Converge at WASSA 2023 Empathy, Emotion and Personality Shared Task: A Transformer-based Approach for Multi-Label Emotion Classification

Aditya Paranjape, Gaurav Kolhatkar, Yash Patwardhan, Omkar Gokhale, Shweta Dharmadhikari
Pune Institute of Computer Technology, Pune, India
adifeb24@gmail.com, gauravk403@gmail.com, yash23pat@gmail.com, omkargokhale2001@gmail.com, scdharmadhikari@pict.edu

Abstract

In this paper, we highlight our approach for the "WASSA 2023 Empathy and Emotion Shared Task". We present our approach for track 3 of the shared task which aims to identify emotions from text. Each sample in the dataset has one or more labels making it a multi-label classification task. We compared multiple transformer-based models by fine-tuning them for multi-label classification. Oversampling was used to overcome the class imbalance in the dataset. Ensembling techniques were used to improve the performance of the system. We obtained a macro F1-score of 0.5649 using XLNet on the test dataset in the official phase and secured rank 6 on the official leaderboard. During the post-competition phase, a threshold-based voting mechanism was performed on three models (Longformer, BERT, BigBird) that yielded the highest overall macro F1-score of 0.6605.

1 Introduction

With a rapid increase in the technological and scientific advancements seen in the domains of Machine Learning and Deep Learning, machines can now easily perform complex tasks at a degree of proficiency comparable to that of humans. However, one aspect where machines fall short in performing human-like tasks is those which require the understanding and contextualization of emotions. Emotions can be broadly divided into primary and secondary emotions (Rodríguez-Torres et al., 2005). Primary emotions include but are not limited to joy, sadness, and anger; while secondary emotions are emotions that are caused by other emotions.

Emotion Classification is an approach that helps in identifying the emotional context of textual data. This classification can serve as a concise summary for the readers. Applications such as recommendation systems also benefit greatly when used in combination with emotion classification approaches. Based on the classification of the user query, potential recommendations (Barrière and Kembellec, 2018) are narrowed down for the user and help the model in finding the best response. Emotion classification plays a crucial role in bridging the gap between human-computer interaction.

Through this paper, we intend to examine the efficacy of several transformer-based models for producing competitive results for emotion classification. The texts on which the models are trained and tested are essays that are responses to news articles. The models perform multi-label classification to identify the emotions expressed in the essays.

2 Related Work

Ekman and Friesen (1986) suggested that there are a set of universal emotions, which include happiness, sadness, anger, fear, disgust, and surprise, that are expressed by all humans through specific facial expressions regardless of their cultural background.Darwin and Prodger (1998)'s investigation into the expression of emotion on the face and through body gestures in both humans and animals marked a pioneering moment in the science of emotion recognition and analysis. Emotions can be recognized primarily through three categories: facial expressions (Goldman and Sripada, 2005), voice (Koolagudi and Rao, 2012), and text (Thakur et al., 2018). The process of automatically tagging a text with an emotion from a list of predetermined emotion labels is known as emotion recognition in text.

Early research concentrated on a lexicon-based methodology (Pradhan et al., 2023) which establishes polarity or sentiment to classify emotions from a text as positive, negative, or neutral. This was followed by the introduction of keyword-based methodology (Tao, 2004; Ma et al., 2005) that involves locating keyword occurrences in a text and tagging each one with an emotion from an emotion dictionary. Subsequently, based on rule-based techniques, rule-based models (Lee et al., 2010; Udochukwu and He, 2015) were presented

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 558–563
July 14, 2023

in which the rules for emotion detection were extracted from the preprocessed dataset and the best rule among them was selected for emotion labeling.

With the emergence of machine learning approaches that categorize text into multiple emotion categories, it has been observed that SVM (Desmet and Hoste, 2013) and Bayesian networks (Liew and Turtle, 2016) consistently produce good results. Several classification algorithms were evaluated for multi-label emotion recognition (Xu et al., 2018) and it was discovered that logistic regression produced the best results on the provided features. As research in the field of deep learning gained traction, various models for multi-label emotion recognition that used CNNs (Wang et al., 2016), DNNs (Du and Nie, 2018), LSTMs (Li et al., 2018) and Bi-LSTMs (Baziotis et al., 2018) were proposed. In addition to other deep learning ideas, transformer models like BERT (Devlin et al., 2018) were employed in a variety of applications to improve performance. The most popular deep learning methods, nevertheless, were those based on LSTM and its subtypes.

