CIC-IPN at SemEval-2025 Task 11: Transformer-Based Approach to Multi-Class Emotion Detection

Abiola O.J.[2], Abiola T.O.[1], Ojo O.E.[1], Kolesnikova O.[1], Sidorov G.[1], H. Calvo[1],
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico.
[2]Federal University Oye-Ekiti, Ekiti, Nigeria.
Correspondence: sidorov@cic.ipn.mx

Abstract

This paper presents a multi-step approach for multi-label emotion classification as our system description paper for the SEMEVAL-2025 workshop Task A using machine learning and deep learning models. We test our methodology on English, Spanish, and low-resource Yoruba datasets, with each dataset labeled with five emotion categories: anger, fear, joy, sadness, and surprise. Our preprocessing involves text cleaning and feature extraction using bigrams and TF-IDF. We employ logistic regression for baseline classification and fine-tune Transformer models, such as BERT and XLM-RoBERTa, for improved performance. The Transformer-based models outperformed the logistic regression model, achieving micro-F1 scores of 0.7061, 0.7321, and 0.2825 for English, Spanish, and Yoruba, respectively. Notably, our Yoruba fine-tuned model outperformed the baseline model of the task organizers with micro-F1 score of 0.092, demonstrating the effectiveness of Transformer models in handling emotion classification tasks across diverse languages.

1 Introduction

Emotions play a crucial role in human communication, shaping our interactions, decisions, and psychological well-being. According to the Oxford English Dictionary, emotion is defined as "a strong feeling deriving from one's circumstances, mood, or relationships with others." In social interactions, emotions are frequently invoked and help individuals navigate complex relationships and make sense of their environments Hwang and Matsumoto, 2016. In text, emotions can be conveyed explicitly or implicitly through linguistic patterns, allowing authors to communicate their mental states. Consequently, emotion classification—identifying and labeling the emotions embedded in text—has become a key research area in Natural Language Processing (NLP), with applications across various domains such as marketing, healthcare, and education.

While sentiment analysis focuses on determining the overall emotional tone (positive, negative, or neutral) of a text, emotion classification goes a step further by identifying specific emotions such as anger, joy, fear, or sadness. This task is particularly challenging when multiple emotions are present simultaneously in a single text, a problem known as multi-label emotion classification. Unlike traditional single-label emotion classification, where only one emotion label is associated with a given statement, multi-label classification assigns multiple emotion labels to a text, reflecting the complex nature of human emotional expression.

Multi-label emotion classification faces numerous challenges, particularly in the realm of social media, where the language evolves rapidly and the context is often ambiguous. To address these challenges, researchers have turned to deep learning models, especially Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms, which have shown promising results in recent years. These models can capture nuanced emotional expressions by learning from large-scale datasets, such as the SemEval-2018 Task-1 and the Blog Emotion Corpus, which contain emotional annotations for a variety of social media texts.

However, despite advancements in deep learning, existing models still face limitations in capturing the full range of emotional nuances in multi-label classification tasks. Moreover, traditional machine learning methods often struggle with feature engineering, making them less adaptable to rapidly changing language used on platforms like Twitter. In light of these challenges, this paper explores Transformer approaches to multi-label emotion detection by leveraging transfer learning and multi-attention mechanisms. By fine-tuning pre-trained

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1609–1615

models such as BERT and XLMRoBERTa, along with incorporating feature engineering to capture emotion-specific features, we aim to enhance the accuracy and robustness of emotion classification systems.

In the following sections, we first review the related work on emotion classification, particularly in multi-label contexts. Then, we describe our methodology, which includes data preprocessing, model architecture, and experimental setup. Finally, we present our experimental results, demonstrating the effectiveness of our approach in English, Spanish and Yoruba emotion classification tasks.

2 Recent Literature

Text detection and classification has taken several forms and gained attention by researchers over the last couple of years in NLP, with the application of different classifiers and models, also some more accurate models being developed by researchers, performing significant roles in the series of experiments that have been undertaken in recent work Abiola et al., 2025b; Kolesnikova and Gelbukh, 2020; Ojo et al., 2024; Adebanji et al., 2022; Abiola et al., 2025a. A variety of traditional ML methods Ojo et al., 2021, 2020; Sidorov et al., 2013 and DL models Aroyehun and Gelbukh, 2018; Ashraf et al., 2020; Han et al., 2021; Hoang et al., 2022; Poria et al., 2015; Muhammad et al., 2025a have been applied in the last few years for text prediction on various domains.

