Using BERT for Checking the Polarity of Movie Reviews

Saad Abdul Rauf
Taiyuan University of Technology
Taiyuan, China

Yan Qiang
Taiyuan University of Technology
Taiyuan, China

Syed Basit Ali
Taiyuan University of Technology
Taiyuan, China

Waqas Ahmad
Taiyuan University of Technology
Taiyuan, China

ABSTRACT
In this era, the data of the user on social media is generating in every millisecond. The importance of data can be noted or observed as these are the reviews, emotions, and opinions of the human being in the form of text. The customer-generated data can be related to events, food, products, etc. This information is a "key to success" for those who do business or are in government and other individuals. As the data is in the form of bulk so, it is evident that to analyze a content that is generated by a user must be complicated to manage as well as it must be time-consuming. For this, we need an intelligent system that helps us to figure out whether the content is positive, negative or neutral in the form of categories. This smart system commonly named sentiment analysis (SA), opinion mining (OM), subjectivity mining, etc. Opinion mining is the systematized mining approach. Through this, we can classify, thoughts and emotions from the text, speech, and databank sources through Natural Language Processing (NLP). The actual purpose of writing this paper is to determine the idea of human emotions with the help of BERT model, where we took a dataset of IMDB movie reviews, which are generated by a users' data. Our experimental methodology is adequate and robust, which in turn describes the quality of sentiment analysis.

Keywords
Text classification, sentiment analysis, natural language processing, Bidirectional Encoder Representations from Transformers.

1. INTRODUCTION
These days the data is growing day by day, and that data is mostly-generated by EWOM (Electric Word of Mouth). Eventually, decision-making influence people a lot. These EWOM are mainly produced in social media platforms such as; Facebook, Twitter, Weibo, WeChat moments, etc. The data generated by social media is primarily in the form of reviews, comments, suggestions about the products. These comments are not only useful for the consumer as well as for producers too. As, when a customer wants to buy/order a product, they often check the other users' comments. Through this, the producers also get to know the strength and weaknesses of their product on the bases of other people's analyses. However, these analyses are very beneficial for manufacturers. The textual data, which is in the form of bulk, is overwhelming. The information is useful for researchers who are doing opinion mining. As we already know that sentiment analysis is an automatic system for human emotions summarization. This computerized system helps to identify a human emotion from text-based data, speech, and other databases (DB) sources through NLP (Natural Language Processing). SA classifies the text in the categories of "positive," "negative," or "neu-tral." Customers always want to know about other person opinions as well as the manufacturer also wants to hear about the feedback of customers for their future growth.

The work which we have focused on is to determine the polarity of the movie reviews by categorizing whether it is positive or negative polarized. The problem can be stood as a multi-label classification task or by binary classification, where it is easy to identify whether the polarity is good, bad or average. As, BERT is Bidirectional Encoder Representations from Transformers, the network architecture presented by google is in transformed in which we can know what is state of the art for NLP tasks. With the help of BERT, we can classify a text, translation, summarization and question answering (Q.A). The best thing is that it is a pre-trained on a dataset (Wikipedia and books-Corpus). With this, we can solve many different NLP tasks. The other specialties are Token level classification (e.g., Part of speech tagging). This pre-trained model includes a source code that's built upon TensorFlow, it is a machine learning (M.L) frame-work, and it also consists of a series of pertained language representation models. BERT differentiates its self with many other models such as it makes use of a transformer, and it shows the contextual relationship with the words and sub words. Here transformer contains two processes in the form of encoder and decod-er, where text input is read by encoder and task is predicted by a decoder. As it is a linguistic model, the encoder is an important element in it. While the unidirectional models read data either from left to right, which is in most cases and right to left in some instances when Urdu or Arabic language characters are processed, but bidirectional model reads data from left as well as right. Before the use of BERT, bidirec-tional RNN[19] has been used for text classification, but its mechanism was different. It did not use the encoder-decoder form, and the mirror-based mechanism was used, which had certain drawbacks.

Here we took the dataset of IMDB movie reviews, which we used for our experimental studies. This data consists of movie reviews that are organized in the form of positive and negative. In this experiment, the BERT model gives its best performance on the IMDB datasets.