In order to produce accurate results for emotion detection tasks, numerous hybrid models (Park et al., 2018; Seol et al., 2008; Shaheen et al., 2014; De Bruyne et al., 2018) combining various strategies were proposed from the pool of methods developed for text-based emotion analysis. In this paper, we compare various transformer-based models for emotion classification and perform experiments on the same.

3 Dataset Description

The dataset provided for this task (Omitaomu et al., 2022; Barriere et al., 2023) comprised essays that were written in response to news articles. The essays vary in length, ranging from 300 characters to 800 characters. The training data had 792 samples of such essays, the development data contained 208 samples, and the test data comprised 100 samples. The training data contained features like the essay, article-id, speaker-id, gender, education, etc. This shared task problem falls under the category of multi-label classification. There are 8 base emotions or labels (Anger, Hope, Sadness, Neutral, Disgust, Surprise, Joy, Fear) and each essay in the dataset is assigned one or more of these labels. The class of 'Sadness' had the highest number of samples in the training data, with 297 samples. Whereas, the class 'Joy' had the least number of samples in the training data, with only 5 samples.

4 Methodology

First, we evaluate and compare the performance of different models on the test dataset based on their Macro F1-score and Micro F1-score metrics. These models are listed and explained below. We finetune these models on the training dataset using the standard procedure for multi-label classification. We use a threshold value of 0.37 to decide whether a label should be assigned to a particular example. If the probability output for a certain label is greater than the threshold, then that label is selected. All the models were trained for 12 epochs (except for Longformer, which was trained for 10 epochs) with a learning rate of 4e-5. The results obtained in the post-competition phase have been showcased in Table 1. The official phase score for XLNet is also mentioned in Table 1.

4.1 Longformer

Longformer (Beltagy et al., 2020) is a transformer-based model that is useful for tasks that require processing long sequences of text. Longformer uses a modified attention mechanism that scales linearly with the input size, as opposed to the quadratic time taken by the traditional attention mechanism. It achieves this by using a combination of local and global attention.

4.2 BERT

BERT is a language representation model. It is used to obtain bidirectional representations of text input, which yield state-of-the-art results on many NLP tasks.

4.3 XLNet

XLNet (Yang et al., 2019)is an autoregressive pretraining technique that improves on the deficiencies of BERT. XLNet uses a Permutation Language Modelling objective, to help understand the bidirectional context. The model outperforms BERT on several NLP tasks.

4.4 BigBird

BigBird (Zaheer et al., 2020) is a BERT-like model that is useful for longer input sequences. It replaces the self-attention mechanism in BERT with a combination of sparse, global, and random attention. This requires much lesser computational power while giving a comparable performance.

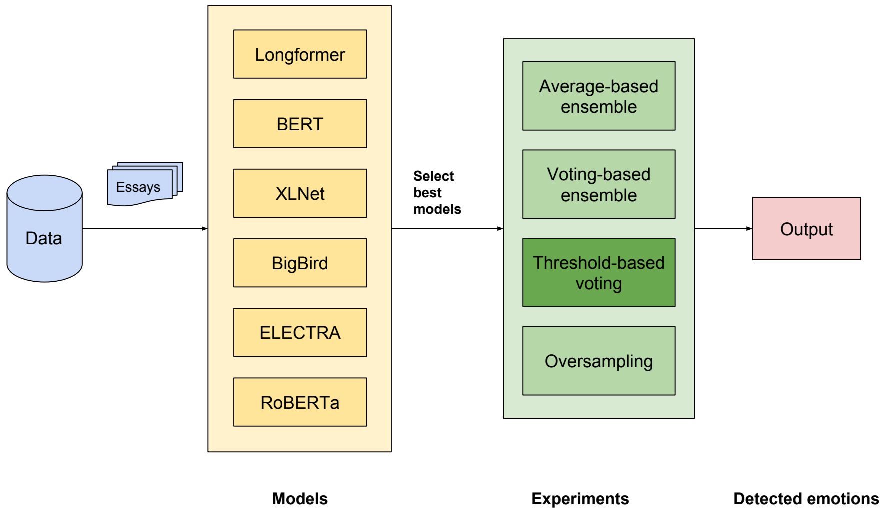

Figure 1: Methodology

4.5 ELECTRA

ELECTRA (Clark et al., 2020) is a pre-training method that aims to use significantly fewer compute resources than an MLM pre-training method. The pre-training stage involves training two transformer models: the generator and the discriminator. The discriminator model is further used on downstream tasks.