Most previous work on emotion detection has focused on deep neural networks such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) networks. However, these models have not utilized multiple attention mechanisms or Transformer-based networks such as XLNet, DistilBERT, and RoBERTa for multi-label emotion classification. Ameer et al., 2023 proposed multiple attention mechanisms to reveal the contribution of each word to each emotion, which had not been investigated before. Their RoBERTa-Multi-Attention (RoBERTa-MA) model achieved 62.4% accuracy, outperforming the previous state-of-the-art accuracy of 58.8% on the SemEval-2018 Task-1C dataset. Similarly, their XLNet-MA model achieved 45.6% accuracy on the Ren-CECps dataset for Chinese, demonstrating the effectiveness of Transformer-based models in multi-label emotion classification.

Shahiki and et al., 2024 conducted a psycholinguistic and emotional analysis of cryptocurrency discussions on social media, focusing on nine major digital assets, including Bitcoin, Ethereum, and Dogecoin. Using advanced text analysis techniques, the study examined linguistic patterns and emotional expressions across different cryptocurrency communities. The authors also analyzed comentions among cryptocurrencies to understand their interrelations. A dataset of 832,559 tweets was collected and refined to 115,899 for analysis, providing insights into the distinct discourse surrounding each coin and the emotional trends shaping cryptocurrency discussions online.

Speech emotion recognition (SER) plays a crucial role in enhancing human-computer interaction (HCI) by enabling machines to understand human emotions from acoustic signals. However, the lack of large-scale datasets remains a major challenge in this field. BanSpEmo: A Bangla Audio Dataset for Speech Emotion Recognition and Its Baseline Evaluation Kusal et al., 2025 addresses this issue by introducing BANSpEmo, a Bangla speech emotion dataset comprising 792 recordings from 22 native speakers, covering six emotions: disgust, happiness, anger, sadness, surprise, and fear. The study evaluates baseline models, including support vector machine (SVM), logistic regression (LR), and multinomial Naïve Bayes, finding that SVM achieves the highest accuracy of 87.18%. Inspired by this work, transformer-based models provide an advanced approach to SER by leveraging deep contextual representations for improved multi-class emotion detection across languages. This study builds upon such datasets to enhance cross-linguistic emotion classification using transformer architectures.

Emotion detection in online communication has been extensively studied, but many approaches focus solely on textual cues, overlooking the role of emojis in conveying emotions. Multimodal Text-Emoji Fusion Using Deep Neural Networks for Text-Based Emotion Detection in Online Communication Kusal et al., 2025 highlights the significance of incorporating emoji analysis to improve sentiment interpretation, especially in cases where text alone may not fully capture emotional intent. The study proposes an emoji-aware hybrid deep learning framework that leverages convolutional and recurrent neural networks for multimodal emotion detection. Inspired by this, transformer-

based models offer a promising approach to cross-linguistic multi-class emotion detection, as they can capture deep contextual relationships in multi-modal data. This study builds on such insights by integrating transformer architectures for improved emotion classification in diverse linguistic settings.

3 Methodology

Our methodology involves a multi-step approach to text preprocessing, feature extraction, and multi-label emotion classification using machine learning and deep learning models. Experiments were conducted on English and Spanish datasets to test our method across different languages, including the low-resource language Yoruba.

3.1 Dataset Preprocessing

We loaded our datasets into Pandas DataFrames from three separate files for training, development (validation), and test datasets as given by the shared task organizers Muhammad et al., 2025b. The datasets contain text samples labeled with five emotion categories: anger, fear, joy, sadness, and surprise. Class 0 depicts no emotion for each class, and class 1 depicts emotion for each class. Table 1, 2 and 3 give insight into the English, Spanish and Yoruba datasets, respectively.

