JUNLP_Sarika at SemEval-2025 Task 11: Bridging Contextual Gaps in Text-Based Emotion Detection using Transformer Models

Sarika Khatun, Dipanjan Saha, Dipankar Das
Jadavpur University, Kolkata, India
{sarikakhatun088, sahadipanjan6, dipankar.dipnil2005}@gmail.com

Abstract

Because language is subjective, it can be difficult to infer human emotions from textual data. This work investigates the categorization of emotions using BERT, classifying five emotions—angry, fearful, joyful, sad, and surprised—by utilizing its contextual embeddings. Preprocessing techniques like tokenization and stop-word removal are used on the dataset, which comes from social media and personal tales. With a weighted F1-score of 0.75, our model was trained using a multi-label classification strategy. BERT has the lowest F1-score when it comes to anger, but it does well when it comes to identifying fear and surprise. The findings demonstrate the difficulties presented by unbalanced datasets while also highlighting the promise of transformer-based models for text-based emotion identification. Future research will use data augmentation methods, domain-adapted BERT models, and other methods to improve classification performance.

1 Introduction

Natural language processing (NLP) has made emotion identification from textual data a crucial problem because of its many uses, which include social media trend prediction, consumer sentiment analysis, mental health monitoring, and human-computer interaction. However, because language is inherently subjective, figurative phrases like sarcasm and metaphors are present, and people express emotions differently, it is still difficult to accurately identify emotions in writing. Text-based emotions, in contrast to structured data, are frequently implicit and need a thorough contextual knowledge in order to correctly categorize various emotional states.

Using hand-crafted features and word embeddings, traditional machine learning techniques like Support Vector Machines (SVM) (Cristianini and Ricci, 2008) and Random Forests have been used to emotion identification. Nevertheless, these techniques frequently have trouble capturing contextual relationships and deeper semantic meaning, which results in less than ideal classification performance. Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019a), have transformed natural language processing (NLP) by offering strong contextual representations that allow more accurate emotion detection thanks to developments in deep learning. BERT is well-suited to handling the complexity of emotional expressions in text due to its bidirectional text processing capability, which allows it to understand both local and global relationships..

In order to enable the model to identify several emotions in a single text instance, we use BERT for multi-label emotion categorization in this work. We test the model's performance using a benchmark dataset drawn from internet forums, social media postings, and personal narratives.This work is submitted under SemEval 2025 Task 11, Track A: English-only multi-label emotion classification. We concentrate on five main emotions: anger, fear, joy, sadness, and surprise. To enhance the quality of textual inputs, the dataset is subjected to preparation procedures such as Named Entity Recognition (NER), tokenization, and stop-word removal.Our method addresses issues with ambiguity and overlapping emotional states by improving classification performance through the use of BERT's contextual embeddings.

2 Problem Statement

Understanding human emotions from textual data is a complex challenge due to the inherent ambiguity and subjectivity of language. The primary objective of this project is to develop a robust model capable of accurately classifying multiple emotions present in a given text snippet. Specifically, we aim to predict the following emotions:

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2116–2120

joy, sadness, fear, anger, surprise. Each emotion will be represented as a binary label, indicating its presence (1) or absence (0) in the text.

3 Dataset Description

We used the dataset provided in SemEval-2025 Task 11 (Muhammad et al., 2025b), (Muhammad et al., 2025a) used for this challenge consists of annotated textual data sourced from personal narratives, social media posts, and various online forums. The Track A English dataset contains only 2769 samples across diverse sources. The dataset comprises five primary features, with preprocessing steps such as tokenization, stop-word removal, and Named Entity Recognition (NER) to effectively extract emotional expressions.

The training and validation datasets contain 5 columns, supporting two tasks: Multi-label emotion classification, where the model predicts the presence of multiple emotions (anger, fear, joy, sadness, surprise) within the text, and Emotion intensity detection, where the model assesses the strength of each identified emotion. In contrast, the test data set includes only the *ID*, *text* and corresponding emotion labels, which require models to infer emotional information without predefined labels. These datasets enable structured prediction of emotional states based on textual expressions.

4 Related Work

Earlier approaches to emotion detection from text used machine learning models like SVMs and Naive Bayes with hand-crafted features. These lacked contextual understanding. With the rise of deep learning, models like RNNs and CNNs improved emotion classification but still had limitations in capturing long-range dependencies.

