Social Media Hate and Offensive Speech Detection Using Machine Learning Method

Girma Yohannis Bade, Olga Kolesnikova , Grigori Sidorov,
José Luis Oropeza
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Mexico City, Mexico
Correspondence : girme2005@gmail.com

Abstract

Even though the improper use of social media is increasing nowadays, there is also technology that brings solutions. Here, improperness is posting hate and offensive speech that might harm an individual or group. Hate speech refers to an insult toward an individual or group based on their identities. Spreading it on social media platforms is a serious problem for society. The solution, on the other hand, is the availability of natural language processing(NLP) technology that is capable to detect and handle such problems. This paper presents the detection of social media's hate and offensive speech in the code-mixed Telugu language. For this, the task and golden standard dataset were provided for us by the shared task organizer (DravidianLangTech@EACL 2024)[1]. To this end, we have employed the TF-IDF technique for numeric feature extraction and used a random forest algorithm for modeling hate speech detection. Finally, the developed model was evaluated on the test dataset and achieved 0.492 macro-F1.

1 Introduction

The growth of communication technology over the past few decades has resulted in a ballooning user active participation on social media. Social media is highly utilized for a wide range of activities, including news, business, advertising, etc. However, it simultaneously raises hate and offensive speech (Saleh et al., 2023). One of the reasons for this prevalence is users post improper information on social media. Hate speech on the social media platform can be in the form of text, images, or videos. The text mode, particularly is the most prevalent type of harmful content on social media (Bade and Seid, 2018). Hate speech refers to an insult that is aimed toward an individual or group based on identities including race, gender, minorities, political parties, religion, nationality, and public figures (Yigezu et al., 2023d). Today, several federal and international organizations pledged to combat hate speech online (Yasaswini et al., 2021; Ghanghor et al., 2021). However, many communities use multiple languages and mix their opinion in text mode, so the identification becomes complicated manually. Telugu, one of the Dravidian languages experiences code-mixed practice and is subjected to this complication. Code mixing is the mingling of two or more languages, and it can be difficult to identify toxicity in the multilingual statements (Priyadharshini et al., 2023; Yigezu et al., 2023c). In this regard, a shared task(DravidianLangTech@EACL 2024) opened a door to participate in the detection of Telugu social media hate and offensive speech by providing golden standard datasets. This shared task offered an opportunity for researchers to come up with solutions leveraging existing technology to identify hate speech and objectionable pieces of information. This study aims to determine whether a given comment in code-mixed Telugu language contains hate and offensive content and anticipated that the study will improve the detecting efficiency and handle all aspects of language.

2 Related Works

Hate speech spreading on the internet is a serious problem for society, and platforms need to identify objectionable information (Okechukwu et al., 2023). Numerous studies have been conducted to identify hate speech using different approaches with different levels of performance measures (Okechukwu et al., 2023). Hate speech recognition work has been modeled in research as a text classification issue and determines a message's classes from its text as hate speech or non-hate (Madhu et al., 2023). The study (Al-Dabet et al., 2023) presents a transformer-based method to deal with the problem of offensive speech detection.

[1] https://codalab.lisn.upsaclay.fr/competitions/16095

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 240–244
March 22, 2024 ©2024 Association for Computational Linguistics

This model was validated using a combination of four benchmark Twitter Arabic datasets annotated for hate speech detection tasks including the workshop (OSACT5 2022) shared task dataset. The demonstrated model was able to recognize offensive speech in Arabic tweets with 87.15% accuracy and 83.6 % F1 score. Similarly, in the work conducted (Okechukwu et al., 2023), hate speech was detected using the Term Frequency-Inverse Document Frequency (TF-IDF) with a majority voting ensemble learning classification Model. The model's accuracy was 95%, and its F-Measure was 95%. It made use of a Kaggle.com dataset that was accessible to the public.

In the study (Abbes et al., 2023), a deep learning method for identifying harmful and hostile content on Arabic social media platforms like Facebook was proposed. The researchers collected 2,000 Facebook comments in the Tunisian dialect and created two models: a Bi-LSTM based on an attention mechanism combining the BERT for Facebook comment classification toward hate speech detection. After evaluating the suggested model, an accuracy of 98.89% was attained. The researchers used a transformer-based model in (Bilal et al., 2023) to categorize hate speech in Roman Urdu. Furthermore, the first Roman Urdu pre-trained BERT model, known as BERT-RU, was created in this work. This research utilized the BERT's capabilities starting from scratch and trained on the biggest Roman Urdu dataset, which consists of 173,714 text messages. With scores of 96.70%, 97.25%, 96.74%, and 97.89% in accuracy, precision, recall, and F-measure, respectively, the created transformer-based model has met the performance metrics.

3 System Description

In this section, we offer thorough information regarding the dataset and the details of experimental tools . Moreover, it dives into the format of datasets, preprocessing, and the experimental details.

