Angeliki Linardatou at SemEval-2025 Task 11: Multi-label Emotion Detection

Angeliki Linardatou **Paraskevi Platanou**
Athens University of Economics and Business (AUEB)
Athens, Greece
Email: angela2003linardatou@gmail.com, platanou@aueb.gr

Abstract

Multi-label emotion detection is challenging due to contextual complexity and irony. Most sentiment models classify text into single categories, missing overlapping emotions.

This study, competing in SemEval 2025 Task 11 - Track A, detects anger, surprise, joy, fear, and sadness, in English texts. We propose a hybrid approach combining fine-tuned BERT transformers, TF-IDF for lexical analysis, and a Voting Classifier (Logistic Regression, Random Forest, SVM, KNN, XGBoost, LightGBM, CatBoost), with grid search optimizing thresholds.

Our model achieves a macro F1-score of 0.6864. Challenges include irony, ambiguity, and label imbalance. Future work will explore larger transformers, data augmentation, and cross-lingual adaptation.

This research underscores the benefits of hybrid models, showing that combining deep learning with traditional NLP improves multi-label emotion detection.

1 Introduction

Sentiment Analysis (SA) and Emotion Detection are key NLP tasks for analyzing emotional tone in text, essential for understanding human behavior. While SA classifies text as positive, negative, or neutral, Emotion Detection identifies specific emotions (joy, anger, sadness) but faces challenges like sarcasm, ambiguity, and overlapping emotions.

Transformer models such as BERT and XLNet have greatly improved sentiment classification, with hybrid approaches emerging. Jlifi et al. (2024) (Jlifi et al., 2024) combined BERT with Random Forest (Ens-RF-BERT), while Danyal et al. (2024) (Danyal et al., 2024) proposed BERT-XLNet for long-text classification. However, multi-label emotion detection remains underexplored.

SemEval 2025 Task 11 addresses this by providing a benchmark for detecting anger, surprise, joy, fear, and sadness. In our study, we develop a fine-tuned BERT model, enhanced with TF-IDF for lexical analysis, and a Voting Classifier. We also explore RoBERTa-large for better predictions.

Our model achieves an F1-score of 0.68, demonstrating the effectiveness of combining transformers with traditional ML techniques. Future improvements include data augmentation, cross-lingual adaptation, and larger transformer models to enhance performance.

2 Related Work and Background

Sentiment Analysis (SA) and Emotion Detection are key NLP tasks, but traditional methods rely on single-label classification, failing to capture multiple coexisting emotions.

SemEval 2025 Task 11 - Track A addresses this by requiring models to detect multiple emotions per text. Multi-label classification is essential, as texts can express mixed emotions (e.g., joy and surprise, sadness and fear). Existing single-label methods struggle with this complexity, necessitating models that learn emotional dependencies.

This task involves five core emotions: joy, sadness, anger, surprise, and fear. Effective multi-label classification requires techniques like binary relevance, classifier chains, and deep learning to improve detection accuracy.

3 Previous Studies

Several studies have explored multi-label sentiment analysis using ML and deep learning approaches:

Jin and Lai (2020) (Jin and Lai, 2020) proposed a hybrid model combining BERT's contextual embeddings with a modified TF-IDF weighting technique. Their approach enhanced key term importance while leveraging deep contextual

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 502–507
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

representations, leading to a 4.2% F1-score improvement over traditional TF-IDF and standalone BERT models.

Ni and Ni (2024) (Ni and Ni, 2024) introduced a sentiment correlation modeling approach to capture interdependencies between emotions. Their correlation-aware mechanism refined sentiment prediction, improving F1-score by 3.8% compared to conventional multi-label models.

These studies highlight the benefits of hybrid and correlation-aware approaches in multi-label sentiment classification.

