Emotion and sentiment analysis of tweets using BERT

Andrea Chiorrini
Università Politecnica delle Marche
Ancona, Italy
a.chiorrini@pm.univpm.it

Claudia Diamantini
Università Politecnica delle Marche
Ancona, Italy
c.diamantini@univpm.it

Alex Mircoli
Università Politecnica delle Marche
Ancona, Italy
a.mircoli@univpm.it

Domenico Potena
Università Politecnica delle Marche
Ancona, Italy
d.potena@univpm.it

ABSTRACT

The huge diffusion of social networks has made available an unprecedented amount of publicly-available user-generated data, which may be analyzed in order to determine people's opinions and emotions. In this paper we investigate the use of *Bidirectional Encoder Representations from Transformers* (BERT) models for both sentiment analysis and emotion recognition of Twitter data. We define two separate classifiers for the two tasks and we evaluate the performance of the obtained models on real-world tweet datasets. Experiments show that the models achieve an accuracy of 0.92 and 0.90 on, respectively, sentiment analysis and emotion recognition.

KEYWORDS

sentiment analysis, emotion recognition, BERT, deep learning, tweet sentiment analysis

1 INTRODUCTION

In the last decade, the great diffusion of social networks, personal blogs and review sites has made available a huge amount of publicly-available user-generated content. Such data is considered authentic, as in the above contexts people usually feel free to express their thoughts. Therefore, the analysis of this user-generated content provides valuable information about the opinion of users about a large variety of topics and products, allowing firms to address typical marketing problems as, for instance, the evaluation of custo-mer satisfaction or the measurement of the impact of a new marke-ting campaign on brand perception. Moreover, the analysis of customers' opinions about a certain product can be a driver for open innovation, as it helps business owners to find out possible issues and can possibly suggest new interesting features. For this reason, in the last years many researchers (e.g., [10], [13], [22]) focused on techniques for the automatic analysis of writer's opinions and emotions, generally referred to as, respectively, sentiment analysis and emotion analysis.

Sentiment analysis is the process of automatic extraction of writer's opinions and their characterization in terms of polarity: positive, negative and neutral. On the other hand, *emotion analysis* has the goal of recognizing the emotion expressed in the text. This task is usually more difficult than sentiment analysis given the greater variety of classes and the more subtle differences between them. Although in literature such tasks have been addressed through both lexicon-based [7] and learning-based approaches [11], the latter have shown better performance in terms of classification. For this reason, recent works have focused on large deep learning models [32] [3]. In order to be accurately trained, such models require large corpora of labelled data, which are usually scarce and expensive to build [19].

As a consequence, pre-trained models that only need a fine-tuning phase with a smaller dataset have been widely used. In particular, many neural networks composed of a task-agnostic pre-trained word embedding layer (e.g., GloVe [21]) and a task-specific neural architecture have been proposed but the improvement of these models measured by accuracy or F1 score has reached a bottleneck [14]. Anyway, recent architectures based on Trasformer [28] have shown further room for improvement.

In the present paper, we investigate the enhancement in terms of classification accuracy of Bidirectional Encoder Representations from Transformers (BERT) [6], one of the most popular pre-trained language models based on Transformer, on both the tasks of senti-ment analysis and emotion recognition. To this purpose, we propose two BERT-based architectures for text classification and we fine-tune them in order to evaluate their performance. In the rest of the work we focus on data collected from microblogging platforms and, in particular, from Twitter. The main reasons of this choice are the wide availability of tweets (as opposed to, for instance, Facebook posts, due to different data policies) and the fact that such data are usually challenging to analyze due to the presence of slang, typos and abbreviations (e.g., "btw" for "by the way") and hence represent a good benchmark for text classifiers.

The rest of the paper is structured as follows: the next section presents some relevant related work on sentiment analysis and emotion recognition. The architecture of the models used for both tasks is proposed in Section 3, while Section 4 reports the results of the experimental evaluation of the models on real-world datasets of tweets. Finally, Section 5 draws conclusions and discusses future work.

