DataBees at SemEval-2025 Task 11: Challenges and Limitations in Multi-Label Emotion Detection

Sowmya Anand, Tanisha Sriram, Rajalakshmi Sivanaiah, Angel Deborah, Mirnalinee TT
Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
sowmya2310543.ssn.edu.in, tanisha2310538.@ssn.edu.in
rajalakshmis@ssn.edu.in, angeldeborahS@ssn.edu.in, mirnalineeTT@ssn.edu.in

Abstract

Text-based emotion detection is crucial in NLP, with applications in sentiment analysis, social media monitoring, and human-computer interaction. This paper presents our approach to the Multi-label Emotion Detection challenge, classifying texts into joy, sadness, anger, fear, and surprise. We experimented with traditional machine learning and transformer-based models, but results were suboptimal: F1 scores of 0.3723 (English), 0.5174 (German), and 0.6957 (Spanish). We analyze the impact of preprocessing, model selection, and dataset characteristics, highlighting key challenges in multi-label emotion classification and potential improvements.

1 Introduction

Emotion detection from text is a crucial NLP task with applications in customer feedback analysis, mental health detection, and social media monitoring. Unlike sentiment analysis, which determines polarity, emotion detection classifies text into **joy, sadness, fear, anger, and surprise**, often requiring **multi-label classification** since a sentence can evoke multiple emotions. Despite advancements in **transformer-based models** like BERT and XLM-RoBERTa, challenges remain due to:

- **Subjectivity:** Different individuals may perceive emotions differently.
- **Contextual Complexity:** Subtle emotional cues require deep contextual understanding.
- **Multi-label Classification:** A single text can express multiple overlapping emotions.

1.1 Competition Overview

Track A (Muhammad et al., 2025b) of the **Multi-label Emotion Detection** competition involved classifying text snippets into one or more of five emotions or as neutral. Our experiments included a range of approaches, starting with traditional machine learning models such as Logistic Regression, Random Forest, and SVM. We also explored transformer-based models, including BERT, DistilBERT, XLM-RoBERTa, and language-specific BERT variants, as well as ensemble models that combined classifiers like KNN, Decision Trees, and Neural Networks. Despite extensive preprocessing, hyperparameter tuning, and model optimization, our overall performance—particularly on English data—fell short of expectations. This paper delves into the key challenges we encountered and the insights gained throughout our approach.

2 Dataset

Our dataset comes from Task 11 of SemEval 2025 (Muhammad et al., 2025a), focusing on multi-label emotion classification in English, German, and Spanish. Sourced from social media, each text snippet is annotated using a binary scheme (1: present, 0: absent), allowing for multi-label classification where multiple emotions can co-occur. Neutral instances contain no marked emotions. The dataset is split into train, dev, and test sets for structured evaluation. Table 1 provides an overview.

3 Related Works

Emotion detection in textual data has been extensively explored in NLP, spanning lexicon-based, machine learning, and deep learning methods. Transformer-based models have recently set new benchmarks, particularly for multi-label and multilingual emotion classification.

3.1 Traditional Approaches to Emotion Detection

Early systems used lexicon-based methods with resources like WordNet-Affect (Strapparava and Valitutti, 2004) and LIWC (Tausczik and Pennebaker, 2010), but lacked context sensitivity. Machine learning models such as Naïve Bayes (Alm et al.,

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 222–227

Language	Data Source(s)	Train	Dev	Test	Total
English (eng)	Social media	2768	116	2767	5651
German (deu)	Social media	2603	200	2604	5407
Spanish (esp)	Social media	1996	184	1695	3875

Table 1: Description of Track A dataset.

2005), SVMs (Wang and Manning, 2012), and Random Forests (Strapparava and Mihalcea, 2007) improved generalization but struggled with semantic ambiguity and multi-label complexity.

3.2 Multi-Label Emotion Detection

Emotion detection is inherently multi-label, as a single text may express multiple emotions (Mohammad and Bravo-Marquez, 2018). Traditional methods addressed this using hierarchical classification (Hatzivassiloglou and McKeown, 2000) or CRFs (Strapparava and Mihalcea, 2007). Deep learning models, like BiLSTMs with attention (Majumder et al., 2019), showed substantial improvements by capturing emotion co-occurrence and contextual dependencies.