4.6 RoBERTa

RoBERTa (Liu et al., 2019) improves on the BERT model by making some important tweaks to the hyperparameters. It removes the next sentence prediction pre-training objective and uses much larger mini-batch sizes and learning rates.

Model name	**Macro F1 score**	**Micro F1 score**
XLNet* (Official)	0.5649	0.7009
XLNet (Post-Competition)	0.5927	0.7018
RoBERTa	0.5716	0.6937
BERT	0.6308	0.7039
BigBird	0.6281	0.7074
Electra	0.5860	0.7167
LongFormer	**0.6360**	**0.7289**

Table 1: Vanilla Model outputs (Post-Competition)
* Official result was submitted on the official leaderboard and was trained with a higher learning rate.

5 Experiments

5.1 Ensemble

Based on our results on the test data, we ensemble the top models by using three strategies as shown in Figure 1.

5.1.1 Voting

We calculate the outputs for each sample using all 3 models. We then take a vote between the models to determine the actual output. If all three models give different outputs, preference is given to the top model. In this case, the top 3 models are Longformer, Bigbird, and BERT, with the highest preference given to Longformer. We repeat this process for the top 5 models which are Longformer, Bigbird, BERT, XLNet, and RoBERTa.

5.1.2 Averaging

We average the individual probability values for each class obtained from the top 3 models and then determine the output label/labels for each sample based on the 0.37 threshold mentioned in 4. We repeat this process with the top 5 models and compare the results.

5.1.3 Threshold-based voting

We observed that the previous voting strategies seem to fail for samples having ground truths that consist of multiple labels. To counter this we implement a threshold-based strategy. This strategy is implemented on an ensemble of the top 3 as well as the top 5 models.

Top 3 models:

We implement voting with an extra stipulation that if a model predicts a label with a confidence higher than 0.55, then its label is retained irrespective of whether it wins or loses the vote.

Top 5 models:

In this ensemble, we add two stipulations to the voting process. First, if two models predict the same label with a confidence higher than 0.5 then that label is retained. Second, if a single model predicts a label with confidence higher than 0.75, then that label is retained.

Experiment	Models used	Macro F1 score	Micro F1 score
Average-based ensemble	Top 3	0.5695	0.6953
Voting-based ensemble	Top 3	0.5683	0.6926
Threshold based voting	Top 3	**0.6605**	**0.7236**
Average-based ensemble	Top 5	0.6098	0.7094
Voting-based ensemble	Top 5	0.561	0.693
Threshold based voting	Top 5	0.6104	0.6917
Oversampling	Long-former	0.4653	0.6784

Table 2: Post-competition results in the test dataset (Top 3: Longformer, BERT, BigBird, Top 5: Top 3 + XLNet, ELECTRA

5.2 Oversampling

There is a significant class imbalance in the data. To counter this we implement oversampling. Here, we duplicate samples from classes having less number of samples. The end goal is to have an equal number of samples for each class. In our dataset, class 'Sadness' has 292 samples which is the highest number of samples. So, we oversample the other classes such that each class has 292 samples.

6 Results

We hereby analyze the results shown in Table 2. We make some key observations regarding the results as follows:

Longformer is the best standalone model:

Out of all the vanilla models we trained, we see that 'Longformer' performs best with a macro F1-score of 0.6360. In the provided dataset, the average number of words per essay is 86. Since Longformer works well for long input sequences, as is seen in the provided training dataset, it outperforms the other models.

Oversampling yields no performance improvement:

We observed that oversampling leads to a significant decrease in macro F1-score, obtaining a score of 0.4653. Further investigation is required to explain this discrepancy.

