HTU at SemEval-2025 Task 11: Divide and Conquer - Multi-Label emotion classification using 6 DziriBERTs submodels with Label-fused Iterative Mask Filling technique for low-resource data augmentation.

Abdallah Saleh
Al Hussein Technical University
22210054@htu.edu.jo

Mariam Biltawi
Al Hussein Technical University
Mariam.Biltawi@htu.edu.jo

Abstract

In this paper, the authors address the challenges of multi-label emotion detection in the Algerian dialect by proposing a novel Label-fused Iterative Mask Filling (L-IMF) data augmentation technique combined with a multi-model architecture. The approach leverages DziriBERT, a BERT variant pre-trained on Algerian text, to generate contextually and label-sensitive augmented data, mitigating class imbalance while preserving label consistency. The proposed method uses six independent classifiers, each trained on a balanced dataset for a dedicated label, to improve performance. The results show significant improvement on the multi-label classification task using Deep Learning, with an official Macro F1 score of 48.6% and a best score of 51.2%.The system ranked 28/41 on the Algerian dialect scoreboard, achieving scores more than 7% to 9% higher than the task baseline using RemBERT.

1 Introduction

Dialectical Arabic, like any low-resource language, offers many challenges in the areas of Natural Language Processing. Lack of High-Quality Annotated datasets for tasks makes it particularly difficult to train Machine and Deep Learning models on downstream tasks like sentiment analysis, machine translation, named entity recognition, and emotion detection. Other challenges when dealing with Dialectic Arabic include increased morphological complexity, and variety of dialects,and variation within dialect itself, in a fairly close geographical space, making it considerably harder to develop a generalizable model(Faheem et al., 2024).

Hence, when dealing with a certain Arabic dialect, such as the Algerian dialect, one ought to be resourceful in both data pre-processing and model selection, leveraging any data pre-processing tool that might increase data size without degrading data quality, which could enhance model performance and generalizability. Furthermore, leveraging certain systems and model architecture techniques might alleviate bottlenecks in tasks like multi-label classification(Tarekegn et al., 2024).

The proposed data augmentation strategy, **L**abel-fused **I**terative **M**ask **F**illing (L-IMF), is a resourceful and contextually-based data augmentation technique that uses DziriBERT, a BERT model variant pre-trained on a large Algerian corpus (Abdaoui et al., 2021). Besides the proposed augmentation strategy, the usage of 6 DziriBERTs is also proposed to train on the given task by breaking down the multi-label technique into a set of binary classification tasks, with each DziriBERT dedicated to resolving one. The proposed system architecture addresses common problems associated with deep learning applications in multi-label classification tasks, primarily imbalanced data.

Despite limited time for optimization, the proposed system achieved strong results. This success was driven by a novel data augmentation technique, which expanded the original 901 training instances into over 15,500 samples, and by a model architecture that addressed multi-label classification using a binary relevance problem transformation. Two sampling strategies were applied. In the undersampling approach, a subset of the augmented samples was strategically selected to mitigate the dataset's intrinsic class imbalance, contributing to improved performance. In the oversampling approach, data from the minority class was increased to achieve class balance.

The remainder of this paper is structured as follows: Section 2 provides a background overview of the task and related work. Section 3 describes the proposed system. Section 4 presents the experimental setup, followed by results in Section 5. Section 6 concludes the paper and outlines potential future work. Finally, Section 7 discusses ethical considerations, while Section 8 includes

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 675–683

acknowledgments.

2 Background

The proposed system was designed for Track A: Multi-label Emotion Detection (Muhammad et al., 2025b), which predicts multiple emotions from text snippets. Each input is assigned a binary label (1 or 0) for six emotions; joy, sadness, fear, anger, surprise, and disgust; indicating their presence or absence. The Algerian dialect dataset includes 901 labeled training instances with gold-standard annotations, 100 validation instances, and 902 test instances (Muhammad et al., 2025a).

Multi-label classification extends beyond single-label approaches, allowing multiple classes to be assigned to the same input, sometimes with varying intensity levels (Tarekegn et al., 2024). This approach is widely used across fields, including healthcare, document classification, and emotion recognition.

