Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

Analyzing the Performance of Bidirectional Transformer and Generalized Autoregressive Permutation Pre-trained Language Models for Sentiment Classification task

D.Deepa [1], A.Suhana Nafais [2], B.Motti Kumar [3], J.M.Ravi Prasath [4], Dr.T.Suba [5], Dr.P.Jenopaul [6]

[1] Assistant Professor (SRG), Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Tamilnadu.Email Id: deepa@kongu.ac.in

[2]Assistant professor, Department of Master of Computer Applications, Sona College of Technology, Salem, Tamilnadu. Email Id: suhana.nafais@sonatech.ac.in

[3,4] UG Scholars, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Tamilnadu.Email Id: mottikumar@gmail.com,raviprasad0116@gmail.com

[5] Assistant Professor, Department of English, M.Kumarasamy College of Engineering, Karur, Tamilnadu,India. Email Id: subat.snh@mkce.ac.in

[6] Professor, Department of Electrical and Electronics Engineering, Adi Shankara Institute of Engineering and Technology, Kerala-683574. Email Id: jeno.eee@adishankara.ac.in

ABSTRACT

With the advancement of deep learning, automatic feature extraction and processing larger data is achievable now days. With the ability of modeling two-way contexts, a new language representation model called Bidirectional Encoder Representations from Transformer (BERT) and Generalized Autoregressive Pretraining for Language Understanding (XLNet) has been introduced to pre-training from larger corpus to understand the linguistic feature for sentiment classification task. These two models learn the context bidirectionally but differ in masking strategy and pre-train-fine-tune discrepancy. In the paper both $BERT_{base}$ and $XLNET_{base}$ models are applied are experimented on IMDB and coursera dataset and compared with RNN. XLNET overcomes the constraints of BERT because of it uses autoregressive.

Keywords: pre-train, language model, encoder, transformer, autoregressive, sentiment classification

1. INTRODUCTION

Sentiment analysis is a natural language processing task to conclude the opinion expressed in the sentence which will be in the form of online review, twits etc., in ecommerce websites, twitter, and blogs and from any other forums. The result of the sentiment analysis task is a polarity classification which may be bi class classification to fine grain classification like very negative, negative, neutral, positive, and very positive.

Sentiment analysis task helps to improve businesses by monitoring the customer's feedback about the product sentiment and helps to understand customer needs. The emerging social media tends a way to the customers to express their thoughts and feelings freely than before. [1] Sentiment analysis is not an easy job to manually categorization millions of everyday feedbacks of stock inverters and customers of online shopping. This kind of analysis involves in critical real-time issues and should help people to take action right away. Also, people involving in sentiment analysis process spent 65% of their time in determining the sentiment from the volume of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

the comment. So, it requires a cost –effective automatic and context effective sentiment analysis model to complete the task.

Sentiment analysis models are built using various machine learning and deep learning algorithms. Dictionary Based methods like WordNet (Miller, Beckwith et al. 1990) contains synonyms and antonyms for every phrase is given with numerical rating which will be used for polarity classification. [2] Using part-of-speech (POS) these phrases can be extracted as unigram, bigram, trigram and syntactic phrases. Machine learning algorithms need manual conversion of the reviews into required format and need more training data. Obtaining training data with label for all the domain is impossible one. From the above, it could be observed that machine learning algorithms need more training data and need manual effort. [3] Later Deep learning algorithms have offered a revolution to reduce the flaw in preprocessing and insufficiency of data. Deep learning algorithms provide automatic word embedding done often pre-trained on text corpus from co-occurrence statistics which reduces the manual preprocessing. word2Vec, GloVe like tools provides pre-trained word representation for modeling analysis.

