TensorTalk@DravidianLangTech 2025: Sentiment Analysis in Tamil and Tulu using Logistic Regression and SVM

K. Anishka[1], Anne Jacika J[1]
[1] SSN College of Engineering , Tamil Nadu, India
anishka2310506@ssn.edu.in, annejacika2310581@ssn.edu.in

Abstract

Words are powerful; they shape thoughts that influence actions and reveal emotions. On social media, where billions of people share their opinions daily. Comments are the key to understanding how users feel about a video, an image, or even an idea. But what happens when these comments are messy, riddled with code-mixed language, emojis, and informal text? The challenge becomes even greater when analyzing low-resource languages like Tamil and Tulu. To tackle this, TensorTalk deployed cutting-edge machine learning techniques such as Logistic regression for Tamil language and SVM for Tulu language , to breathe life into unstructured data. By balancing, cleaning, and processing comments, TensorTalk broke through barriers like transliteration and tokenization, unlocking the emotions buried in the language.

1 Introduction

The modern world is rapidly advancing communication and networking technology. Users express opinions in social media comment sections. Therefore, understanding and analyzing the sentiments from these textual data can significantly improve classifying and recommending products and other services by understanding the general public opinion as observed by the authors of (Taboada, 2016). It is also observed that comments are not often in the same language or in its original form i.e, there are slang words, transliterated words, and words in native script as well. Therefore, developing efficient methods to analyze such diverse textual forms of data becomes imperative. Sentiment Analysis is a field of Natural Language Processing (NLP) that focuses on analyzing and interpreting the emotions, opinions, and sentiments expressed in textual data. This generally includes identifying the emotions as positive, negative or neutral based on certain factors. The authors of (Wankhade et al., 2022) interpret sentiment analysis as identifying and extracting subjective information from text using natural language processing and text mining, and discuss methods to complete the given sentiment analysis task and its applications. In this paper, TensorTalk addresses the task of analyzing sentiments of multifaceted Dravidian languages such as Tamil and Tulu, by analyzing textual data obtained across various social media platforms' comment sections. The data that has been analyzed was found to be highly imbalanced, which is as expected of general public opinions and takes varying forms as expressed earlier featuring transliterated words, original text form, slang/dialect words and other language words as well. TensorTalk has thus used the logistic regression and SVM models to address the problem at hand. The proposed solutions for the task: Sentiment Analysis in Tamil and Tulu were presented to, and evaluated by DravidianLangTech@NAACL2025 for the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025 (Durairaj et al., 2025) . In addition to the academic contributions, the proposed solution can help in practical implications as mentioned earlier for product/service recommendation purposes, etc. The detailed analysis of dataset ,data preprocessing, oversampling techniques, vectorization methods, Logistic Regression, SVM model implementations and results are discussed in the subsequent sections. The code for the classification tasks discussed in this paper can be accessed through this link: [1]

2 Related Work

With rapid digital transformation, user-generated content from social networks, blogs, and forums have become a valuable information source. Sentiment analysis, a subfield of Natural Language Pro-

