ABCD team at EMit: Ensemble Approach for Categorial Emotion Detection in Social Media Messages

Nguyen Ba Dai[1,3,4], Nguyen Ngoc Phuong Uyen[2,3,5] and Dang Van Thin[1,3]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]University of Economics and Law, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[5]Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam

Abstract

This paper presents a submission system for the EMit (Emotions in Italian) shared task at EVALITA 2023, which focuses on categorial emotion detection in social media messages related to TV shows, TV series, music videos, and advertisements domain. We employ an ensemble approach, leveraging the BERT (Bidirectional Encoder Representations from Transformers) models known for its advanced language understanding capabilities. The BERT model is fine-tuned using domain-specific data to enhance its performance in emotion detection. The ensemble architecture combines multiple pre-trained models and utilizes a soft voting technique for robust decision-making. The results demonstrate the effectiveness of the team's ensemble model, achieving a Top 3 ranking in task 1 with 49.94% of the F1-score.

Keywords

Ensemble Approach, Categorical Emotion Detection, BERT, Sentiment Analysis, NLP, Ensemble Model

1. Introduction

Within several fields, including Natural Language Processing (NLP), Artificial Intelligence (AI) has made significant contributions in offering efficient answers for crucial societal and human issues. Two critical areas of NLP are sentiment analysis and emotional recognition [1]. While sentiment analysis usually classifies data into three main categories (positive, negative, neutral), emotional recognition extracts distinct human emotions such as disgust, fear, joy, and more. Emotion detection has been studied and applied extensively in computational and linguistic techniques to help computers understand and, at times, generate human languages. We can understand its significance because emotions play vital roles in the existence or the complete make-up of individuals. EVALITA 2023 [2] provides a shared framework for the evaluation of different systems and approaches on the EMit (Emotions in Italian) shared task [3]. In this shared task, two sub-tasks, both designed as multilabel classification problems, were proposed for participants. The first challenge called Categorial Emotion Detection, aims to identify emotions in social media messages or the absence of emotions. The EMit of EVALITA offered 8 emotional labels defined by Plutchik (anger, anticipation, disgust, fear, joy, sadness, surprise, trust, and love) [4]. In this sub-task, we detect emotions from data provided by task organizers in the social media domain related to TV shows, TV series, music videos, and advertisements. The study of emotions in social media, employed as a social signal for capturing the emotional reactions of the Italian audience, can offer artists or broadcasters highly valuable finer-grained information in assessing the delivered content.

2. Related Work

Detection of emotions is not entirely novel, for more than a decade, there have been several international evaluation campaigns launched for example the Emotion Classification shared task at WASSA 2022 [5], EmoContext at SemEval 2019 [6], The Affective Text shared task at SemEval 2007 [7],... However, further research is still needed, especially on identifying emotions through messages related to TV shows, TV series, music videos, and advertisements.

In EVALITA 2018, the Hate Speech Detection (HaSpeeDe) task was conducted specifically for Italian, with the objective of automatically labeling messages from Twitter and Facebook as either containing or not containing hate speech. To develop robust hate speech detection systems, [8] utilized the dataset released for the HaSpeeDe shared task, which combined an English dataset and a German dataset distributed for the Identification of Offen-

EVALITA 2023, September 7–8, 2023, Parma, Italy
21521914@gm.uit.edu.vn (N. B. Dai); uyennnp21411@st.uel.edu.vn (N. N. P. Uyen); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0009-0008-8559-3154 (N. B. Dai); 0009-0001-8604-1991 (N. N. P. Uyen); 0000-0001-8340-1405 (D. V. Thin)

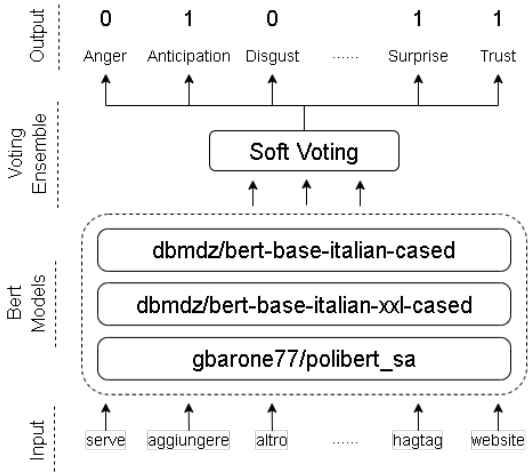

Figure 1: Overview of our ensemble model for the Task A.

sive Language shared task organized at Germeval 2018. Their findings identified a recurrent neural architecture that demonstrates stability and high performance across various languages. They also evaluated the impact of several components commonly used in the task, including the type of embeddings, the incorporation of additional features (text-based or emotion-based), the role of hashtag normalization, and the influence of emojis. [9] introduced FEEL-IT, a novel benchmark corpus of Italian Twitter posts annotated with four basic emotions: anger, fear, joy, and sadness. Another relevant research by [10], paper evaluated BERT's performance in emotion recognition using the EmotionLines dataset from the Friends Television Sitcom series and the EmotionPush dataset from Facebook Messenger chat.