Table 1: Binary Emotion Label Distribution in the Dataset

Emotion	1 (Present)	0 (Absent)
Anger	333	2435
Fear	1611	1157
Joy	674	2094
Sadness	878	1890
Surprise	839	1929

Table 2: Binary Emotion Label Distribution in the Spanish Dataset

Emotion	1 (Present)	0 (Absent)
Anger	492	1504
Disgust	654	1342
Fear	317	1679
Joy	642	1354
Sadness	309	1687
Surprise	421	1575

Our preprocessing involved the removal of special characters, non-word tokens, extra whitespace, and lowercasing the text with regex expressions.

Table 3: Binary Emotion Label Distribution in the Yoruba Dataset

Emotion	1 (Present)	0 (Absent)
Anger	195	2797
Disgust	81	2911
Fear	77	2915
Joy	272	2720
Sadness	836	2156
Surprise	254	2738

Stopword removal was not applied initially, as certain emotion-indicating words might be filtered out inadvertently. The frequency distribution of emotions was analyzed to understand class imbalances. This analysis was performed using Pandas.

3.2 Bigram Feature Augmentation

To enhance text representations, we extracted top bigrams using `CountVectorizer` module from `sklearn`. The bigram extraction process involved transforming the text into a bag-of-words model with bigram tokens, then we computed their frequencies across the dataset and select the most frequent bigrams. These bigrams were then appended to the original text, improving contextual information while preserving sequence structures.

We made a helper function that iterated over each text entry and appended bigram tokens that appeared in the sample, concerning their frequencies. Our additional feature engineering improved the model's ability to capture co-occurring patterns that signal emotional expression.

3.3 TF-IDF Feature Extraction and Logistic Regression Model

We converted the preprocessed text data into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization through `TfidfVectorizer` from Scikit-learn. This transformation helped quantify the importance of words relative to the entire corpus.

Our multi-label classification approach was adopted using `OneVsRestClassifier`, which trains separate Logistic Regression classifiers for each emotion label. The model was trained with `max_iter=1000` to ensure convergence. Predictions were evaluated using precision, recall, and F1-score metrics from `sklearn.metrics`.

3.4 Fine-tuning Transformer Models for Emotion Classification

Beyond traditional machine learning, we implemented a deep learning approach using the base BERT model (`bert-base-uncased`) for the English dataset and (`xlm-roberta-base`) for the Yoruba and Spanish datasets from the Hugging Face Transformers library. The text was tokenised using `BertTokenizer`, truncating longer texts to a maximum of 512 tokens.

We fine-tuned the models for classification on the dataset with the problem type set to `multi-label classification`, using a binary cross-entropy loss function (`BCEWithLogitsLoss`). Class imbalance was mitigated using weighted loss, computed based on the proportion of positive and negative instances for each emotion label.

The dataset was converted into a Hugging Face Dataset format and tokenised for efficient batching. Training was performed using the Hugging Face `Trainer` API, with a batch size of 16, 20 epochs, and weight decay of 0.01. The model was evaluated based on the macro-F1 score and fine-tuned on an RTX 3080 Nvidia 16GB GPU.

3.5 Evaluation Metrics and Performance Analysis

Model performance was assessed using precision, recall, and F1-score. The BERT model predictions were converted into probabilities using the sigmoid function with a threshold for label assignment.

4 Results

4.1 Performance of Logistic Regression

The performance metrics of the logistic regression model on the dev dataset give a micro F1 score of 0.50. The model achieved a reasonable F1-score on fear detection on the English dataset since it takes the majority of the present emotion class in the dataset but struggled with minority emotion classes due to data imbalance.

4.2 Performance of Fine-Tuned BERT Model

We performed the final test set predictions on the Transformer models since it outperformed the logistic regression model on the development dataset for the three languages we worked on. As a blind grading requested by the organizers of the workshop, the predicted labels of the Transformer models were stored in a CSV file, where each row contained an ID and predicted binary emotion labels.

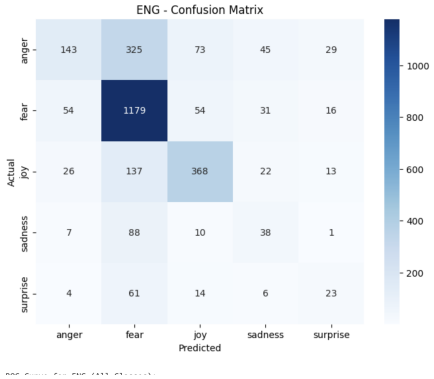

Figure 1: Confusion Matrix of the English Dataset Prediction

Post-processing ensured no missing values, and labels were converted to binary (0 or 1) to align with the dataset format.

