2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) | 978-1-6654-0505-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCWAMTIP51612.2020.9317379

COMPARATIVE ANALYSES OF BERT, ROBERTA, DISTILBERT, AND XLNET FOR TEXT-BASED EMOTION RECOGNITION

ACHEAMPONG FRANCISCA ADOMA[1], NUNOO-MENSAH HENRY[2], WENYU CHEN[1]

[1]Computational Intelligence Lab, School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu, China
[2]Connected Devices Lab, Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
E-MAIL: francaadoma@gmail.com, hnunoo-mensah@knust.edu.gh, cwy@uestc.edu.cn

Abstract:

Transformers' feat is attributed to its better language understanding abilities to achieve state-of-the-art results in medicine, education, and other major NLP tasks. This paper analyzes the efficacy of BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from texts. The paper undertakes this by analyzing each candidate model's output compared with the remaining candidate models. The implemented models are fine-tuned on the ISEAR data to distinguish emotions into anger, disgust, sadness, fear, joy, shame, and guilt. Using the same hyperparameters, the recorded model accuracies in decreasing order are 0.7431, 0.7299, 0.7009, 0.6693 for RoBERTa, XLNet, BERT, and DistilBERT, respectively.

Keywords:

Natural Language Processing; Transfer Learning; Emotion Detection; BERT; DistilBERT; RoBERTa; XLNet

1. Introduction

The advent of social media and its gains have led to exponential increases in social media users. The number of active social media users, notwithstanding the global effect of the COVID-19 pandemic, has increased by over a hundred million representing a growth of more than 10 percent from the previous year-2019. Harnessing these social media users' profiles for polarity assignment serve as a bedrock for sentiment analysis (SA). However, the coarse granular attribute of SA in representing user profiling makes it ineffective. A subtler granular method has been shown to portray users' detailed view and thus more suited for user profiling. The detection or recognition of emotions happens to be an extraction of finer-grained user sentiments. Text-based emotion recognition is a sub-branch of emotion recognition (ER) that focuses on extracting fine-grained emotions from texts. Though research in the field is fast gaining traction, the challenge of identifying appropriate embedding techniques for extracting the relationship between long term dependent texts and parallel processing of text sequence has for long inhibited the pace of attaining state-of-the-art results. The proposal of transformers [1] and the transformer language model [2] provided a breakthrough in solving these limitations.

The Bidirectional Encoder Representations from Transformers (BERT) pre-trained model [3], using the vanilla transformer language model [2] released by Google in 2018 as a substructure, has been described as the rediscovery to the Natural Language Processing (NLP) pipeline due to the improved level of language understanding it offers [4]. However, the BERT model suffers from fixed input length size limitations, wordpiece embedding problems, and computational complexities [5]. The Generalized Auto-regression Pre-training for Language Understanding (XLNet), Robustly optimized BERT pre-training Approach (RoBERTa), and DistilBERT pre-trained models were necessary proposals for mitigating different underpinning problems associated with BERT. While BERT and its variants are being used actively in question answering (QA), natural language inference (NLI), text summarization (TS), and other NLP tasks to solve human-related and environmental problems, little has been seen in recognizing emotions from texts.

This paper aims to shed light on the efficacy of the BERT, RoBERTa, DistilBERT, and XLNet models in recognizing emotions from the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset [6]. The experiments and results obtained from the four models are comparatively discussed concerning their accuracy, precision, and recall on the ISEAR dataset's emotion classes. It is worth stating that no work has comparatively analyzed the efficacy of BERT, RoBERTa, DistilBERT, and XLNet on the ISEAR dataset to the best of our knowledge.

The organization of the paper is as follows; Section 2

discusses related works; Section 3 highlights the emotion detection pipeline and model implementation. The model experiments are outlined in Section 4. The results obtained are presented and discussed in Section 5. In Section 6, the conclusion and future works are highlighted

2. Related Work

Alotaibi [7] proposed a supervised logistic regression approach to detect emotions from texts. The data they obtained from ISEAR was divided for training and testing. In the training process, sentences carrying emotions were fed into their logistic regression model and their emotion labels. Only the unseen emotion labeled sentences were passed through the trained classifier for prediction in their testing process. They evaluated their model's performance using precision, recall, and F1-score. They reported an F1-score of 0.76, 0.64, 0.73, 0.62, and 0.57 for joy, fear, sadness, shame, and guilt emotion classes. They hinted that a deep learning model could perform better on engineering the features for classification.

