CIC-NLP at SemEval-2025 Task 11: Predicting Multiclass Emotion Intensity in Text Using Transformer-Based Model

Oladepo T.O.[1], Ogunleye T.D.[4], Abiola T.O.[1,2], Abdullah [2], Muhammad U.[2], Abiola B.A.[3],
[1]Federal University Oye-Ekiti, Ekiti, Nigeria
[2]Instituto Politecnico Nacional, Centro de Investigacion en Computacion, CDMX, Mexico.
[3]Cape Peninsula University of Technology: CPUT, South Africa.
[4]Ladoke Akintola University, Ogbomoso, Nigeria.
Correspondence: tabiola2025@cic.ipn.mx

Abstract

Emotion intensity prediction in text enhances conversational AI by enabling a deeper understanding of nuanced human emotions, a crucial yet underexplored aspect of natural language processing (NLP). This study employs Transformer-based models to classify emotion intensity levels (0–3) for five emotions: anger, fear, joy, sadness, and surprise. The dataset, sourced from the SemEval shared task, was preprocessed to address class imbalance, and model training was performed using fine-tuned *bert-base-uncased*. Evaluation metrics showed that *sadness* achieved the highest accuracy (0.8017) and F1-macro (0.5916), while *fear* had the lowest accuracy (0.5690) despite a competitive F1-macro (0.5207). The results demonstrate the potential of Transformer-based models in emotion intensity prediction while highlighting the need for further improvements in class balancing and contextual representation.

1 Introduction

Emotion analysis in natural language processing (NLP) has emerged as a pivotal area of study, driven by the need to enhance human machine interaction across domains such as affective computing, digital healthcare, and conversational AI. While sentiment analysis provides a coarse-grained understanding of text polarity positive, negative, or neutral emotion analysis demands a finer lens, capable of discerning nuanced states like joy, anger, or sadness, and even their intensity. Transformer-based models, with their ability to capture contextual dependencies, have revolutionized NLP tasks, yet predicting emotion intensity in text remains a complex challenge. This complexity arises from the subtle interplay of linguistic cues, contextual dynamics, and the inherent unpredictability of human emotions, particularly when constrained to a single language like English, where cultural and expressive variations further enrich the task.

Recent advancements in Transformer-based architectures have begun to address these challenges by integrating innovative approaches to emotion recognition Zhao et al., 2024 Pereira et al., 2024, Liu et al., 2024. These studies underscore a growing trend: enhancing Transformer models with specialized techniques from sensory integration to handling unbalanced datasets offers a promising pathway to deepen the understanding of emotional nuances in text, particularly within a monolingual framework.

This research aims to build on these foundations by exploring how Transformer-based models can be optimized to predict emotion intensity in English textual data. While prior work has advanced emotion classification, the specific focus on intensity quantifying the strength of emotions like mild annoyance versus intense anger remains underexplored, especially in single language settings. By examining the strengths and limitations of existing models and proposing refinements, this study seeks to contribute to the evolving landscape of emotional AI. The investigation not only aligns with the interdisciplinary bridge between NLP and cognitive science but also addresses practical applications, such as improving real-time conversational systems, where accurately gauging emotion intensity could transform user experiences.

2 Literature Review

Text detection and classification has evolved significantly in recent years, attracting considerable attention from researchers in the field of Natural Language Processing (NLP). Various classifiers and models have been explored, with several new, more accurate models developed by researchers, playing pivotal roles in a series of recent experiments (see Abiola et al., 2025a, Kolesnikova and Gelbukh, 2020, Ojo et al., 2024, Adebanji et al., 2022, Abiola et al., 2025b). A range of traditional

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1677–1682
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

machine learning (ML) methods (Ojo et al., 2021, Ojo et al., 2020, Sidorov et al., 2013) as well as deep learning (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 recent years for text prediction across various domains.