The model has a micro-F1 score of 0.7061, 0.7321 and 0.2825 for the English, Spanish and Yoruba datasets, respectively; our English and Spanish models underperformed slightly to the SEMEVAL Baseline model that has the micro-F1 score of 0.7083 and 0.7744 but our Yoruba fine-tuned model outperformed the baseline model that has micro-F1 score of 0.0922 The result analysis for the Transformer models prediction on the test dataset in this languages are displayed in figures 1 to 6.

The low micro-F1 score observed for the Yoruba dataset in the multiclass emotion detection task can be attributed to several challenges associated with low-resource languages. Primarily, the limited availability of annotated training data in Yoruba likely hindered the model's ability to generalise well across different emotion classes. Additionally, the pretrained Transformer models' exposure to Yoruba during pretraining, results in weaker language representations. Cultural and linguistic nuances in emotion expression, which are often context-dependent and idiomatic in Yoruba, also contribute to the difficulty in accurately detecting emotions. These factors combined likely led to the model's reduced performance compared to English and Spanish.

5 Conclusion

The experiment demonstrated the effectiveness of Transformer models over traditional machine learning for multi-label emotion classification. The addition of bigram features enhanced feature representation for logistic regression and the deep learning

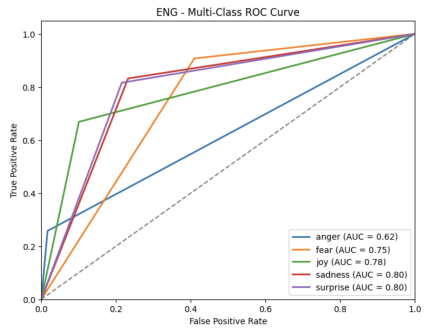

Figure 2: ROC of the English Dataset Prediction

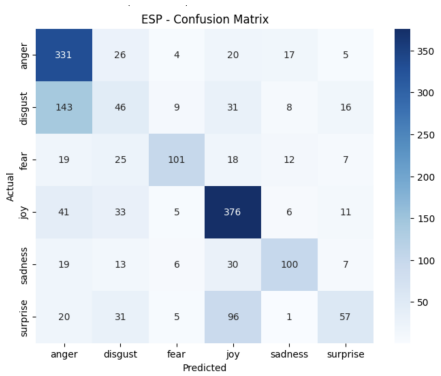

Figure 3: Confusion Matrix of the Spanish Dataset Prediction

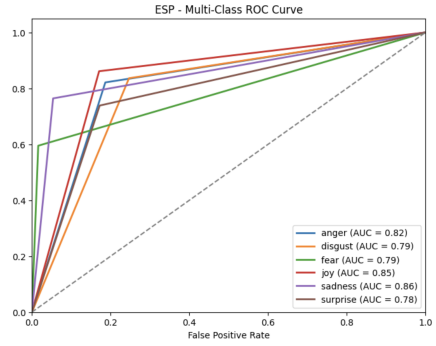

Figure 4: ROC of the Spanish Dataset Prediction

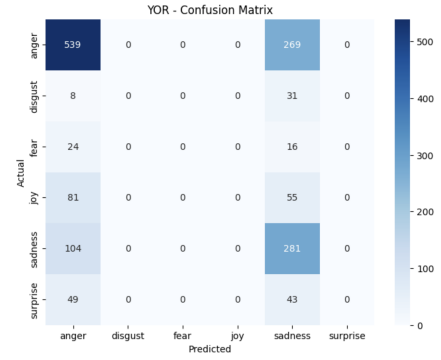

Figure 5: Confusion Matrix of the Yoruba Dataset Prediction

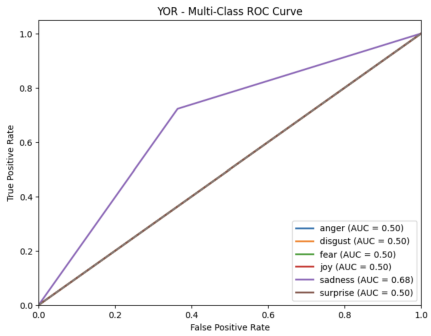

Figure 6: ROC of the Yoruba Dataset Prediction

model used as we discovered during the development stage with dev dataset that it improves the result of each of the models between +3 to +10%, but deep learning provided a more robust approach to capturing contextual meaning. Future work will explore cross-lingual emotion classification and domain-specific fine-tuning for improved performance.

