Enhancing Sentiment Analysis using BERT-Hybrid Model for Detection of Irony and Sarcasm in Code-Mixed Social Media

Pinaki Sahu , Dr. Nagaraja S R (Asst.Prof)

School of Computer Science Engineering

Presidency University Bangalore, Karnataka

Abstract.

Sarcasm is a common verbal irony that can be difficult to apply in code-mixed social media, as people fluently switch between languages. The difficulty of sarcasm detection in these kinds of language environments is the subject of this study. Conventional models find it difficult to capture the complex interactions between different languages, slang terms, and cultural quirks. As a result, we suggest an Hybrid model based on BERT that efficiently detects sarcasm by utilizing the contextual knowledge of BERT embeddings. Combining different BERT versions, the Hybrid model performs better at detecting and interpreting sarcasm in code-mixed social media. Our method greatly advances the state of natural language processing through extensive experimentation, especially when it comes to managing the intricacies of multilingual and multicultural online communication.

Keywords: Sarcasm detection, code-mixing, BERT Hybrid model, linguistic complexity.

1.Introduction

1.1 Background

Social media sarcasm detection is a sentiment analysis difficulty. This task is more complicated with code-mixed content. Code-mixing, where two or more languages are mixed in a statement or discourse, is becoming more common in globalized and multilingual cultures. Social media users blend languages to communicate thoughts, feelings, and sarcasm, highlighting this trend.

Sarcasm detection in a linguistically varied environment is difficult. Sarcasm, a sort of linguistic sarcasm, requires context and culture to interpret. Many NLP models focus on sarcasm detection in monolingual text, however standard methods fail with code-mixed content. Sarcasm identification requires knowledge of various languages' linguistic structures, nuanced clues, and cultural contexts, which makes it difficult.

Social media material is informal and dynamic, adding to its complexity. Social media messages utilize slang, acronyms, and non-standard grammar, making sarcasm detection harder. Social media's code-mixed language is difficult for traditional NLP models, which were built for formal and monolingual text.

2. Objective

Given these constraints, this project aims to create an Hybrid model that uses several Transformer-based architectures to identify sarcasm in code-mixed text. Transformer models, which excel in NLP tasks, help digest complex language patterns and comprehend contextual nuances. The suggested methodology uses these characteristics to analyse sarcasm more sophisticated and accurately in a multilingual and culturally diverse digital context.

The Hybrid technique improves model performance and resilience. Ensemble learning, which combines predictions from many models, improves accuracy and reduces overfitting to training data linguistic traits or biases. This method is useful for code-mixed sarcasm detection, where linguistic and cultural components might vary greatly.

This research seeks to improve sentiment analysis in multilingual and multicultural environments by developing NLP models that can handle code-mixed social media content. A model like this has major implications for computational linguistics research, social media analytics, marketing, and cross-cultural communication.

3.Models

NLP pioneer BERT (Bidirectional Encoder Representations from Transformers). BERT, created by Google, transformed machine language understanding. Instead of processing words individually, its Transformer design lets it process sentences as a whole. BERT's bidirectional comprehension lets it capture a word's context from its surrounding text, improving question answering, language inference, and named entity recognition. BERT's versatility and accuracy come from pre-training on a vast corpus of text and fine-tuning for individual jobs.

The BERT-LSTM model combines the strengths of BERT and LSTM networks. Recurrent neural networks (RNNs) like LSTM are good at handling sequential data like text. Its long-term memory helps it recognize language context. BERT-LSTM models combine BERT's bidirectional context understanding with LSTM's sequence modelling. This combination helps the model process and understand text data where word sequence and context are critical, delivering a more detailed understanding than either model alone.

The BERT-GRU model combines BERT with a Gated Recurrent Unit (GRU), a simpler RNN like LSTM. Since GRUs overcome the vanishing gradient problem in regular RNNs, they are good at modelling sequences. BERT's contextual awareness and GRUs' efficient sequence processing are coupled in a BERT-GRU model. GRUs use gating strategies to control information flow and capture sequence dependencies without as many parameters as LSTM. BERT-GRU models are faster and more efficient while still comprehending language sequences well.

The fusion of BERT's bidirectional context comprehension and Bidirectional Long Short-Term Memory (BiLSTM) networks in the BERT-BiLSTM model allows for a comprehensive understanding of sequential text data. The inclusion of BiLSTMs effectively captures the context from both past and future steps, enhancing the model's ability to comprehend sequential language patterns. Similarly, the BERT-BiGRU model merges BERT with Bidirectional Gated Recurrent Units (BiGRUs), offering a fast and streamlined approach to sequence processing. This integration maintains contextual awareness, allowing for swift and insightful comprehension of language sequences. With a harmonious balance between computational efficiency and sequence modeling, BERT-BiGRU proves to be an ideal choice for tasks that require both speed and nuanced context interpretation.