Various approaches have been proposed to address challenges in multi-label learning. Traditional problem transformation techniques include binary relevance, classifier chain, and label Powerset.

In binary relevance, the multi-label problem is divided into multiple binary classification tasks, where each class is predicted independently and later combined into the final multi-label output. While simple to implement, this method ignores label dependencies, co-occurances, and correlations. Classifier chain addresses these limitations by modeling binary label predictions sequentially, allowing the model to learn relationships from earlier predictions. Label Powerset treats each unique label combination as a distinct class, meaning a six-class binary output results in 64 possible label combinations. However, both classifier chain and label Powerset are computationally expensive and struggle to capture high-order label correlations (Tarekegn et al., 2024).

With the rise of Deep Learning, researchers have explored neural architectures to bypass these traditional transformations. (Yang et al., 2018) used a LSTM layer as a decoder in multi-label document classification tasks where the LSTM layer produces labels sequentially and predicts the upcoming label from previous ones, thus allowing the high order capturing of label relationships. Transformer based model also were used in multi-label classification.

Despite Deep Learning models' ability to overcome the need for traditional problem transformation, many issues have been noticed in the usage of Deep Learning systems in multi-learning classification tasks. One of those limitations include difficulty in capturing high-order label dependencies, where it has been noted that Deep Learning has yet to overcome that inability to effectively address more that two labels simultaneously. There is also an issue with difficulty in addressing class imbalance in multi-label classification tasks. It has been noted that, compared to traditional problem transformations' approaches to resolve class imbalance, Deep Learning approaches are still underdeveloped(Tarekegn et al., 2024).

Several studies have attempted to address these challenges in Arabic multi-label text classification.(Aslam et al., 2024) used three BERT models, MarBERT, AraBERT and ArabicBERT for embedding extraction of a preprocessed arabic 2018 SemEval multi-label dataset where the word embeddings are then concatenated then passed to a meta learner made of a BI-LSTM layer that produces the multi-label outputs. The system was trained using a custom hybrid loss function that leverages label correlation matrix, contrastive learning, and class weighting, the use of three BERT models as extraction embeddings enhanced generalization, and the use of the custom hybrid loss function reduced to the negative effects of class imbalance present in the data along with promoting label dependency and correlation present in the data.

(Taha and Tiun, 2016) have noted the limited attention allocated by researchers towards multi-label classification for Arabic text. Binary classifier, their work proposed a binary relevance model that consists of different traditional machine learning classifiers. They also used multi-label classification for Arabic news articles. This is where each traditional machine learning classifier was trained on a distinct dataset to be able to categorize a news article as belonging to a class or not. They used different combinations of KNN, SVM and NB to be trained on different pre-processed text on different feature selection techniques, they found that a heterogeneous system of the different traditional machine learning model training on text preprocessed on chi-squared as feature selection performed the best results.

Proposed systems for Arabic multi-label classification using Deep Learning include those highlighted by Al-Smadi. (Al-Smadi, 2024) proposed DeBERTa-BiLSTM, which is a DeBERTa model

with a BiLSTM layer that receives the hidden state of the pre-trained DeBERTa model for the purpose of labeling FAQ Covid-19 multi-label dataset released from Arabic digital health platform Altibbi. The author trained the model using a Binary Cross-Entropy with Logits Loss as a loss function, thus treating each label independently as a binary classification problem.

Contrary to the belief that Deep Learning forgoes the need for traditional problem transformation, (Yang and Emmert-Streib, 2024) proposed a Deep Learning system that also successfully integrate traditional problem transformation; Yang and Emmert-Streib developed BR-CNN(Binary-relevance Convolutional neural network) with a custom weight scaling factor in the Binary Cross-Entropy loss function, BR-CNN achieves state-of-the-art performance on AAPD and MIMIC-III datasets, also outperforming models that leverage label dependencies.

While Deep Learning advancements have improved multi-label classification, challenges such as data scarcity and class imbalance persist. To address these issues, data augmentation techniques are employed to synthetically expand datasets, enhancing model performance, generalizability, and robustness while reducing overfitting.

