Wit Hub@DravidianLangtech-2024:Multimodal Social Media Data Analysis in Dravidian Languages using Machine Learning Models

Anierudh H S [1], **Abhishek R**[2], **Ashwin V Sundar**[3], **Amrit Krishnan**[4]&**B. Bharathi**[5]
Department of Computer Science And Engineering
Sri Sivasubramaniya Nadar College of Engineering,
Tamil Nadu, India
anierudh2210395@ssn.edu.in[1]
abhishek2210170@ssn.edu.in[2]
ashwin2210412@ssn.edu.in[3]
amrit2210213@ssn.edu.in[4]
bharathib@ssn.edu.in[5]

Abstract

The main objective of the task is categorised into three subtasks. Subtask 1 Build models to determine the sentiment expressed in multimodal posts (or videos) in Tamil and Malayalam languages, leveraging textual, audio, and visual components. The videos are labelled into five categories: highly positive, positive, neutral, negative and highly negative. Subtask 2 Design machine models that effectively identify and classify abusive language within the multimodal context of social media posts in Tamil. The data are categorized into abusive and non-abusive categories. Subtask 3 Develop advanced models that accurately detect and categorize hate speech and offensive language in multimodal social media posts in Dravidian languages. The data points are categorized into caste, offensive, racist and sexist classes. In this session, the focus is primarily on Tamil language text data analysis. Various combination of machine learning models have been used to perform each tasks and do oversampling techniques to train models on biased dataset.

1 Introduction

The digital age has fundamentally transformed our information landscape, with social media emerging as a dominant force shaping how we interact and engage with the world. While its benefits are undeniable, the rapid spread of online hate has become a pressing concern, posing significant threats to trust, democracy, and societal cohesion. Unrestricted access to post any data that may be offensive or abusive is a very important con of social media. This issue is particularly acute in Dravidian languages, where the lack of dedicated tools and resources exacerbates the impact of negativity in social media.

To address this challenge, significant research efforts have been directed towards developing advanced models capable of effectively detecting and categorizing various forms of harmful content in online spaces. This paper delves into the development of such models within the context of Tamil text data, focusing on three critical tasks:

1. Multimodal Sentiment Analysis : This allows for a nuanced understanding of expressed sentiment, ranging from highly positive to highly negative, offering valuable insights into online interactions and fostering constructive dialogue. It is trained with movie reviews and a supervised learning mode.

2. Multimodal Abusive Language Detection : Recognizing the prevalence of online abuse, this task focuses on building robust models that accurately identify and classify abusive language within Tamil text contexts. The models aim to improve detection accuracy and create a safer online environment by combating harmful interactions.

3. Multimodal Hate and Offensive Language Detection : Expanding beyond binary classification, this task delves into the complexities of offensive language by developing sophisticated models capable of accurately identifying and categorizing diverse forms of harmful content in Tamil text data. This includes nuanced distinctions between subtle categories like caste-based discrimination, general offensiveness, racism, and sexism, ultimately paving the way for a more inclusive and respectful online experience.

Through these tasks, the research presented in this paper highlights the immense potential of multimodal NLP and deep learning techniques in analyzing the complexities of communication within Tamil text data. The developed models offer practical solutions for combating misinformation, fostering trust, and promoting healthy online spaces.

The model traverses through different models employed for each model, such as LSTM, K-nearest neighbors, Linear Regression, Multinomial Naive Bayes and others and explains the purpose

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 229–233

for each method. Firstly the methodology and data is analysed, which includes model description, previous models and disadvantages,then an overview of obtained results, limitations, and conclusion with the findings. The datasets that have been used are (Premjith et al., 2023) , (Premjith et al., 2022) , (Chakravarthi et al., 2021). The overview of the shared task is given in Findings of the Shared Task on Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) (B et al., 2024).

2 Related work

Banerjee (Banerjee et al., 2020), has used an autoregressive XLNet model to perform sentiment analysis on code-mixed Tamil-English and Malayalam-English datasets.

DravidianMultiModality: A Dataset for Multimodal Sentiment Analysis in Tamil and Malayalam is a paper (Chakravarthi et al., 2021), where the product or movies review videos were downloaded from YouTube for Tamil and Malayalam. Next, the captions were created for the videos with the help of annotators and the videos were labelled for sentiment.

In the paper (Ofli et al., 2020), they are doing analysis on social media data using multi modal deep learning for disaster response. They have used CNN, image modality and others. By using these models, they performed 2 tasks, Informativeness classification task and humanitarian classification task with F1 score 84.2 and 78.3 respectively.

3 Methodology and Data

3.1 Multimodal Sentiment Analysis

The fundamental goal of sentiment analysis using machine learning (ML) classification is to create reliable and accurate models that can distinguish between positive, negative, highly negative, highly positive and neutral comments. The model is restricted to testing and training only on Tamil dataset. The objective is to use labeled datasets with examples ranging from highly positive to highly negative to train the models. The primary goal of the model is to develop algorithms that can apply the trained features onto any Tamil text content and determine the sentiment accurately.