3.3 Transformer-Based Approaches

Transformers like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019) have achieved state-of-the-art results. Lighter models such as DistilBERT (Sanh et al., 2019) offer faster inference, while EmotionBERT (Saravia et al., 2018) improves emotion-specific learning. Prompt-based models (Gao et al., 2021) enable zero-shot emotion detection.

Recent work has extended transformers to mental health detection. (Sivanaiah et al., 2024) compared BERT, RoBERTa, and traditional models for suicide and self-harm classification, with RoBERTa achieving the highest F1-score (99%). (Yenumulapalli et al., 2023) explored depression detection via BERT in LT-EDI-2023, achieving a macro F1 of 0.407, highlighting the capability of transformer models in capturing nuanced emotional and psychological cues.

4 Methodology

The steps taken are laid out in the methodology section in detail, with the inclusion of preprocessing, exploratory data analysis (EDA), model selection, evaluation, and the implementation of recommendations received while conducting the study.

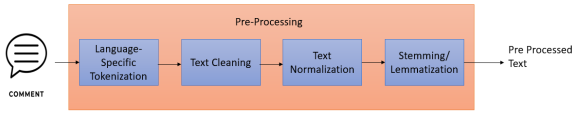

Figure 1: Pre-Processing Steps.

4.1 Preprocessing and Exploratory Data Analysis (EDA)

Preprocessing ensures clean, tokenized input for training machine learning models. We used language-specific tokenizers: BETO for Spanish, bert-base-german-cased for German, and the default BERT tokenizer for English. These tokenizers helped capture the linguistic nuances of each language. Our preprocessing pipeline involved removing stopwords, special characters, and punctuation to prevent noise during training, as illustrated in Figure 1. Additionally, we applied lowercasing and contraction expansion (e.g., I'm to I am) to maintain consistency. Stemming and lemmatization were deliberately excluded from our preprocessing pipeline to preserve the rich morphological and contextual information inherent in the text, which is often critical for accurately detecting emotions. In emotion recognition tasks, subtle variations in word forms—such as verb tenses or pluralizations—can convey important affective cues; for example, "crying" may carry a stronger emotional weight than "cry." Reducing words to their base or root form risks stripping away these distinctions, potentially leading to loss of emotional intensity or misinterpretation. Furthermore, the transformer-based models employed in our study, such as BETO, BERT, and their variants, are pretrained on large corpora of raw text and are inherently capable of understanding and disambiguating word forms in context. Thus, applying stemming or lemmatization could disrupt the linguistic patterns these models have learned to leverage, ultimately impairing performance rather than enhancing it. For exploratory data analysis (EDA), we checked the distribution of emotions across the datasets for all languages. This included looking at the proportion of various emotion classes (like joy, sadness, fear, etc.) and

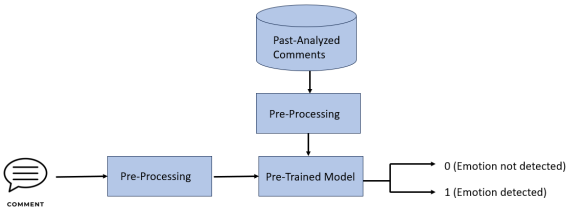

Figure 2: Methodology.

checking if there were any data biases.

4.2 Model Selection and Explanation

To tackle the task of emotion detection, we utilized both common machine learning classifiers as shown in Figure 2 and state-of-the-art transformer-based models.

We began by utilizing **BERT-base-uncased**, a pre-trained transformer model known for its strong performance in various NLP tasks, including sentiment and emotion classification. BERT's bidirectional nature allows it to understand context from both directions, making it particularly effective for detecting emotions expressed through subtle language cues. BERT was applied to English, Spanish, and German datasets. While it performed adequately on English and Spanish, its performance on German was notably weaker—likely due to its English-centric training data, making it less effective for other languages without fine-tuning.