The various architectures of Deep Leaning models like CNN, RNN, LSTM, Bi-LSTM learns the context in a better way. These pre-trained Word embeddings are applied in a context free manner. The solution to this problem is shifting from Pre-trained Word Embeddings to Pre-trained Language Models which learns contextual representations from text corpus (Garcia-Silva, Berrio et al. 2019) ELMo (Peters, Neumann et al. 2018) introduced in 2017 is a Deep Contextual Word Embeddings provides such a representation.

A model needs to be trained only on the task-specific dataset both to understand the language and the task using atleast smaller dataset. The goal of language modelling is to learn and estimate the probability distribution of sequence of words. [4] [5] these learnt probability distributions used as understood rules of a language. Since the language model pre-trained using larger corpus greater language understanding is possible.

Deep learning with increased parameter it is possible to train much larger dataset and can avoid over fitting. Though, building a especially for syntax and semantically related tasks model with large-scale labeled datasets is a great challenge for most NLP tasks since annotating the sentence with label is extremely expensive. In contrast, building a model with large-scale unlabeled corpora is relatively easy to construct.[6] To hold this property the context representation of the unlabeled corpus is learnt from them and then use these representations for other downstream tasks. Bidirectional Encoder Representations from Transformer (BERT) and Generalized Autoregressive Pretraining for Language Understanding (XLNET) is a Pre-trained language model provides deeper context language understanding which can be used for many of the downstream tasks involving in Natural Language Processing (NLP).

1.1 BERT MODEL

The problem with the language models is that it learns only from left context or right context, but language understanding is bidirectional. A new language representation model called Bidirectional Encoder Representations from Transformer (BERT) developed by Google (Devlin, Chang et al. 2018) outperforms in eleven natural language processing tasks since it is trained from larger corpus.. Even though BERT suffer with random 15% of masking and prediction. BERT is pre-trained on a large corpus of unlabeled text including the entire Wikipedia and Book Corpus of 3,300 million words. BERT is a “deeply bidirectional” model. BERT uses Random masking pre-training model. BERT learns the context of a word left and right at same instant.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

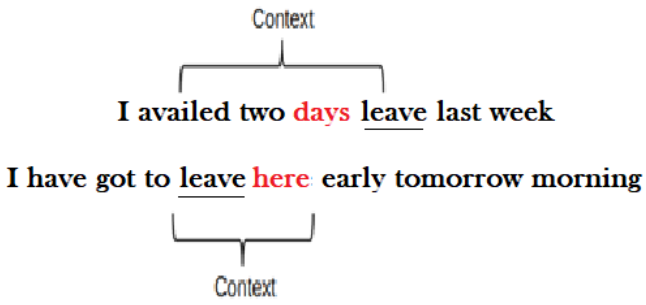

Figure 1.1. BERT learning the context

When the nature of the word "leave" is being predicated by only taking either the left or the right context, then an error will be made in at least one of the two given examples. One way to deal with this is to consider both the left and the right context before making a prediction. That's exactly done by BERT

The BERT architecture builds on top of Transformer. The two variants of BERT are BERT Base: 12 layers (transformer blocks), 12 attention heads, and 110 million parameters BERT Large: 24 layers (transformer blocks), 16 attention heads and, 340 million parameters. All of these Transformer layers are Encoder-only blocks.

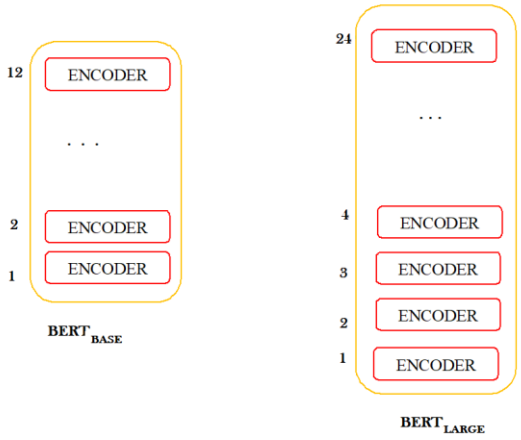

Figure1.2. BERT Variations

I <MASK> two days leave last week

I have <MASK> to leave here <MASK>tomorrow morning

Figure 1.3. BERT masking (15% words in each input sentence during Pre-training)