Transformer-based models, especially BERT(Devlin et al., 2019b), brought significant improvements by capturing deep contextual relationships. Recent studies (Muhammad et al., 2025b), have shown that fine-tuning BERT on emotion datasets improves multi-label classification performance. This motivates our use of BERT for the current task.

5 Methodology

In this study, we propose a methodology for emotion classification from English text using a neural network architecture built upon the BERT (Bidirectional Encoder Representations from Transformers) model. Our approach leverages the contextual embeddings of BERT to effectively understand the semantics of textual data and classify them into five emotion categories: *anger*, *fear*, *joy*, *sadness*, and *surprise*. This methodology integrates state-of-the-art natural language processing techniques with a deep learning model optimized for multi-label emotion classification.

5.1 Data Preparation and Preprocessing

The dataset used in this study comprises English textual inputs labeled with multiple emotions. Each sample can express more than one emotion, necessitating a multi-label classification approach. Columns = ID, text, and 5 emotion columns = 7 columns total. A value of *1* indicates the presence of a particular emotion, while *0* indicates its absence. To ensure the quality of the input data, we first remove any incomplete or noisy records using the **`dropna()`** function in pandas. This step eliminates missing values and ensures consistency in model input. The textual inputs are then extracted as a NumPy array, and the corresponding emotion labels are stored as a separate array of shape $N \times 5$, where N is the number of samples.

The dataset is divided into training and validation sets using an $80 - 20$ split. We employ the **`train_test_split()`** function from scikit-learn, ensuring a randomized distribution while maintaining the relative proportion of each class. This is critical for preventing overfitting and ensuring that the model generalizes well to unseen data.

5.2 Text Tokenization and Encoding

To convert the textual inputs into numerical form suitable for BERT, we utilize the `BertTokenizer` from the Hugging Face Transformers library. The tokenizer encodes each text by breaking it down into subword tokens and mapping them to unique input IDs from the BERT vocabulary. Additionally, it generates attention masks to distinguish real tokens from padded elements. Each text is padded or truncated to a maximum sequence length of 200 tokens to maintain consistency across inputs. The tokenization process employs the encode_plus method, which adds special tokens `[CLS]` at the beginning and `[SEP]` at the end of each sequence. These tokens help BERT understand the start and end of the input. The encoded input consists of:

- **input_ids :**Tokenized numerical representation of the text.

- **attention_mask:**Binary mask indicating real tokens and padded elements.

The inputs are then wrapped in a custom `EmotionDataset` class, which inherits from PyTorch's Dataset module. This class facilitates efficient data loading and batching using the `DataLoader`, which iterates through the dataset in mini-batches of size 8. We use **`shuffle=True`** for the training set to introduce randomness and prevent model overfitting.

5.3 Model Architecture

We employ a neural network architecture built on the pre-trained BERT model (**`bert-base-uncased`**). BERT's bidirectional attention mechanism enables it to learn complex contextual relationships in text. Specifically, we use the `[CLS]` token's hidden representation as the aggregate sequence embedding for classification purposes. The model architecture consists of two primary components:

- **BERT Feature Extractor:**The base BERT model is used to obtain contextualized embeddings for each input sequence. The hidden state corresponding to the `[CLS]` token is extracted as it captures the overall meaning of the sentence.

- **Fully Connected Layer:** A linear layer projects the hidden state to five neurons, corresponding to the five emotion categories. This layer computes the emotion logits as follows:

$$Logits = W \cdot H_{CLS} + b \tag{1}$$

  where W and b are the weights and biases of the fully connected layer.

5.4 Loss Function and Optimization

Given the multi-label nature of the problem, we use the Binary Cross Entropy with Logits Loss (`BCEWithLogitsLoss`) as the loss function. This function applies a sigmoid activation to the logits and computes the binary cross-entropy for each emotion category independently.

To optimize the model, we use the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 1×10^{-5} and weight decay of 0.01. `AdamW` is known for its adaptive learning rates and weight decay regularization, which enhances generalization.

5.5 Training and Evaluation

The model is trained for 5 epochs, iterating through mini-batches from the `DataLoader`. During each epoch, the model's parameters are updated using *backpropagation* and *gradient descent*. The training loss is accumulated and averaged across batches to monitor learning progress. This is represented as:

$$Loss_{train} = \frac{1}{B} \sum_{k=1}^{B} L_k \tag{2}$$

where B is the total number of batches. After each epoch, the model is evaluated on the validation set in inference mode using **`torch.no_grad()`** to reduce memory usage. The validation loss is calculated similarly to the training loss. To generate predictions, the model's logits are passed through a sigmoid activation, and a threshold of 0.5 is applied to determine the presence of each emotion:

$$\hat{y} = I(\sigma(z) > 0.5) \tag{3}$$

where I is the indicator function. Performance is measured using precision, recall, and F1score, computed using scikitlearn's classification_report. This comprehensive evaluation ensures that the model accurately identifies multiple emotions per text.