4 Comparison with the Proposed Model

Unlike prior studies relying solely on transformers or traditional ML, our approach combines both for improved accuracy. We fine-tune BERT for contextual understanding while integrating TF-IDF for lexical features. A Voting Classifier (Logistic Regression, Random Forest, SVM, KNN, XGBoost) optimizes classification via ensemble learning. Grid search refines probability thresholds, balancing precision and recall. To handle overlapping emotions and class imbalance, we fine-tune decision boundaries. Training efficiency is improved through reduced learning rates, minimal epochs, and cross-validation, ensuring robust predictions.

5 Innovative Aspects of Our Approach

Our proposed method differentiates itself from previous studies in several key ways:

Our approach stands out by integrating deep learning with traditional NLP techniques for multi-label emotion detection. By combining BERT for contextual understanding with TF-IDF for lexical emphasis, we ensure that important words are not downweighted, enhancing interpretability and classification accuracy. Additionally, we incorporate ensemble learning through a voting classifier, which combines multiple models to improve stability and reduce variance. To further refine classification performance, we apply grid search for optimal probability threshold selection, maximizing the F1-score. This hybrid methodology effectively balances semantic understanding and lexical precision, leading to a more robust and accurate emotion detection system.

6 Dataset and Examples

The dataset used in our experiments consists of **English** textual data annotated with multiple emotional labels. The data is provided by the SemEval 2025 Task 11 competition (Track A), which focuses specifically on English-language emotion detection. Below, we present representative examples from the dataset along with their detected emotions:

Text	Emotions Detected
"I can't believe I did it!"	**Joy, Surprise**
"What happened was truly terrifying."	**Fear**
"I feel so alone right now..."	**Sadness**
"How dare you say that to me?!"	**Anger**

Table 1: *Representative examples from the dataset with their detected emotions.*

These examples demonstrate the necessity of multi-label classification, as a single sentence can express multiple emotions simultaneously (refer to Fig. 2). Capturing such nuances is crucial for improving the accuracy of emotion detection models. Our approach introduces an effective fusion of modern NLP techniques and traditional text analysis, demonstrating the potential of hybrid methodologies for improving multi-label emotion detection.

The following figure illustrates the frequency distribution of detected emotions in the dataset, highlighting which emotions appear more frequently in the text samples.

7 System Overview

Our system for multi-label emotion detection combines state-of-the-art transformer models with traditional feature engineering and ensemble learning to maximize classification performance. Below, we describe the key components and methodological choices that define our approach.

8 Key Algorithms and Modeling Decisions

Our hybrid approach integrates deep learning, feature engineering, and ensemble learning for multi-

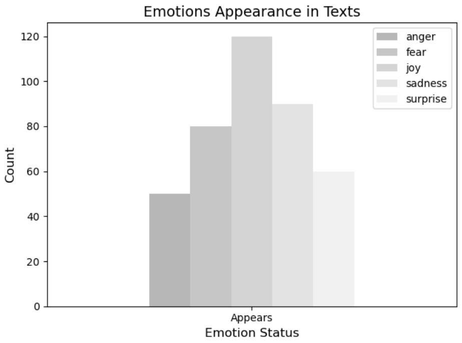

Figure 1: *The X-axis represents the detected emotions, while the Y-axis shows the number of texts in which each emotion appears. The chart illustrates how frequently each emotion is detected, highlighting the most commonly recognized emotion.*

label emotion detection. We fine-tune BERT and RoBERTa for contextual embeddings, complemented by TF-IDF for lexical features. A soft voting ensemble (Logistic Regression, Random Forest, SVM, KNN, XGBoost, LightGBM, CatBoost) enhances classification, leveraging each model's strengths. Grid search-based threshold optimization refines decision boundaries, maximizing the F1-score. To prevent overfitting, we use a reduced learning rate and minimal training epochs for BERT fine-tuning.

9 Data Preprocessing and Feature Extraction

The dataset consists of English text snippets labeled with five emotions: anger, fear, joy, sadness, and surprise. Some data samples are multi-labeled, while others contain only one label.