2 RELATED WORK

2.1 Sentiment analysis

With the ever increasing amount of user generated content available online, the field of automatic sentiment analysis has become a topic of increasing research interest. As in many other field, deep learning techniques are being widely used for sentiment analysis, as demonstrated by the presence of various surveys regarding the subject over the last years [12, 16, 24]. The first complex task that sentiment analysis must tackle is the vector representation of words, which is typically performed thorough word embeddings: a technique which transforms the words in a vocabulary into vectors of continuous real numbers.

The most commonly used word embeddings are Word2Vec [1] and Global Vector (GloVe) [21].

Word2Vec is a neural network that learns the word embeddings from text, and contains both the *continuous bag-of-words* (CBoW) model [17] and the *Skip-gram* model [18]. Given a set of context words (e.g. "the girl is _ an apple," where "_" denotes the target word) the CBoW predicts the target word (e.g., "eating"), conversely the *Skip-gram* model, given the target word, predicts the context words.

GloVe is trained on the non-zero entries of a global word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus.

Subsequent works have focused on further refining the idea of embedding. In [26, 27] the authors proposed models that learn sentiment-specific word embeddings (SSWE). In these embeddings, the senti-ment information is embedded in the learned word vectors as well as the semantic. The authors of [29] designed and trained a neural network that learns a sentiment-related embedding representation through the integration of sentiment supervision both at document level and at word level. A further refinement of semantics-oriented word vectors has been proposed in [33] which integrates the word embedding model with standard matrix factorization through a projection level.

2.2 Emotion analysis

Though there is not universal agreement over which are the primary emotions of human being, the scientific community is giving ever increasing attention to the specific problem of emotion recognition.

In [30] a bilingual attention network model has been proposed for code-switched emotion prediction. In particular, a document level representation of each post has been built using a Long Short-Term Memory (LSTM) model, while the informative words from the context have been captured through the attention mechanism.

In [1], the authors used distant supervision to automatically build a dataset for emotion detection and trained a fine-grained emotion detection system using Gated Recurrent Unit (GRU) network.

Another approach [8] focused on learning a better representations of emotional contexts by using millions of emoji occurrences in social media to pre-train neural models.

Even more recently a Bidirectional Encoder Representations from Transformers (BERT) model has been proposed in [5]. This pre-trained BERT model has provided, without any substantial task-specific architecture modifications, state of the art performances over various NLP tasks.

In the sentiment analysis field, BERT has been mostly used in aspect-based sentiment analysis such as in [15, 25, 31], while few authors focused on emotion analysis.

In [2], the authors performed a comparative analysis of various pre-trained transformer model, including BERT, for the text emotion recognition problem. However, our work differs from the previous as we evaluate the performance of the emotion classification when applied to social content, which is usually more challenging.

[1] https://code.google.com/archive/p/word2vec/ https://code.google.com/archive/p/word2vec/

3 MODEL

In this section we describe the proposed model for the tasks of emotion and sentiment analysis. The model is built by fine-tuning BERT on specific datasets of tweets developed for such tasks. Since tweets usually contain words that are irrelevant for text classification, a text preprocessing phase is needed in order to remove:

- mentions: users often cite other Twitter usernames in their tweets through the character '@' in order to direct their messages;
- urls: urls are very common in tweets, both for media (i.e., pictures and videos) and links to other webpages;
- retweets: users often resend tweets they consider relevant to their followers. Retweets are usually marked with the prefix "RT" and hence are easily identifiable.

After the preprocessing phase, data can be used as input to train task-specific BERT-based models. The architecture of a generic BERT model consists of a series of bi-directional multi-layer encoder-based Transformers. Nowadays, several pre-trained BERT models are available. Table 1 shows the main BERT models as a function of the number of layers L (i.e. the number of encoders) and the number of hidden units H. Smaller models are intended for environments with limited computational resources, since bigger models have a large number of trainable parameters: a model of average size like BERT-Base has approximately 110 million trainable parameters, while BERT-Large has more than 340 million parameters.