Research on emotion detection has been conducted across various languages and tested on multiple technological models: an attention-based methodology for identifying and categorizing emotions in textual interactions [11], using the Bi-LSTM to classify emotions in textual and emoji utterances [12], detecting text emotions in social networks with a novel ensemble classifier based on the Parzen Tree Estimator (TPE) [13], applying K-NN and NB ML techniques in the detection of emotions [14]. To further contribute to emotion detection experiments and conduct research on identifying emotions in Italian social media audiences, we employ an ensemble approach for Categorical Emotion Detection in Social Media Messages.

3. Approach

Figure 1 shows the overview of our ensemble model for task A. The combined model integrates multiple pre-trained models to leverage their domain-specific expertise and enhance performance. By using the soft voting method, the model aggregates the predictions from each model to generate accurate results.

The integration process involves obtaining independent outputs from each pre-trained model and combining them using the soft voting method. This method considers the confidence scores associated with each prediction, assigning higher weights to more reliable predictions for a robust final decision. The model architecture facilitates seamless communication and coordination between the pre-trained models, enabling efficient information exchange and fusion of predictions. It can be easily adapted and extended to incorporate new models and emerging techniques for continuous improvement.

In this paper, we utilize the pre-trained BERT model [15] for the following reasons: (1) First, BERT is a state-of-the-art model renowned for its exceptional language understanding capabilities. It effectively captures the semantics and nuances of text due to its deep contextualized representations. (2) Second, by leveraging transfer learning, BERT provides a significant advantage. It has been pre-trained on extensive and diverse datasets, allowing us to benefit from the knowledge it has acquired. This saves computational resources and time compared to training a model from scratch. (3) Finally, BERT's adaptability to different domains is a key factor. Its pre-training covers a wide range of domains, making it flexible for various tasks. By fine-tuning BERT with domain-specific data, we can enhance its performance for our specific task.

Table 1
The general information of official datasets in our experiments.

Information	Training set	Test set
Number of samples	6 359	1 000
Number of tokens	93 621	15 713
Number of unique targets	23 028	6 215
The average length	102	102
The maximum length	304	302

Table 2
Performance Rankings of Teams in F1-Score.

Ranking	Team	Run ID	F1-Score
Top 1	extremITA	2	60.28
Top 2	extremITA	1	50.86
Top 4	EmotionHunters	1	48.35
Top 5	App2Check	2	37.41
Ours (Top 3)	**ABCD Team**	**1**	**49.94**

4. Experimental Setup

4.1. Dataset

We only use the official training set [3], which is provided by the organizers, to train our models for task A in the shared task.

Table 1 presents the general statistics of the training and testing set. As depicted in Table 1, it is evident that the training and test sets in this shared-task experiment are balanced in terms of average length. These statistics provide a foundation for understanding the data and designing appropriate models for the given task.

4.2. Pre-processing

Pre-processing steps are essential in classification-type tasks to improve the quality of the data and facilitate effective analysis. Additionally, the provided dataset is collected from social media messages, therefore, it is necessary to design the list of pre-processing steps. Based on the analysis of the training dataset, we design the list of pre-processing steps in our system as follows:

- **Step 1**: We removed the "@USER" and the placeholder "_" carried after it
- **Step 2**: We removed the redundant symbolic expressions and keep only one symbol, for example, "!!!", "???" transformed to "!", "?".
- **Step 3**: We transformed the emoji from the dataset to text because of the large amount of emoji in sentences.
- **Step 4**: We tagged the link, phone number, and hashtag to its related token, for example, "0123456789" transformed to "<phone>".
- **Step 5**: Finally, we removed the extra space symbols from the text.

4.3. Models

After investigating various models, we have concluded that the dbmdz/bert-base-italian-cased, gbarone77/polibert_sa, and dbmdz/bert-base-italian-xxl-cased BERT models are the most suitable choices for this shared task. We have identified several reasons to support our decision: (1) These models were trained on the Italia corpus, which aligns well with the requirements of this shared task. (2) All three models have demonstrated outstanding performance for this task, as indicated in Table 3.

4.4. System setting

We evaluated various models using the HuggingFace Transformer library [16]. Each model was trained with a fixed number of epochs, specifically 5 epochs. The learning rate used for dbmdz/bert-base-italian-cased and gbarone77/polibert_sa was set to 3e-5, while for dbmdz/bert-base-italian-xxl-cased, it was set to 2e-5. The batch size for all pre-trained language models was set to 16. No development data was utilized for model tuning. We employed the AdamW optimizer to optimize our models. Additionally, we added an extra sigmoid layer at the output for each model. To ensure reproducibility, we set a fixed random seed of 42 for training the models.

5. Main results

The official results and the results of the top systems are shown in Table 2. Our best model achieves the Top 3 ranking in the challenge during the final round. Our model achieved a result of 49.94% in terms of F1-score,

Table 3
Comparison of Models and F1-Score Scores.