3.1 Datasets

In the real world, the problems are always existing until the solutions are investigated. To investigate solutions for computational linguistic challenges, the availability of data is crucial (Bade, 2021; Bade and Afaro, 2018). The dataset for this particular task was provided on Codalab by the Shared_task (DravidianLangTech@EACL 2024) organizer (Premjth et al., 2024). The dataset is arranged in three different lists training, development, and test set. The training and development data sets are made available when we register for the competition on the Codalab and the test set was released when ten days left for the run submission deadline.

Table 1: Sample data statistics in both training and test data of Telugu language

Text	Label	Dataset lists	# of records
Jagan meeda jaganke visvasam ledu anduke Students tho adukovtam thappu Gudivada king true leader	hate non-hate non hate	training	4000
Anna gurinchi chili excellent ga cheppindi Arey budder khan nuvvu asalu	— —	Testing	500

Table 1 shows sample instances of both training and testing data, the class feature or label, and the record size in both lists. Telugu uses Arabic scripts in addition to Latin but the table skipped the Arabic text to sample due to Unicode issue.

3.2 Preprocessing

Preprocessing is the process of preparing raw data for machine learning algorithms by cleaning, converting, and organizing the data rendering it to the machine. It is the vital stage that fills in the gaps between raw data and useful insights because raw data is rarely in an ideal state (Tonja et al., 2022). During the data preparation phase of machine learning tasks, there are typical or standard activities that we should use. The following are some among others.

Importing dependency libraries:- There are two libraries that we must always bring in. A library containing mathematical functions is called NumPy and the library used to import and manage the 'CSV' data sets is called Pandas.

Loading the data set:- In most cases, data sets are offered in a csv format. Tabular data is stored in plain text in a CSV file. In a file, every line represents a data record. To read a local CSV file as a data frame, the pandas library's (read_csv) function was utilized.

Handling Missing Data:- In real-world datasets, handling missing data is a prevalent difficulty. Preprocessing methods like imputation and the removal of missing data or null values ensure that the model is fed accurate and comprehensive data. For a variety of reasons, data may be missing, and it must be handled to prevent our machine-learning model from performing worse (Tash et al.). In addition, we used "raw['category'].fillna(0, inplace=True)" to handle empty strings of class label.

Data Cleaning:- is finding and fixing inaccuracies or flaws in the data (Yigezu et al., 2023b). In this regard, researchers explored the dataset listing and applied all needed.

Handling Outliers:- Anomalies that drastically depart from the average might cause distortions in learning. Preprocessing techniques such as transformation or scaling lessen the negative effects of outliers on model performance (Shahiki-Tash et al., 2023).

Data Encoding:- Since machine learning algorithms usually operate on numerical data, it is necessary to properly encode our text inputs in numerical equivalent (Yigezu et al., 2023a). To do so we have specifically used the TF-IDF text vectorization technique. It preserves the semantics and instance positions in addition to converting the provided text into a numeric representation (Yigezu et al., 2023e). However, in the case of converting 'class label', we used the "to_numeric()" function as "raw['category'] = pd.to_numeric(raw['category'], errors='coerce')".

3.3 Model Selection and Experimentation

The selected machine learnig model for this study is random forest.This is because several decision trees are combined in random forest, an ensemble learning technique to produce predictions that are more reliable and accurate. In a random forest, every decision tree is trained using a random subset of features and a random subset of the data (bootstrap samples). The diversity among the individual trees is increased and overfitting is lessened by this randomization (Yigezu et al., 2023b). During prediction, the ultimate result is established by combining all of the trees' predictions, either by average (for regression) or by majority voting (for classification). The capacity to manage complicated datasets, high-dimensional data, and non-linear interactions is a well-known feature of random forests. They are also frequently utilized in machine learning applications and are less prone to overfitting than a single decision tree (Destaw et al., 2022).

Experimental setup:- This section discusses the details of the developmental tool and the dependency libraries we used. For this research, we used Jupyter Notebook3 which is the Integrated Development Environment(IDE) of Python. After the tool setup was finished, we imported the four basic dependency libraries known as pandas, TfidfVectorizer, RandomForest, Joblib. Among those, the first three(pandas, TfidfVectorizer, RandomForest) are found in the Sklearn module. At the usage level, Pandas library is used to read CSV files from the local drive to a Python-run environment, TfidfVectorizer is for converting text data inputs into a numerical representation, and RandomForest is the principal algorithm to train the input data based on the predefined class. Finally, joblib which is a standalone module for saving the trained model for later use.

4 Result and Discussion

The Random Forest algorithm-based model was developed and classified the test dataset into two classes as they are presented in training data.

Table 2: Class label test data overview of manually or by annotator classified and machine or our model classified classification distribution.

Class	Manually classified	**Machine classified**
Non-hate	250	**375**
Hate	250	**125**
Total	500	**500**

As we can see from Table 2, our model classified 125 instances of the class label "Hate" as a "Non-hate" incorrectly. It also indicates that the model is more biased toward the 'non-hate' category. The Figure 1 shows in more detail below.