Ensembling significantly improves results:

Both the approaches provided competitive results, however threshold-based voting with three models(Longformer, BERT, BigBird) gives the best overall macro F1-score score of 0.6605. Average-based Ensemble with five models(Longformer, BERT, BigBird, XLNet, ELECTRA) also provides good results with a macro F1-score of 0.6098.

7 Conclusion

In this paper, we compared the performance of six transformer-based models (Longformer, BERT, BigBird, XLNet, ELECTRA, RoBERTa) for emotion classification on the test dataset. Our official macro F1-score in the official phase was 0.5649, which was obtained on XLNet. Further, many improvements were made in the scores in the post-competition phase. It was observed that Longformer outperformed all other models with a macro F1-score of 0.636. We conducted multiple experiments by employing ensembling and oversampling techniques which concluded that the threshold-based voting method yields the best performance with a macro F1-score of 0.6605. In the future, we plan to improve our oversampling score and combine it with threshold-based voting.

References

Valentin Barrière and Gérald Kembellec. 2018. Short review of sentiment-based recommender systems. In *Proceedings of the 1st International Conference on Digital Tools & Uses Congress*, pages 1–4.

Valentin Barriere, Shabnam Tafreshi, Jo ao Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis*.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2018. Lt3 at semeval-2018 task 1: A classifier chain to detect emotions in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 123–127.

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pan Du and Jian-Yun Nie. 2018. Mutux at semeval-2018 task 1: exploring impacts of context information on emotion detection. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 345–349.

Paul Ekman and Wallace V Friesen. 1986. A new pancultural facial expression of emotion. *Motivation and emotion*, 10:159–168.

Alvin I Goldman and Chandra Sekhar Sripada. 2005. Simulationist models of face-based emotion recognition. *Cognition*, 94(3):193–213.

Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology*, 15:99–117.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 45–53.

Meng Li, Zhenyuan Dong, Zhihao Fan, Kongming Meng, Jinghua Cao, Guanqi Ding, Yuhan Liu, Jiawei Shan, and Binyang Li. 2018. Isclab at semeval-2018 task 1: Uir-miner for affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 286–290.

Jasy Suet Yan Liew and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL student research workshop*, pages 73–80.

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. *arXiv preprint arXiv:1907.11692*.

Chunling Ma, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion estimation and reasoning based on affective textual interaction. In *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1*, pages 622–628. Springer.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and Jo ao Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Ji Ho Park, Peng Xu, and Pascale Fung. 2018. Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and# hashtags. *arXiv preprint arXiv:1804.08280*.

Anima Pradhan, Manas Ranjan Senapati, and Pradip Kumar Sahu. 2023. Comparative analysis of lexicon-based emotion recognition of text. In *Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021*, pages 671–677. Springer.

Ramón Rodríguez-Torres, Jacques Philippe Leyens, Armando Rodríguez Pérez, Verónica Betancor Rodriguez, María N Quiles del Castillo, Stéphanie Demoulin, and Brezo Cortés. 2005. The lay distinction between primary and secondary emotions: A spontaneous categorization? *International Journal of Psychology*, 40(2):100–107.

Yong-Soo Seol, Dong-Joo Kim, and Han-Woo Kim. 2008. Emotion recognition from text using knowledge-based ann. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, pages 1569–1572.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, pages 383–392. IEEE.

Jianhua Tao. 2004. Context based emotion detection from text input. In *Eighth International Conference on Spoken Language Processing*.

Priyanka Thakur, Dr Rajiv Shrivastava, and A DR. 2018. A review on text based emotion recognition system. *International Journal of Advanced Trends in Computer Science and Engineering*, 7(5).

Orizu Udochukwu and Yulan He. 2015. A rule-based approach to implicit emotion detection in text. In *Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings 20*, pages 197–203. Springer.

Yaqi Wang, Shi Feng, Daling Wang, Ge Yu, and Yifei Zhang. 2016. Multi-label chinese microblog emotion classification via convolutional neural network. In *Web Technologies and Applications: 18th Asia-Pacific Web Conference, APWeb 2016, Suzhou, China, September 23-25, 2016. Proceedings, Part I*, pages 567–580. Springer.

Huimin Xu, Man Lan, and Yuanbin Wu. 2018. Ecnu at semeval-2018 task 1: Emotion intensity prediction using effective features and machine learning models. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 231–235.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.