To better handle Spanish, we used **BETO**, a BERT variant fine-tuned on a large Spanish corpus. BETO outperformed both BERT-base-uncased and mBERT on the Spanish dataset, especially in identifying anger, disgust, and fear. Its improved performance is due to its specialization in Spanish syntax and semantics.

We also tested **mBERT**, a multilingual BERT trained on 104 languages. Its ability to handle all three languages made it efficient for a multilingual dataset. While mBERT performed reasonably well on English and Spanish, it struggled with German, likely because its generalized training across languages made it less effective for those with more complex grammar, like German.

To address this, we employed **Google-BERT/bert-base-german-cased**, fine-tuned specifically for German. This model significantly improved emotion classification on the German dataset, particularly for anger and disgust, thanks to its training on a large German corpus.

In addition to transformer models, we experimented with traditional machine learning classifiers: **Multinomial Naive Bayes (NB)**, **Support Vector Classifier (SVC)**, **Logistic Regression**, and **Random Forest**. Naive Bayes, while effective for simpler emotions like fear and disgust (especially in Spanish and German), struggled with more nuanced emotions like joy and surprise in English. **SVC** outperformed Naive Bayes—especially for fear—but still lagged behind transformers in handling complex emotions. **Logistic Regression** performed reasonably well on fear and sadness but underperformed on joy and surprise. **Random Forest**, despite being strong in ensemble learning, was less effective across all datasets, particularly for surprise and anger.

To enhance model performance, we implemented **hyperparameter tuning**, adjusting the learning rate to 2e-5, increasing training epochs to five, and reducing batch size to 8. While this improved model stability and convergence, it did not lead to notable improvements in F1 scores for the English dataset.

We also incorporated **lexicon-based testing** using sentiment lexicons from the NLTK library. Although helpful for validating predictions and estimating overall correctness, these approaches could not match the performance of transformer models, which better capture the contextual subtleties of language.

4.3 Feedback-Based Potential Improvements

Throughout the course of the research, a number of useful suggestions were made that might have otherwise increased the quality of this research. A possible recommendation was to try models without using tokenization, lemmatization, or stemming. These preprocessing operations tend to cause quite a loss of information, as indicated in the feedback. By skipping these methods, the models could have preserved more useful linguistic features, which could have resulted in improved performance at emotion detection. Looking back, refraining from these inductive text processing methods might have retained more of the original sentence meaning and structure, which could have helped with emotion detection. The other recommendation was to move away from having separate binary classifiers for every emotion to a **single multi-class classifier**. By representing the emotion labels as integers (e.g., Anger = 0, Joy = 1, Fear = 2, etc.), the model might have been able to better differentiate between emotions, instead of being trained to

predict the occurrence or non-occurrence of each emotion separately. This would have most likely resulted in improved performance, as the model would have been in a position to comprehend the association between various emotions and classify them more holistically. We tried implementing this recommendation and discovered that it worked to yield a more coherent classification of emotion, especially when applied to the case of very complex or unclear cases.

4.4 Combined Model Performance

Following the assessment of the performance of single models, we chose to aggregate the predictions of various models to form an ensemble model. The ensemble strategy was effective in improving performance, especially for the English dataset. By aggregating models like KNN, Random Forest, XGBoost, and Logistic Regression, we managed to improve the accuracy of recognizing emotions such as joy and surprise, which were more difficult for single models to identify. The current research shows that transformer-based models such as BERT, BETO, and Google-BERT are greatly effective for multilingual emotion detection. Although mBERT had potential for multilinguality, it performed poorly for German. Baselines were created by the classic machine learning classifiers but were dominated by the advanced transformer models. While there were some aspects to be improved, especially in preprocessing, model structure, and hyperparameter optimization (decreasing in learning rate to 2e-5, bumping up training epochs to five, and cutting the batch size to 8), the ensemble of these models produced encouraging outcomes for emotion recognition in multilingual text.

