Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

Artificial Intelligence-Based Sentiment Analysis

[1]Aymen Samir, [2]Saleh Mesbah and [1]Magda Madbouly
[1]Institute of Graduate Studies and Research, Alexandria University, Sharqi, Egypt
[2]Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

Key words: BERT, latest technology, neural network, accuracy

Corresponding Author:
Saleh Mesbah
Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

Page No.: 18-22
Volume: 16, Issue 1, 2021
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: Bidirectional Encoder Representations from Transformers (BERT) represents the latest technology of pre-trained language models which have recently advanced a wide range of natural language processing tasks. This study aims to investigate how BERT can be usefully applied in sentiment analysis tasks with fully connected neural network. The proposed model is developed using simple tips preventing it from over-fitting and enabling it to be fine-tuned easily on such down stream task. BERT performs much better as a Strong text embedding Model. Using such procedures successfully provide a better accuracy than the expensive machine learning procedures.

INTRODUCTION

Transfer learning has been widely used through these golden NLP years. Leakage of labeled data is forming a huge difficulties in many machine learning tasks. Language models emerged as a rescue for such problems. Its core power is concentrated on learning the context, syntax, semantics of such language. In an unsupervised way, it is just predicting the next word, or in some other language models cases predicting masked words[1]. From such point word 2vec[2] and glove[3] and other models, emerged mapping word identity to a continuous representation in high dimensional space with intuition that such high dimensional vector captures the meaning of such word. The main problems with such models are concerned around the fact that such vector is being fixed not considering its context. ELMO[4] is a state of the art model that takes such point in to consideration, in order to be more concerned with context, using a 2back-boned LSTM layers left-to-right and right-to-left embedding, then concatenating such embedding in a single vector to be the token embedding. Bidirectional Encoder Representations from Transformers (BERT)[1] on the other hand builds a deep bidirectional embedding using stack of transformer encoders[5] connected in a bidirectional way. Using such models as an encoder for many NLP tasks as sentiment analysis, question answering and text summarizing, for fine-tuning on a given down-stream task, can easily reach the state of the art results with few training epochs. Although fine-tuning can be little burden some, by which the fact of large models being easily over-fit, due to its complexity and capacity of learning. While using some regularization techniques like SMART[6] could be a good practice in training phase and enabling models to be generalize don unseen data.

This study fine-tuned BERT with single layer added on top for the sentiment analysis task. Using few tips and recommendations from BERT, the proposed model provided 91.4% validation accuracy for just 4 epoch sona data set that consists of 1.5+Million Tweet using Tesla P100GPU.

MATERIALS AND METHODS

This study explains the main concepts of BERT.

Pre-Trained Language Models: Pre-trained language models[4, 7, 1, 8] have recently emerged as a key technology for achieving impressive gains in a wide variety of natural language tasks. These models extend the idea of word embedding by learning contextual representations from large-scale text data using a language modeling objective. BERT[1] is a language representation model which is trained with a masked language modeling and Next Sentence prediction task on a large corpus of 3,300 M words. Each token is assigned three kinds of embedding: token embedding indicate the meaning of each token, segmentation embedding are used to discriminate between two sentences and position embedding indicate the position of each token within the text sequence. These three embedding methods are summed to a single input vector and fed to a bidirectional transformer with multiple layers:

$$\hat{h}^{l} = LN(h^{l-1} + MHAtt\left(h^{l-1}\right) \quad (1)$$

$$h^{l} = LN(\hat{h}^{l} + FFN\left(\hat{h}^{l}\right) \quad (2)$$

Where:

$h_0 = x$ = The input vectors
LN = The layer normalization operation[9]
MHAtt = The multi-head attention operation[5]
Superscript l = The depth of the stacked layer

On the top layer, BERT will generate an output vector for each token with rich contextual information. Pre-trained language models are usually used to enhance performance in language understanding tasks. Recently, there have been attempts to apply pre-trained models to various generation problems. When fine-tuning for a specific task, unlike ELMO who SE parameters are usually fixed. Parameters in BERT are jointly fine-tuned with additional task-specific parameters.

