Avalanche@DravidianLangTech:Abusive Comment Detection in Code Mixed Data Using Machine Learning Techniques with UnderSampling

Rajalakshmi S, Rajasekar S, Srilakshmisai K, Angel Deborah S, Mirnalinee T T
Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110, Tamil Nadu, India
rajalakshmis@ssn.edu.in, rajasekar2110857@ssn.edu.in,
srilakshmisai2110452@ssn.edu.in, angeldeborahs@ssn.edu.in,
mirnalineett@ssn.edu.in

Abstract

In recent years, the growth of online platforms and social media has given rise to a concerning increase in the presence of abusive content. This poses significant challenges for maintaining a safe and inclusive digital environment. In order to resolve this issue, this paper experiments an approach for detecting abusive comments. We are using a combination of pipelining and vectorization techniques, along with algorithms such as the stochastic gradient descent (SGD) classifier and support vector machine (SVM) classifier. We conducted experiments on an Tamil-English code mixed dataset to evaluate the performance of this approach. Using the stochastic gradient descent classifier algorithm, we achieved a weighted F1 score of 0.76 and a macro score of 0.45 for development dataset. Furthermore, by using the support vector machine classifier algorithm, we obtained a weighted F1 score of 0.78 and a macro score of 0.42 for development dataset. With the test dataset, SGD approach secured 5^{th} rank with 0.44 macro F1 score, while SVM scored 8^{th} rank with 0.35 macro F1 score in the shared task and to improve the macro F1 score, we used SVC and got a macro F1 score as 0.39.

1 Introduction

In recent times, social media has emerged as a prominent platform for discussions due to its wide reach and accessibility. It has granted individuals the power to express themselves, but unfortunately, it has also become a breeding ground for attacks based on characteristics such as race, gender, sexual orientation, or even threats of violence towards others.

According to the recent survey conducted by Economic Times, India in 2023 [1], 8 out of 10 urban women are using the Internet for various purposes. Nearly 83% of the people surveyed said that the safety measures for the usage of Internet need be enhanced. It states that, "Key concerns of urban Indian women when using the Internet include online sexual harassment, trolling, abuse, extortion and fraud". Hence it is a pressing need to identify the abusive content in Internet and take necessary actions for that.

To address the issue of data imbalance, sampling techniques are employed, and feature extraction is performed using count vectorizer and TF-IDF. Various machine learning classifiers are applied in the process of classifying the text as abusive or not and find the category.

The paper is organized as follows: Section 2 provides an overview of the relevant research conducted in the field. Section 3 examines the provided dataset, and Section 4 outlines the methodology employed for the task. The results obtained are presented in Section 5, and the paper concludes with a summary in the final section.

2 Related work

Nobata et al., published in 2016 [1] specifically addressing the detection of abusive comments. This seminal work introduced a methodology for identifying abusive language in various online platforms and utilized machine learning techniques for classification. It is widely recognized as one of the pioneering contributions in the field of abusive comment detection.

Waseem and Hovy, published in 2016 [7] focused on identifying predictive features for detecting hate speech on the Twitter platform. The objective was to understand the characteristics of hate speech and develop effective detection models.

Davidson et al. (2017) [13] addressed the challenge of automated hate speech detection and offensive language. Their work involved constructing a dataset of Twitter posts annotated for hate speech

[1] https://economictimes.indiatimes.com/news/india

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 166–170, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_022

and employing machine learning algorithms to classify offensive content. The research aimed to develop robust models capable of identifying hate speech in social media.

In 2018, Founta et al [11] conducted large-scale crowdsourcing to characterize abusive behavior on Twitter. By collecting and analyzing a substantial amount of data, they sought to understand the prevalence and nature of abusive content. This study significantly contributed to the understanding of abusive behavior patterns and provided valuable insights for the development of detection systems.

Badjatiya et al. (2020) [5] explored the application of deep learning techniques for hate speech detection in tweets. Their research employed deep neural network architectures, including convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, to classify tweets into different categories of hate speech. The study aimed to harness the power of deep learning for accurate hate speech detection.

Earlier we have worked on offensive language and misogyny detection for English language. Offensive content is recognized in English tweets using deep learning techniques and machine learning techniques in [15]. Misogyny detection from the multimodal data with English language is done in [16]. Now we are experimenting our work on the low resource languages and code mixed data.

Anusha Gowda [17] Spreading positive vibes or hope content on social media may help many people to get motivated in their life. To address Hope Speech detection in YouTube comments, this paper presents the description of the models submitted by our team-MUCIC, to the Hope Speech Detection for Equality, Diversity, and Inclusion (HopeEDI) shared task at Association for Computational Linguistics (ACL) 2022. This shared task consists of texts in five languages,namely: English, Spanish (in Latin scripts), and Tamil, Malayalam, and Kannada.