Models	F1-Score
dbmdz/bert-base-italian-cased	**43.22**
dbmdz/bert-base-italian-xxl-cased	**47.07**
gbarone77/polibert_sa	**47.24**
bert-base-multilingual-cased	39.03
Geotrend/bert-base-it-cased	40.37
mgrella/autonlp-bank-transaction-classification-5521155	38.02
nlptown/bert-base-multilingual-uncased-sentiment	40.20
Babelscape/wikineural-multilingual-ner	38.02

which is lower than the F1 scores of the Top 1 and Top 2 teams, which are +10.34% and +0.92%, respectively.

Table 3 shows the overall results of our submission model and other variants on the test set on the challenge. Overall, it can be seen that the performance of the model is improved when we preprocessed data input, fine-tuned, and ensemble those models. These techniques, mentioned in previous work [17], have consistently shown their effectiveness in enhancing model performance. Preprocessing improves data quality, fine-tuning tailors models to the task, and ensembling combines strengths for better results. These findings reinforce the value of these techniques for improving model performance in similar domains or tasks. Our ensemble methods have significantly improved the F1-Score, increasing it from 2.7% to 6.42%. This demonstrates the effectiveness of ensembling in enhancing model performance on the task at hand.

On the other hand, Table 3 shows that the models dbmdz/bert-base-italian-cased, dbmdz/bert-base-italian-xxl-cased, and gbarone77/polibert_sa outperformed the other models. Among them, the gbarone77/polibert_sa model achieved the best performance due to its utilization of Italian language data and its specific training for sentiment analysis, which aligns well with the task requirements. The superior performance of these three models can be attributed to their training on a large amount of data specific to the Italian language.

In contrast, the models bert-base-multilingual-cased, Babelscape/wikineural-multilingual-ner, and nlptown/bert-base-multilingual-uncased-sentiment were trained on multilingual language data, which resulted in comparatively lower performance when applied to Italian-specific tasks. To address this issue, the Geotrend/bert-base-it-cased model was introduced as an enhancement to the bert-base-multilingual-cased model by training it on Italian language data. This targeted training on Italian data resulted in a performance improvement of 1.34% compared to the original multilingual model. However, despite this improvement, the Geotrend/bert-base-it-cased model still exhibits lower performance than the top three models mentioned earlier. One possible reason for this is that the amount of training data used for the Geotrend/bert-base-it-cased model was relatively smaller compared to the extensive data used for the top-performing models. The mgrella/autonlp-bank-transaction-classification-5521155 models, while trained on Italian language data, were specifically tailored for the ”bank transaction” field. As a result, their performance is lower for sentiment analysis tasks, which is the focus of our study.

6. Conclusion

In this paper, we present our submission system for the EMit at EVALITA 2023: The Categorial Emotion Detection Task, where our approach achieved a Top 3 ranking. Instead of relying on a single model, we adopt an ensemble method approach using the pre-trained BERT model [15] to tackle the task. Through extensive experimentation and analysis, we have found this approach to be highly effective and believe it can be applied to other domains for sentiment analysis as well. While we acknowledge that our familiarity with the Italian language was limited, we recognize the importance of effective text pre-processing techniques in enhancing performance on the test set. Although we were unable to leverage these techniques to their full potential, we firmly believe that incorporating appropriate text pre-processing techniques can further improve the performance of our model.

References

[1] F. A. Acheampong, C. Wenyu, H. Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities, Engineering Reports 2 (2020) e12189.

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th

evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] L. A. Camras, R. Plutchik, H. Kellerman, Emotion: Theory, research, and experience. vol. 1. theories of emotion, American Journal of Psychology 94 (1981) 370.

[5] V. Barriere, S. Tafreshi, J. Sedoc, S. Alqahtani, Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories, in: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 2022, pp. 214–227.

[6] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, Semeval-2019 task 3: Emocontext contextual emotion detection in text, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 39–48.

[7] E. Agirre, L. Márquez, R. Wicentowski, Proceedings of the fourth international workshop on semantic evaluations (semeval-2007), in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 2007.

[8] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata, A multilingual evaluation for online hate speech detection, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–22.

[9] F. Bianchi, D. Nozza, D. Hovy, et al., Feel-it: Emotion and sentiment classification for the italian language, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2021.

[10] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu, Y.-S. Chen, Emotionx-idea: Emotion bert–an affectional model for conversation, arXiv preprint arXiv:1908.06264 (2019).

[11] W. Ragheb, J. Azé, S. Bringay, M. Servajean, Attention-based modeling for emotion detection and classification in textual conversations, arXiv preprint arXiv:1906.07020 (2019).

[12] L. Ma, L. Zhang, W. Ye, W. Hu, Pkuse at semeval-2019 task 3: emotion detection with emotion-oriented neural attention network, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 287–291.

[13] F. Ghanbari-Adivi, M. Mosleh, Text emotion detection in social networks using a novel ensemble classifier based on parzen tree estimator (tpe), Neural Computing and Applications 31 (2019) 8971–8983.

[14] M. Suhasini, B. Srinivasu, Emotion detection framework for twitter data using supervised classifiers, in: Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, Springer, 2020, pp. 565–576.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[17] Y. Xu, X. Qiu, L. Zhou, X. Huang, Improving bert fine-tuning via self-ensemble and self-distillation, arXiv preprint arXiv:2002.10345 (2020).