In response to Alotaibi's recommendation, Polignano *et al*. [8] designed a model that implemented Bi-LSTM, Self-Attention, and Convolutional Neural Networks (CNN) together. They focused on the extraction of word embeddings as a fundamental feature to improving the recognition of emotions from texts. Thus, they compared the Google word embedding performance, the GloVe embedding, and the Fast-Text embedding using the Bi-LSTM, Convolutional Neural Network (CNN) ensemble, and a Self Attention model. They evaluated their model on the ISEAR dataset, the SemEval-2018 Task 1 dataset, and the SemEval-2019 Task 3 dataset. They reported improved performance on all the datasets with the FastText embedding. They, therefore, recommended that a robust pre-trained word embedding could enhance the model performance.

Kazameini *et al*. [9] extracted contextualized word embeddings from text data using the BERT pre-trained model and used the bagged-SVM classifier to predict the authors' personality traits automatically. The input to their model was essays. The essays were divided into sub-documents, preprocessed, and fed into a BERT base model. Feature vectors for the document were extracted and fed into ten SVM classifiers to produce a prediction. The final prediction was obtained by majority voting. They obtained an increased performance of 1.04% in comparison with baseline methods.

3. Methodology

This section elucidates the various blocks of the machine learning pipeline used for the detection task, as illustrated in Figure 1. The dataset was acquired, preprocessed, and fed to the various candidate models. The candidate models were all fine-tuned on the data before final predictions were carried out.

3.1. Data Acquisition

The ISEAR dataset [10] is a publicly available dataset constructed through cross-culture questionnaire studies in 37 countries. It contains 7666 sentences classified into seven distinct emotion labels: joy, anger, sadness, shame, guilt, surprise, and fear. Its balanced class feature makes it ideal for making generalized predictive inferences; hence, its use for this study. Table 1 presents the data distribution of the ISEAR dataset.

Fig.1 The Machine Learning pipeline for the detection task

Table 1 Data Distribution of the ISEAR Dataset

Emotion Labels	Quantity
Anger	1096
Disgust	1096
Sadness	1096
Shame	1096
Fear	1095
Joy	1094
Guilt	1093
Total	7666

3.2. Data Preprocessing

The obtained data contained several columns; the columns containing individuals' responses and the emotion labels were the columns of interest to this work. These two columns were, therefore, extracted for further processing. It was also realized that some columns contained emotion labels but no textual responses. These were again removed and the total amount of data reduced from 7666 to 7589. Special characters, double spacing, tags, and other irregular expressions found in the remaining data were removed. They were noticed to affect recognition performance negatively. Stop words were further removed, and the seven emotion labels encoded to a numerical scale (i.e., 0, 1, 2, 3, 4, 5, 6). All the above preprocessing schemes were carried out before the final data samples were split into two groups, i.e., 80% for training and 20% for testing purposes. Investigations revealed that the longest sentence in the dataset had 178

words; this helped set the maximum sentence length of 200 for the tokenizer. The decision was to ensure that no sentence was truncated, and all sentences had the same length. As a result, all sentences were padded to attain a length of 200. The training and test samples were tokenized to generate the tokens, which were then fed to the fine-tuning candidate models.

3.3. Model Fine-Tuning and Classification

The generated tokens were converted to vector representations and fed to the pre-trained models during the fine-tuning process. Thus, the models were trained on the input vector transformations and their outputs generated. The output was then evaluated using the designated test data, and results were obtained. Emotions were then classified into joy, sadness, fear, anger, guilt, disgust, and shame for each of the pre-trained models in the emotion classification process.

4. Model Experiments

Experiments were carried out using Google Colab's GPU hardware accelerator platform. The extracted features for the training set were fed to the input of the tuned models. The batch size and learning rate were set to 16 and $4*10^{-5}$, respectively. The models were optimized using the Adam optimizer, and the loss parameter was set to *sparse_categorical_crossentropy*. The models were trained for ten epochs.