Leveraging pre-trained language models has also been noted to be an effective technique that enhances model performance; Hence, it was successfully used in english classification tasks. One of the successful usages of pre-trained langauge models for english data augmentation is outlined by (Kesgin and Amasyali, 2023). Kesign and Amasyali proposed Iterative Mask Filling, a BERT augmentation technique that iteratively replaces words using masked language modeling to produce a single final augmented sentence for each input sentence. The proposed method outperforms traditional data augmentation approaches like synonym replacement, back-translation, and random modifications in topic classification. By using probabilistic word replacements and confidence based filtering, Iterative Mask Filling significantly improves model robustness. It was noted by the authors that the approach limited effectiveness in sentiment analysis, due to some words being critical in determining sentiment, when masked, allowing for the chance for the filled word to change the entire sentiment of the subsequent fully augmented sentence.

Another english data augmentation technique through language models is sketched by (Wu et al., 2019). Wu et al. proposed Conditional BERT (C-BERT), which is a fine-tuned version of BERT that performs context-aware word filling of randomly masked tokens. C-BERT outperforms other data augmentation techniques mentioned like random synonym replacement from WordNet and Contextual augmentation proposed by (Kobayashi, 2018).

Arabic literature, comparatively, has limitedly explored data augmentation techniques in the field of Arabic text classification. (Sabty et al., 2021) employed methods such as modified Easy Data Augmentation (EDA), back-translation, and word embedding substitution for Arabic Named Entity Recognition (NER) tasks, achieving positive but varying results. Similarly, (Abuzayed and Al-Khalifa, 2021) observed substantial improvements in classification performance through label augmentation.

(Refai et al., 2023) were one of the few researchers to use Transformer-based data augmentation in Arabic text classification, they leveraged AraGPT-2 and AraBERT for generative data augmentation, demonstrating significant performance gains in Arabic sentiment analysis. Additionally, (Carrasco et al., 2021) utilized Generative Adversarial Networks (GANs) to generate dialectal Arabic datasets for sentiment analysis, reporting enhanced model performance when training on the augmented dataset.

3 System overview

3.1 Label-fused iterative mask filling

3.1.1 Augmentation

In low-resource settings, any trained feature space that can be utilized for data augmentation or model enhancement has the potential to significantly improve performance on downstream tasks. The Algerian dialect presents challenges for data augmentation due to limited linguistic resources. For instance, synonym replacement is not a viable augmentation technique, as no comprehensive lexical database comparable to WordNet exists. More advanced techniques, such as back-translation, are also impractical since most Arabic-English translation models are designed for Modern Standard Arabic (MSA) and do not effectively handle dialectal variations.

Techniques leveraging pre-trained language models, such as iterative masking or Conditional BERT, pose additional challenges for multi-label

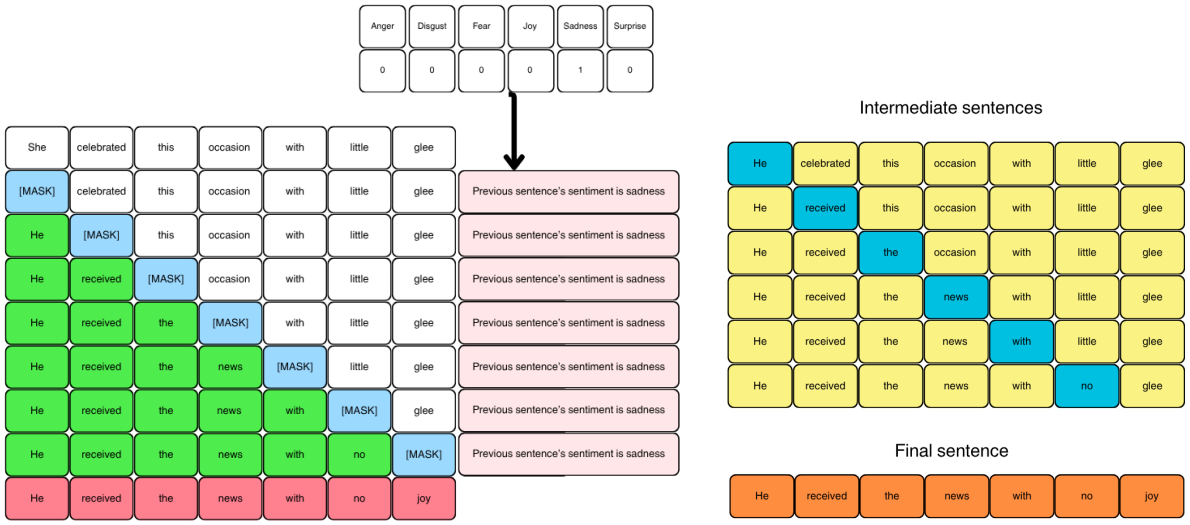


Figure 1: Left figure displays the process Label-fused Iterative Mask Filling . Right figure displays the output of the process.