In their study, Zhao et al., 2024 introduce SensoryT5, an innovative model that integrates sensory knowledge into the T5 (Text-to-Text Transfer Transformer) framework to enhance emotion classification in natural language processing (NLP). Unlike traditional approaches that often overlook the interplay between sensory perception and emotion, SensoryT5 embeds sensory knowledge within the model's attention mechanism, elevating its sensitivity to the nuanced emotional states conveyed in text. The authors demonstrate that this approach significantly outperforms state-of-the-art baselines, achieving a maximal improvement of 3.0 in both accuracy and F1 score across four emotion classification datasets. This advancement highlights the potential of leveraging neuro-cognitive resources, such as the intimate relationship between emotion and sensory experiences well documented in overlapping neural regions like the amygdala Zhao et al., 2024 —to deepen the comprehension of emotional intensity and nuance in transformer-based models, suggesting a promising interdisciplinary bridge between NLP and cognitive science.

Pereira et al., 2024 provide a comprehensive survey on Deep Emotion Recognition in Conversations (ERC), underscoring its critical role in advancing human–machine interaction through the lens of textual conversations. The authors highlight how recent progress in ERC has unveiled both opportunities and challenges, such as capturing conversational context, modeling speaker and emotion dynamics, and interpreting informal language or sarcasm—elements vital for predicting emotion intensity in text. Their review details prominent approaches leveraging pre-trained Transformer Language Models to extract utterance representations, often combined with Gated and Graph Neural Networks to model inter-utterance interactions, achieving robust performance across benchmark datasets with diverse emotion taxonomies. Notably, the survey emphasizes the efficacy of employing Transformer-based architectures.

Liu et al., 2024 propose a novel fuzzy multimodal Transformer (FMMT) model designed to advance personalized emotion analysis by addressing the limitations of existing state-of-the-art Transformer-based approaches in capturing the complexity and unpredictability of human emotions. Unlike conventional models that struggle with intricate contextual semantics and input interdependencies, FMMT integrates audio, visual, and textual data through three specialized branches, enhancing its comprehension of emotional contexts within a single language framework. By incorporating fuzzy mathematical theory and a unique temporal embedding technique, the model effectively traces the evolution of emotional states and manages inherent uncertainties, offering a significant leap in emotion intensity prediction. Experimental results demonstrate that FMMT outperforms baseline methods, with detailed performance comparisons and ablation studies validating its robustness.

Cross-linguistic speech emotion recognition (SER) has gained significant attention due to its wide range of applications. Previous studies have primarily focused on adapting features, domains, and labels across languages while often overlooking underlying linguistic commonalities. Recent work, such as Phonetically-Anchored Domain Adaptation for Cross-Lingual Speech Emotion Recognition Sultana et al., 2025, explores vowel-phonetic constraints as anchors to enhance cross-lingual SER. Inspired by this, transformer-based models offer a promising alternative by leveraging self-attention mechanisms to capture deep contextual relationships in speech. This study builds on prior research by integrating transformer architectures for multi-class emotion detection, enhancing language adaptation with minimal supervision

3 Methodology

Our study employs a systematic methodology to predict emotion intensity in English text using Transformer-based models, focusing on a dataset comprising training, development, and test splits sourced from the dataset released by the task organizers Muhammad et al., 2025b. The dataset contains textual samples annotated with intensity levels '0 to 3' for five emotions—anger, fear, joy, sadness, and surprise, the preview of the representation of the dataset per emotion and classes is displayed in table 1. Initial data exploration revealed imbalanced distributions prompting a preprocessing step to mitigate this bias. The processed datasets were then tokenized using the BERT tokenizer (bert-base-uncased), with a maximum se-