The BERT-base-uncased model consisting of twelve layered transformer blocks with each block containing twelve head self-attention layers and 768 hidden layers resulting in approximately 110 million parameters, was used. A single sentence was fed into the model at a time. The input sentences were split into tokens and mapped to their indexes using the BERT tokenizer library, indicated as *input_ids*. The [CLS] (classification token) and [SEP] (separate segment token) were appended at the beginning and end of every sentence, respectively. An input attention mask of fixed length with 0 indicating padded tokens, and 1, indicating unpadded tokens was applied. Each of the transformers indicated received a list of token embeddings and produced a feature vector of the same length at the output. The output of [CLS] for the 12*th* transformer layer containing vector transformations of prediction probabilities was used as aggregated sequence representation from which classifications were made.

The RoBERTa-base model was made up of twelve transformer layers with 768-hidden layers, twelve attention heads, and 125 million parameters used in the experiment. The RoBERTa tokenizer was used to encode the input texts into tokens and designated them as the *input_ids.* These ids were padded to a fixed length to avoid variations per row. Features were then extracted from these tokens from which sentence pair classification was made.

The DistilBERT-uncased model was used in this experiment. It contained six transformer layers, 768-hidden layers, and twelve attention heads. After tokenizing the input texts and converting the tokens into *input_ids*, they were padded and fed into the DistilBERT model for the multiclassification task.

The XLNet-base-cased model was made up of twelve transformer layers with 768 hidden layers. Twelve attention head layers were used. The XLNet tokenizer was used to split the sequences into tokens. The tokens were then padded, and classifications were made.

5. Results and Discussion

This section discusses in detail the results obtained for the individual models. The time taken for each model to run to completion, model accuracy, performance in detecting the seven emotion labels, and the revealed difficulties are elucidated.

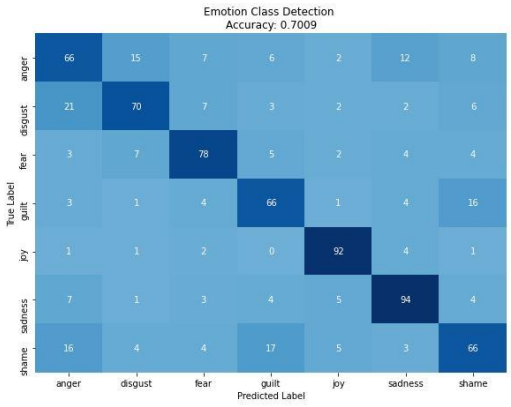

Fig.2 Confusion Matrix for BERT

The confusion matrices attained after the experiments are presented as Fig.2, Fig.3, Fig.4, and Fig.5 for BERT, RoBERTa, DistilBERT, and XLNet. The classification report presented in Table 2 indicates the various precision, recalls, and F1-scores for the individual emotion classes after testing the various candidate models on the test data.