sentiment-based tasks. Iterative mask filling has a detrimental impact on sentiment-based tasks, while Conditional BERT requires a large dataset for fine-tuning and does not support multiple simultaneous labels.

Data analysis revealed a significant class imbalance in certain instances, with 5 out of 6 emotion labels showing a data imbalance of more than 60% toward one of the labels, with only the sadness label having 44.84% labeled as 1 and 55.16% as 0. More than 57% of instances were annotated with at least two emotions present. The most common co-occurring label pairs were anger and disgust, anger and sadness, fear and surprise, and fear and sadness, with co-occurance percentages of 15.5%, 14.7%, 13.1%, and 12.8%, respectively. A detailed breakdown of label distribution and co-occurance patterns is provided in Appendix A.

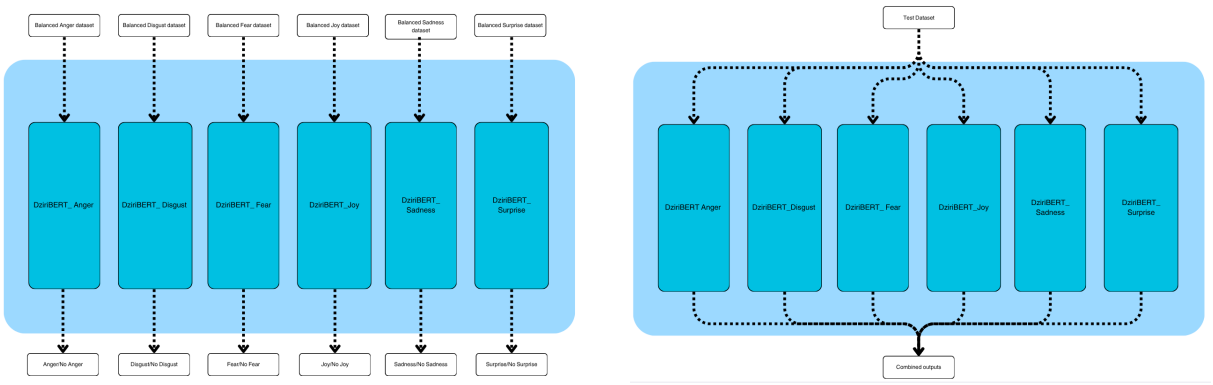
To address this data imbalance, the L-IMF method was proposed, which uses a pre-trained BERT model, or any other language model with Masked Language Modeling (MLM). This approach iteratively masks and replaces words, generating new sentences while also producing intermediate versions. To prevent the model from altering key words that could change the meaning of the sentence, a prompt-like sentence is inserted to guide the model. This prompt helps maintain label-aware token to be predicted. Using this algorithm, shown in Figure 1, around 14,500 intermediate sentences were created, along with 901 fully augmented, or "final", sentences.

Different Label-Fused prompts were tested. The simplest prompt used during augmentation, considered the baseline, involved simply appending the emotion present in the sentence. To minimize confusion, this prompt will be referred to as the *Simple Prompt*. Another prompt involved appending "The previous sentence's sentiment is {emotion(s)}" in Algerian Arabic; This will be referred to as the *Elaborate Prompt*.

To contrast the effectiveness of the proposed L-IMF, simple augmentation techniques were employed namely, random insertion, deletion, and swap. The algorithm that employed these augmentation techniques processed each sentence and randomly selected one of the three. If random insertion was selected, a random word collected from the entire dataset was inserted into the sentence at a random position. If random deletion was chosen, a random word was removed from the sentence. If random swap was selected, two random words within the sentence were swapped.