[1] https://github.com/Anishka-K556/Dravidian-Sentiment-Analysis

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 688–693

cessing (NLP), plays a crucial role in understanding public opinions, emotions, and attitudes from such textual data. The authors of (Chakravarthi et al., 2020) remark that comments from social networks do not follow strict rules of grammar, contain more than one language, and are often written in nonnative scripts, thereby making sentiment classification processes much more difficult. In order to effectively analyze the underlying sentiments and prove insights on current trends and various decision-making applications especially in Dravidian languages is in trend according to the authors of (Sambath Kumar et al., 2024). Sentiment analysis in low-resource Dravidian languages like Tamil and Tulu, complicated by code-mixed data, is still in its early stages (Hegde et al., 2023). One of the most significant roles in prediction models is exploration of the data distribution as remarked by the authors of (Bailly et al., 2022) who illustrate the working of models like Logistic Regression and the ways training dataset size and interactions affect the performance of these prediction models. Many significant improvements have been made in the fields of sentiment analysis of textual data in Tamil and Tulu using traditional models like logistic regression as illustrated by authors of (Ponnusamy et al., 2023). According to the authors of (Fang and Zhan, 2015) major components of sentiment analysis include processing the data, vector generation, feature extraction etc. Despite being low resourced languages, the evolution of these Dravidian languages into much more modern forms needs to be understood and researched upon from the perspective of the authors of (Hegde et al., 2022). Datasets obtained from various social media platforms, or in general datasets for various real-time problems are mostly imbalanced. To handle these imbalanced datasets, either oversampling or undersampling techniques are applied. According to the authors of (Elreedy et al., 2024), the SMOTE method generates new synthetic data patterns by performing linear interpolation between minority class samples and their K nearest neighbors, which need not always conform to the original distribution. These can thus, boost a model's performance.Representation of textual data in a form that is compatible with the training models also determines the model's performance. In (Qorib et al., 2023),the authors have extensively tested and implemented various vectorization methods paired with different models to achieve higher performance. Higher performance and model performance generally follow from thoroughly cleaned data. Among multiple vectorization methods, the TF-IDF method known for its robustness and efficiency has been implemented. Textual data are better represented by vectorization methods such as TF-IDF methods that aid the models in training. The authors of (Liu et al., 2018) remark that in addressing problems, such as ignoring contextual semantic links and different vocabulary importance in traditional text classification techniques, vectorization techniques such as word2vec, TF-IDF methods improve traditional and deep learning methods' performances. Vectorization techniques adapted from word2vec such as fasttext embeddings have proven to represent Tamil and Tulu texts better as observed by the authors of (K et al., 2023). Pre trained transformer models have also been used to explore and understand the underlying sentiment in these texts coupled with feature extraction techniques by the authors of (Balaji et al., 2024). Many developments in textual sentiment analysis have been made using deep learning models as illustrated by the authors of (Tang et al., 2015) who have remarked that deep learning approaches emerge as powerful computational models as they discover intricate semantic representations of texts automatically from data without feature engineering.

3 Dataset and Task Description

The problem at hand is to detect and classify the sentiments in textual data obtained from various social platfoms such as comments/posts in Tamil and Tulu languages. Due to the diversity and complexity of the Dravidian languages due to code-mixed data there is a growing demand for improving and developing models for sentiment analysis in these Dravidian languages. The given dataset instances of comments and posts are such that they may contain more than one sentence but the average length is 1. The data that has been supplied is found to have classes such as positive, negative, mixed feelings and unknown state for Tamil data and positive, negative, neutral, mixed and not Tulu for Tulu data and is highly imbalanced, in accordance with real world scenarios. Refer Table 1 that shows the distribution of the textual data that has been provided for the given problem. It is noted that highly imbalanced classes thus lead to biasing of the model i.e. overfitting (latching onto irrelevant patterns) of the majority class and poor generalization of the minority classes.

Task	Labels	No. of instances
Tamil Sentiment Analysis	Positive	18145
	Negative	4151
	Mixed feelings	3662
	Unknown state	5164
Tulu Sentiment Analysis	Positive	3769
	Negative	843
	Mixed feelings	1114
	Neutral	3175
	Not Tulu	4400