The following dataset contains various movie reviews of Tamil movies in Tamil text. The dataset consists of 2 attributes namely the TextContent and the corresponding sentiment label.

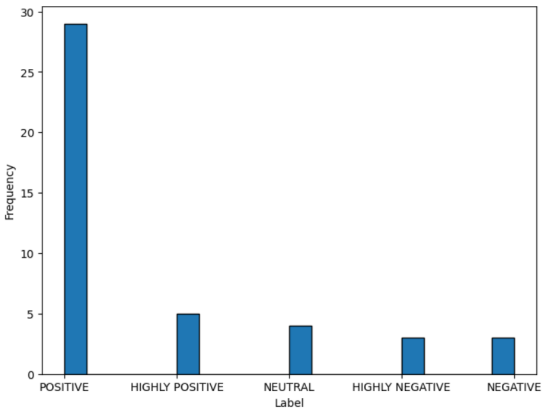

Figure 1: Training Data Bias

The dataset was very biased with high datapoints corresponding to positive reviews and very few for the others as shown in Figure-1.

With such a biased dataset, the two available options to train the model were to use OverSampling such as Smote Analysis and test on development data to test the best model, or to concatenate training and development data and obtain a train test split to train the model and test it.

3.1.1 Trial 1 (Random Forest with SMOTE)

Methods: 1. Text preprocessing: removing punctuation and converting to lowercase. 2. TF-IDF vectorization for text features. 3. SMOTE for oversampling imbalanced classes. 4. Random Forest Classifier for prediction. In this model some basic text preprocessing on train and development data has been applied. Oversampled training data has been used to train. Next using TF-IDF vectorizer text features are extracted. SMOTE is applied onto the training data to oversample. Then a random-forest classifier has been used and the model has been trained and tested on development data to find F1 score. This method is likely suitable for simple text classification tasks but may not capture long-range dependencies or word order. And SMOTE has its disadvantages in classification tasks. SMOTE involves creating synthetic examples, which increases the size of the dataset. This larger dataset can lead to increased computational complexity, especially for algorithms that scale poorly with the number of instances.

3.1.2 Trial 2 (K-Nearest Neighbors with SMOTE)

Methods: 1. Text preprocessing: removing punctuation and converting to lowercase. 2. TF-IDF vectorization for text features. 3. SMOTE for over-

sampling imbalanced classes. 4. K nearest neighbor for prediction. This method is very similar to Trial 1 and thus follows a few disadvantages of Trial-1 such as not being able to capture word order. SMOTE focuses on generating synthetic instances for the minority class. While this helps balance class distribution, it does not address potential imbalances within the majority class, and synthetic instances may not accurately capture the characteristics of the majority class. And also KNN can be sensitive to noisy data and may lack interpretability compared to other models.

3.1.3 Final Method (LSTM,KNN and Linear Regression)

Methods: 1.Text preprocessing 2.Tokenization and padding for LSTM input 3.K-Means clustering of LSTM model predictions to extract high-level features. 4.KNN classifier trained on clustered features for added robustness. 5.Linear Regression on clustered features for another prediction perspective. 6.Model saving and loading for prediction on new data. In this model, considering the limitations of Smote in classification tasks, training and development data are merged and train test split is done onto the concatenated dataset. This model combines LSTM for capturing complex text patterns with K-Means clustering for identifying latent features and Linear Regression for class prediction. LSTM offers an unfair advantage over other models as it is implemented using neural network and takes into account word order for determining patterns in the data. So it is best for feature extraction. Then the outputs are clustered and offered into a linear model for accurate prediction of test data.

3.1.4 Comparison of F1 scores

Table 1 shows the F1 scores for different models on training and development data. The results are evident to prove that the chosen model is advanced in performance.

Model used	F1 Score
Trial 1	0.228
Trial 2	0.228
Final Model	0.603

Table 1: Output Comparison

3.2 Multimodal Abusive Language Detection

In this subtask machine models are built such that they effectively identify and classify abusive language within the multimodal context of social media posts in Tamil. The dataset is classified into abusive and non-abusive texts. Figure 2 shows the provided dataset from codalab which is unbiased.

