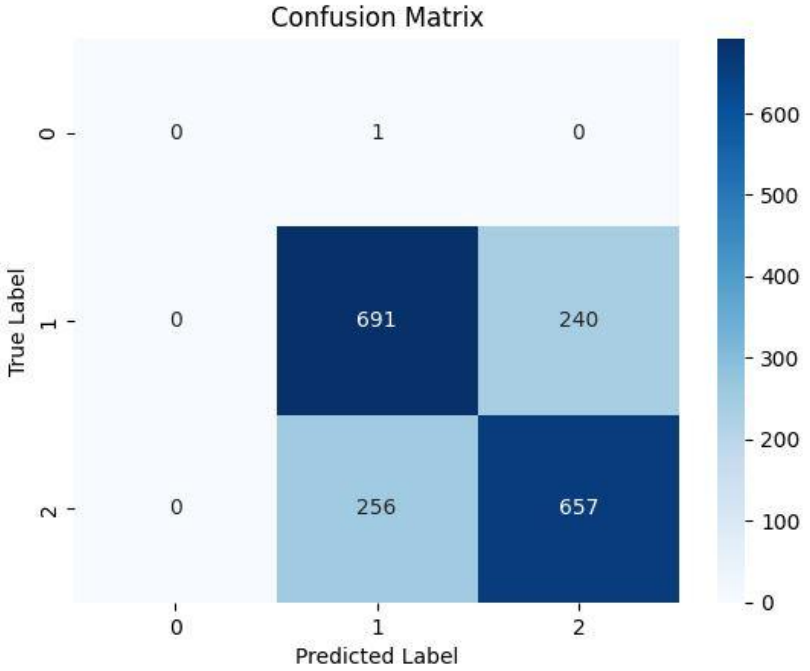

Figure 1: Confusion Matrix for Model 1

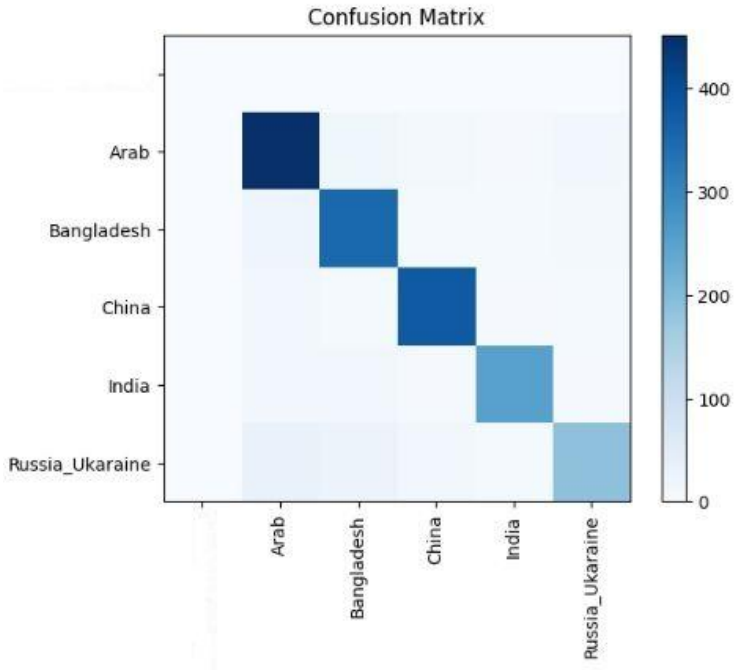

Figure 2: Confusion Matrix for Model 2

Limitations

Limited Dataset: Having risk of overfitting

Limited Scope: Model's performance may have been limited when applied to texts in other languages.

No linguistic analysis: Linguistic analysis such as part-of-speech tagging, sentiment analysis, or named entity recognition

Free Speech Right: Another import aspect is that we don’t consider the free speech right of a human being.

Conclusion

- Our study demonstrates the potential of using geographical information to improve hate speech detection.
- Our study also shows that we can develop generalized hate speech detector.
- 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.

Future Works

- Use larger and more diverse datasets
- Perform linguistic analysis
- Consider ethical considerations


Any
Question