Table 1: Dataset Distribution

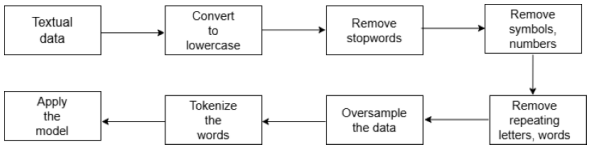

Figure 1: Text Preprocessing and Model Application Workflow

4 Methodology

Given the linguistic diversity and grammatical uniqueness of the language, TensorTalk refined traditional textual analysis methods and applied vectorization techniques. Additionally, TensorTalk has utilized models such as Logistic Regression to effectively identify patterns within the dataset, which was enhanced through vectorization and oversampling. Analyzing the distribution of the given data revealed that the data were highly imbalanced, as observed in the Tamil data set, where positive-labeled data occur predominantly throughout. By creating synthetic samples of the minority class, oversampling increases the representation of the minority class, helping the model learn its patterns. Thus, a better representation of the minority class can be achieved using the SMOTE technique compared to undersampling the majority class, which may lead to data loss and poor generalization, or simple synthetic generation, which can cause overfitting by duplicating existing samples without adding meaningful variation. To tackle this problem, TensorTalk experimented with classical machine learning models, including Logistic Regression and SVM, on textual data that underwent various preprocessing steps along with tokenization methods used to represent text in vectorized forms. These models are implemented along with SMOTE to handle the imbalanced classes in the dataset. These models were chosen due to their robustness. These are discussed in detail below to illustrate their role in improving model training. Refer Figure 1. The decision to choose ML models listed above over deep learning models aligns with the requirements of low computational costs, availability of dataset, need for interpretability, and simplicity. Taking into consideration that the given dataset is on the smaller side TensorTalk decided to opt for traditional models like SVM and Logistic Regression over deep learning models such as LSTM. After adequate testing and hyperparameter tuning, the development data set was used to train the model to improve the model training accuracy.

4.1 Preprocessing

Data generated from most real-time problems contain noise, missing values, physically impossible values, and format that might not be compatible with the models. Thus, to overcome these issues, preprocessing is performed on the data which significantly increases the efficiency of any model. The preprocessing techniques that TensorTalk has used include removing null values and unifying the spaces between words. Considering the fact that comments usually contain various other punctuations, emojis and symbols, these have been removed as well. Some traditional cleaning methods such as removing stopwords ,lemmatization have also been applied. In addition to these, some unique cleaning techniques such as removing duplicate words in an entry, removing repeating letters, timestamps, etc. have also been implemented. To address the issue of class imbalance SMOTE techniques have been applied on the dataset to oversample the minority class. This method has allowed us to generate samples to balance the dataset across all classes. In comparison, to undersampling the majority class, TensorTalk preferred oversampling to avoid the risk of potential loss of information. Finally, the dataset was subject to tokenization using the method of TF-IDF Term Frequency-Inverse Document Frequency. This feature extraction technique allowed us to represent the textual data in a vectorized format by measuring how often a term has appeared in a document. By computing the importance of the term by this weight and reducing the dimensionality, the dataset is converted to a form apt for models to learn from.

4.2 Task: Sentiment Analysis in Tamil

The first task to classify the given textual data was implemented using the Logistic Regression model into multiple classes for varying degrees of sentiment polarity. Logistic Regression, being a linear model, was used to effectively classify the

given textual data into multiple classes in the given problem, where the data had been subject to TF-IDF techniques prior along with the preprocessing techniques mentioned above, by using the logistic/sigmoid function in determining the probability of a given text belonging to a specific class. TensorTalk chose to implement this model for the given problem due to its ability to act as a good baseline model for most text classification tasks. The model proved to be much simpler when compared to the other models, along with its ability to adapt for considerably large datasets, higher interpretability, speed and scalability.

4.3 Task: Sentiment Analysis in Tulu

The task is to classify the given textual data was implemented using the machine learning model Support Vector Machine (SVM). Support Vector Machine (SVM) which is known for it's high dimensional data handling, binary and multi-class classification and also for it's robust to overfitting. This model proved the strength of the SVM in handling complex classification tasks , even in the presence of noisy and unbalanced textual data. TensorTalk chose SVM because one of it's key advantages is it's ability to work well with sparse and high-dimensional feature spaces, which are common in textual data. SVM can capture complex patterns and relationships within text features. It also performs well even with relatively small training datasets, making it a strong choice when labeled data are limited. In this implementation, SVM demonstrated its strength in handling complex classification tasks, even in the presence of noisy and imbalanced data.

4.4 Results and Discussions

The performance of machine learning models such as Logistic Regression and Support Vector Machine (SVM) in sentiment analysis in Tamil language and Tulu language respectively was evaluated on key metrics such as accuracy, precision, recall, and F1 score. The evaluation of the model was mainly based on the macro-averge F1 score, during the training of the model, which was found to be 46% for Tamil and 55% for Tulu, mentioned in table 2. It was found that among all the other models that were trained and tested, the Logistic Regression model gave the best performance, especially when combined with the SMOTE and TF-IDF techniques. The underlying reason could be attributed to the TF-IDF's inverse frequency technique and better generalization of minority classes. In the case of sentiment analysis in Tulu, the SVM model showed better performance compared to other models, which could likely have resulted as SVM performs well in high-dimensional spaces by finding an optimal separating hyperplane. In addition, the test scores for both tasks were found to be 24% for Tamil and 53% for Tulu, refer to table 3. The Logistic regression model's F1 score of 47% indicates that it performed reasonably well on validation data. However, the sharp drop to 24% in the test data suggests that the model struggles to generalize. The few plausible reasons for these could be overfitting of the model, excessive tuning to validation dataset, etc. Although the exact cause remains unknown, TensorTalk believes that better preprocessing techniques could still improve the model's performance. Future work may focus on refining the models by incorporating deep learning models.