3.1.2 Sampling strategies

The original training dataset for the task contained 901 sentences, with approximately 15,500 additional sentences generated through L-IMF, resulting in a final training dataset of around 16,400 samples. Despite the increased dataset size, the class distribution exhibited severe imbalance, maintaining proportions comparable to those observed in the original dataset. To address this issue, undersampling and oversampling were performed using the Python-based Scikit-learn library. Each class was processed separately, with an output size



(a) Training process

(b) Inference process

Figure 2: A comparison of training and inference processes.

of 4,000 samples per class for the undersampled dataset and 30,000 samples per class for the oversampled dataset. The resulting balanced datasets were then stored for subsequent training.

With regard to the simple augmentation techniques, namely randomly insertion; deletion; and swap, the final dataset with the original and augmented sentences had a total of 1802 sentences. To avoid discarding any data from the limited dataset size, only oversampling was employed to create 4,000 samples per class for this simple data augmentation technique.

3.2 Model

To mitigate the challenges of class imbalance in multi-label classification with deep learning models, six independent classifiers were trained separately on tailored datasets. Each classifier was responsible for predicting a single class, and their outputs were concatenated only during inference, as illustrated in Figure 2. The model architecture included six pre-trained DziriBERT models, each equipped with a multi-layer perceptron (MLP) head. The MLP consisted of two linear layers with an input size of 768, a hidden size of 768, and an output dimension of 2. A ReLU activation function and dropout layers were applied between each linear layer.

To provide a performance comparison against the proposed model and its problem transformation strategy, a standalone DziriBERT model was evaluated. This model employed an MLP classification head identical to that of the proposed system, with the sole difference being that the output vector dimension is 6 instead of 2. This adjustment effectively transformed the multi-label classification task into a Powerset-based classification problem.

4 Experimental Setup

During training, the proposed model processed six tokenized text inputs, each corresponding to a label from the six datasets. Each submodel employed a separate Cross-Entropy loss function and was optimized using the Adam optimizer. Most layers in the submodels were frozen, except for the pooler layer and the MLP head. Dropout probabilities of 0.1, 0.2, and 0.3, alongside learning rates of 1.5×10^{-5}, 2.5×10^{-5}, and 3.5×10^{-5} were evaluated; however, the batch size was set at 32. Training was conducted for 15 epochs, with performance evaluated on the dedicated test set at the end of each epoch. The epoch with the best Macro F1 score was selected. A Grid Search hyperparameter optimization algorithm was used to tune the hyperparameter for each model. The baseline single DziriBERT model was trained on the original data with no augmentation. Despite almost the same training hyperparameters as the proposed model, the only difference was that Binary Cross Entropy with Logits Loss was set as the loss function. This loss function enables simultaneous prediction of multiple non-mutually exclusive labels by applying a sigmoid activation to each output logit. A detailed breakdown of the computational resources used is provided in Appendix B.

5 Results

The proposed system, combining the model with the L-IMF augmentation technique, achieved its highest performance with a Macro F1 score of 51.2% when trained on an undersampled dataset generated using the L-IMF *Simple Prompt*, while using oversampling on the same L-IMF *Simple Prompt* augmented data achieved 41.6%. Under

alternative settings, training on an undersampled dataset produced via the L-IMF *Elaborate Prompt* resulted in a Macro F1 score of 42.1% For comparison, the task's baseline, RemBERT, attained a Macro F1 score of 41.4%, while the standalone DziriBERT model, that provided a contrast to the proposed problem transformation, achieved only 25.1% under its best hyperparameters. While the proposed problem transformation proved effective, the same could not be said for the L-IMF augmentation technique. When the model was trained on data generated by the simple data augmentation techniques, the model achieved a Macro F1 score of 53.4%. Nonetheless, the proposed system, incorporating the problem transformation and L-IMF, achieved an improvement of nearly 10% over the task's baseline model and over 26% compared to the DziriBERT model. It ranked 28th out of 41 teams on the Algerian dialect task leaderboard with an official score of 48.6%.A more detailed breakdown of the scores along with the tuned hyperparameters for the best performing models is provided in Appendix C.