5.6 Implementation and Deployment

The model is implemented using PyTorch and Hugging Face's `Transformers` library. The training is accelerated using GPU support for faster computation. After training, the model's state dictionary is saved in the **`.pth`** format for future inference and fine-tuning.

The proposed methodology demonstrates the effectiveness of leveraging BERT's contextual embeddings for multi-label emotion classification. This approach can be extended to other emotion recognition tasks and adapted for different languages or text domains.

6 Results and Discussions

Our model's classification performance on the validation dataset is shown in Table 1, which is assessed using precision, recall, and F1-score for each of the five emotion classes. With the greatest F1-score of 0.84, and recall of 0.90, the model shows excellent performance in identifying fear. Similarly, with F1-scores of 0.74 and 0.71, respectively, surprise and sadness show respectably strong

categorization results. With the lowest F1-score of 0.55, rage is the emotion that the model finds most difficult to differentiate from other emotions. The F1-score of 0.67 indicates that the joy class performs somewhat as well.

	precision	recall	f1-score	support
anger	0.61	0.50	0.55	34
fear	0.78	0.90	0.84	168
joy	0.66	0.69	0.67	48
sadness	0.76	0.67	0.71	84
surprise	0.73	0.75	0.74	83
micro avg	0.74	0.77	0.75	417
macro avg	0.71	0.70	0.70	417
weighted avg	0.74	0.77	0.75	417
samples avg	0.70	0.73	0.68	417

Table 1: Classification Report on Validation Dataset

Overall, the weighted and micro-averaged F1-scores are 0.75, indicating that the model does well when label distribution is taken into account. However, the macro-averaged F1-score is somewhat lower at 0.70, suggesting that overall performance is impacted by some class imbalance. Inconsistencies across occurrences are further reflected in the sample-average F1-score of 0.68. According to these findings, the model needs more refinement, especially for underrepresented emotions like anger, even while it successfully categorizes dominating emotions like surprise and fear. To increase model resilience, future developments may use sophisticated feature extraction and data augmentation strategies.

7 Limitations

Notwithstanding the encouraging outcomes of text-based emotion categorization using BERT, our work has certain drawbacks. One significant issue is the dataset's class imbalance, which has an impact on the model's capacity to fairly identify all emotions. The model has the lowest F1-score when it comes to anger, but it does well when it comes to fear and astonishment. Due to this imbalance, the model could perform poorly on less common emotion classes while favoring dominating ones. Furthermore, BERT is less suited for real-time or low-resource applications, as it is a transformer-based model that requires substantial computing resources for both training and inference.

The categorization of emotions based only on textual input is another drawback. A solely text-based method lacks multimodal signals like tone, pitch, and facial expressions, which are frequently more effective in eliciting emotions. The model's capacity to distinguish between emotions that exhibit comparable textual patterns—like sarcasm or neutral statements with emotional undertones—is hence limited. Additionally, even while BERT does a good job of capturing contextual meanings, it may still misread emotions in texts that include casual language, slang, or cultural differences. To increase classification robustness, future studies should investigate domain-specific fine-tuning and the use of multimodal data.

8 Conclusion and Future Work

In this work, we used text data to classify emotions using BERT, and we assessed how well it performed across five different emotion classes : anger,fear,joy,sadness and surprise. According to the results, anger performed the worst, whereas BERT successfully categorizes emotions like fear and surprise, obtaining high F1-scores. Although the somewhat lower macro-average F1-score of 0.70 indicates that there is potential for improvement in addressing class imbalances, the overall weighted F1-score of 0.75 emphasizes the model's strong generalization capabilities.

In order to improve classification performance, especially for underrepresented emotions, our future research will investigate domain-adapted BERT models and data augmentation strategies. In order to enhance emotion recognition, we also intend to look into multimodal techniques by combining text with audio and visual data. It may also be possible to obtain more balanced categorization across all emotion categories by experimenting with ensemble models and further optimizing BERT through emotion-specific pretraining.

References

Nello Cristianini and Elisa Ricci. 2008. *Support Vector Machines*, pages 928–932. Springer US, Boston, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

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 Zhangand Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages.

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.