2. RELATED WORK
Till now, several studies have done on text mining and sentiment analysis at different levels. The process can be described in simple words, textual data, which is used to analyze the human emotions in it. In the past, there are many S.A models are being for this task. In one paper, they analyze the problem from numeric data, which is related to drug reviews. We need to understand the importance of sentiments, which we are going to target [22]. In 2013, more than 40 companies and 150 market and social research consultancies provided services using online surveys [21]. They also found the polarity by using a fuzzy set theory [20]. These all researches are on the document-level [1],[6],[7], sentence-level [4],[5], and phrase or word [2],[3]. In other studies, they are also on a user level [8],[9]. If we talk about the world level

sentiment analysis, this describes the word into text, which leads to the emotions. Whereas, sentence-level gives us results of others opinion and provide an overview of its operation. Document-level is where it shows the searches for the possibility where all users are connected on social media. Who has the same opinion [11]. There are to most popular methods in literature reviews are hot encoding and BOW (Bag of Words) [18]. To improve text mining methods, in past decades, many researchers used Wikipedia for ease of correcting words.[26],[27],[28]. The most important thing in opinion mining is to collect a data, cleaning of a data and then mining process, later this data need to be evaluated, the results obtained from this data are needed to apply in major steps which are already used in many applications, which is related to social media [23]. To make an appropriate taxonomy for getting the desired result, which will help us in sentiment analysis, we need to understand how we will discover implicit aspects [24]. If we look moreover, its difficult to analyze reviews, some are biased reviews, and some are direct reviews, there are some im-proved methods in 2-class reviews and 3-class reviews in sentiment analysis. Where we can easily judge that neutral reviews are very difficult to judge in most of the opinion mining methods [25]. Positive or negative emotions can easily be identified if we see financial success, which is an important factor [10]. Sentiment analysis of a text is also done by a multilingual system, in that the author used a lexical resource with the dataset of amazon movie reviews [12]. Using single multiplicative LSTM, outperform NB and SVM with small datasets in all classification accuracies. They took a lot of power in training this model [13]. For text classification, data augmentation techniques are also being used in the form of Deep-learning to recurrent neural networks [14]. To improve NLP tasks, an effective pre-training of the language model is done too [15],[16]. Through token, segment and position embedding, the Next sentence prediction (NSP) can be made very useful by using a google model as it works in a bi-direction [17]. The bi-directional architecture is very efficient in pretraining tasks of NLP [17].

3. METHODOLOGY

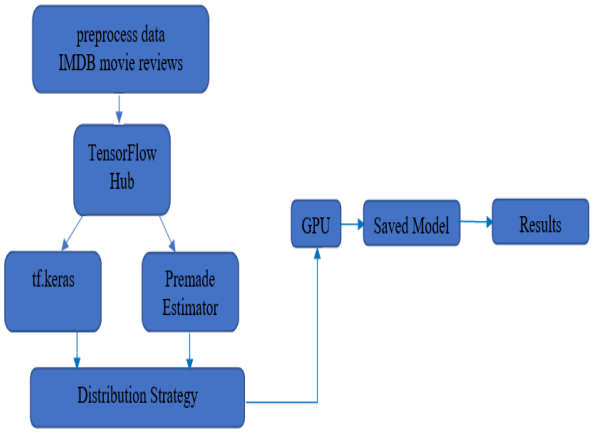
The purpose of this research is to find the polarity of the movie reviews in which we analyze the sentiments of the reviewers. For this, we followed the procedure, which is mentioned below.

3.1 Data

Here, the data used is batched data, which is hosted by Stanford, which entitled IMDB Movie Reviews. This data has a 5, 000 movie reviews for training, and testing data is of 5,000 too.

3.2 Procedure

Detailed analyses of the process, as depicted in (fig.1) include pre-processing, which takes place through the built-in fine-tuning techniques of TF-Hub. After that, the data is subjected to TensorFlow Keras, which in turn divides the with respect to columns. The next step is data distribution in a strategic manner that takes place after data is passed to tf-hub. After the distribution strategy, data can either be passed to CPU or GPU, which in this case, is GPU, where the actual data mining of such significant and complex data takes place. In the GPU, data is processed and classified in the saved model which in this project is BERT (Bi-directional encoding and transformer). BERT performs a detailed and precise analysis of the given data through both the directions and results are stored. Results are displayed in terms of accuracy, precision, recall, f1-score, Area under the curve (AUC), e value and confusion matrix, which is shown in tabular and graphical form in the subsequent sections.