6 Conclusion

This study introduced the Label-fused Iterative Mask Filling (L-IMF) augmentation technique alongside a multi-model approach for multi-label classification in the Track A: Multi-label Emotion Detection challenge for Algerian dialect. The approach addressed key challenges in multi-label emotion detection, including label dependencies, class imbalance, and limited linguistic resources. By leveraging L-IMF, contextually and label-sensitive augmented data were generated, mitigating class imbalance while maintaining label consistency.

To further tackle the challenges of Deep Learning in multi-label classification, a system of six independent classifiers was implemented, with each DziriBERT-based sub-model specializing in predicting a single emotion. This design allowed the reduction of class imbalance by creating six balanced datasets for model training. Moreover, undersampling and oversampling helped ensure a more equitable distribution of classes, preventing the dominance of majority classes.

Results from the conducted experiments demonstrated the effectiveness of the proposed methodology in enhancing emotion classification performance in low-resource language settings. The integration of a pre-trained Algerian dialect model, L-IMF augmentation, and independent classifiers contributed to strong performance in multi-label emotion detection for dialectal Arabic that more than doubled the Macro F1 score obtained by using a single model via a Powerset problem transformation.

The experimental results indicate that the proposed model and binary problem transformation had the greatest impact on performance. Specifically, using six DziriBERT models as binary classifiers doubled the Macro F1 score compared to a single DziriBERT model with a output vector of dimension 6. In contrast, the L-IMF technique had a slightly negative effect on the model's performance when compared to a simpler augmentation method. However, the authors still argue that L-IMF holds significant potential for future improvements and remains a promising area for further research. Future studies may investigate the effectiveness of the L-IMF technique in augmenting datasets of different types, such as topic-based or sentiment-based datasets. Future work could also explore how the type and positioning of prompts influence performance on downstream tasks.

7 Ethical considerations

As a Pre-trained Language Model was used during both data augmentation and finetuning on downstream tasks, ethical challenges are posed due to the potential reinforcement of cultural bias and stereotypes found in the pre-training data used to train the language model.

8 Acknowledgments

The authors express gratitude to Al Hussein Technical University for its support and for fostering a research environment committed to innovation.

Research supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Abeer Abuzayed and Hend Al-Khalifa. 2021. Sarcasm and sentiment detection in arabic tweets using bert-based models and data augmentation. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 312–317.

Bushra Salem Al-Smadi. 2024. Deberta-bilstm: A multi-label classification model of arabic medical questions using pre-trained models and deep learning. *Computers in Biology and Medicine*, 170:107921.

Muhammad Azeem Aslam, Wang Jun, Nisar Ahmed, Muhammad Imran Zaman, Li Yanan, Hu Hongfei, Wang Shiyu, and Xin Liu. 2024. Improving arabic multi-label emotion classification using stacked embeddings and hybrid loss function. *arXiv preprint arXiv:2410.03979*.

Xavier A Carrasco, Ashraf Elnagar, and Mohammed Lataifeh. 2021. A generative adversarial network for data augmentation: The case of arabic regional dialects. *Procedia Computer Science*, 189:92–99.

Mohamed Faheem, Khaled Wassif, Hanaa Bayomi, and Sherif Abdou. 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14.

Himmet Toprak Kesgin and Mehmet Fatih Amasyali. 2023. Iterative mask filling: An effective text augmentation method using masked language modeling. In *International Conference on Advanced Engineering, Technology and Applications*, pages 450–463. Springer.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

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.

Dania Refai, Saleh Abu-Soud, and Mohammad J Abdel-Rahman. 2023. Data augmentation using transformers and similarity measures for improving arabic text classification. *IEEE Access*, 11:132516–132531.

Caroline Sabty, Islam Omar, Fady Wasfalla, Mohamed Islam, and Slim Abdennadher. 2021. Data augmentation techniques on arabic data for named entity recognition. *Procedia Computer Science*, 189:292–299.

Adil Yaseen Taha and Sabrina Tiun. 2016. Binary relevance (br) method classifier of multi-label classification for arabic text. *Journal of Theoretical & Applied Information Technology*, 84(3).

Adane Nega Tarekegn, Mohib Ullah, and Faouzi Alaya Cheikh. 2024. Deep learning for multi-label learning: A comprehensive survey. *arXiv preprint arXiv:2401.16549*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: sequence generation model for multi-label classification. *arXiv preprint arXiv:1806.04822*.

