KEC_AI_NLP@DravidianLangTech:Sentiment Analysis using Hybrid Model

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[1], **VetriVendhan S**[1], **Pramoth Kumar M**[1], **Karthickeyan S**[1], **Kavin Vishnu N**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{vetrivendhans.21aim, pramothkumarm.21aim}@kongu.edu
{karthickeyans.21aim, kavinvishnun.21aim}@kongu.edu

Abstract

Sentiment Analysis is a process that involves analyzing digital text to determine the emotional tone, such as positive, negative, neutral, or unknown. In this study, we obtained the dataset from the CodaLab website by participating in a competition and accessing code-mixed format train and development data. Later, on May 10th, the test data was provided, including an unlabeled class. Sentiment Analysis of code-mixed languages presents challenges in natural language processing due to the complexity of code-mixed data, which combines vocabulary and grammar from multiple languages and creates unique structures. The scarcity of annotated data and the unstructured nature of code-mixed data are major challenges. To address these challenges, we explored various techniques, including Machine Learning (ML), Deep Learning, and Transfer Learning. ML models such as Decision Trees, Random Forests, Logistic Regression, and Gaussian Naïve Bayes were employed. Deep Learning models, such as Long Short-Term Memory (LSTM), and Transfer Learning models like BERT, were also utilized. The results demonstrated promising performance in sentiment analysis of code-mixed text. Overall, this study contributes to the field of sentiment analysis by addressing the challenges posed by code-mixed language and employing diverse ML and Deep Learning techniques for accurate sentiment classification.This dataset was taken in the competition in Codalab with dataset description of (Chakravarthi et al., 2020)and (Hegde et al., 2022). Our team participated in the shared task organized by (Hegde et al., 2023)

Key Words Sentiment Analysis, Emotional tone, Natural language processing, Tokenizer, padded sequence, Machine Learning, BERT, LSTM

1 Introduction

Sentiment Analysis, also referred to as opinion mining, is a computational process that utilizes natural language processing (NLP), text analysis, and computational linguistics to uncover the emotional sentiment expressed in each text. It aims to categorize and determine opinions regarding a product, service, or idea, by extracting polarity (positivity or negativity), subject matter, and the opinion holder within the text. Sentiment Analysis can be performed on various levels, including full documents, paragraphs, sentences, or even smaller units. It finds applications in diverse domains such as product reviews, social media analysis, and market research. Different types of Sentiment Analysis exist, such as aspect-based Sentiment Analysis, grading Sentiment Analysis, multilingual Sentiment Analysis, and emotion detection. Approaches for Sentiment Analysis include knowledge-based techniques, statistical methods, and Machine Learning algorithms. Social media platforms serve as significant sources of data for Sentiment Analysis, as they generate vast and interconnected information in the form of user-generated content.

2 Literature Survey

(Pang et al., 2008) Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends in Information Retrieval. This seminal work provides an overview of sentiment analysis techniques, including both traditional machine learning approaches and early deep learning methods.(Kim, 2014) Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). The author introduces a simple yet effective model architecture that utilizes multiple parallel convolutional filters with different kernel sizes

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 300–305, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_044