3.2.1 Data Preprocessing

All we have done is that we transformed the data which our model can understand. For preprocessing, we did a couple of things. That is one of the reasons we use python. As, we are using BERT lower case model, so we need to lower case text. As the model we are using it is a fine-tuning model, where on a broad suite, it can perform token level and sentence level tasks. In most of the models, fine-tuning is done, but the hyperparameters are the same, but here we have the batch size, learning rate, and some training epochs. The good thing about this model is that it can use in bi-directional tasks. The text can be read-out in both ways “left to right or right to left” this technique can be work in both ways whether the text is in a sentence form or word form. Another task is to break the words into a chunk of pieces. The mapping of words can be done on the index by using a model library. In BERT model, we use three input sequences, Token ids, Mask ids, and Segment ids. The token id is that which is used in every sentence, which is restored from BERT dictionary (book of corpus and Wikipedia). To create the sequence of the same length, we use Mask ids, whereas, for one sentence sequence, we use 0 and 1 for two-sentence sequences by using segment ids. Unique tokens have already added in it, SEP[] and CLS[]. Whereas SEP[], is a sentence separator and CLS[] is a classifier. Which is introduce in a pre-training period. CLS[] token used at the starting of the sentence whereas, SEP[] used at the end of the sentence. Index and segment tokens are used at the input stage.

3.2.2 TF hub

Tf hub is also known as a TensorFlow hub, which is a reusable ML module. It has the ability to transfer learning among different tasks. Where transfer learning is a gate for taking weights and variables (assets) for our obtaining model as it is already trained on different data. We used a TF-Keras, which is a high-level API of TF. This helps us in integrating with different TF-features. TF-estimator is also included in TF-hub, which provides the functionality of training, eval and predictions. The amazing thing about TF-hub is port-ability with the CPUs and GPUs. Where we use google GPUs for this research.

3.2.3 Creating Model

As we have already prepared our data, the next step is to build a model “create_modle.” First, I will call module BERT with tf hub again. It will create a single layer. Which will help to train and adapt BERT to sentiment tasks (which are, to find the polarity of movie reviews). This strategy is also known as

fine-tuning. It's time to wrap our model in a "model_fn_builder" this model will adapt our three things, i.e. training, evaluation, and adaption.

3.2.4 Training data

In this step, data is already trained, where we use a google GPU; it took only almost 4 minutes to prepare our data.

3.3 Hyperparameters

The parameters which use during our experiment. We took a batch size of 32, the learning rate is 2e-5, and there are four epochs. We also use a warmup proportion with 0.1 value. Later we configure our model with 500 checkpoints steps and summary steps of a hundred. Following are the parameters we used are,

3.3.1 Learning rate

Here we took the learning rate of 2e-5 which means the during learning period this will work between 0 – 2e-5.

3.3.2 Batch size

This is the sample size of the data which we use in one iteration that is 32.

3.3.3 Epochs

This means that samples of a data have being passed. Here we are using 4 epochs.

3.3.4 Precision

Precision is a matric which is used to identify the cost of the false Positive is high. Which help us in fake reviews. It can be calculated from the following formula.

Precision = True Positive/ (True Positive + false Positive)

True Positive + False Positive = Total Predicted Positive

Precision = True Positive/Total predicted positive

3.3.5 Recall

Recall logic is use for to find out the how many actual positive are in our model. It can be calculated as,

Recall = True Positive/ (True Positive + false Negative)

Actual Positive = True Positive + False Negative

Recall = True Positive/Total Actual positive

3.3.6 F1 Score

F1 score is a function, which is calculated by the help of precision and recall.

F1 = 2 * (Precision *Recall)/ (Precision +Recall)

3.3.7 Accuracy

Accuracy is a mathematical value which is calculated by the ratio of true predictions and to the total predictions. Which can be seen as,

Accuracy = True Positive + True Negative / True Positive + True Negative + False Positive + False Negative

3.3.8 Area Under Curve

Area under curve is commonly known as AUC which is the integral value of each class vs time which means in our case, True Positive, True Negative, False Positive and False Negative have their respective areas under curve.

4. RESULTS

After subjecting our data through BERT model, the following results were obtained. Table-1 shows all the accuracy parameters of evaluation. Accuracy parameters, which include auc, val and f-1 score, are depicted in Figure-2, Figure-3 describes the confusion matrix in graphical form and other values such as Loss, Precision and Recall are shown in Figure-4.