1.2 XLNET MODEL

XLNet (Yang, Dai et al. 2019) is a large bidirectional transformer with improved training methodology and more computational power than BERT.[7] XLNET pre-trained with larger corpus prediction metrics on 20 language tasks. To improve the training, XLNet introduces permutation language modeling, where all tokens are predicted but in random order. Like BERT, XLNET also capable of modeling bidirectional contexts, by denoising the issue related

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

to the input with masking, that BERT fails to care the dependency between the masked positions. XLNet, a generalized autoregressive Pretraining method allows learning bidirectional contexts by adapting permutations of the factorization order.

Permutations: Given a sentence sequence x, an auto-regressive (AR) model calculates the probability P. The AR objective in this case could be seen as Pr (xi) = Pr (xi | x>i).

$$\max_{\theta} \quad \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=1}^{T}\log p_\theta(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}})\right] \quad \text{--- Equation -1}$$

Equation.1 Permutation Language Modeling

2. Literature Review

(Sun, Qiu et al. 2019) investigated different fine-tuning methods of BERT on text classification task for BERT fine-tuning. They have proposed language model pre-training and proved that which will be useful to learn the deeper contextual language information and to achieve state of art performance on various downstream on NLP tasks.

(Munikar, Shakya et al. 2019) used transfer learning with BERT model for fine-grained sentiment classification task using SST dataset and showed outperforming results. (Li, Fu et al. 2020), predict the Aspect-based sentiment analysis for the given aspects or targets by introducing a context-aware embeddings layer which contains the most correlated information for the selected in the context.[8] The gating mechanism is used to control the propagation of sentiment features from BERT output with context-aware embeddings. The proposed model achieved state-of-the-art results test F1 of 88.0 and 92.9 on the SentiHood and SemEval-2014 datasets respectively.

An ensemble of BERT models was proposed by (Lehečka, Švec et al. 2020) using self-collected Czech movie reviews dataset and distilled the knowledge. The model was improved with pooling layer architecture of BERT and showed better result.

(Abuzayed and Al-Khalifa 2021) proposed sarcasm and sentiment detection for Arabic language by augmenting the shared task data set to identify the sentiment of a tweet or to detect if a tweet is sarcasm. [9] They proved assured results for sarcasm detection and sentiment identification by MARBERT model with data augmentation. (Narayanaswamy 2021) experimented BERT and RoBERTa models for aspect-based sentiment analysis to capture the context from the unstructured data. Then compared the results with the traditional model. BERT and RoBERTa show the improved results.

(Chriqui and Yahav 2021) developed a BERT model HebEMO for the Modern Hebrew text, which trained on a Covid-19-related UGC dataset. The model HebEMO yields prediction capability with F1-score of 0.96 for polarity classification. (Azeemi and Waheed 2021) modeled RoBERTa for 200,000 COVID-19 tweets for downstream tasks.

(Mustapha, Krasnashchok et al. 2020) proposed a classification model using XLNet without fine-tuning on domain-specific data which are trained on the GDPR dataset. [10] This model has got improved F1 by 1–3%. Since the model not used fine tuning, which reduces the training time and complexity, compared to the BERT-based model? The pre-trained language models such as Bert and XLNet hold implicit semantics only relying on surface information between words in corpus. So the model is improved with pretraining by explicit knowledge facts from knowledge graph (KG) and then add a knowledge injunction layer to transformer directly without changing its architecture. (Mustapha, Krasnashchok et al. 2020) experimented this model on various datasets for different downstream tasks and showed improved performance.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

XLNet model is finetuned by (Alshahrani, Ghaffari et al. 2020) to accurately predict optimism and pessimism of Twitter messages related to individual health in social media.