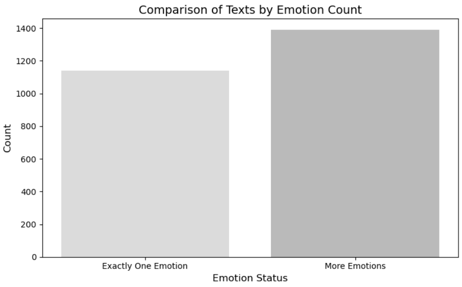

Figure 2: *The figure shows the number of texts that contain only one emotion versus those that contain multiple emotions.*

The dataset is split into 80% training and 20% testing. Text is preprocessed using the **BERT tokenizer** for subword embeddings with padding and truncation. **TF-IDF** (vocabulary size: 8000) extracts numerical features, which are concatenated with BERT embeddings to form an enriched feature space. **Lemmatization, stemming, and stopword removal** are not applied.

10 Model Training

The transformer model is fine-tuned for **two epochs** using the AdamW optimizer and Binary Cross-Entropy Loss.This choice was made to *minimize overfitting* and ensure *efficient training*. To support this decision, we conducted additional experiments by training BERT for 1, 2, 3, and 4 epochs and monitoring the macro F1-score. The model achieved its highest performance at epoch 2, with F1-scores plateauing or slightly decreasing afterward, indicating potential overfitting.This empirical finding aligns with **Jin and Lai (2021)** (**Jin and Lai, 2020**), who observed that *2–3 epochs were optimal* when fine-tuning BERT on emotion-labeled datasets of similar size. After training, sentence embeddings are extracted from the final hidden layer of the BERT model. These are then concatenated with TF-IDF features to form a hybrid representation, which is used to train an ensemble classifier. Binary Cross-Entropy Loss was selected as it is well-suited for multi-label classification, treating each label as an independent binary task. It effectively handles overlapping labels and produces well-calibrated probability estimates. Finally, we perform grid-based hyperparameter tuning to adjust the number of estimators, learning rates, and maximum depths of the ensemble classifiers (e.g., Random Forest, XGBoost, LightGBM, CatBoost), ensuring the best possible configuration for robust performance.

11 Threshold Optimization

Since our task involves multi-label classification, we needed to decide how to convert the model's probabilistic outputs into binary predictions for each emotion. To do this, we optimized the probability thresholds using a grid search approach. We tested a range of threshold values from 0.30 to 0.70, increasing by 0.05 each time. This process was carried out separately for each emotion, allowing the model to adjust its sensitivity depending on the characteristics of each label — which is especially helpful in cases of class imbalance. For example, emotions like *surprise* or *fear*, which appear less frequently in the data, may benefit from a lower threshold. Our goal was to find the combination of thresholds that would maximize the macro F1-score across all five emotions. To make the results more reliable, we used 5-fold cross-validation on the training set. The best thresholds were then applied to the test predictions. This per-label tuning led to a better balance between precision and recall, especially for texts expressing subtle or overlapping emotions. By avoiding a "one-size-fits-all" threshold (like the default 0.5), we were able to make the model more adaptable to the specific behavior of each emotion category.

12 Model Evaluation

Our final model achieves:

Metric	Value
Macro Precision	0.6814
Macro Recall	0.7094
Macro F1-Score	0.6864
Accuracy	0.3827
SemEval Baseline	0.7083

Table 2: *Performance Metrics showing the results of model evaluation for macro precision, recall, F1-score, and accuracy, with a comparison to the baseline.*

The results confirm that combining deep learning with traditional machine learning improves multi-label emotion classification. LightGBM and CatBoost further boost performance alongside XGBoost.

The model's low accuracy (0.3827) reflects the difficulty of multi-label classification, where texts express multiple emotions. The macro F1-score (0.6864) suggests an imbalance between precision (0.6814) and recall (0.7094), likely due to class imbalance, overlapping emotions, or suboptimal threshold selection. Additionally, language subjectivity and ambiguity pose challenges, as the same phrase can convey different emotions depending on context.