3 Dataset analysis and preprocessing

The provided dataset consists of comments extracted from social media platforms, primarily YouTube, and the train data-set contains dimensions of 5948 rows and 2 columns and test dataset contains dimensions of 1856 rows and 1 column, and is available in both Tamil and English languages [16]. These comments are categorized into different classes, namely **Misogyny, Misandry, Xenophobia, Transphobia, Homophobia, Counter-speech, and not abusive**. Initially, the data-set contains unwanted special characters and emojis. Most of the comments in the dataset are short, typically consisting of a single sentence, with an average sentence count close to 1. For reference, the number of classes in training and development dataset with their count have been listed in Table 1.

It is important to note that the dataset exhibits a significant class imbalance, with some categories being more dominant than others. This class imbalance can potentially lead to biased predictions favoring the majority class during model training.

To address the above mentioned issues, it is necessary to preprocess the raw dataset. The preprocessing step involves cleaning the data by removing special characters, punctuation, and irrelevant words that do not contribute significantly to the overall category or meaning of each comment.

4 Methodology

The methodology involves the steps of data preprocessing, class balancing, encoding, feature extracting, model building, evaluating and fine tuning the model. After extracting the necessary features from the cleaned dataset, we used classifier algorithms namely Stochastic Gradient Descent (SGD) and Support Vector Machine (SVM) to train the model and to predict the results i.e type of comment from the comments given in the dataset.

4.1 Encoding

Label Encoding is utilized in this task to handle the categorical features in machine learning. It transforms the categorical data into numerical labels, enabling effective processing by algorithms. In this task, we experimented with the use of label encoder.

4.2 Resampling

Resampling techniques help to balance the class distribution in the dataset which can improve the performance of machine learning models. It involves creating a new dataset by either undersampling or oversampling. Here, for this model, we used undersampling, since the label **none of the above** is significantly over represented compared to other labels.

Undersampling is done to reduce the size of the datasamples of a particular class to match the num-

S.no	Labels	Train dataset	Dev Dataset
1	None-of-the-above	3720	918
2	Misandry	830	218
3	Counter-speech	347	94
4	Xenophobia	297	70
5	Hope-Speech	213	53
6	Misogyny	211	50
7	Homophobia	172	43
8	Transphobic	157	40
9	Counter-speech	1	1

Table 1: Class label distribution of the dataset

ber of samples in other classes. When we are trying to over sample the number of samples of imbalance classes in the range of 1 - 830 to 3720, we are losing the importance features of those particular classes. But when we are under sampling the 3720 samples of non-abusive class to 500, then we are not losing much information. Hence it is planned to use under sampling techniques for balancing the data.

4.3 Feature extraction

Feature extraction involves quantifying or measuring unique properties of a text, reducing the complexity of the dataset used for model training. As part of this process, the text is numerically encoded.

4.3.1 Feature Extraction using Count Vectorizer

Count Vectorizer is employed to tokenize a set of texts by converting them into a vector representation based on token counts. This approach encompasses tokenization, counting, and normalization, collectively known as the n-gram representation.

4.3.2 Feature Extraction using TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) is a method for quantifying a sentence based on the words it contains. Each row is vectorized using a scoring technique that evaluates the importance of each word in the text. The scores for commonly used words are decreased, while the scores for rare words are increased.

4.4 Model Building

The machine learning models used for experimenting this task, includes **Stochastic Gradient Descent** (SGD) and **Support Vector Machine** (SVM) classifiers with Pipelining. These experiments are conducted on Tamil-English code-mixed data. The models are built using the training dataset and evaluated and fine tuned using the development dataset. We selected the best-performing models to generate performance scores for the test dataset.

5 Observation Results

For reference, the models under consideration for the Tamil-English dataset have been listed in Table 3 with the evaluation metrics like precision, recall, F1-score and accuracy.

In the study conducted on the Tamil-English dataset, we employed two different classifiers, namely Stochastic Gradient Descent (SGD) and Support Vector Machine (SVM) along with a simple transformer that involved pipelining. To convert categorical data into numerical labels, label encoding was applied. Count vectorizer and TF-IDF vectorizer are used for extracting features from text data. We evaluated the performance of various models and selected the best ones to generate performance scores for the test dataset.

Using the SGD classifier with both count vectorizer and TF-IDF features, our model achieved a F1 score of 0.45 and an accuracy of 0.73 for the development dataset. In the case of the SVM classifier with TF-IDF vectorizer, the model attained a a F1 score of 0.42 and an accuracy of 0.72 for the development dataset. For the test dataset SGD classifier achieved 0.44 F1 score and SVM achieved 0.35 F1 score. Our submission achieved the 5^{th} rank in the test evaluation for SGD and 8^{th} rank for SVM.

