wangkongqiang at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Kongqiang Wang
School of Information Science and Engineering, Yunnan University,
Kunming 650500, Yunnan, China
wangkongqiang60@gmail.com

Abstract

This paper presents our system developed for the SemEval-2025 Task 11:Bridging the Gap in Text-Based Emotion Detection, on Track A: Multi-label Emotion Detection.(Muhammad et al., 2025b)Given a target text snippet, predict the perceived emotion(s) of the speaker. Specifically, select whether each of the following emotions apply: joy, sadness, fear, anger, surprise, or disgust. To this end, we focus on English source language selection strategies on four different pre-trained languages models: google-bert,FacebookAI-roberta,dccuchile-bert and distilbert-multi.We experiment with 1) the training set data is analyzed visually, 2) multiple numbers of single models are trained on the training set data, and 3) multiple number of single models for voting weight ensemble learning. We further study the influence of different hyperparameters on the integrated model and select the best integration model for the prediction of the test set. Our submission achieved the good ranking place in the test set.Emotion Macro F1 Score 0.6998 and Emotion Micro F1 Score 0.7374. For the final ranking, organizers will use the Macro F1 score.Even so, my approach has yielded good results.

1 Introduction

Emotions are simultaneously familiar and mysterious.(Vaidya et al., 2024) On the one hand, we all express and manage our emotions every day. Yet, on the other hand, emotions are complex, nuanced, and sometimes hard to articulate. We also use language in subtle and complex ways to express emotion.Further, people are highly variable in how they perceive and express emotions (even within the same culture or social group).Thus, we can never truly identify how one is feeling based on something that they have said with absolute certainty.Emotion recognition is not one task but an umbrella term for several tasks such as detecting the emotions of the speaker, identifying what emotion a piece of text is conveying and detecting emotions evoked in a reader. Based on the predictive task background of predictive emotion text, We propose an ensemble learning method based on pre-trained language model. The code of this method is available on my GitHub website.[1]

2 Related Work

SemEval in previous years has introduced tasks focusing on Multi-label text classification and text binary classification (Wang et al., 2024)(Su and Zhou, 2024)(Tran and Tran, 2024)(Brekhof et al., 2024)to evaluate Internal potential elements and potential content of the text.These tasks provided datasets with human labeled similarity scores, which have been extensively utilized for training sentence embedding models and conducting semantic evaluations.

2.1 Sentence Embeddings

Word embedding models such as BERT, GloVe, RoBeRTa and Word2Vec are frequently employed to assess the semantic distance between words.They are also some of the more commonly used methods in text classification tasks.Sentence embeddings with a fixed length are often generated via mean/max pooling of word embeddings or employing CLS embedding in BERT. The semantic distances are commonly measured using the cosine similarity of embeddings of two expressions.Siamese or triplet network architectures are frequently employed in sentence embedding training. For example, models such as Sentence-BERT utilize a dual-encoder architecture with shared weights for predicting sentence relationships (e.g., semantic contradiction, entailment, or neutral labeling) or for similarity score prediction using regression objectives, e.g., the difference between human

[1] https://github.com/WangKongQiang

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 160–164
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

annotated similarity score (sim) of two sentences and the cosine of two sentence embeddings.

2.2 Ensemble Learning

In previous studies, ensemble learning presents several advantages. The ensemble approach can reduce the errors from individual models by amalgamating results from multiple sources or can make the system more robust. In our study, using multiple pre-trained models can also save a substantial amount of computation while making use of information from the large data during pre-training. Previous research has demonstrated that ensemble learning can achieve remarkable success.

In our study, we aim to integrate multiple pre-train learning models to assess semantic relatedness.When models are trained on diverse datasets with different architectures, they may produce varied predictions on semantic relatedness, and combining them may improve overall performance.We use sentence embeddings mainly from the following models.Multilingual BERT (cased, uncased),RoBERTa,BETO (cased, uncased),DistilBERT.

3 Methodology

3.1 overall architecture

The pursued approach involves using a weighted voting system of ensembles composed of different transformers. We trained several state-of-the-art NLP(Natural language processing) models on a large dataset of annotated tweets to create ensembles of classifiers with different architectures and configurations.We then combined the predictions of these ensembles using a weighted voting system to produce the final predictions.We have used the following transformers for the ensembles:Multilingual BERT (cased, uncased),RoBERTa,BETO (cased, uncased) DistilBERT.

For each instance, the final classification decision is based on the weighted sum of outputs of these models. The novel weighted-voting system presented involves using each (normalized) transformer's metric score in the ensemble (F1-score or RMSE, depending on the task) to assess the importance of these in the final outputs of the ensemble (as opposed to the arithmetic mean typically used in conventional voting systems).

3.2 Implementation step

First,The simpletransformers Python library that will be used below requires the data to be presented in a specific form.The following data cell adapts each split to contain only two columns: text and labels, where the latter is an array equal in size to the number of labels.

Second,Models' definition.In this section,the different transformers that will be evaluated are gathered.For this purpose,the implementation mainly relies in the simpletransformers Python library, which allows to train and test transformers within few steps.