Table -1 Evaluated Results

Parameters	Values
auc	0.89553244
Eval_ accuracy	0.895481336
F1_score	0.898974554
False_negatives	319
False_possitives	213
Global_steps	468
loss	0.4856
Precision	0.91744186
recall	0.881236039
True_negatives	2191
True_possitives	2367

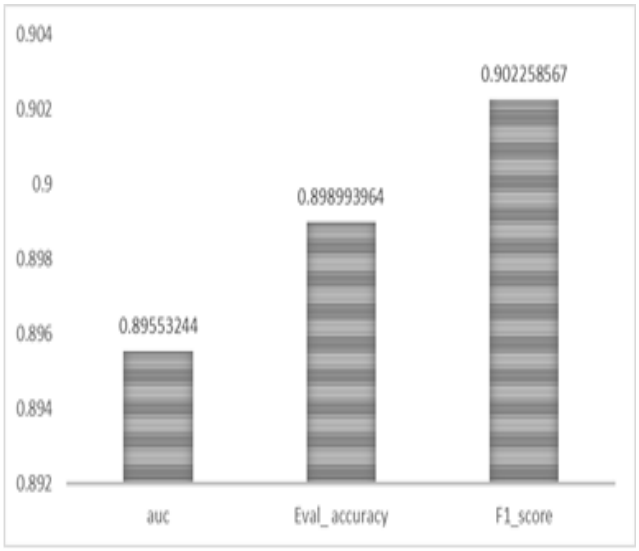


Figure-2 Accuracy Parameters

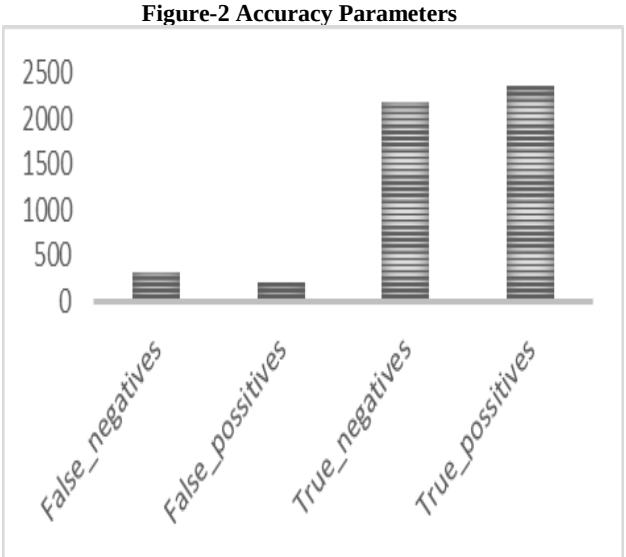


Figure-3 Confusion Matrix

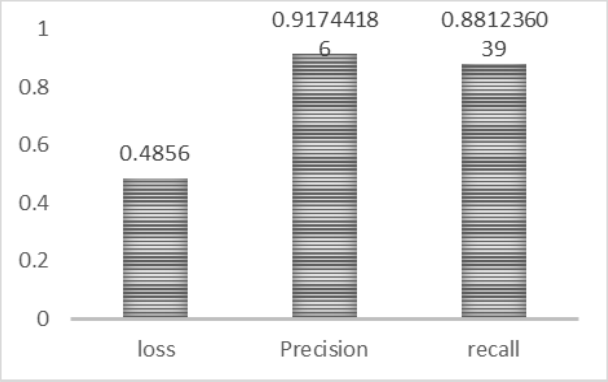


Figure-4 Loss, Precision and Recall

4.1 Result of Classification of Words

The below bar chart (Fig.5) describes the total number of words that can show the average of the aggregate data.

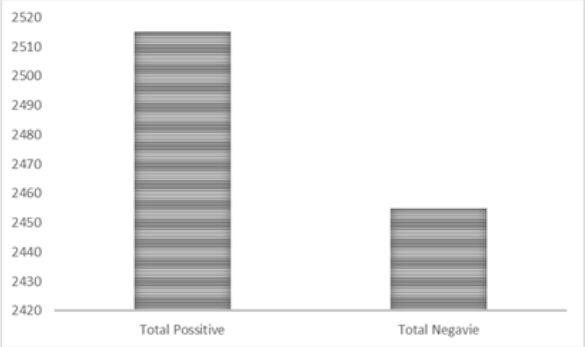


Figure-5 Representation of Complete words

5. CONCLUSION

This paper was based on checking the polarity of IMDB movie reviews by using the concept of sentiment analysis. Where we use a BERT classifier. In our proposed model, we use the data set of IMDB movie re-views. Here we downloaded the dataset from the Standford sentiment analysis database. Later we use the TF-Hub with BERT model and the distribution strategy, where GPU worked, was started, later that orga-nized data is transfer to our model where all parameters are already mentioned. The parameters which we used in our model are Eval_ accuracy, F1_score, False_negatives, False_possitives, Global_steps, Loss, Precision, True_negatives, and True_possitives. Where our batch size is 32, and the learning rate is 2e-5, 4 epochs. This modle gave us a marvellous result. In future work, we are looking forward to using this model and technique with the stream or online data.