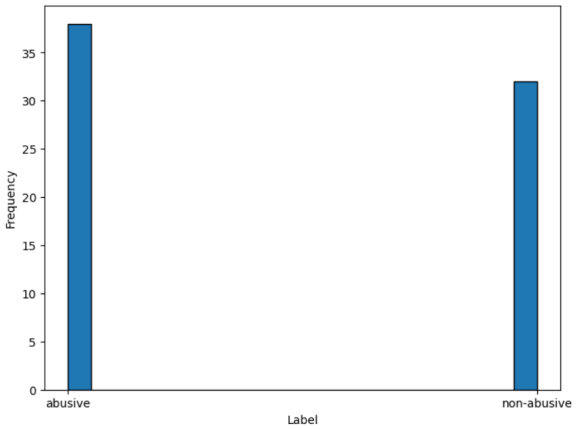

Figure 2: Task2 dataset

3.2.1 Model using LSTM, KMeans, KNN, Logistic Regression

In this method, LSTM (Long short term memory) has been used to identify sequence of words that may be abusive, the features have been clustered using K-Means and logistic regression has been applied as it is a binary classification task. This is a pretty decent model but does have a few limitations. The model might be effective for complex abusive language patterns but requires significant training data and computation. Since the training dataset is small, it may not be the perfect model. The model is best suited for large datasets for its faster adaptability.

3.2.2 Final Model using MNB

MultiNomial Naive Bayes is used for this task. It is a lightweight probabilistic classifier based on word frequency in different classes. The reason this model is chosen is because the main reason a statement can be found abusive is due to the presence of a single or a few offensive words and not on the sequence of words. Thus a model using MNB is better for this scenario. It can be easily integrated into online systems because of its lightweight nature.It also provides insights into the words and phrases that contribute to the classification through word frequency analysis.

3.2.3 Output classification

Table 2 will provide an insight onto the obtained f1 scores of training and testing after a train test

split and shows why the chosen model is better for a small dataset

.

Model used	F1 Score
Trial model	0.590
Final Model	0.791

Table 2: Output Comparison

3.3 Multimodal Hate and Offensive Language Detection

In this task advanced models that accurately detect and categorize hate speech and offensive language in multimodal social media posts in Dravidian languages are developed. The data points in the dataset are categorized into caste, offensive, racist and sexist classes.

3.3.1 Trial 1 (Multinomial Naive Bayes)

In this task, MNB is used in same pattern as in Subtask 2. The F1 score obtained is very less, around 0.16. The reasons for the inefficiency can be looked upon to the facts that the model may not capture complex language patterns. Thus this model is discarded and new model is used.

3.3.2 Final Code (Random Forest with combined TF-IDF and Count vectors)

In this model we use Random Forest with two feature sets: TF-IDF for term weighting and Count vectors for word frequency. The pros of the model lies in the capability to capture both term importance and word frequency through combined features. Though the F1 score obtained on train test split on the training data was not great, reflecting to the size and variability of dataset. Thus this model is preferred despite the difficulty in its complex training procedures.

3.3.3 Output Classification

Table 3 shows the classification of outputs of trial model and final model, hence proving its efficiency over others

Model	F1 score
Trial Model	0.166
Final Model	0.371

Table 3: Output Classification

Tamil Sentimental Analysis			
Team	Run	F1score(score)	Rank
WitHub	3	0.244	1
Tamil Hate Speech Detection			
Team	Run	F1score(score)	Rank
WitHub	3	0.288	1
Tamil Abusive Language Detection			
Team	Run	F1score(score)	Rank
BinaryBeasts	1	0.714	1
WitHub	1	0.415	2

Table 4: Results

4 Experimental Result and Performance Analysis

The prescribed models had been submitted to codalab and the runs on the test data had been submitted. The submissions were evaluated and and the results are as in Table 4. SubTask 1 got a F1 score of 0.244. Subtask 2 got a F1 score of 0.288 and Subtask 3 got a score of 0.415. From these scores and corresponding ranks, we infer that the prescribed models are very effective and adapted to the dataset.

5 Limitations

The model is trained only over a small sample of training data. There may be various other data that should be included for better performance of the model. And also, as linguistic trends change, the model may be ineffective over time. We need measures to prevent that too.

6 Conclusions

In conclusion, various models have been trained and tested for each subtask. These models have been trained only under a specific small dataset and are adapted to it. The models prescribed are best adapted to the small training datasets and are proven to produce a good F1 score for each task. The model can be improved by including online learning techniques and reinforcement learning to adapt to new data and trends and thus have an enhanced performance.

References

Premjth B, Jyothish Lal G, Sowmya V, Bharathi Raja Chakravarthi, Nandhini K, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Span-

dana Reddy Mekapati. 2024. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl). In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Shubhanker Banerjee, Arun Jayapal, and Sajeetha Thavareesan. 2020. Nuig-shubhanker@ dravidiancodemix-fire2020: Sentiment analysis of codemixed dravidian text using xlnet. *arXiv preprint arXiv:2010.07773*.

Bharathi Raja Chakravarthi, KP Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, John P McCrae, et al. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*.

B Premjith, Bharathi Raja Chakravarthi, Malliga Subramanian, B Bharathi, Soman Kp, V Dhanalakshmi, K Sreelakshmi, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260.

B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.