Third,Training.Each of the aforementioned models is trained separatedly with the entire training set.This training is directly performed in the previously defined dictionary for convenience.

Fourth,Ensembles' definition.The ensembles of transformers that can be defined with the previously trained models are created.A dictionary is create for convenience, univocally identifying each ensemble.

Fifth,Evaluation.Firstly, each transformer is individually evaluated using the validation split. Subsequently, the main evaluation metrics (accuracy, F1-score, precision and recall) are stored.Secondly, the predictions of each ensemble for the validation set instances are derived. After calculating their metrics, it is possible to determine which ensemble obtained the best F1-Score. This will be the final ensemble used for the test dataset.Regarding the ensembles' predictions, these are obtained through a hard voting system: after computing the output that each of the ensemble's models produces for a given instance, the most-voted class turns out to be the ensemble result.The voting system can be non-weighted or weighted. In the latter, the prediction of each individual transformer is weighted according to their normalized F1-score, thus providing a greater importance to the best model without disregarding the outputs of the other transformers.

Sixth,The vote function determines the ensembler prediction based on the outcomes of its transformers.Its arguments are:predictions, list of transformers' (raw) outputs.weighted,bool that determines if a weighted voting system must be used.weights,list of weights (normalized weighted F1-scores).

Seventh,Selecting the best ensemble Once the predicted labels for each validation instance are calculated for each ensemble, their metrics can be computed. Given that it is a multi classification

training set text	value
count	2768.000000
mean	17.581286
std	11.701499
min	3.000000
25%	9.000000
50%	15.000000
75%	23.000000
max	90.000000

training set label	value
Anger	333
Fear	1611
Joy	674
Sadness	878
Surprise	839

Table 1: The text data situation and the number of emotional labels are described

Hyperparameter	Values
Optimizer	AdamW, Adafactor
Learning rate	2e-05, 4e-05, 8e-05

Table 2: Experimentation configuration hyperparameters

task, the best ensemble will be that with a maximum F1-score.

Eighth,Predictions on test set Finally, the ensemble which obtained a higher F1-score can be used to predict the label of each test instance.

Further, these results will be used to portray some evaluation plots, including the Confusion Matrix and the ROC curve.

4 Results and Analysis

4.1 Training set analysis

The text and label of training set is described in Table 1.The length and quantity distribution of training text data are analyzed in Figure 1.Distribution of the size of texts for each class in Figure 2.It shows the number of percentages relative to each class for various cases.

4.2 Experimentation configuration

For the sake of completeness and in an attempt to improve the results obtained by the transformer assemblers, each run was repeated a total of 6 times with the different combinations of the following hyperparameters:See Table 2.

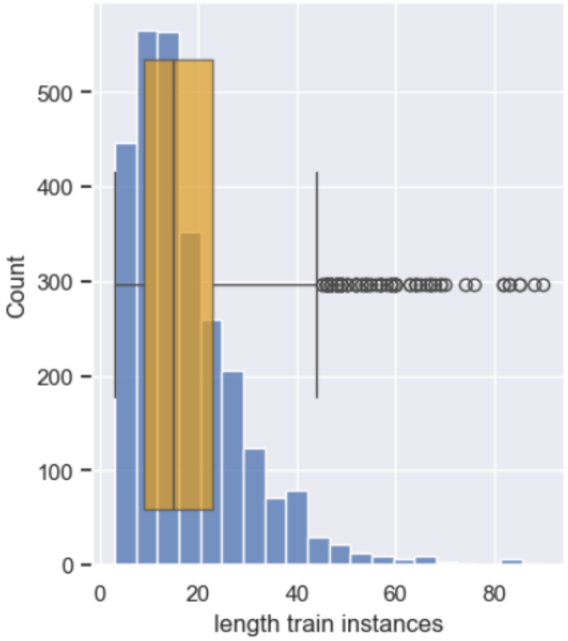

Figure 1: The length and quantity distribution of training text data are analyzed.

Dev set Emotion	Score
Macro F1	0.7068
Micro F1	0.7304
Anger	0.6667
Fear	0.7794
Joy	0.625
Sadness	0.7647
Surprise	0.6984

Table 3: The Dev data situation detailed results described

4.3 Dev set result

The following Table 3 records the official results of SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, on Track A: Shared task of multi-label Emotion Detection. The metrics recorded by the best (winning) approach in the evaluation task of the development set.

4.4 Test set result

The following Table 4 records the official results of SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, on Track A: Shared task of multi-label Emotion Detection. The metrics recorded by the best (winning) approach in the evaluation task of the test set.