to capture various n-gram features for sentence classification.(Socher et al., 2013) Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP). The authors propose a recursive deep model that captures the hierarchical nature of sentiment in sentences, achieving improved accuracy by leveraging fine-grained sentiment information.[3] Tang, Duyu, et al. "Learning sentiment-specific word embedding for Twitter sentiment classification." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. This study focuses on learning sentiment-specific word embeddings for Twitter sentiment classification, demonstrating the effectiveness of incorporating sentiment information into word representations.(Zhang et al., 2015) Zhang, Lei, et al. "Character-level convolutional networks for text classification." Advances in neural information processing systems. The authors propose character-level convolutional networks for text classification tasks, showcasing their effectiveness in sentiment analysis by capturing local and compositional features at the character level.[6] Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805. This paper introduces BERT, a pre-training approach based on bidirectional transformers, which has shown remarkable performance in various natural language processing tasks, including sentiment analysis.[7] Tang, Duyu, et al. "Attention-over-attention neural networks for reading comprehension." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). The authors propose attention-over-attention neural networks for reading comprehension tasks, demonstrating their effectiveness in capturing intricate relationships within sentences for sentiment analysis.(Zhang et al., 2015) Zhang, Xu, et al. "Character-level attentive recurrent neural networks for sentiment analysis." Proceedings of the AAAI Conference on Artificial Intelligence. Vol.This work introduces character-level attentive recurrent neural networks for sentiment analysis, showcasing the significance of character-level attention mechanisms in capturing fine-grained sentiment information.(Young et al., 2018) Yang, Zichao, et al. "Hierarchical attention networks for document classification." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. The authors propose hierarchical attention networks that capture contextual dependencies at different levels of granularity, achieving state-of-the-art results in document classification, which is applicable to sentiment analysis.[10] Xing, Wei, et al. "Recurrent convolutional neural networks for text classification." Proceedings of the 2015 conference on empirical methods in natural language processing. This study presents recurrent convolutional neural networks and machine learning techniques for text classification tasks, combining the strengths of both recurrent and convolutional neural networks to capture long-term dependencies.

3 Methodology

Machine learning models have revolutionized the field of artificial intelligence by enabling computers to learn and make predictions or decisions without being explicitly programmed. In this introduction, we will explore several popular machine-learning models and their applications in sentiment analysis. The proposed system workflow is presented in Figure 1.

Decision trees are a powerful supervised learning technique that can be used for both classification and regression tasks. They create a tree-like model where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome or prediction. Decision trees are intuitive, easy to interpret, and can handle both numerical and categorical data. They are often used in sentiment analysis to classify text into positive, negative, or neutral sentiments. Random forests are an ensemble learning method that combines multiple decision trees to make more accurate predictions. It works by constructing a multitude of decision trees, each trained on a different subset of the training data and using random feature subsets. The final prediction is made by aggregating the predictions of individual trees. Random forests are known for their robustness, ability to handle high-dimensional data, and resistance to overfitting. They are commonly used in sentiment analysis to improve classification accuracy. Logistic regression is a popular classification algorithm used to predict the probability of a binary outcome based on input variables. Despite its