6. REFERENCES

[1] Abbasi, A., France, S., Zhang, Z. and Chen, H., 2010. Selecting attributes for sentiment classification using feature relation networks. IEEE Transactions on Knowledge and Data Engineering, 23(3), pp.447-462.

[2] Tetlock, P.C., Saar- Tsechansky, M. and Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance, 63(3), pp.1437-1467.

[3] Wilson, T., Wiebe, J. and Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level senti-ment analysis. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.

[4] Yu, H. and Hatzivassiloglou, V., 2003, July. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguis-tics.

[5] Tan, L.K.W., Na, J.C., Theng, Y.L. and Chang, K., 2011, October. Sentence-level sentiment polarity classification using a linguistic approach. In International Conference on Asian Digital Libraries (pp. 77-87). Springer, Berlin, Heidelberg.

[6] Das, S.R., 2011. News analytics: Framework, techniques and metrics. In The Handbook of News Analytics in Finance (Vol. 2). John Wiley & Sons Chichester.

[7] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[8] Melville, P., Gryc, W. and Lawrence, R.D., 2009, June. Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1275-1284). ACM.

[9] an, C., Lee, L., Tang, J., Jiang, L., Zhou, M. and Li, P., 2011, August. User-level sentiment analysis incorporating social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1397-1405). ACM.

[10] Mishne, G. and Glance, N.S., 2006, March. Predicting movie sales from blogger sentiment. In AAAI spring symposium: computational approaches to analyzing weblogs (pp. 155-158).

[11] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C., 2011, June. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 142-150). Association for Computational Linguistics.

[12] Denecke, K., 2008, April. Using sentiwordnet for multilingual sentiment analysis. In 2008 IEEE 24th International Conference on Data Engineering Workshop (pp. 507-512). IEEE.

[13] Radford, A., Jozefowicz, R. and Sutskever, I., 2017. Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444.

[14] Wei, J.W. and Zou, K., 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.

[15] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

[16] Howard, J. and Ruder, S., 2018. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

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

[18] Ramos, J., 2003, December. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).

[19] Adelia, R., Suyanto, S. and Wisesty, U.N., 2019. Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit. Procedia Computer Science, 157, pp.581-588.

[20] Yazdavar, A.H., Ebrahimi, M. and Salim, N., 2017. Fuzzy based implicit sentiment analysis on quantitative sentences. arXiv preprint arXiv:1701.00798.

[21] Dolnicar, S., Grün, B. and Yanamandram, V., 2013. Dynamic, interactive survey questions can increase survey data quality. Journal of Travel & Tourism Marketing, 30(7), pp.690-699.

[22] Menner, T., Höpken, W., Fuchs, M. and Lexhagen, M., 2016. Topic detection: identifying relevant topics in tourism reviews. In Information and Communication Technologies in Tourism 2016 (pp. 411-423). Springer, Cham.

[23] Schmunk, S., Höpken, W., Fuchs, M. and Lexhagen, M., 2013. Sentiment analysis: Extracting decision-relevant knowledge from UGC. In Information and Communication Technologies in Tourism 2014 (pp. 253-265). Springer, Cham.

[24] Alaei, A.R., Becken, S. and Stantic, B., 2019. Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 58(2), pp.175-191.

[25] García-Pablos, A., Duca, A.L., Cuadros, M., Linaza, M.T. and Marchetti, A., 2016. Correlating languages and sentiment analysis on the basis of text-based reviews. In Information and Communication Technologies in Tourism 2016 (pp. 565-577). Springer, Cham.

[26] Wang, P., Hu, J., Zeng, H.J. and Chen, Z., 2009. Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems, 19(3), pp.265-281.

[27] Hu, X., Zhang, X., Lu, C., Park, E.K. and Zhou, X., 2009, June. Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 389-396). ACM.

[28] Boubacar, A. and Niu, Z., 2014. Conceptual clustering. In Future Information Technology (pp. 1-8). Springer, Berlin, Heidelberg.