BERT: BERT is built from stack of transformer encoders connected in bidirectional way. That is how its name is formed. One of the strongest language models that could be used across many down stream tasks by simple fine-tuning. BERT is different from GPT[7] on the way it's structured. GPT is an auto-regressive language model approach. It is built on stack of transformer decoders connected in a unidirectional way (left to right) where embedding of such token is based on previous tokens only (not all tokens like BERT). BERT is learned with masked language model due to its bidirectional nature as context can be visible to the next layers. So, a [MASK] token is replacing 15% of the text, by which model should predict. The other task is Next Sentence Prediction (NSP) where model is tuned to learn the semantics of language and sentences structures, whether pair of sentences could be in the same context or not.

GPT on the other hand could be a better generative model while BERT is better at embedding due to its deep bidirectional embedding across transformer layers.

Experiments on BERT output whether to consider the [CLS] token embedding as the sentence embedding, or include the 12 hidden layers output. After some trials on taking just last 4 hidden layers as sentence embedding found that [CLS] or‘ pooler-output‘ converges quicker and train more stable for such down stream task.

The overall pre-training and fine-tuning procedures for BERT. [CLS] is a special token added before every input example, mostly Its embedding used in classification tasks) and [SEP] is a special separator token (e.g., separating questions/answers).

Related work: The researcher by Rathor *et al*.[10] discussed the utility of machine learning algorithms (SVM, NB and ME) to identify online feedback by guided methods of learning. Customer reviews are categorized in to optimistic, negative and neutral reviews that allow not only the customers to purchase but also to consider the market reaction of the companies to the particular goods. During the course of training the algorithm, the author used weighted unigrams and unigrams. The researcher used Amazon API.

By Al Amarani *et al*.[11] specifically stated the key variances that can be rendered useful for establishing the correct rules for classification techniques between Random Forest (RF) and Support Vector Machine (SVM). The finding soft his experiment indicated that Amazon user ratings, based on datasets are given by Amazon are better described as the Random Forest Support Vector Machine algorithm (RFSVM). It uses both of the classification approaches for SVM and RF to achieve stronger results in hybrid classification.

By Rathor *et al*.[10], developed a better-automated way to identify the feelings of Twitter messages. The feeling in Twitter tweets is identified according to the questionnaire grade das optimistic or bad. The data set for Twitter messages is composed of emoticons that are considered to be noisy labels. These data are adequately available for training. In Naïve Bayes, Maximum Entropy and SVM when emoticon data are trained, the accuracy is above 80%. This study out lines specifically the required pre-processing measures for high precision. Twitter data are stated as unstructured, heterogeneous and messages are optimistic, negative, or neutral. Naïve Bayes, Max Entropy and Help Vector Machine use the emotion analysis of Twitter data and its problems are clearly outlined in this study.

By Govindarajan[12], the hybrid approach was tested using the data from film reviews and ultimately graded by NB and GA for reduced results. The NB, Genetic Algorithm (GA) and NB-GA hybrid versions are also designed to be simple classifications. The developer has developed a hybrid program that better integrates and combines the base classification and hybrid solution in the best possible way. The NB-GA hybrid model gives a higher degree of precision classification compared with the base classifier and increases the test period due to are duction in data dimensions.

Description ensembles and lexicons for the automated classification of the tweet feeling are developed and described by Silva *et al*.[13]. The tweets are marked as either positive or negative with regard to the message. Sentimental analyzes may be utilized by business extracting their brand's market perceptions, customers searching their goods and more. The opinion study of microblogging systems such as Twitter has not been carried out with the use of grouping ensembles and lexicons. The experiment demonstrates the classification accuracy of classifier ensembles made up of models like SVM, Logistic Regression, the Multinomial Naïve Bayes and Random Forest on feeling tweeter data sets. SVM models shape the basis of the Ensemble Emotion Classification. The theory of plurality votes among different classification methods, along with Bayesian Network, Random Forest, C4.5 Decision Tree, Naïve Bayes. The suggested Ensemble Sentiment Classification and the six classification models are evaluated to verify the classifiers with a 10-fold assessment for the 12864 Tweet data set. The suggested Ensemble Sentiment Classification method reaches individual classification for the chosen Twitter airline service data set. This strategy also increases the overall quality of the Twitter ranking for certain services.