Specifically, the reference model used in this work is the BERT-Base, both in the *uncased* and the *cased* version. The *uncased* version implies that text is converted to lowercase before the word tokeniza-tion process (ex. *Michael Jackson* becomes *michael jackson*) and accents are ignored. The architecture of the BERT-Base model consists of 12 encoders, each composed of 8 layers: 4 multi-head self-attention layers and 4 feed forward layers. We extended such model by adding a fully connected layer and a softmax layer for classification, as reported in Figure 1. The architecture is common to both the sentiment and emotion classifiers: the only difference between the two models is represented by the last softmax layer, in which the number of neurons is equal to the number of classes (i.e., 3 for sentiment analysis and 4 for emotion recognition).

4 EXPERIMENTS

In this section we present some experimental results aimed at evaluating the performance of the proposed BERT-based models. The results for the emotion analysis and the sentiment analysis tasks are discussed separately.

4.1 Experimental setting

The proposed models have been evaluated through two different datasets, namely Go et al. [9] for sentiment analysis and the Tweet Emotion Intensity dataset [20] for emotion recognition. The same criteria have been used for the experiments: in particular, each dataset has been split through a stratified sampling into train (80%), dev (10%) and test (10%) set. Moreover, we tested both the uncased and the case version of BERT. Experiments have been performed on a laptop with 2x2.2GHz CPU, 8GB RAM and a Nvidia Geforce 740M graphics card: execution times reported in the following subsections refer to such hardware configuration.

We evaluated the model by means of two metrics: classification accuracy and F_1 score. Let x_{ij} be the number of data belonging

Table 1: BERT pre-trained models

BERT Models	H=128	H=256	H=512	H=768	H=1024
L=2	BERT-Tiny	–	–	–	–
L=4	–	BERT-Mini	BERT-Small	–	–
L=8	–	–	BERT-Medium	–	–
L=12	–	–	–	BERT-Base	–
L=24	–	–	–	–	BERT-Large

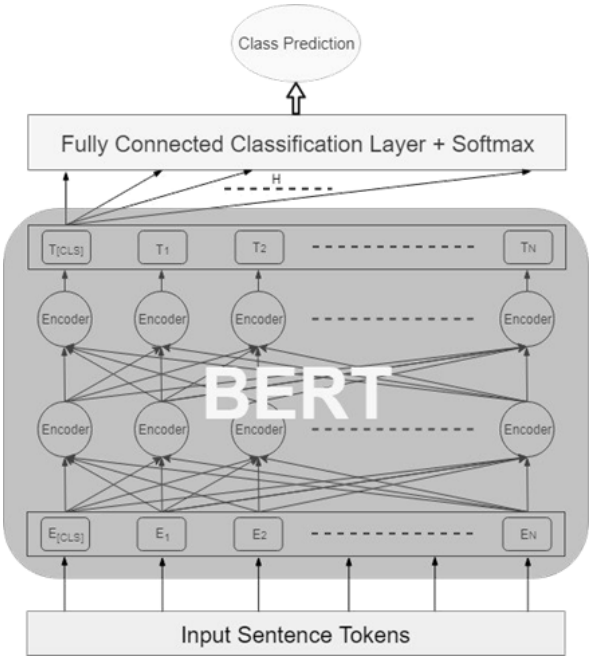

Figure 1: The architecture of the proposed classification model.