5 Results and Analysis

5.1 Results

The results of emotion detection across Spanish, German, and English datasets show varying performance across different models. In tables 2, 3 and 4, the models are represented as follows- LR- Logistic Regression, RF- Random Forest, SVM- Support Vector Machine, BERT- BERT base uncased, wt BERT- weighted BERT, Ensmbl-Ensemble of KNN, RF, DT and LR, distil- DistilBERT, XLM-R- XSLM-RoBERTa, MNB- Multinomial Naive Bayes, SVC- Support Vector Classifier, g-BERT-Google BERT. The emotions An, Di, Fe, Jo, Sa and Su are Anger, Disgust, Fear, Joy, Sadness and Surprise respectively.

Model	An	Di	Fe	Jo	Sa	Su
BERT	0.71	0.73	0.85	0.72	0.77	0.54
BETO	0.75	0.80	0.86	0.80	0.78	0.72
mBERT	0.72	0.77	0.82	0.76	0.75	0.70
MNB	0.55	0.68	0.80	0.67	0.53	0.47
SVC	0.52	0.70	0.81	0.70	0.74	0.37
LR	0.55	0.71	0.85	0.70	0.64	0.47
RF	0.49	0.63	0.87	0.62	0.65	0.47

Table 2: Model Performance Metrics for Emotion Detection in Spanish.

Model	An	Di	Fe	Jo	Sa	Su
BERT	0.67	0.59	0	0.41	0.28	0
g-bert	0.7	0.65	0.26	0.56	0.57	0.28
MNB	0.66	0.59	0	0.16	0.22	0
SVC	0.61	0.54	0.19	0.48	0.48	0
LR	0.58	0.53	0.07	0.43	0.46	0
RF	0.38	0.34	0	0.17	0.13	0

Table 3: Model Performance Metrics for Emotion Detection in German.

Model	An	Fe	Jo	Sa	Su
LR	0.31	0.65	0.44	0.24	0.57
RF	0.10	0.67	0.37	0.13	0.52
SVM-Lin	0.32	0.65	0.43	0.25	0.56
SVM-RBF	0.20	0.67	0.41	0.18	0.56
BERT	0.54	0.62	0.11	0.57	0.69
wt bert	0.48	0.42	0.59	0.69	0.63
Ensmbl	0.83	0.53	0.77	0.67	0.68
distil	0.7	0.73	0.18	0.43	0.26
XLM-R	0	0.75	0.34	0.46	0.49

Table 4: Model Performance Metrics for Emotion Detection in English.

For **Spanish**, BETO, a language-specific transformer, outperforms all other models, particularly in detecting disgust (0.80) and fear (0.86). BERT-base-uncased and mBERT perform well but are slightly less effective than BETO, as seen in table 2. Traditional models like Naive Bayes and SVC show moderate performance, with SVC achieving the highest F1 score for fear (0.81), but struggle with emotions like joy and surprise. We were placed 37th with our results. In **German**, Google-BERT (fine-tuned for German) surpasses BERT-base-uncased but still struggles across all categories, especially fear, joy, and surprise as demonstrated by table 3. Traditional models such as Naive Bayes and SVC also show weak performance, highlighting the challenges in detecting emotions in German. We secured 41st place in this run. For **English**, ensemble models (KNN, Random Forest, XGBoost, etc.) achieve the best results, especially