6 Limitations

Despite the promising results of our multilingual emotion classification approach, several limitations must be acknowledged. Data imbalance, particularly in low-resource languages like Yoruba, affects model performance, leading to biased predictions where underrepresented emotions are harder to detect. While weighted loss functions and data augmentation techniques were applied, challenges in balancing class distributions persist. Finally, linguistic and cultural variations across English, Spanish, and Yoruba affect emotion representation. Future research should explore cross-lingual adaptation strategies, improved data augmentation for low-resource languages, and adaptive thresholding mechanisms to enhance classification performance.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope O. Abiola, Tewodros A. Bizuneh, Oluwatobi J. Abiola, Temitope O. Oladepo, Olumide E. Ojo, Adebanji O. O., Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tolulope O. Abiola, Tewodros A. Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide E. Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Olaronke Oluwayemisi Adebanji, Irina Gelbukh, Hiram Calvo, and Olumide Ebenezer Ojo. 2022. Sequential models for sentiment analysis: A comparative study. In *Advances in Computational Intelligence, 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Proceedings, Part II*, Monterrey, Mexico. Springer.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

S.T. Aroyehun and A. Gelbukh. 2018. Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

N. Ashraf, R. Mustafa, G. Sidorov, and A.F. Gelbukh. 2020. Individual vs. group violent threats classification in online discussions. In *Companion of The 2020 Web Conference*, pages 629–633, Taipei, Taiwan.

W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.P. Morency, and S. Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15, New York, NY, USA. Association for Computing Machinery.

Thang Ta Hoang, Olumide Ebenezer Ojo, Olaronke Oluwayemisi Adebanji, Hiram Calvo, and Alexander Gelbukh. 2022. The combination of bert and data oversampling for answer type prediction. In *Proceedings of the Central Europe Workshop*, volume 3119.

H. C. Hwang and D. Matsumoto. 2016. Emotional expression. In C. Abell and J. Smith, editors, *The Expression of Emotion: Philosophical, Psychological and Legal Perspectives*, pages 137–156. Cambridge University Press, Cambridge.

O. Kolesnikova and A. Gelbukh. 2020. A study of lexical function detection with word2vec and supervised machine learning. *J. Intell. Fuzzy Syst.*, 39.

Sheetal Kusal, Shruti Patil, and Ketan Kotecha. 2025. Multimodal text-emoji fusion using deep neural networks for text-based emotion detection in online communication. *Journal of Big Data*, 12.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

OE Ojo, A Gelbukh, H Calvo, and OO Adebanji. 2021. Performance study of n-grams in the analysis of sentiments. *Journal of the Nigerian Society of Physical Sciences*, pages 477–483.

O.E. Ojo, A. Gelbukh, H. Calvo, O.O. Adebanji, and G. Sidorov. 2020. Sentiment detection in economics texts. In *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, 12–17 October 2020, Proceedings, Part II*, pages 271–281, Berlin, Heidelberg. Springer-Verlag.

Olumide E Ojo, Olaronke O Adebanji, Hiram Calvo, Alexander Gelbukh, Anna Feldman, and Ofir Ben Shoham. 2024. Doctor or ai? efficient neural network for response classification in health consultations. *IEEE Access*.

S. Poria, E. Cambria, and A. Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*.

Tash Shahiki and et al. 2024. Psycholinguistic and emotional analysis of cryptocurrency discussions on social media: Focusing on nine major digital assets. *Journal of Computational Linguistics and Social Media*, 14:2703.

G. Sidorov et al. 2013. Empirical study of machine learning based approach for opinion mining in tweets. In *MICAI 2012. LNCS (LNAI)*, volume 7629, pages 1–14, Heidelberg. Springer.