Zhen Yang and Frank Emmert-Streib. 2024. Optimal performance of binary relevance cnn in targeted multi-label text classification. *Knowledge-Based Systems*, 284:111286.

A Data analysis and label co-occurance

A.1 Emotion class distribution

Table 1 displays the severe class imbalance that most emotions have, with the most severe class imbalance being present in the joy label. Only around 17% of sentences have a positive joy label.

A.2 Number of labels per data point distribution

Almost 90% of the dataset has been labeled positively with atleast one emotion, with around 10% of the text instances present having no positive labels. The most occuring numbers of labels per text instance is 2 at around 34% of the entire dataset.

Emotion	Count (0)	Count (1)	Ratio (1:0)	% of 1s
Anger	605	296	1:2.04	**32.85%**
Disgust	695	206	1:3.37	**22.86%**
Fear	678	223	1:3.04	**24.75%**
Joy	748	153	1:4.89	**16.98%**
Sadness	497	404	1:1.23	44.84%
Surprise	588	313	1:1.88	**34.74%**

Table 1: Training dataset emotion class distribution, with ratios exceeding a 60/40 split highlighted in bold.

Number of Labels	Count	Percentage
0	91	10.10%
1	294	32.63%
2	303	33.63%
3	160	17.76%
4	50	5.55%
5	3	0.33%

Table 2: Distribution of the number of labels per instance in the training dataset.

Only 3 sentences have been labeled with 5 emotions present. No text instances have been labeled with all 6 emotions, as shown in table 2.

A.3 Co-occurance of emotions

Table 3 shows the co-occurance of emotions with each other. The least co-occurring emotions are joy with disgust, joy with fear, and joy with anger, with co-occurance percentages of 1.11%, 1.33%, and 1.44%, respectively.

B Computational Resources

T4 GPU and TPU v4-8 were used for the set of experiments. TPU v4-8 was used to train models that used either the oversampled data from L-IMF with the *Simple Prompt*, or undersampled data from L-IMF with the *Elaborate Prompt*. The rest of experiments were conducted on a enviroment with T4 GPU as the hardware accelerator.

The T4 GPU environment was configured with PyTorch version 2.5.0 and Transformers version 4.46.2. The TPU v4-8 environment employed PyTorch version 2.6.0, PyTorch XLA version 2.6.0, and Transformers version 4.49.0.

C Results breakdown

Table 4. shows a breakdown of optimal hyperparameters best performing model for each model type, augmentation technique employed and sampling strategy.

	anger	disgust	fear	joy	sadness	surprise
anger	32.85%	15.54%	8.44%	1.44%	14.65%	10.54%
disgust	15.54%	22.86%	4.22%	1.11%	13.32%	5.33%
fear	8.44%	4.22%	24.75%	1.33%	12.76%	13.10%
joy	1.44%	1.11%	1.33%	16.98%	3.55%	6.10%
sadness	14.65%	13.32%	12.76%	3.55%	44.84%	12.09%
surprise	10.54%	5.33%	13.10%	6.10%	12.09%	34.74%

Table 3: Label co-occurance percentage for the multi-label emotion training dataset. Diagonal values represent the individual label's prevalences in the training dataset.

Problem Trans.	**Augmentation**	**Sampling Strat.**	**Dropout**	**Learning Rate**	**Epoch**	**Macro F1**
Binary Relevance	L-IMF simple	Undersample	0.1	3.5×10^{-5}	8	51.2%
Binary Relevance	L-IMF simple	Oversample	0.2	1.5×10^{-5}	4	41.6%
Binary Relevance	L-IMF elaborate	Undersample	0.1	2.5×10^{-5}	11	42.1%
Binary Relevance	Simple augment.	Oversample	0.1	3.5×10^{-5}	10	53.4%
Powerset	None	None	0.2	3.5×10^{-5}	15	25.1%
Powerset - Baseline	None	N/A	N/A	N/A	N/A	41.4%

Table 4: Performance comparison across problem transformation methods, augmentation strategies, sampling strategies, and hyperparameters. Last row refers to RemBERT task baseline's performance