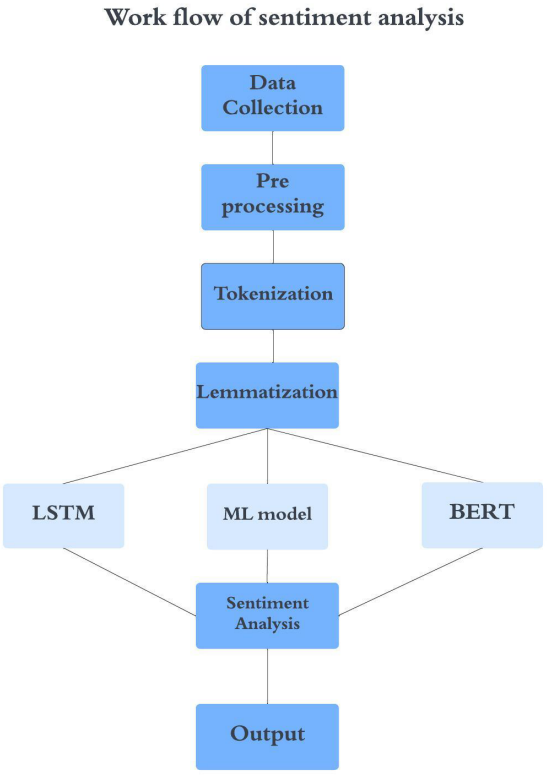

Figure 1: proposed system workflow

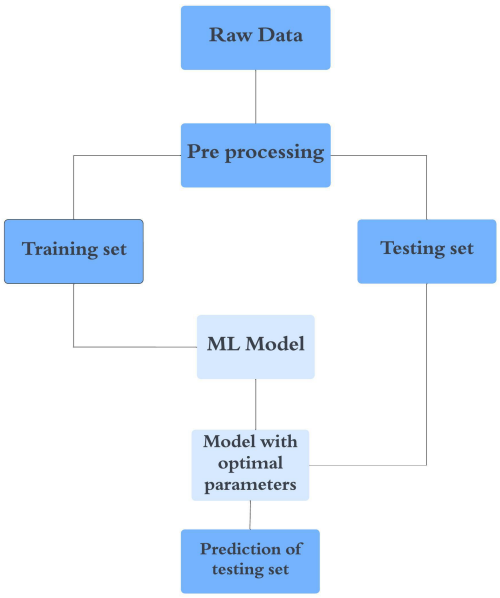

Figure 2: Sentiment Analysis by Machine Learning Models

name, logistic regression is a linear model, but it applies a logistic function to the linear output to transform it into a probability. It is widely used in sentiment analysis to determine the likelihood of a given text belonging to a specific sentiment class. Logistic regression is computationally efficient and can provide interpretable results. Gaussian Naïve Bayes is a probabilistic classifier based on Baye's theorem and assumes independence between features. Gaussian Naïve Bayes specifically assumes that the features follow a Gaussian distribution. It is a simple yet effective algorithm that works well with high-dimensional data and requires a small amount of training data. Gaussian Naïve Bayes is often used in sentiment analysis to classify text based on the likelihood of belonging to a particular sentiment class. Performing sentiment analysis using machine learning models are depicted in Figure 2.

Deep learning models, such as LSTM, are a type of artificial neural network [7-11] with multiple layers that can learn complex patterns and dependencies in data. LSTM networks are specifically designed to capture long-term dependencies by using memory cells and gates that regulate the flow of information. LSTMs have been highly successful in natural language processing tasks, including sentiment analysis. They can learn contextual information, understand sequential patterns, and effectively model text sentiment over longer sequences. Performing Sentiment Analysis by LSTM is represented in Figure 3.

Transfer Learning - BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art deep learning model for natural language processing tasks. It is based on the Transformer architecture and is pre-trained on a large corpus of text data. BERT has achieved remarkable results in various NLP tasks, including sentiment analysis, by leveraging its ability to understand contextualized word representations. Transfer learning with BERT involves fine-tuning the pre-trained model on a specific sentiment analysis task, which can significantly improve performance even with limited training data. Sentiment analysis by BERT is presented in Figure 4.

4 Performance Evaluation

Whether utilising machine learning or deep learning techniques, performance evaluation for sentiment analysis models often incorporates multiple indicators to gauge the model's efficacy. Following are a few typical evaluation measures for sentiment analysis: Accuracy: The ratio of examples that