13 Experimental Setup

Our experimental setup consists of carefully designed steps to ensure reliable model training and evaluation. Below, we describe the dataset splitting strategy, preprocessing steps, hyperparameter tuning, and external tools utilized.

13.1 Data Splitting Strategy

The dataset was divided into three subsets: the Training Set (80%), which was used to train the models, and the Testing Set (20%), which was used for final evaluation. No separate validation set was used, as hyperparameter tuning was conducted using internal cross-validation techniques.

13.2 Preprocessing Steps

To optimize model performance, several preprocessing steps were applied. **BERT tokenizer** converted text into subword tokens, while **TF-IDF transformation** (vocabulary size: 8000) extracted lexical features. These were combined with **BERT embeddings** to form a hybrid feature set.

Data normalization was applied via standard scaling. Text cleaning included **lowercasing, removing digits, punctuation, and stopwords** (NLTK). These steps ensured consistency across the dataset, where text served as input (X) and emotion labels as output (Y).

13.3 Hyperparameter Tuning

For model training and evaluation, we employed a combination of individual classification models and an ensemble classifier.

13.3.1 Individual Classification Models

The study utilizes multiple classifiers: **Logistic Regression (LR)** with 1500 iterations (`max_iter=1500`) and class balancing (`class_weight='balanced'`); **Random Forest (RF)** with 200 trees (`n_estimators=200`) and depth 15 (`max_depth=15`); **Support Vector Machine (SVM)** with a linear kernel (`kernel='linear'`) and probability estimation (`probability=True`); **K-Nearest Neighbors (KNN)** using 5 neighbors (`n_neighbors=5`); and **XGBoost (XGB)** with 200 estimators (`n_estimators=200`), disabled label encoding (`use_label_encoder=False`), and `mlogloss` as the evaluation metric.

13.3.2 Ensemble Classifier

To enhance performance, an ensemble learning strategy was applied using the following models:

Random Forest with 200 trees and a maximum depth of 15, **XGBoost** with a learning rate of 0.1 and 200 estimators, **LightGBM** with 150 estimators and a maximum depth of 10 and **CatBoost** with a learning rate of 0.05 and 100 estimators.

Additionally, a grid search was applied to optimize probability thresholds for multi-label classification.

13.4 External Tools and Libraries

The following external tools and libraries were utilized for model training and evaluation: **Transformers Library** was used for BERT tokenization and training, while **Scikit-learn** handled TF-IDF extraction, model training, and evaluation. **XGBoost, LightGBM, and CatBoost** contributed to ensemble learning, with **PyTorch** used for BERT fine-tuning and embedding extraction. **Joblib** ensured efficient model storage. These components structured our experimental setup for multi-label emotion detection.

14 Results

In this section, we analyze the performance of our model based on official evaluation metrics, competition ranking, and additional qualitative observations.

14.1 Overall Performance

As shown in Table 2 in Section 12, these results indicate that our approach performs well in detecting multiple emotions simultaneously, with a balanced trade-off between precision and recall, as demonstrated by the F1 score of 0.6864.

Model	Macro F1-Score
TF-IDF only (Logistic Regression)	0.5141
BERT only	0.5594
Hybrid (BERT + TF-IDF + Ensemble)	0.6814
RoBERTa-large (planned future work)	0.7437

Table 3: Comparison of Macro F1-scores across different configurations: traditional methods, deep learning, hybrid models, and future improvements.

Table 3 summarizes the macro F1-scores achieved by various configurations. The TF-IDF model combined with Logistic Regression achieves limited performance due to its lack of contextual understanding. The BERT-only setup improves performance but still lacks lexical precision. Our proposed hybrid model significantly outperforms both baselines by combining the strengths of deep embeddings and lexical features.

Preliminary experiments with RoBERTa-large show further improvements, indicating promising directions for future work in enhancing semantic representations.