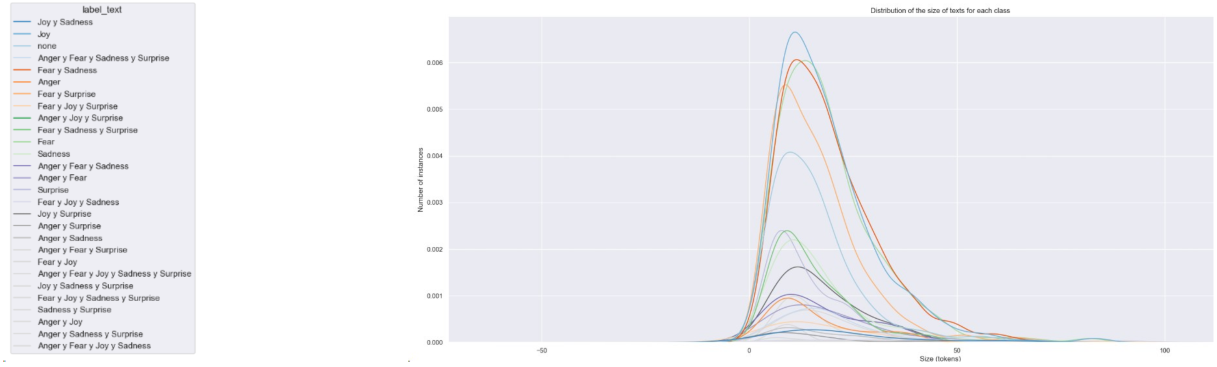

Figure 2: Distribution of the size of texts for each class.

Test set Emotion	Score
Macro F1	0.6998
Micro F1	0.7374
Anger	0.5812
Fear	0.8152
Joy	0.7032
Sadness	0.7104
Surprise	0.6891

Table 4: The Test data situation detailed results described

4.5 Biased Performance

From Figure 1 of the visual analysis, we can observe that 75% of tweets in training set data, either in the chart or in the previous input column, have no more than 25 words. This information could be useful in determining the size of a network of neurons, or when a sentence length limit needs to be set.

SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, on Track A: Shared task of multi-label Emotion Detection. This task is multi-label sorting. Each instance can have 0 to n(n=5,6) categories, and you need to predict which category each instance belongs to. In the specific cases(English language) we focus on, there may be up to five different categories: anger, fear, joy, sadness, surprise. For this task, we will look at the quantity distribution followed by each category, as shown in Table 1. In this case, percentages cannot be assessed because of the intersection.

5 Conclusion

Our system employs an ensemble approach to estimate semantic relatedness(Eneko Agirre and Wiebe, 2014),integrating results from multiple systems:google-bert-base-multilingual-uncased and FacebookAI-roberta-base.The hyperparameter is following: eval-batch-size is 8,num-train-epochs is 5,learning-rate is 4e-05,optimizer is AdamW,use-early-stopping is True.The dataset usage is shown in Table 5. Our findings suggest that semantic relatedness can be deduced from a variety of sources. Although some features (e.g., lexical overlap ratio)may not perform as strongly as models specifically designed to obtain sentence representations, the results demonstrate that these features, when used in a combined manner, can outperform many individual systems and collaboratively achieve a better correlation with human judgment on semantic relatedness.(Siino, 2024)

6 Limitation and Future Work

Our experiments are based on English language data sets only. Constrained by the size of the training data and the availability of pre-trained language models, it is regrettable that we did not offer insights into other Asian and African languages.In future research,studies on low-resource languages will be valuable, including tasks such as data collection, annotation,and pre-training models tailored to these languages.

Acknowledgments

We are very grateful for the assistance and discussions provided by Semeval-2025 Task11 leader and organizer.

References

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on*

Dataset input	description	Use or not
(Muhammad et al., 2025a)	`Datasets for 28 Languages.`	yes
(Belay et al., 2025)	`Amharic, Oromo, Somali, and Tigrinya`	no
other Dataset	use external or additional corpora	no

Table 5: Use dataset supported by Semeval-2025 Task11 on Track A. The style is based on raw data.

Computational Linguistics, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Thijs Brekhof, Xuanyi Liu, Joris Ruitenbeek, Niels Top, and Yuwen Zhou. 2024. Groningen team D at SemEval-2024 task 8: Exploring data generation and a combined model for fine-tuning LLMs for multidomain machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Claire Cardie Daniel Cer Mona Diab Aitor Gonzalez-Agirre Weiwei Guo Rada Mihalcea German Rigau Eneko Agirre, Carmen Banea and Janyce Wiebe. 2014. *SemEval-2014 task 10: Multilingual semantic textual similarity. In Proceedings of the 8th Interna tional Workshop on Semantic Evaluation (SemEval 2014), pages 81–91.* Association for Computational Linguistics, Dublin, Ireland.

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.

Marco Siino. 2024. All-mpnet at SemEval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 379–384, Mexico City, Mexico. Association for Computational Linguistics.

Lianshuang Su and Xiaobing Zhou. 2024. NLP_STR_teamS at SemEval-2024 task1: Semantic textual relatedness based on MASK prediction and BERT model. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 337–341, Mexico City, Mexico. Association for Computational Linguistics.

Bao Tran and Nhi Tran. 2024. NewbieML at SemEval-2024 task 8: Ensemble approach for multidomain machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 354–360, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. CLTeam1 at SemEval-2024 task 10: Large language model based ensemble for emotion detection in Hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 365–369, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.