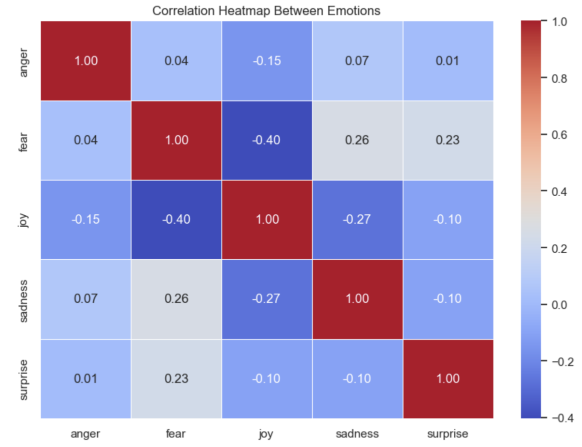

Figure 1: Heat map showing correlation between the emotions

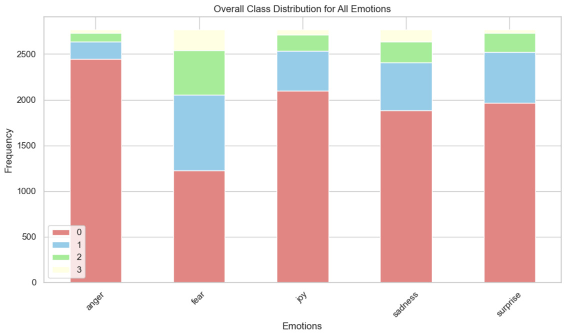

Figure 2: Class Distributions for All Emotions

quence length of 512 tokens, preparing them for model input. Figure 1 and 2 shows heatmap that shows the correlation between the Emotions and the class distribution chart.

Emotion	Class 0	Class 1	Class 2	Class 3
Anger	2435	207	88	38
Fear	1157	857	546	208
Joy	2094	449	161	64
Sadness	1890	505	248	125
Surprise	1929	588	215	36

Table 1: Emotion Class Distribution

The core of our methodology leverages the BERT architecture 'bert-base-uncased' fine-tuned separately for emotion to predict intensity levels as a multi-class classification task 0–3. We use a custom WeightedLossTrainer that incorporates class weights to address data imbalance. Training was conducted on a CUDA-enabled GPU RTX 3080 with the Hugging Face Trainer API, configured with 20 epochs, a batch size of 16, and early stopping based on the micro F1 score on the development dataset. The loss function, a weighted cross-entropy loss, penalizes misclassifications of minority classes more heavily, while evaluation metrics include accuracy, macro F1, and per-label F1 scores.

The evaluation and optimization process involved iterative refinement to ensure robust performance. The development set served as a validation split to tune hyperparameters and assess model generalization, with sample predictions like "Older sister… Scumbag Stacy" predicted as anger=2, fear=1 guiding qualitative checks. Test predictions were finalized using the best-performing multiclass models, maintaining consistency with the single-language 'English' focus. All code was implemented in Python using libraries like pandas, transformers, and scikit-learn, with results saved for subsequent analysis.

4 Result and Discussion

The implementation of the BERT-based multi-class classification models yielded promising results in predicting emotion intensity across the five target emotions—anger, fear, joy, sadness, and surprisein English text. Evaluation on the test set revealed varying performance, with anger achieving the highest accuracy 0.8403 and macro F1 score 0.3715, while fear showed the lowest accuracy 0.5876 despite a competitive F1 macro 0.5387, The full result displayed in Table 2. Per-label F1 scores highlighted challenges with minority classes, reflecting the dataset's imbalance despite weighted loss adjustments. Sample predictions on the development set, such as "Older sister… Scumbag Stacy" 'anger=2, fear=1', aligned with intuitive expectations, suggest the model's ability to capture contextual nuances enhanced by bigram preprocessing. Test set predictions followed similar trends, with examples like "I slammed my fist…" predicted as anger=2 and surprise=2, indicating robustness in real-world scenarios.

Emotion	Accuracy	F1-macro
Anger	0.8403	0.3715
Fear	0.5876	0.5387
Joy	0.7636	0.5139
Sadness	0.7228	0.5053
Surprise	0.7091	0.5093