6 Inferences

Based on the observation, it can be noted that the datasets used in the study are relatively small, resulting in a limited number of training samples. Since the dataset is small, we identified that deep learning methods are not giving good results when

S.no	Feature extraction	Classifier	Precision	Recall	F1-Score	Accuracy
1	Count vectorizer	SGD	0.75	0.73	0.45	0.73
2	TF-IDF vectorizer	SGD	0.75	0.73	0.45	0.73
3	TF-IDF vectorizer	SVM	0.71	0.71	0.42	0.72
4	TF-IDF vectorizer	SVC	0.68	0.71	0.39	0.73

Table 2: Performance of the selected classifier models on Tamil-English using development data (With Re-sampling)

S.no	Feature extraction	Classifier	Precision	Recall	F1-Score	Accuracy
1	Count vectorizer	SGD	0.73	0.72	0.43	0.71
2	TF-IDF vectorizer	SGD	0.73	0.72	0.46	0.72
3	TF-IDF vectorizer	SVM	0.71	0.71	0.40	0.72
4	TF-IDF vectorizer	SVC	0.72	0.73	0.38	0.73

Table 3: Performance of the selected classifier models on Tamil-English using development data (Without Re-sampling)

compared to ML models. Furthermore, it is evident that both the Count and TF-IDF vectorizers exhibit a comparable accuracy rate. In summary, when comparing the SGD classifier and the SVM classifier, it is observed that the SGD classifier consistently achieves higher scores. Consequently, the SGD classifier can be considered as yielding the best results.

7 Conclusion and Future Work

In this study, we have performed a comprehensive analysis of different models for the Dravidian-LangTech@RANLP 2023 shared task focused on detecting abusive comments. We investigated the effectiveness of multiple classifiers on the preprocessed data by extracting relevant features. Our findings indicated that the SGD classifier produced comparable results using both vectorizers. Furthermore, we observed that the SVM classifier achieved a similar level of accuracy as the SGD classifier. In future, we have planned to increase the accuracy and F1-score by involving other feature extraction techniques and augmentation techniques. The potential challenges for further research in this field includes Multilingual and Multimodal Settings, Adversarial Attacks and Domain and Cultural Variations. The directions for further research includes User-Adaptive Models, Continuous Learning, Explainable AI and Real-Time Detection.

References

[1] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web (pp. 145-153). ACM.

[2] Fortuna, P., Nunes, M. D., & Cardoso, N. (2018). An analysis of machine learning approaches for abusive language detection on Twitter. In Proceedings of the 9th International Conference on Social Media and Society (pp. 1-5). ACM.

[3] Djuric, N., Zhou, J., Morris, R. R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web (pp. 29-30). ACM.

[4] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web (pp. 1391-1399). ACM.

[5] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web (pp. 759-760). ACM.

[6] Park, S., & Fung, P. (2017). One-step and two-step classification for abusive language detection on Twitter. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018) (pp. 1041-1048).

[7] Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop (pp. 88-93).

[8] Dadvar, M., Trieschnigg, D., de Jong, F., & Ordelman, R. (2013). Improving web search by detecting and assigning sentiment to query terms. Information Retrieval Journal, 16(5), 565-586.

[9] Qian, Y., Zhang, H., Li, J., Li, S., & Sun, X. (2018). Offensive tweet detection using convolutional neural networks. Future Generation Computer Systems, 88, 656-663.

[10] Bandyopadhyay, S., Malakar, S., Ganguly, N., & Mitra, P. (2019). Deep learning based abusive language detection in online social media platforms. arXiv preprint arXiv:1904.05772.

[11] Fortuna, P., Nunes, C., Sarmento, L. (2018). Abusive language detection on social media using lexicon-based approaches. In Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 291-299). INCOMA Ltd.

[12] Mathew, B., & D'Cunha, C. (2020). Hybrid feature-based detection of hate speech on social media using machine learning. IEEE Access, 8, 240354-240366.

[13] Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International AAAI Conference on Web and Social Media (pp. 512-515). AAAI.

[14] Sivanaiah, R., Angel, S., Rajendram, S. M., & Mirnalinee, T. T. (2022, July). TechSSN at semeval-2022 task 5: Multimedia automatic misogyny identification using deep learning models. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 571-574).

[15] Suseelan, A., Rajalakshmi, S., Logesh, B., Harshini, S., Geetika, B., Dyaneswaran, S.,& Mirnalinee, T. T. (2019, June). TECHSSN at SemEval-2019 task 6: Identifying and categorizing offensive language in tweets using deep neural networks. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 753-758).

[16] Priyadharshini, Ruba and Chakravarthi, Bharathi Raja and Cn, Subalalitha and Durairaj, Thenmozhi and Subramanian, Malliga and Shanmugavadivel, Kogilavani and U Hegde, Siddhanth and Kumaresan, Prasanna, Overview of Abusive Comment Detection in Tamil-ACL 2022, Association for Computational Linguistics (2022)

[17] Anusha Gowda, Fazlourrahman Balouchzahi, Hosahalli Shashirekha, and Grigori Sidorov. 2022. MUCIC@LT-EDI-ACL2022: Hope Speech Detection using Data Re-Sampling and 1D Conv-LSTM. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 161–166, Dublin, Ireland. Association for Computational Linguistics.