3. System

The Experiment is performed by comparing the performance of the two pre-trained language models BERT and XLNET with its default pre-training corpus and without changing the architecture.[11][12] Two different datasets are used analyse the performance of BERT and XLNET. The experiment includes the following execution steps.

.

3.1 Execution Steps

- Setting up Hyper parameters
- Loading pretrained BERT and XLNet model
- Train step function
    - ❖ Load the data from the Data Loader
    - ❖ Pass the data to the model
    - ❖ Calculate loss
    - ❖ Calculate accuracy
    - ❖ Back propagate and update weights
- Evaluation step function
    - ❖ Load the data from the Data Loader
    - ❖ Pass the data to the model
    - ❖ Calculate loss
    - ❖ Calculate accuracy
- Fine-tuning the model
    - ❖ Training Accuracy
    - ❖ Train Loss
    - ❖ Validation Accuracy
    - ❖ Validation Loss

3.2 Dataset

The dataset for this experiment has acquired from Kaggle. The first dataset is IMDB dataset having 50K movie reviews for binary sentiment classification for various natural language processing or Text analytics. It provides a set of 25,000 movie reviews for training and 25,000 for testing. [13] The second dataset used for the experiment is Coursera dataset contains a collection of 100k review from all available cause videos in their learning platform. 50k reviews were used from the given dataset. Where reviews containing label 1 and 2 are considered negative reviews and 3,4,5 are considered as positive reviews. Which is then fed to the model to train and evaluate.25k of reviews containing 1,2 label (negative)&25k of reviews containing 3,4,5 (Positive)Totally:[14]50k were used for the experiment. Among this 50 % reviews used for training and 50% for testing.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

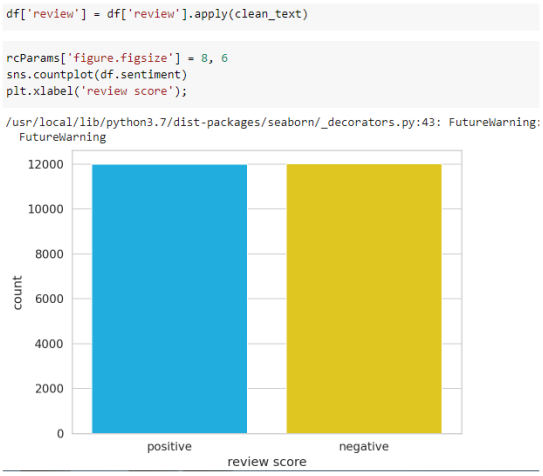

Fig 3.1 Number of positive and Negative reviews

4. RESULTS AND DISCUSSION

Table 4.1 shows that for the data set XLNET gives the better accuracy. Recurrent Neural Network (RNN) is a pre-trained word embedding deep learning model, but it learns the context only from left to right.[15] [16] The other two pre-train language model learns the contexts from the largest corpus and bidirectional. In addition to BERT uses transformers and attention heads to capture the contexts deeper. XLNET covers the discrepancy of BERT by masking each word and permuting to learn the context. So compare to the pre-trained embedding mode, the pre-trained language models are better classify the reviews.

Model/ Dataset	IMDB	Coursera
RNN	71.1	78.7
BERT	91.2	89.8
XLNET	95.6	96.1

Table 4.1. Accuracy of the models RNN, BERT, XLNET

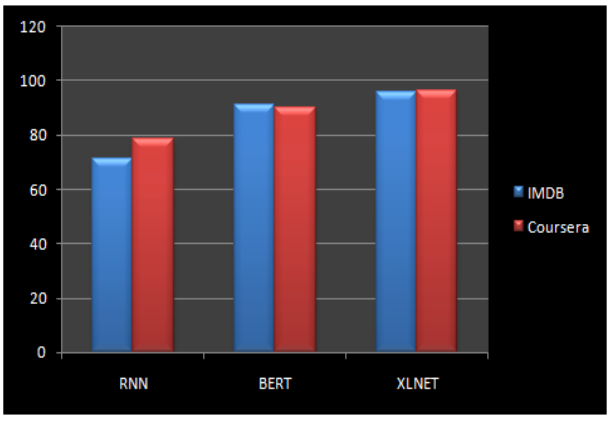

Figure 4.1 Comparisons of Results: RNN, BERT, XLNET on IMDB and Coursera dataset

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7598 - 7604
Received 25 April 2021; Accepted 08 May 2021.