By Wang *et al*.[14], proposed a model consisting of both CNN and LST M that would predict texts for the Valencia Arousal (VA). The typical CNN would use the whole text as the model's content. Growing single sentence is taken as a region in this model that divides the input in to many regions that contribute to the extraction and weighing of the affective information that contributes to VA prediction. The researchers integrated the CNN and LSTM regional and reveals that the new model succeeds. The models focused on lexicon, regression and NN.

By Wan and Gao[15], a joint CNN and RNN architecture including long-distance dependence learned from RNN and coarse-grain local characteristics created from CNN is represented for sentimental analysis of the short texts. The three-pronged group, SST1, SST2, MR has an accuracy of 51.50 and 89.95% and increases the modernity obtained from experimental tests. Sentiment Analysis of Twitter Data[15] using Hybrid approach with

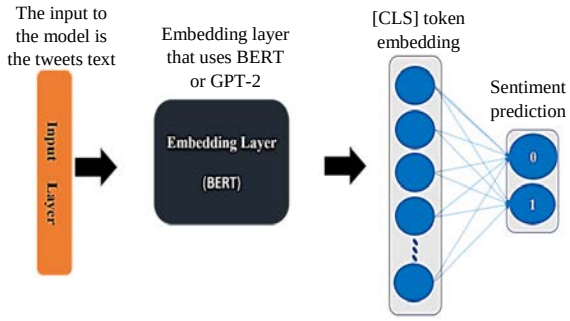

Fig. 1: Model architecture

an expensive approach using sequence of the following machine learning models for obtaining the token embedding:

- Pre-processing text
- N-Gram
- Tf-IDF
- Singular value decomposition
- Principle component analysis
- Random forest
- Naïve Bayes
- Principle component analysis

Model architecture: This research paper aims to prove that such process could be more efficient by building a deep bidirectional embedding model as BERT.

In this study, we propose a NLU Model based on fine-tuning BERT Model, taking the' CLS' token as the sentence embedding on a fully connected neural network produce the sentiments core, we the such sentence is Positive (1) or Negative (0) as illustrated in Fig. 1.

After BERT embedding, we add a Dropout layer with 10% probability as a regularization technique; preventing neural network from over fitting.

Gradient Clipping[16] is a crucial technique for stable training and preventing from over fitting as large models such as BERT can easily trapped in gradient explosion which make model oscillate rigorously in the loss space. Variant of Adam Optimizer the Decoupled weight decay regularization used as the optimizer for our model with small Learning rate of 2e-5.

Experimental: In this study, we describe the Twitter data set used for such task and discuss the implementation in details.

RESULTS AND DISCUSSION

Dataset and modelling analysis: The proposed model is trained on 1.5+Million tweets collected via. Stanford sentiment 140 data set project[17]. This data is collected via twitter API.

Table 1: Model data sets and accuracy

Variables	Training	Testing	Amazon	Hachette
Dataset size	1+million	450 K	200	200
Accuracy (%)	87.5	86.5	91	94.5

Table 2: Accuracy comparison

Variables	Amazon (%)	Hachette (%)
The proposed model	91.0	94.1
ML[18]	95.6	92.4

First, text cleaning and pre-processing are carried out by removing: links, mentioning(@user), Symbols (except[',?,%]) and emojis. Digits and stop-words were not removed as they can be good indicators for such task.

Secondly, we began with choosing the maximum sequence length to be as an input of the model, to be of length 30 tokens as maximum sequence length. This step also includes truncating longer sequences and padding others with '100' token as it is the ignore index in cross-entropy loss function.

Tokenizer is based on BERT as it is introduced by transformers package through which it provides encoding, padding, truncating tweet. It also returns the attention mask and the input tokens encoded or being processed by BERT for embedding.

Third, deciding batch size of 128, learning rate of 2e-5 as has been recommended by Devlin *et al*.[1]. Binary cross entropy is used as the loss function:

$$\text{Loss} = -\left(y\log\left(\bar{y}\right) + \left(1-y\right)\log\left(1-\bar{y}\right)\right) \quad (3)$$

The model performs table on training process and after few epochs reached a high score of 87% on the training set and 86% on the validation set.

Batch Scheduler is used for decaying the learning rate in a linear fashion without warm-up steps before decaying learning rate of optimizer each batch.