14.2 Error Analysis

To investigate misclassifications, we analyzed false positives and false negatives. Sentences containing sarcasm or irony were often misclassified. Additionally, short text samples with ambiguous wording had lower prediction confidence. Some labels also overlapped significantly, making a clear distinction difficult. To address these issues, future research could focus on improving semantic representations through larger models such as RoBERTa-large and integrating emotion lexicons.

14.3 Observations About the Data

Our analysis revealed class imbalance, with emotions like *surprise* being underrepresented. Some samples contained conflicting emotions, making labeling challenging. More data and augmentation could improve generalization.

Our approach—combining transformers, ensemble learning, and feature engineering—proved effective. Future work should focus on class balancing and integrating external emotion lexicons for better accuracy.

15 Conclusion

We proposed a hybrid approach for multi-label emotion detection, combining transformer models (BERT, RoBERTa) with TF-IDF and ensemble learning. This method leveraged deep embeddings and classical ML models, achieving an F1-score of 0.6864.

Challenges like ambiguous text, sarcasm, and label imbalance persist. Future work includes larger transformers (e.g., RoBERTa-large) with potential F1-score improvements to 0.74, alongside emotion lexicons, sentiment-aware embeddings, and advanced augmentation.

Our results highlight the benefits of integrating deep learning with traditional NLP, with further refinements promising enhanced performance.

16 Limitations

Our work presents several limitations. First, the dataset used in our experiments is limited to English texts. This may restrict the generalizability of the model to multilingual texts.

Second, although the hybrid approach combining BERT embeddings and TF-IDF features improves performance, it remains sensitive to subtle linguistic phenomena, such as sarcasm, irony, and cultural differences in emotional expression. Misclassifications frequently occur in cases involving multiple or ambiguous emotions.

Third, due to limited computational resources, we restricted the fine-tuning of large transformer architectures (such as RoBERTa-large) and did not apply extensive data augmentation techniques, which could have further improved performance.

Finally, class imbalance in the dataset affects the detection of rare emotions, such as ”surprise.” Future work could focus on improving class imbalance handling, applying multilingual transfer learning, and exploring sentiment-aware pretrained embeddings.

17 Ethical Considerations

The proposed model presents ethical concerns regarding misuse, bias, and applicability.

17.1 Misuse and Bias

It could be exploited for manipulating individuals or spreading misinformation. Additionally, biases in training data may impact performance across demographics and languages, leading to uneven recognition of emotions.

17.2 Use Cases and Risks

The model should be used responsibly in sentiment analysis and mental health applications with consent, avoiding privacy violations. It is unsuitable for high-risk fields like law or medicine, where misclassification could have severe consequences. Safeguards are necessary to prevent misuse.

(Muhammad et al., 2025a) (Muhammad et al., 2025b) (Jlifi et al., 2024) (Danyal et al., 2024) (Jin and Lai, 2020) (Ni and Ni, 2024)

References

Muhammad Danyal, Rahim Khan, and Samia Latif. 2024. Proposing sentiment analysis model based on bert and xlnet for movie reviews. *ResearchGate*. Available at: https://www.researchgate.net/publication/377410754_Proposing_sentiment_analysis_model_based_on_BERT_and_XLNet_for_movie_reviews.

Liang Jin and Mei Lai. 2020. Multi-label sentiment analysis based on bert with modified tf-idf. In *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, pages 2674–2680. IEEE.

Moez Jlifi, Mohamed Yahya Sayadi, and Faiez Gargouri. 2024. Beyond the use of a novel ensemble-based random forest-bert model (ens-rf-bert) for the sentiment analysis of the hashtag covid19 tweets. *Social Network Analysis and Mining*.

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, and 29 others. 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, and 2 others. 2025b. Semeval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1–12, Vienna, Austria. Association for Computational Linguistics.

Li Ni and Hao Ni. 2024. Emotion correlation-aware multi-label classification using deep learning techniques. *Frontiers in Psychology*, 15:1490796.