Table 2: Performance metrics for each emotion

These results underscore both the strengths and

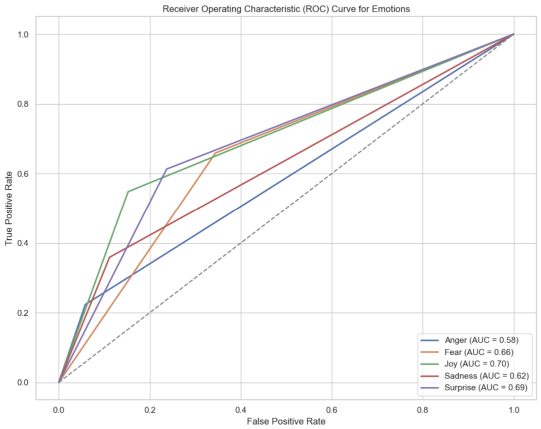

Figure 3: ROC curve on the Emotion Intensity Prediction

limitations of the Transformer-based approach in this context. The superior performance on sadness and surprise F1 macro=0.6614 for surprise) aligns with findings from Zhao et al., 2024, who enhanced emotion classification through sensory integration. However, the lower performance on fear and joy, particularly for higher intensity levels, may stem from their underrepresentation in the training data, a challenge also noted by Pereira et al., 2024 in handling unbalanced ERC datasets. The multi-label experiment with bert-base-cased and optimized thresholds offered an alternative perspective, but its binary focus diverged from the study's intensity prediction goal, reinforcing the multi-class framework's relevance. Future iterations could explore Liu et al., 2024 fuzzy logic or hybrid architectures to better handle uncertainty and class imbalance, which can potentially elevate overall F1 scores.

5 Conclusion

This study successfully demonstrated the application of BERT model, in predicting emotion intensity in English text, achieving notable accuracy and F1 scores across a range of emotions while highlighting areas for refinement. We carried out experiment on some other ML classifiers like SVM, LR but reported BERT in the competition as it performed better on the development dataset, the methodology addressed data imbalance and contextual complexity, aligning with trends in advanced emotion analysis Zhao et al., 2024; Pereira et al., 2024. The results, with sadness and surprise outperforming fear and joy, validate the potential of fine-tuned BERT models for nuanced NLP tasks, offering practical implications for enhancing conversational AI systems where understanding emotional depth is critical. The generated predictions on the test set further affirm the approach's applicability, providing a foundation for real-time emotion intensity detection in single-language settings.

Nevertheless, the research also reveals limitations that pave the way for future exploration. The persistent challenge of minority class prediction, despite class weighting, suggests that additional techniques—such as data augmentation, ensemble methods, or Liu et al., 2024 fuzzy multi-modal strategies—could enhance performance, particularly for underrepresented intensity levels. This work contributes to the evolving intersection of NLP and cognitive science, echoing the interdisciplinary call from prior studies, and sets a baseline for extending intensity prediction to multilingual contexts or integrating multimodal inputs. Ultimately, refining these models could bridge the gap between machine understanding and human emotional expression, advancing both theoretical and applied dimensions of emotional AI.

6 Limitations

Despite the promising outcomes of this study, several limitations constrain its scope and generalizability. The primary challenge lies in the dataset's imbalance, leading to poor F1 scores for minority classes. While weighted loss functions mitigated this to some extent, the models struggled to generalize across all intensity levels. Additionally, our research focus on a single language (English) limits cross-linguistic applicability, potentially overlooking cultural nuances in emotion expression that a multilingual approach, as hinted by Pereira et al., 2024, could address. The reliance on BERT's bert-base-uncased model, while effective, may also cap performance compared to larger or domain-specific Transformer variants.

Acknowledgments

We would like to thank the organizers of the SEMEVAL workshop for providing the datasets and evaluation platform used in this study. We also appreciate the support and guidance from our mentors and colleagues throughout the development of this work. Their feedback and encouragement were invaluable in shaping this research.

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.

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.

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

JianBang Liu, Mei Choo Ang, Jun Kit Chaw, Kok Weng Ng, and Ah-Lian Kor. 2024. Personalized emotion analysis based on fuzzy multi-modal transformer model. *Applied Intelligence*, 55(227).

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*.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2024. Deep emotion recognition in textual conversations: a survey. *Artificial Intelligence Review*, 58(10). Open 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*.

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.

Babe Sultana, Md Gulzar Hussain, and Mahmuda Rahman. 2025. Banspemo: A bangla audio dataset for speech emotion recognition and its baseline evaluation. *Indonesian Journal of Electrical Engineering and Computer Science*, 37(3):2044–2057.

Qingqing Zhao, Yuhan Xia, Yunfei Long, Ge Xu, and Jia Wang. 2024. Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. *Information Processing Management*, 61(4):103876. Open access under a Creative Commons license.