Task	Macro Avg F1 score
Tamil Sentiment Analysis	0.46
Tulu Sentiment Analysis	0.55

Table 2: Sentiment Analysis Cross Validation Score

Task	Macro Avg F1 score
Tamil Sentiment Analysis	0.24
Tulu Sentiment Analysis	0.53

Table 3: Sentiment Analysis Test Score

5 Conclusion

As communication continues to grow exponentially, so must techniques and models for machines to analyze them. Through this paper, TensorTalk addresses sentiment analysis in low-resource Dravidian languages, Tamil and Tulu, using classical machine learning. TensorTalk has employed Logistic Regression for Tamil and SVM for Tulu, overcoming challenges such as code-mixed data and transliterations. Our proposal includes preprocessing, tokenization, and oversampling that improve model performance. Future work will explore deep learning and domain-specific embeddings to enhance classification. This research advances Dravidian language processing.

6 Limitations

The proposed models have faced challenges in training to classify the given data into the required labels. This is primarily because low-resource languages, especially in their transliterated form, do not easily conform to specific spellings or distinct word boundaries. As a result, identifying and removing stopwords becomes particularly difficult. The presence of diverse slang and dialectal variations in textual data from social media is found to be less effective than tokenizing text in a standardized language such as English. Due to the complex morphology of Dravidian languages, sentiment analysis becomes more challenging as words can take different forms based on tense, gender, and other grammatical variations. Sentences often mix between English and native languages which makes it difficult to predict labels accurately.The informal nature of social media text introduces noise in the form of misspellings, abbreviations, and emojis, further complicating preprocessing. The same word or phrase can express different sentiments based on context, and traditional models like SVM or Logistic Regression struggle to capture this dynamic contextuality. Support Vector Machine (SVM) and Logistic Regression struggle with contextual and sematic meanings in text, making them less effective. Logistic Regression and SVM are sensitive to the scale of input features, requiring feature normalization for optimal performance. SVM and Logistic Regression can perform poorly when the data is imbalanced, especially in real-world scenarios where some sentiment labels may be underrepresented, as observed in this case. Although oversampling technique has been applied, the model still struggles and the potential reason for this could be overfitting. This emphasizes the need for advanced language models and preprocessing techniques to enhance sentiment analysis in low-resourced languages.

References

Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. 2022. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Computer Methods and Programs in Biomedicine*, 213:106504.

Shreedevi Balaji, Akshatha Anbalagan, Priyadharshini T, Niranjana A, and Durairaj Thenmozhi. 2024. WordWizards@DravidianLangTech 2024: Sentiment analysis in Tamil and Tulu using sentence embedding. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 218–222, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Dina Elreedy, Amir F Atiya, and Firuz Kamalov. 2024. A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Machine Learning*, 113(7):4903–4923.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big data*, 2:1–14.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Rachana K, Prajnashree M, Asha Hegde, and H. L Shashirekha. 2023. MUCS@DravidianLangTech2023: Sentiment analysis in code-mixed Tamil and Tulu texts using fastText. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Cai-zhi Liu, Yan-xiu Sheng, Zhi-qiang Wei, and Yong-Quan Yang. 2018. Research of text classification based on improved tf-idf algorithm. In *2018 IEEE international conference of intelligent robotic and control engineering (IRCE)*, pages 218–222. IEEE.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. VEL@DravidianLangTech: Sentiment analysis of Tamil and Tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, and Paul Cotae. 2023. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on covid-19 vaccination twitter dataset. *Expert Systems with Applications*, 212:118715.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Maite Taboada. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.