IMPROVING HATE SPEECH DETECTION USING MACHINE LEARNING: A PRELIMINARY STUDY

Jawaid Ahmed Siddiqui[1], Siti Sophiayati Yuhaniz[2], Zulfiqar Ali Memon[3], Yumna Amin[3]

[1] Sukkur IBA University Airport road, Sukkur (Sindh) Pakistan [2]Razak Faculty of Technology and Informatics Universiti Teknologi Malaysia [3] FAST National University of Computer and Emerging Sciences, Pakistan

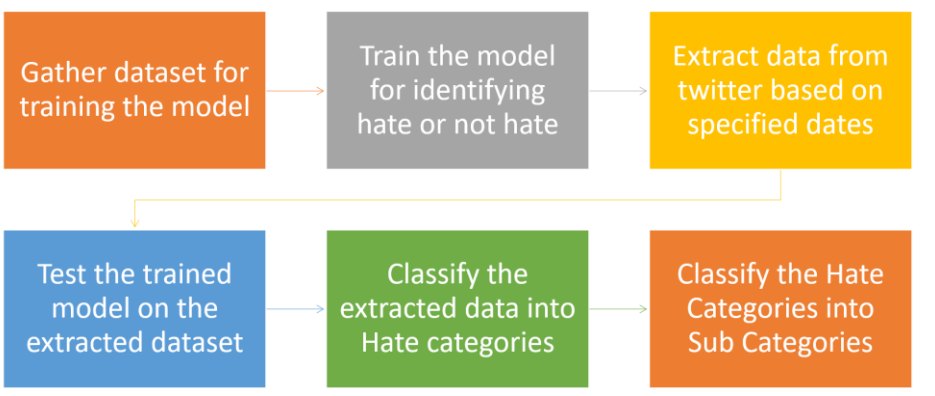
ABSTRACT

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/deep learning algorithms to automatically detect the hate speech messages on different datasets. However, to the best of our knowledge, most of the studies classify the hate speech related message using existing feature engineering approaches and suffer from the low classification results. This is because, the existing feature engineering approaches suffer from the word order problem and word context problem. In this research work we will identify hateful content from latest tweets of twitter and classify them into these top Categories: Ethnicity, Nationality, Religion, Gender, Sexual Orientation, Disability and Other. These categories are further classified to identify the targets of hate speech e.g. Black, White, Asian belongs to Ethnicity and Muslims, Jews, Christians can be classified from Religion Category. An evaluation will be performed among the hateful content identified using deep learning model LSTM and traditional machine learning models which includes Linear SVC, Logistic Regression, Random Forest and Multinomial Naïve Bayes to measure their accuracy and precision and their comparison on the live extracted tweets from twitter which will be used as our test dataset.

LITERATURE REVIEW

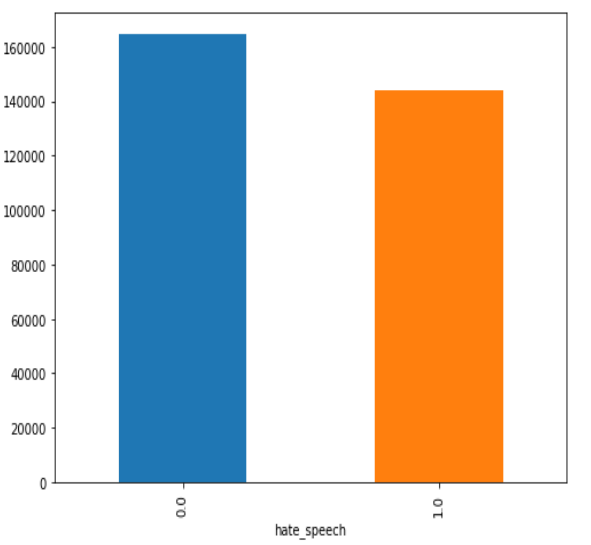
Research Paper	Proposed Technique	Results
Vega, L. E. A., Reyes-Magaña, J. C., Gómez-Adorno, H., & Bel-Enguix, G. (2019, June). MineriaUNAM at SemEval-2019 Task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 447-452).	• Spanish & English Tweets • SVM & RF(Target)	• Spanish Dev: 0.73 Evaluation: 0.596 • English Dev: 0.54 Eva: 0.368
Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759-760).	• comparison of deep learning methods • 16K annotated tweets • TF-IDF and N-gram • LSTM + Random Embedding + GBDT	• LSTM + Random Embedding + GBDT 0.93 • TF-IDF + SVM 0.816

METHODOLOGY



RESULTS

Model	Score
Linear SVC	0.92
Random Forest	0.95
Logistic Regression	0.88
Multinomial Naïve Bayes	0.85



CONCLUSIONS

In our research, tweets are extracted from the twitter by specifying the date range and some vocabulary list. The extracted tweets are more than 235k on which different experiments are performed to detect hate speech which includes deep learning model i.e. LSTM and traditional machine learning models including linear Support Vector Classification (LSVC), Random Forest (RF), Logistic Regression (LR) and Multinomial Naïve Bayes (MNB). For hate speech detection Logistic Regression performed the best 0.74 followed by Random Forest 0.72. For Hate speech Categorization LSVC performed the best and gave 0.85 accuracy but RF gave slight better categorization results. For sub categorization, Random Forest gave best result i.e. 0.95 accuracy.