5. Conclusion

The main purpose of pre-trained language models is to learn linguistic features, which are very useful in downstream tasks like question answering, next sentence prediction, sentiment analysis, named entity reorganization etc., In our Experiment The two such models BERT and XLNET to understand the context of the language. BERT uses 15% masking approach with bidirectional way by suffering from the dependency between the masking word and the other words. Whereas the XLNET model uses Autoregressive permutation with factorization order which covers masking all the words. Among the two model, that we have experimented XLNET outperforms BERT for both IMDB and coursera dataset.

Reference

[1] Abuzayed, A. and H. Al-Khalifa (2021). Sarcasm and Sentiment Detection In Arabic Tweets Using BERT-based Models and Data Augmentation. Proceedings of the Sixth Arabic Natural Language Processing Workshop.
[2] Alshahrani, A., M. Ghaffari, K. Amirizirtol and X. Liu (2020). Identifying Optimism and Pessimism in Twitter Messages Using XLNet and Deep Consensus. 2020 International Joint Conference on Neural Networks (IJCNN), IEEE.
[3] Azeemi, A. H. and A. J. a. p. a. Waheed (2021). "COVID-19 Tweets Analysis through Transformer Language Models."
[4] Chriqui, A. and I. J. a. p. a. Yahav (2021). "HeBERT & HebEMO: a Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition."
[5] Devlin, J., M.-W. Chang, K. Lee and K. J. a. p. a. Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding."
[6] Garcia-Silva, A., C. Berrio and J. M. Gómez-Pérez (2019). An empirical study on pre-trained embeddings and language models for bot detection. Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019).
[7] Lehečka, J., J. Švec, P. Ircing and L. Šmídl (2020). BERT-Based Sentiment Analysis Using Distillation. International Conference on Statistical Language and Speech Processing, Springer.
[8] Li, X., X. Fu, G. Xu, Y. Yang, J. Wang, L. Jin, Q. Liu and T. J. I. A. Xiang (2020). "Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis." **8**: 46868-46876.
[9] Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross and K. J. J. I. j. o. l. Miller (1990). "Introduction to WordNet: An on-line lexical database." **3**(4): 235-244.
[10] Munikar, M., S. Shakya and A. Shrestha (2019). Fine-grained sentiment classification using bert. 2019 Artificial Intelligence for Transforming Business and Society (AITB), IEEE.
[11] Mustapha, M., K. Krasnashchok, A. Al Bassit and S. Skhiri (2020). Privacy Policy Classification with XLNet (Short Paper). Data Privacy Management, Cryptocurrencies and Blockchain Technology, Springer**:** 250-257.
[12] Narayanaswamy, G. R. (2021). "Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis."
[13] Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. J. a. p. a. Zettlemoyer (2018). "Deep contextualized word representations."
[14] Sun, C., X. Qiu, Y. Xu and X. Huang (2019). How to fine-tune BERT for text classification? China National Conference on Chinese Computational Linguistics, Springer.
[15] Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. J. a. p. a. Le (2019). "Xlnet: Generalized autoregressive pretraining for language understanding."
[16] H. Muthukrishnan, C. P. Thamil Selvi, Dr. M. Deivakani, V. Subashini, Savitha N. J., S. Gowdham Kumar, Aspect-Based Sentiment Analysis for Tourist Reviews, Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 3, 2021, Pages. 5183-5194