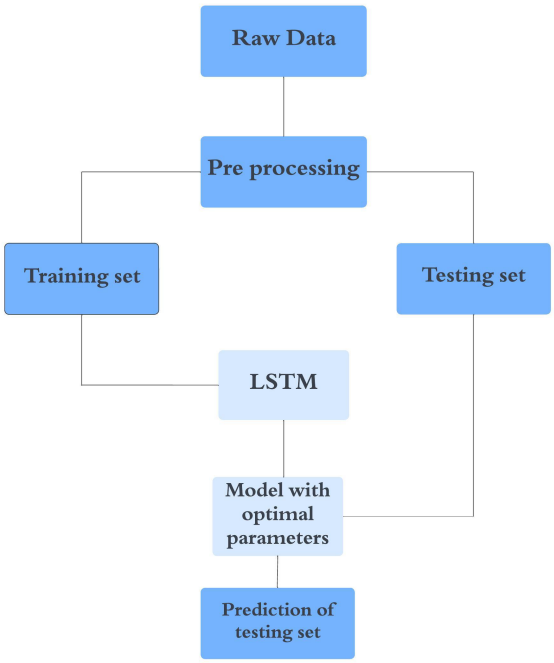

Figure 3: Sentiment Analysis by LSTM

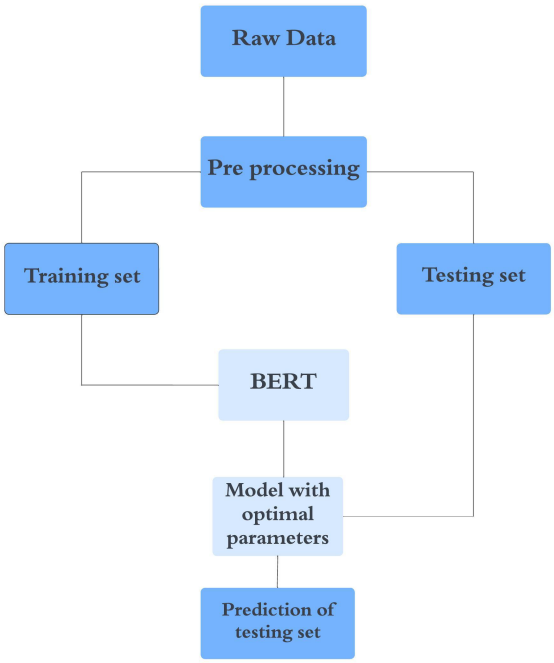

Figure 4: Sentiment Analysis by BERT

were successfully classified to all instances is used to determine the overall accuracy of the model's predictions. However, accuracy by itself could not give a full picture, particularly when dealing with datasets that are unbalanced. Precision: Out of all instances anticipated as positive (or negative), precision represents the percentage of accurately predicted positive (or negative) cases. The accuracy of positive (or negative) predictions is the main focus, and it aids in determining the model's capacity to prevent false positives. Confusion Matrix: A confusion matrix displays the predictions of the model in comparison to the actual classes in a tabular format. Insights into true positives, true negatives, false positives, and false negatives are provided, assisting in the identification of certain areas that require improvement. Table 1 lists the accuracy of each proposed model, and it is evident from the findings that the LSTM model offers superior accuracy with a score of 0.61.Tables 2, 3, and 4 show the Precision, Recall, and f1-Score for each suggested model, respectively. The proportion of accurately predicted positive (or negative) cases out of all actual positive (or negative) instances is determined by recall (also known as sensitivity or true positive rate). It demonstrates how the model can recognise positive (or negative) cases and prevent false negatives. F1-Score: The F1-score is a balanced indicator of a model's performance because it is the harmonic mean of precision and recall. It is helpful when the dataset is unbalanced since it takes precision and recall into account.

Model	Accuracy
Random Forest	0.54
Decision Tree	0.42
Logistic Regression	0.60
GaussianNb	0.14
LSTM	0.61
BERT	0.59

Table 1: Accuracy of proposed models

Model	Precision	Recall	F1-score
GNB	0.62	0.07	0.12
LR	0.36	0.10	0.15
DT	0.14	0.15	0.14
RF	0.22	0.15	0.18

Table 2:
Classification Report of Mixed Feeling Class Label

Model	Precision	Recall	F1-score
GNB	0.13	1.00	0.23
LR	0.14	0.0	0.0
DT	0.15	0.16	0.15
RF	0.22	0.12	0.16

Table 3: Classification Report of Negative Class Label

Model	Precision	Recall	F1-score
GNB	0.35	0.02	0.03
LR	0.42	0.01	0.03
DT	0.18	0.19	0.18
RF	0.23	0.10	0.81

Table 4: Classification Report of Unknown State Class Label

Model	Precision	Recall	F1-score
GNB	0.0	0.0	0.0
LR	0.61	0.98	0.75
DT	0.64	0.59	0.62
RF	0.63	0.82	0.71

Table 5:
Classification Report of Positive Class Label

5 Conclusion

In this research work, sentiment analysis is performed using machine learning, deep learning and transfer learning techniques, specifically Gaussian Naive Bayes, Decision Trees,Random Forests,BERT, and LSTM. For the machine learning approach, we employed logistic regression, decision tree, and random forest algorithms, while for the deep learning approach, we utilized LSTM models and for transfer learning approach, we utilized BERT. After evaluating the models, we obtained an accuracy of 42 percent for the decision tree and 61 percent for the LSTM. In summary, the LSTM model outperformed the decision tree algorithm's precision, achieving a higher accuracy of 61 percent for sentiment analysis. However, the performance can be further improved by fine-tuning hyperparameters, exploring different architectures, or incorporating ensemble methods.

References

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

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.

Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 69–78.

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.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.