2. Related Works

2.1 Sarcasm detection in monolingual contexts

Sarcasm identification in monolingual contexts has been thoroughly explored, laying the groundwork for comprehending it in complex language contexts. Rule-based and machine learning methods dominated early approaches. M Bedi. et al [2]. employed prosodic elements in speech to detect sarcasm, while E Troiano al [1]. used exaggeration and unexpectedness in text. Deep learning has led researchers to neural network models. Sarcasm identification improved significantly with convolutional and recurrent neural networks.

2.2 Code-Mixed Language Processing

Social media has made code-mixed language processing popular. A Pratapa et al [3]. and KC Raghavi et al. studied code-mixed language identification and POS tagging. Standard NLP techniques struggle to handle code-mixed text's syntactic and semantic inconsistencies.

2.3 NLP Transformer Models

Transformer models like BERT (A Gillioz et al. and IV Tetko, [5][6]) advanced NLP. Self-attention models are good at capturing contextual information, making them good at sarcasm detection. NLP applications like sentiment analysis and language translation use BERT's bidirectional context understanding and GPT's generating capabilities.

2.3 NLP ensemble learning

Ensemble learning has been used in NLP tasks to increase model performance. Ensemble approaches worked well for sentiment categorization in J Fattahi et al. Ensemble techniques to sarcasm detection, especially in code-mixed settings, are underexplored. Using many models to represent the complexity of sarcasm in mixed-language literature seems promising [7].

2.4 Detecting Sarcasm in Coded Content

There is little research on code-mixed content sarcasm detection. S Swami et al. used rules to detect sarcasm in Hindi-English code-mixed data. Recently, deep learning has led to the use of neural networks for this job. These studies often struggle with data paucity and simulating sarcasm across languages and cultures.

2.5 The Research Gap

Sarcasm detection and code-mixed language processing have evolved, but the intersection of the two, especially employing advanced Transformer models and ensemble techniques, has not. A sophisticated Hybrid model that uses Transformer architectures to detect sarcasm in code-mixed social media content is the goal of this research.

3. Methodology

3.1 Data Gathering

The dataset for this study includes code-mixed Twitter posts. Popular platforms with different linguistic content are picked. Common code-mixing scenarios are represented by posts in, Hindi-English [9].

3.1.1 Process of annotation

A team of language experts in each language pair manually annotates each post for sarcasm. Sarcasm is identified both explicitly and contextually during annotation. Multiple annotators analyse each post and address disagreements to ensure annotation reliability.

3.2 Model Structure

Model Selection for Transformers

BERT, BERT-LSTM, BERT-GRU,BERT-BILSTM and BERT-BIGRU are Transformer-based designs in the Hybrid model. These models are chosen for their text context and semantics understanding success. Each model has strengths: BERT's bidirectional context understanding, BERT-LSTM's generative abilities, and BERT-GRU's permutation-based training.

3.2.1 Tuning Process

Transformer models are uniquely customized for sarcasm detection. Training models on the annotated dataset lets them adapt to code-mixed content sarcastically. The fine-tuning process is monitored to avoid overfitting and maintain model generalization.

3.2.2 Strategic Ensemble

Voting or stacking incorporates Transformer model results in the Hybrid model. Voting ensembles forecast sarcasm by majority vote of individual models. A meta-model is trained to predict the ultimate outcome from the outputs of the individual models in a stacking ensemble. To select the optimum ensemble methodology, experiments are done.

3.3 Training and Assessment

3.3.1 Training Process

The Hybrid model is trained by training the Transformer models and ensemble strategy. The models are trained on a split dataset including a validation part. This phase optimizes learning rate, batch size, and training epochs.

3.3.2 Performance Measures

The Hybrid model is evaluated using accuracy, precision, recall, and F1-score. These measurements reveal the model's sarcasm detection performance. While accuracy assesses prediction accuracy, precision and recall reflect the model's ability to recognize sarcastic posts, and the F1-score balances these.

3.3.3 Comparative Analysis

The Hybrid model is contrasted to Transformer models and standard sarcasm detection algorithms to prove its efficacy. The ensemble approach improves code-mixed sarcasm detection, as shown in this comparative analysis.