to j-th class which have been classified as i-th class. Let C be the number of classes and N be the total number of data. The accuracy achieved by a classifier is computed as:

$$accuracy = \frac{1}{N}\sum_{i=1}^{C} x_{ii} \tag{1}$$

Precision and recall of i-th class are determined as follows:

$$precision_i = \frac{x_{ii}}{\sum_{j=1}^{C} x_{ij}} \tag{2}$$

$$recall_i = \frac{x_{ii}}{\sum_{j=1}^{C} x_{ji}} \tag{3}$$

F_1 score of i-th class is equal to:

$$F_{1i} = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i} \tag{4}$$

Therefore, the F_1 score achieved by a classification model is defined as the average of F_{1i}:

$$F_1 = \frac{1}{C}\sum_{i=1}^{C} F_{1i} \tag{5}$$

4.2 Emotion analysis

In order to evaluate the performance of the proposed architecture on the emotion analysis task we considered the Tweet Emotion Intensity dataset, which consists of 6755 tweets labelled with respect to the following four emotions: anger, fear, happiness, sadness. Since samples in the original dataset were not equally distributed among classes, we balanced the training set by applying the undersampling technique. In particular, we randomly chose 1300 tweets from each class. We also filtered out 974 meaningless tweets, e.g. tweets only containing non-ASCII characters or very short tweets. As a result, we obtained a training+dev set of 5200 equally distributed tweets and a test set of 581 tweets with the class distribution reported in Figure 2.

Figure 2: Tweet Emotion Intensity dataset: class distribution of the test set.

Each occurrence in the dataset is associated not only with a label emotion but also with a parameter called *intensity*, that represents the intensity of the emotion. Specifically, this parameter is a value between 0 and 1 that indicates the degree of intensity with which the author of the tweet felt that emotion. In Figure 3 it is shown the histogram of the occurrences for different lengths (in characters) of tweets; the length of 452 characters represents the upper-bound of lengths present in the dataset and is actually an isolated case given by a single tweet, while the average length ranges between 9 and 50 characters. Since the model requires defining a maximum length for the sequence of input characters (the *max_seq_length* parameter), analyzing the histogram we decided to set max_seq_length=95.

After a preliminary phase of hyperparameter tuning aimed at determining the best values for the hyperparameters of our model, we trained our classifier by using the values reported in Table 2.

Training required about 5'30"/epoch, while the prediction of the emotion related to a tweet in test set took approximately 0.4 seconds. We trained the model for a variable number of epochs, ranging from 1 to 6. The reason for choosing such a small number

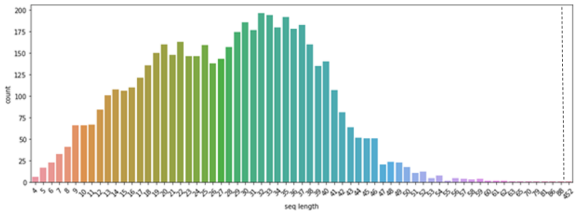

Figure 3: Histogram of the occurrences for different lengths (in characters) of tweets. Source: saifmohammad.com

Table 2: Optimal hyperparameters for the emotion recognition task.

Hyperparameter	Value
learning_rate	2e-5
train_batch_size	8
eval_batch_size	8
max_seq_length	95
adam_epsilon	1e-8

of epochs is the fact that pre-trained models usually need a short fine-tuning phase in order not to overfit the data.

We evaluated both the uncased and the cased version of BERT-Base by using the same hyper-parameter configuration. Training and validation loss in function of the number of epochs are reported, respectively, in Figure 4 for the uncased version and in Figure 5 for the cased one. It has to be noticed that,in line with expectations, in both cases the optimal training is reached in only 2 epochs. In fact, starting from the third epoch, even if the training error diminishes, the validation loss begins to increase: a phenomenon which is usually correlated with overfitting.