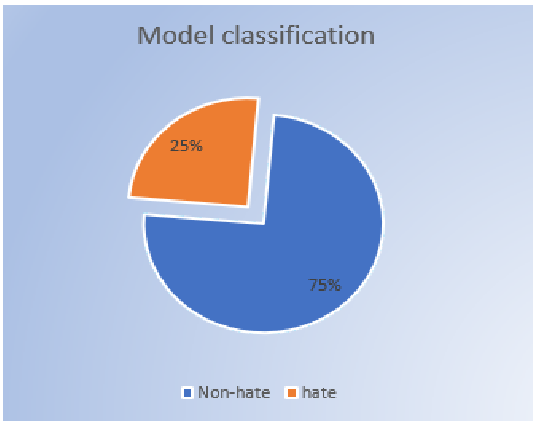

Figure 1: The diagrammatical representation of how the model classified the given test datasets.

According to the result published by the organizer, the model has also been evaluated in terms of macro-F1 scores to assess its performance and performed a 0.4921 macro-F1 score on the test dataset.

5 Conclusion

In this particular task, we have developed a model to classify social media posts into two binary classes hate and non-hate. The model has used the Random Forest algorithm method. The numeric features are extracted using TF-IDF techniques. The newly developed model has been evaluated with the new unseen test dataset and less performed on a selected algorithm for Telugu language text data.

6 Future work

Since social media posts that detect the posts of improper speech are critical, the jobs ought to be transferred into other various languages. Furthermore, by offering additional algorithms for the languages utilized here and expanding the number of dataset sizes, the performance of the suggested model in this study should be enhanced.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20231567, and 20232080 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

7 Limitation and Ethics Statement

Finding words outside of one's lexicon or linguistic occurrences that were not taken into consideration during preprocessing are limitations. Code-mixing can bring linguistic variances that the current language processing algorithms may not be able to handle well enough, which could result in incorrect classifications. Future studies could improve the model's performance and generalization capacities by addressing these linguistic issues. Notably, out of all the participating systems, our method achieved the 24^{th} rank in the shared job. Our model performs well in classifying hate and offensive comments in code-mixed text, even in the face of competition from other participants and obstacles in the competition. Furthermore, our work obeyed the computational ethics[2].

References

Mariem Abbes, Zied Kechaou, and Adel M. Alimi. 2023. Deep learning approach for Tunisian hate Speech detection on Facebook. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 739–744.

Saja Al-Dabet, Ahmed ElMassry, Ban Alomar, and Abdullah Alshamsi. 2023. Transformer-based Arabic Offensive Speech Detection. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6.

Girma Yohannis Bade. 2021. Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia. *Journal of Computer Science Research*, 3(4):26–30.

Girma Yohannis Bade and Akalu Assefa Afaro. 2018. Object Oriented Software Development for Artificial Intelligence. *American Journal of Software Engineering and Applications*, 7(2):22–24.

Girma Yohannis Bade and Hussien Seid. 2018. Development of Longest-Match Based Stemmer for Texts of Wolaita Language. *vol*, 4:79–83.

Muhammad Bilal, Atif Khan, Salman Jan, Shahrulniza Musa, and Shaukat Ali. 2023. Roman Urdu hate speech detection using transformer-based model for cyber security applications. *Sensors*, 23(8):3909.

Tadesse Destaw, Seid Muhie Yimam, Abinew Ayele, and Chris Biemann. 2022. Question answering classification for Amharic social media community based questions. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 137–145, Marseille, France. European Language Resources Association.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITK@ LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203.

Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

Chukwuemeka Okechukwu, I Idris, JA Ojeniyi, Morufu Olalere, et al. 2023. Hate and Offensive Speech Detection Using Term Frequency-Inverse Document

[2] https://www.aclweb.org/portal/content/acl-code-ethics

Frequency (TF-IDF) and Majority Voting Ensemble Machine Learning Algorithms. 4th International Engineering Conference (IEC 2023).

Premjth, Bharathi Raja, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and Janakiram Chandu. 2024. Findings of the Shared Task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu). In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Bharathi Raja andS Malliga andCN SUBALALITHA Priyadharshini, Ruba andChakravarthi, Premjith and-Murugappan Abirami S V, Kogilavani andB, and Prasanna Kumar Kumaresan. 2023. Overview of Shared-task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. Word Level Language Identification in Code-mixed Kannada-English Texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@ DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Multilingual Hope Speech Detection using Machine Learning.

Mesay Gemeda Yigezu, Selam Kanta, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Habesha@ DravidianLangTech: Abusive Comment Detection using Deep Learning Approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 244–249.

Mesay Gemeda Yigezu, Tadesse Kebede, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023c. Habesha@ DravidianLangTech: Utilizing Deep and Transfer Learning Approaches for Sentiment Analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 239–243.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023d. Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification.

Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Olga Kolesnikova, Tadesse Kebede Guge, Alexander Gelbukh, and Grigori Sidorov. 2023e. Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media. In *2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pages=171–175*. IEEE.