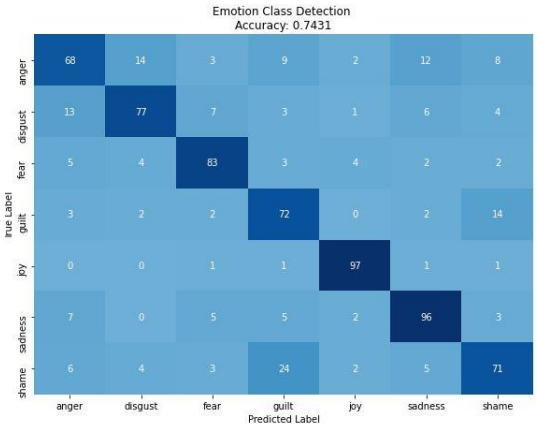

Fig.3 Confusion Matrix for RoBERTa

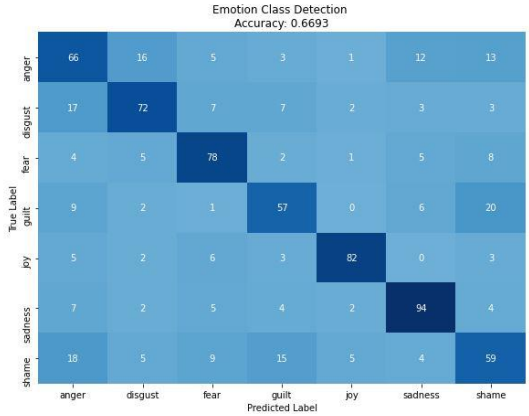

Fig.4 Confusion Matrix for DistilBERT

Fig.5 Confusion Matrix for XLNET

As shown in the confusion matrices, the order of emotion recognition accuracy in decreasing order are RoBERTa, XLNet, BERT, and DistilBERT, with a recognition accuracy of 0.7431, 0.7299, 0.7009, and 0.6693, respectively. Since all models were capable of recognizing emotions from the data, we posit that not only are these models efficient in other NLP tasks but are also efficient in recognizing emotions from texts. Secondly, from the results, we posit that under the same conditions, the RoBERTa pre-trained model outperforms the other pre-trained models under investigation in this work. Observations made during this work showed that even though the DistilBERT yielded the least accurate results, it was the fastest computationally. The XLNet model, on the other hand, was computationally the slowest. RoBERTa slightly outperformed the BERT model in speed. The order of computational resource demand in decreasing order can be given as XLNet, BERT, RoBERTa, DistilBERT. It is worth mentioning that the pre-trained models under discussion in this paper were all fine-tuned on the ISEAR dataset to obtain results. It makes us believe that the RoBERTa pre-trained model can be highly effective in recognizing emotions from texts when thoroughly optimized for the purpose. The classification report presented in Table 2 clearly shows that RoBERTa responded well to most emotion classes. The report further buttresses that RoBERTa is an optimal candidate for detecting emotions on the ISEAR dataset. XLNet also demonstrated some level of efficacy in some aspects of the seven classes. DistilBERT and BERT, on the other hand, could not achieve any high scores in any of the seven emotion classes for the precision, recall, and F1-score. The lower computational complexity of RoBERTa over XLNet also reinforces the recommendation of RoBERTa for emotion recognition in text.

6. Conclusion and Future Work

In conclusion, the paper set out to assess the efficacy of the BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from the ISEAR dataset. The models proved efficient in detecting emotions from the text, with RoBERTa attaining the highest recognition accuracy. The precision, recall, and F1-scores further proved the efficacy of RoBERTa over the other candidate models in recognizing emotions on the ISEAR dataset. In the future, an ensemble of the model would be considered to improve recognition performance. Also, strategies to inculcate commonsense knowledge into the model would be considered to improve its generalization ability.

Table 2 Comparison of Precision, Recall And F1-scores of BERT, RoBERTa, DistilBERT, and XLNET on the ISEAR Dataset

Models	Anger			Disgust			Fear			Guilt			Joy			Sadness			Shame		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT	0.56	0.57	0.57	0.71	0.63	0.67	0.74	0.76	0.75	0.65	0.69	0.67	0.84	0.91	0.88	0.76	0.8	0.78	0.63	0.57	0.6
RoBERTa	**0.67**	**0.59**	**0.62**	**0.76**	0.69	**0.73**	**0.8**	**0.81**	**0.8**	0.62	**0.76**	0.68	0.9	**0.96**	**0.93**	**0.77**	**0.81**	**0.79**	**0.69**	**0.62**	**0.65**
DistilBERT	0.52	0.57	0.55	0.69	0.65	0.67	0.7	0.76	0.73	0.63	0.6	0.61	0.88	0.81	0.85	0.76	0.8	0.78	0.54	0.51	0.52
XLNET	0.59	0.57	0.58	0.69	**0.73**	0.71	0.76	**0.81**	0.78	**0.7**	0.72	**0.71**	**0.91**	0.93	0.92	0.76	**0.81**	**0.79**	**0.69**	0.57	0.63

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, pp. 5998–6008, 2017.

[2] R. Al-Rfou, D. Choe, N. Constant, M. Guo, and L. Jones, "Character-level language modeling with deeper self-attention," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3159–3166, 2019.

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

[4] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical NLP pipeline," arXiv preprint arXiv:1905.05950, 2019.

[5] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," Engineering Reports, vol. 2, no. 6, pp. 1–24, 2020.

[6] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning.," Journal of personality and social psychology, vol. 66, no. 2, p. 310, 1994.

[7] F. M. Alotaibi, "Classifying text-based emotions using logistic regression," VAWKUM Transactions on Computer Sciences, vol. 16, no. 2, pp. 31–37, 2019.

[8] M. Polignano, P. Basile, M. de Gemmis, and G. Semeraro, "A comparison of word-embeddings in emotion detection from text using BiLSTM, CNN and self-attention," in Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 63–68, 2019.

[9] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged SVM over BERT word embedding ensembles," 2020.

[10] K. Scherer and H. Wallbott, "International survey on emotion antecedents and reactions (ISEAR) (1990)," 2017