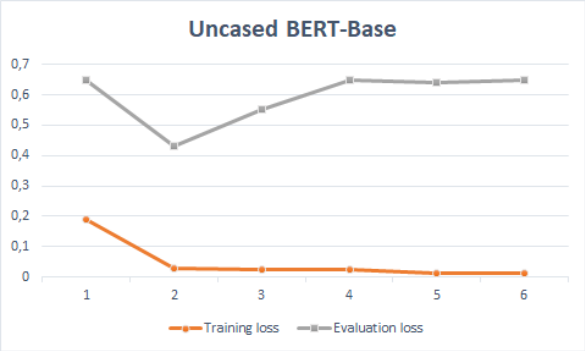

Figure 4: Uncased BERT for emotion recognition: Training and validation loss over epochs

The confusion matrices for the uncased and cased version are respectively shown in Table 3 and 4. The uncased BERT has accuracy = 0.89 and F_1 = 0.89, while the cased version has accuracy = 0.90 and F_1 =0.91: hence, the cased version shows slightly higher performance. Table 4 shows that the *happiness* class has the highest precision, while the highest recall is reached by the *sadness* class, which has also the best average metrics. Happy tweets seems to be the most difficult to be detected, since the *happiness* class has the lowest recall (0.85). Anyway, the difference with other classes is rather small. Generally speaking, the performance of the classifier seems promising, especially considering

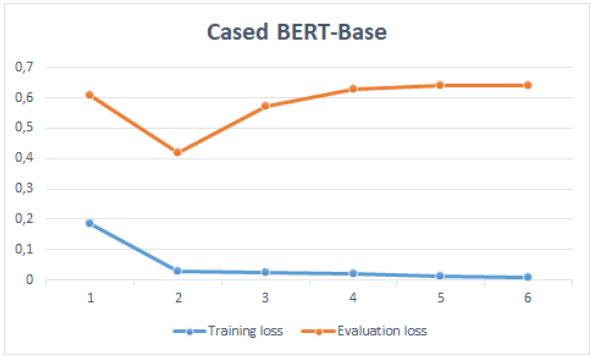

Figure 5: Cased BERT for emotion recognition: Training and validation loss over epochs

that the classification of tweets is often very challenging (see Section 1).

4.3 Sentiment analysis

The performance of the sentiment analysis classifier has been evaluated on the dataset proposed by Go et al. [9]. Such dataset is composed of a training set of 1,600,000 tweets annotated through distant supervision (by considering emoticons in text) and a test set of 430 manually-annotated tweets. We only considered the latter, since that dataset has been annotated by humans and hence it is more reliable. Each tweet has been annotated with respect to its polarity (i.e., positive, negative or neutral); the class distribution is reported in Table 5. The dataset is slightly imbalanced and the *neutral* is the minority class. However, it does not represent a problem since we are more interested in detecting emotion-bearing tweets.

Coherently with the approach proposed in Section 3, we preprocessed the dataset in order to remove noisy words like links, hashtags, retweets and mentions. Similarly to Section 4.2, we analyzed the length (in terms of characters) of each tweet. In this case, we set max_seq_length=82, since tweets were shorter than those in the Tweet Emotion Intensity Dataset on average. We performed a phase of hyperparameter tuning through a grid search and we determined the best configuration (see Table 6).

Due to the smaller size of the dataset, the time required by training was smaller: in particular, it took about 1'15"/epoch. We trained the model for a variable number of epochs - from 1 to 6 - and we noticed a behavior similar to the emotion recognition task. As it can be observed in Figures 6 and 7, the validation loss reached its minimum value after a single epoch, then started to increase, probably due to overfitting. A possible explaination of such phenomenon is that the dataset was small if compared to the number of parameters of the model and hence the classifier rapidly overfitted. Anyway, further investigations with larger datasets are required.

The confusion matrices for the uncased and cased version are respectively reported in Table 7 and 8. In this case, the uncased and cased BERT have similar performance, both in terms of accuracy (0.92) and F_1 (0.92), hence the cased version provides no improvement over the cased one.