4. Results

Five machine learning models—BERT-LSTM, BERT-BiLSTM ,BERT-GRU, BERT-BiGRU, and an Hybrid Model integrating BERT-LSTM ,BiLSTM GRU and BiGRU—are carefully contrasted using various criteria. The comparison table.1 shows BERT-LSTM ,BiLSTM GRU and BiGRU models have similar accuracy of 89.5%, precision and recall of 0.89-0.90, and ROC-AUC scores of 0.96. The Hybrid Model performs somewhat better with 90% accuracy, 0.88–0.92 precision, and 0.88–0.92 recall for the two classes. The F1-score for both courses is 0.90. Hybrid Model's ROC-AUC score is 0.90, somewhat lower than the individual models, but its greater accuracy and balanced precision-recall imply more consistent performance across both classes. This shows that the Hybrid Model is the best of the three models reviewed in the text since it generalizes and balances classifications.

Model	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score	Accuracy	ROC-AUC Score
BERT-LSTM	0.89	0.90	0.90	0.89	0.89	89.59%	0.96
BERT-GRU	0.89	0.90	0.90	0.89	0.89	89.26%	0.96
BERT-BILSTM	0.89	0.90	0.90	0.89	0.89	89.45%	0.96
BERT-BIGRU	0.89	0.90	0.90	0.89	0.89	89.37%	0.96
Hybrid (BERT-LSTM+BiLSTM+GRU+BiGRU)	0.88	0.92	0.92	0.88	0.90	90.00%	0.90

Table.1. Comparison Table

5. Conclusion

The document's extensive study highlights the new approach and noteworthy findings in code-mixed social media sarcasm detection. The study's major goal was to create an Hybrid model using several Transformer-based architectures to identify sarcasm in code-mixed text. Given the specific constraints of code-mixed information, where linguistic and cultural nuances are critical to interpreting context and purpose, our technique is pioneering.

The study findings are noteworthy. The Hybrid Model, which combines BERT-LSTM and BERT-GRU, worked well. With an accuracy of 90% and precision and recall rates of 0.88 to 0.92 for the two classes, the Hybrid Model balances precision and recall better, making it better at handling code-mixed text sarcasm. Although its ROC-AUC score is slightly lower at 0.90 than the BERT-LSTM and BERT-GRU models (both at 0.96), the Hybrid Model's accuracy and balanced precision-recall demonstrate its efficiency and reliability in a complex linguistic environment.

This NLP breakthrough shows Hybrid approaches can handle linguistically diverse and culturally complicated material and establishes a new standard for code-mixed content sarcasm detection. Effective implementation of this approach allows for more nuanced and culturally sensitive AI tools in social media analytics, marketing, and cross-cultural communication. This research highlights the importance of context, cultural awareness, and linguistic diversity in intelligent system development, shedding light on computational linguistics and AI.

6.References

[1] Troiano E, Strapparava C, Özbal G, Tekiroğlu SS. A computational exploration of exaggeration. InProceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2018 (pp. 3296-3304).

[2] Bedi M, Kumar S, Akhtar MS, Chakraborty T. Multi-modal sarcasm detection and humor classification in code-mixed conversations. IEEE Transactions on Affective Computing. 2021 May 26.

[3] Pratapa A, Choudhury M, Sitaram S. Word embeddings for code-mixed language processing. InProceedings of the 2018 conference on empirical methods in natural language processing 2018 (pp. 3067-3072).

[4] Raghavi KC, Chinnakotla MK, Shrivastava M. " Answer ka type kya he?" Learning to Classify Questions in Code-Mixed Language. InProceedings of the 24th International Conference on World Wide Web 2015 May 18 (pp. 853-858).

[5] Gillioz A, Casas J, Mugellini E, Abou Khaled O. Overview of the Transformer-based Models for NLP Tasks. In2020 15th Conference on Computer Science and Information Systems (FedCSIS) 2020 Sep 6 (pp. 179-183). IEEE.

[6] Tetko IV, Karpov P, Van Deursen R, Godin G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nature communications. 2020 Nov 4;11(1):5575.

[7] Fattahi J, Mejri M. SpaML: a bimodal ensemble learning spam detector based on NLP techniques. In2021 IEEE 5th international conference on cryptography, security and privacy (CSP) 2021 Jan 8 (pp. 107-112). IEEE.

[8] Swami S, Khandelwal A, Singh V, Akhtar SS, Shrivastava M. A corpus of english-hindi code-mixed tweets for sarcasm detection. arXiv preprint arXiv:1805.11869. 2018 May 30.

[9]. Shah, A., & Maurya, C. K. (2022). How Effective is Incongruity? Implications for Code-mix Sarcasm Detection. arXiv preprint arXiv:2202.02702.

Dataset: - https://github.com/likemycode/codemix