The proposed model is evaluated using two different datasets, one for Amazon and the other for Hachette, using Twitter API for collecting 200 Tweets for each dataset, labeling done manually, Model performs pretty well on Amazon and Hachette datasets with score 91 and 94.5%, respectively using Eq. 4:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (4)$$

Table 1 shows the dataset size sand accuracy across each dataset. Table 2 shows the accuracy obtained using the proposed model in a comparison with the generalized model[18]. Difference is concerned on the pre-processing, training and evaluation procedures.

The data set are of size 200 on both Amazon and Hachette tweets. Also tweets are not the same (due to time difference) for straight forward comparison.

CONCLUSION

This study proved that transfer learning is successfully playing a crucial role in NLP tasks. Fine-tuning BERT to be a strong embedding layer capturing the sentence semantics, using frame work of Drop-out layer on top of BERT.

BERT performs much better as a strong text embedding Model. Using such procedures, successfully made a better accuracy than other machine learning procedures.

REFERENCES

01. Devlin, J., C. Ming-Wei, K. Lee and K. Toutanova, 2019. BERT: Pre-training of deep bidirectional transformers for language under standing. Proceedings of the 2019 Conference on Computational Linguistics: Human Language Technologies Vol. 1, June 2-7, 2019, ACL., Minneapolis, Minnesota, pp: 4171-4186.
02. Mikolov, T., K. Chen, G. Corrado and J. Dean, 2013. Efficient estimation of word representations in vector space. Comput. Lang., 1: 1-12.
03. Pennington, J., R. Socher and C.D. Manning, 2014. Glove: Global vectors for word representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), October 25-29, 2014, Association for Computational Linguistic, Doha, Qatar, pp: 1532-1543.
04. Peters, M.E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, 2018. Deep contextualized word representations. Proceedings of the 2018 Conference on Human Language Technologies, Vol. 1, June 1-6, 2018, Association for Computational Linguistics, New Orleans, Louisiana, pp: 2227-2237.
05. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez and I. Polosukhin, 2017. Attention is all you need. Adv. Neural Inf. Process. Sys., 30: 5998-6008.
06. Jiang, H., P. He, W. Chen, X. Liu, J. Gao and T. Zhao, 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. Adv Neural Inf. Process. Sys., Vol. 1,
07. Radford, A., K. Narasimhan, T. Salimans and I. Sutskever, 2018. Improving language understanding by generative pre-training. CoRR., Vol. 1,
08. Dong, L., N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang and H.W. Hon, 2019. Unified language model pre-training for natural language understanding and generation. Proceedings of the Advances in Neural Information Processing Systems, December 10-12, 2019, Vancouver, Canada, pp: 13063-13075.
09. Ba, J.L., J.R. Kiros and G.E. Hinton, 2016. Layer normalization. Mach. Learn., 1: 1-14.

10. Rathor, A.S., A. Agarwal and P. Dimri, 2018. Comparative study of machine learning approaches for Amazon reviews. Procedia Comput. Sci., 132: 1552-1561.
11. Al Amrani, Y., M. Lazaar and K.E. El Kadiri, 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput. Sci., 127: 511-520.
12. Govindarajan, M., 2013. Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. Int. J. Adv. Comput. Res., 3: 139-145.
13. Silva, N.F.D., E.R. Hruschka and E.R. Hruschka, 2014. Tweet sentiment analysis with classifier ensembles. Decis. Support Syst., 66: 170-179.
14. Wang, J., L.C. Yu, K.R. Lai and X. Zhang, 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 2, August 7-12, 2016, ACL, Berlin, Germany, pp: 225-230.
15. Wan, Y. and Q. Gao, 2015. An ensemble sentiment classification system of twitter data for airline services analysis. Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW), November 14-17, 2015, IEEE, New Jersey, USA., pp: 1318-1325.
16. Pascanu, R., T. Mikolov and Y. Bengio, 2013. On the difficulty of training recurrent neural networks. Proceedings of the International Conference on Machine Learning, February 17-19, 2013, Carnegie Mellon University, Pittsburgh, Pennsylvania, pp: 1310-1318.
17. Go, A., R. Bhayani and L. Huang, 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford University, Stanford, pp: 1-12.
18. Srivastava, A., V. Singh and G.S. Drall, 2019. Sentiment analysis of twitter data: A hybrid approach. Int. J. Healthcare Sys. Inf. (IJHISI.), 14: 1-16.