It can be observed that the largest part of misclassified tweets is composed by emotion-bearing text that are, instead, classified as neutral. This phenomenon can be justified by considering that there are sentences that are weakly polarized (e.g., for the lack of strongly polarized adjectives, such as "wonderful" or "ugly") and

Table 3: Uncased BERT: confusion matrix for the emotion recognition task

	Actual Happiness	Actual Anger	Actual Sadness	Actual Fear
Predicted Happiness	131	3	0	10
Predicted Anger	10	127	3	10
Predicted Sadness	6	3	122	0
Predicted Fear	11	5	2	138
Recall	**0.83**	**0.92**	**0.96**	**0.87**
Precision	**0.91**	**0.85**	**0.93**	**0.88**

Table 4: Cased BERT: confusion matrix for the emotion recognition task

	Actual Happiness	Actual Anger	Actual Sadness	Actual Fear
Predicted Happiness	135	2	0	6
Predicted Anger	7	121	3	4
Predicted Sadness	9	2	122	1
Predicted Fear	7	2	2	147
Recall	**0.85**	**0.88**	**0.96**	**0.93**
Precision	**0.94**	**0.90**	**0.91**	**0.93**

Table 5: Class distribution of the test set proposed by Go et al.

Class	Occurrences
Positive	157
Neutral	117
Negative	156
Total	430

Table 6: Optimal hyperparameters for the sentiment analysis task.

Hyperparameter	Value
learning_rate	1e-5
train_batch_size	8
eval_batch_size	8
max_seq_length	82
adam_epsilon	1e-7

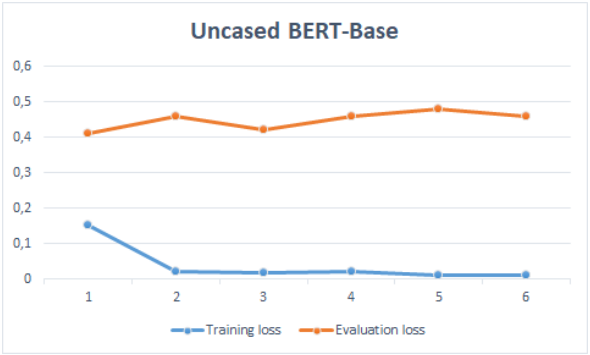

Figure 6: Uncased BERT for sentiment analysis: Training and validation loss over epochs

sentences containing slang, which are really difficult to properly classify. It is remarkable that polarity inversions, i.e. positive sentences classified as negatives and vice versa, are quite rare (1.8%). In fact, polarity inversions are usually more costly in terms of classification error as they correspond to completely misrepresent the user's opinion. Such performance can be compared to those presented in [9], where traditional machine learning algorithms are applied to the same dataset. It can be noticed that the best model proposed in [9], i.e., SVM, has an accuracy of 0.82. Therefore, the use of BERT leads to a remarkable 0.10 improvement in terms of accuracy.

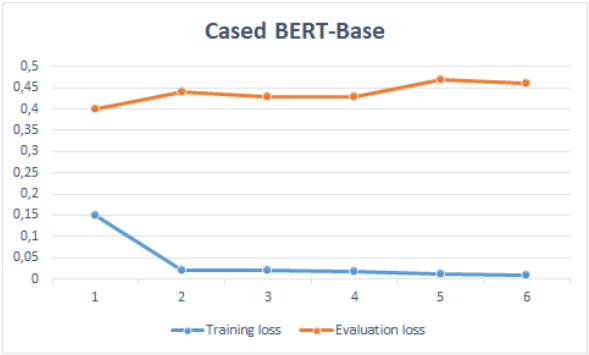

Figure 7: Cased BERT for sentiment analysis: Training and validation loss over epochs

5 CONCLUSION

The goal of this work was the evaluation of the use of *Bidirectional Encoder Representations from Transformers* (BERT) models for both sentiment analysis and emotion recognition of Twitter data. We defined an architecture composed of BERT-Base followed by a final classification stage and we fine-tuned the model for the above-mentioned tasks. We measured the performance of our classifiers by considering two datasets of tweets and we obtained a remarkable 92% accuracy for sentiment analysis and a 90% accuracy for emotion analysis, from which it was possible to deduce that BERT's language modeling power significantly contributes to achieve a good text classification.