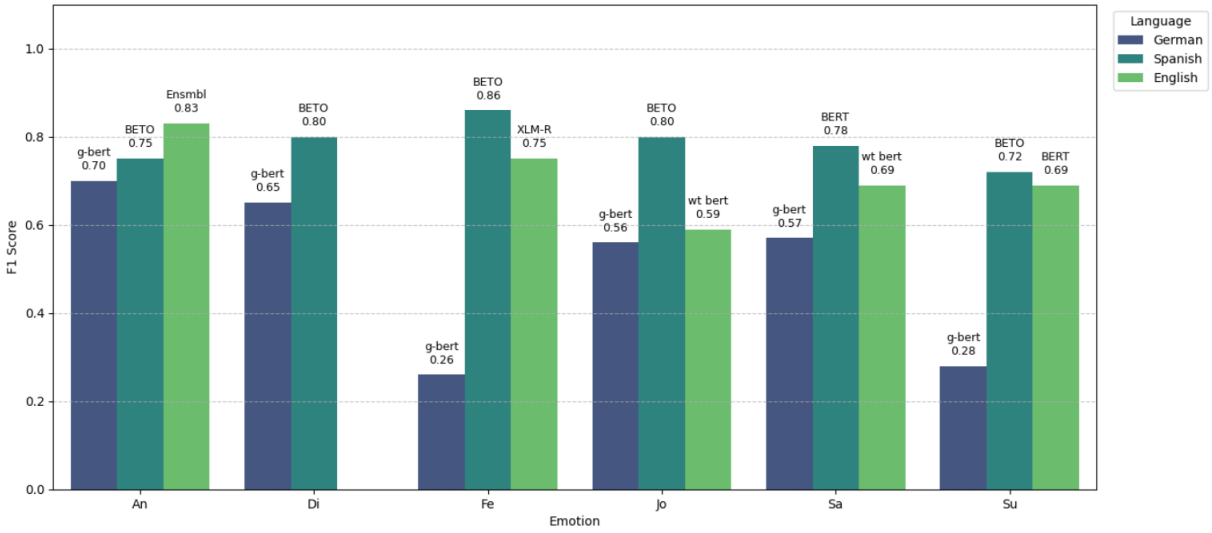

Figure 3: Best F1 Scores by Language.

for anger (0.83) and joy (0.77) as showcased by table 4. BERT-base-uncased shows decent results but struggles with joy and surprise. DistilBERT and XLM-RoBERTa show some promise but are outperformed by the ensemble models. However, the results were poor and could be improved further by implementing the suggestions mentioned above. We were placed 91st due to our results.

The comparative analysis highlights that language-specific models excel in low-resource settings, while ensemble methods perform better in high-resource languages, though all models struggle with nuanced emotions like joy and surprise.

5.2 Analysis

Transformer models, particularly language-specific ones like BETO and Google-BERT, consistently outperform traditional machine learning models, highlighting the importance of fine-tuning for specific languages as shown in Figure 3. In English, ensemble methods offer a strong alternative, outperforming individual models. Overall, transformer models excel in capturing complex emotions, but ensemble methods remain competitive, especially in English. The choice of separate binary classifiers for each emotion may have hindered the model's ability to distinguish between overlapping emotional expressions; by converting the multi-label emotion annotations into unique integer labels that represent specific emotion combinations, and modifying the final classification layer of the transformer models to output softmax probabilities over these combined classes with categorical cross-entropy loss, the model can better capture relationships between emotions, potentially improving overall classification coherence and performance.

6 Conclusion

Our experiments highlight key challenges in multi-label emotion detection. The lower performance in English suggests that pre-processing techniques may have removed valuable contextual information. Additionally, the choice of separate binary classifiers for each emotion may have hindered the model's ability to distinguish between overlapping emotional expressions. Our findings suggest that future research should focus on retaining more textual information during pre-processing, implementing multi-class classification rather than binary classifiers for each emotion and exploring larger, domain-specific pre-trained transformer models with better fine-tuning strategies. By addressing these factors, we can improve the accuracy and reliability of emotion detection in text across multiple languages.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*.

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

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021.

Making pre-trained language models better few-shot learners. In *Proceedings of ACL*.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 2000. Predicting the semantic orientation of adjectives. In *Proceedings of ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Soujanya Poria, Rada Mihalcea, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI*.

Saif Mohammad and Felipe Bravo-Marquez. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of SemEval*.

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)*, Vienna, Austria. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *arXiv preprint arXiv:1910.01108*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Chen Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of EMNLP*.

Rajalakshmi Sivanaiah, Sushmithaa Pandian, S Subhankar, Samyuktaa Sivakumar, R Rohan, and S Angel Deborah. 2024. Self-harm detection from texts: A comparative study utilizing bert, machine learning, and deep learning approaches. In *International Conference on Computational Intelligence in Data Science*, pages 110–123. Springer.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: An affective extension of wordnet. In *Proceedings of LREC*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of ACL*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NeurIPS*.

Venkatasai Ojus Yenumulapalli, Vijai Aravindh R, Rajalakshmi Sivanaiah, and Angel Deborah S. 2023. TechSSN1 at LT-EDI-2023: Depression detection and classification using BERT model for social media texts. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 149–154, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.