In future work, we plan to improve the performance of our classifiers by determining the best number of layers and neurons in the final classification layers (i.e., fully connected layers) . We also intend to extend the experimentation by considering

Table 7: Uncased BERT: confusion matrix for the sentiment analysis task

	Actual Negative	Actual Neutral	Actual Positive
Predicted Negative	141	3	4
Predicted Neutral	11	112	10
Predicted Positive	3	3	143
Recall	**0.91**	**0.95**	**0.91**
Precision	**0.95**	**0.84**	**0.96**

Table 8: Cased BERT: confusion matrix for the sentiment analysis task

	Actual Negative	Actual Neutral	Actual Positive
Predicted Negative	141	2	5
Predicted Neutral	12	112	9
Predicted Positive	3	3	143
Recall	**0.90**	**0.96**	**0.91**
Precision	**0.95**	**0.84**	**0.96**

larger datasets, such as the SemEval 2017 Task 4 [23] dataset for sentiment analysis and the EmoBank [4] dataset for emotion analysis. This is particular important for the sentiment analysis task, in which we observed a repentine increment of the validation loss after the first epoch, probably due to overfitting. Although the models reach high accuracy and the approach seems promising, a comparison with other state-of-the-art classifiers will be useful to thoroughly evaluate the performance of our approach. We also intend to investi-gate the impact of BERT-Base by replacing it with other BERT distributions (e.g., BERT-Large) or traditional word embeddings, such as Word2Vec [17] or GloVe [21].

ACKNOWLEDGMENTS

The authors would like to thank the students Federico Filipponi, Leonardo Lucarelli and Alessandrino Manilii for their help in imple-menting the architecture for emotion recognition.

REFERENCES

[1] Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 718–728.

[2] Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 117–121.

[3] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A Rehman. 2017. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl* 8, 6 (2017), 424.

[4] Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 578–585.

[5] 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* (2018).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] C. Diamantini, A. Mircoli, D. Potena, and E. Storti. 2015. Semantic disambiguation in a social information discovery system. In *2015 International Conference on Collaboration Technologies and Systems, CTS 2015*. 326–333.

[8] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[9] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.

[10] Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband. 2018. Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications* 23, 1 (2018), 11.

[11] Maha Heikal, Marwan Torki, and Nagwa El-Makky. 2018. Sentiment analysis of Arabic Tweets using deep learning. *Procedia Computer Science* 142 (2018), 114–122.

[12] Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences* 30, 4 (2018), 330–338.

[13] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. 2018. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access* 6 (2018), 23253–23260.

[14] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883* (2019).

[15] Xinlong Li, Xingyu Fu, Guangluan Xu, Yang Yang, Jiuniu Wang, Li Jin, Qing Liu, and Tianyuan Xiang. 2020. Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access* 8 (2020), 46868–46876.

[16] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).

[19] A. Mircoli, A. Cucchiarelli, C. Diamantini, and D. Potena. 2017. Automatic emotional text annotation using facial expression analysis. In *CEUR Workshop Proceedings*, Vol. 1848. 188–196.

[20] Saif M Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696* (2017).

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[22] Ana Reyes-Menendez, José Ramón Saura, and Cesar Alvarez-Alonso. 2018. Understanding# WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach. *International journal of environmental research and public health* 15, 11 (2018), 2537.

[23] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 502–518. https://doi.org/10.18653/v1/S17-2088

[24] Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 28, 3 (2015), 813–830.

[25] Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588* (2019).

[26] Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2015. Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering* 28, 2 (2015), 496–509.

[27] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment

classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1555–1565.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[29] Leyi Wang and Rui Xia. 2017. Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 502–510.

[30] Zhongqing Wang, Yue Zhang, Sophia Lee, Shoushan Li, and Guodong Zhou. 2016. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 1624–1634.

[31] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232* (2019).

[32] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1253.

[33] Wei Zhang, Quan Yuan, Jiawei Han, and Jianyong Wang. 2016. Collaborative multi-Level embedding learning from reviews for rating prediction.. In *IJCAI*, Vol. 16. 2986–2992.