Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)
IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2

A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models

Leeja Mathew
School of Computer Sciences
Mahatma Gandhi University
Kottayam
Kerala
leejarejibpc@gmail.com

Bindu V R
School of Computer Sciences
Mahatma Gandhi University
Kottayam
Kerala
binduvr@mgu.ac.in

***Abstract*— Research and industry are becoming highly interested in automatically analyzing the opinion of general public from social networks with respect to a particular subject. Extracting the polarity from these data is always remaining as a significant bottleneck. Pre-trained models built on deep learning architecture can achieve this task in an effective manner by using transfer learning approach. Since it is difficult to develop a model from scratch, due to time constraints or computational limits, pre-trained models with vast potential and possibilities were introduced. They provide a benchmark to either improve the existing model or test the developed model against it. This paper discusses about various word embedding methods used for sentiment analysis followed by an overview on state-of-the-art pre-trained models used for natural language processing, which is commonly used in the process of sentiment analysis. Experimental results of two state-of-the-art pre-trained models are also analyzed.**

Keywords*— *Sentiment analysis; deep learning; pre-trained models; natural language processing; transfer learning

I. INTRODUCTION

Sentiment analysis is widely used for analyzing opinions of people towards various entities such as products, services, topics etc. and there by evaluating customer satisfaction. It is also named as opinion mining which consists of large problem space. Sentiment analysis includes Natural Language Processing (NLP), text analysis and computational linguistics for identifying and extracting subjective information in source materials – for example, checking whether a review is positive or negative. Analysis of sentiments, extraction of opinion, mining sentiments and opinion, subjectivity analysis, emotion analysis, review mining etc. is different names used for slightly different tasks.

Natural language processing was initially introduced for text pre-processing [1]. NLP applications have become very popular in these days. Transfer learning, in NLP context [2], is capable of training a model on one task and then transforms this for performing various NLP functions on different tasks. It is implemented through pre-trained models. A pre-trained model can be implemented in our task instead of building a new model from beginning [3]. A bit of fine-tuning of this model will save computational resources and time. Various pre-trained models are available which perform different tasks.

This paper is organized as follows. Section II covers the related work. Section III includes brief description of various pre-trained models. In section IV, a comparative study of pre-trained model is performed and conclusion is included in section V.

II. RELATED WORK

Natural language processing of Sentiment Analysis using the pre-trained models has become widely popular. This survey includes word embedding techniques in first phase and pre-trained models in second phase for sentiment classification.

In sentiment analysis, accuracy of sentiment classification improves by including word embedding methods. Word2Vec and GloVe are currently used methods for this purpose which converts words into vectors. However, these methods are not considering sentiment information of texts. There needed large corpus of texts for generating and training exact vectors. These are taken as inputs of deep learning models. Even though the corpus is small sized, researchers often use the pre-trained word embedding. Training has done on other large text corpus. This leads to an increasing accuracy. Rezaeinia et al.[4] propose a novel method named improved word vectors (IWV). IWV is based on Part-of-Speech (POS) tagging, lexicon-based and Word2Vec/GloVe methods. The method is effective with various deep learning models using sentiment dataset.

Semantic and syntactic information from context is widely used in NLP tasks. The method for context based word

embedding is not sufficient to extract better sentiment information. Words having an opposite sentiment polarity (e.g. happy and sad) with similar vector representations reduce sentiment analysis performance. Yu Liang-Chih et al. [5] propose a modified word vector model which is applicable to any pre-trained word vectors. Experimental results based on Stanford Sentiment Treebank (SST) shows that the proposed method can increase conventional word embedding.

The fact that semantically similar vectors for expensive annotations can be retrieved from large un-annotated corpora is the major benefit of word embedding. Various word embedding models like Continuous bag of words and Skip gram which is included in word2vec, Glove (Global Vectors for word representation) and Hellinger PCA (Principal Component Analysis) are compared by Bhoir Snehal et al. [6]. It is performed based on different parameters like training data size, basic over-view, relation of target and context words etc.

Text pre-processing is the preliminary phase in a Natural Language Processing (NLP) system which improves the performance. Camacho-Collados Jose and Mohammad Taher Pilehvar [7] examine the impact of simple text pre-processing decisions. The dataset for evaluation is based on text categorization and sentiment analysis. The importance of inconsistency through pre-processing is mentioned in the experiment. This shows the importance of pre-processing step in the pipeline. They develop the best pre-processing practices for training word embedding.

There is a dramatic growth in the popularity of word embedding which has the ability to capture semantic information from massive amounts of textual content. So many tasks in NLP have tried to take advantage of this technique. Iacobacci Ignacio et.al [8] study how word embedding can be used in word sense disambiguation (WSD). They propose different methods in a newly supervised WSD system architecture, and evaluate the performance in a deep analysis with different parameters.

Jianqiang Zhao et al. [9] had introduced an unsupervised word embedding on a large twitter corpora. In this method, co-occurrence statistical characteristics and latent contextual semantic relationships between words in tweets were used. The sentiment classification labels are predicted from deep convolution neural network by applying sentiment feature set which is formed by integrating n-grams and word sentiment polarity score of tweets for training. Comparison is made with baseline model on five Twitter datasets.

Polarity detection for applications ranging from analysis of product feedback to understanding of user statement has been a hot task in NLP. Convolutional neural network(CNN) architecture perform efficiency in NLP tasks compared to machine learning approaches like SVM, Naive Bayes, recursive neural network and auto-encoders. The performance of CNN classifier has increased by adding the width of convolutional filter functions in N-grams model. Gao Yazhi et al. [10] check the advantage of using different filter lengths and thereby implement an Adaboost method which combines different classifiers with respective filter sizes.

Detection of sarcasm; which change the polarity of positive sentence and vice versa; is an important task in natural language processing especially in sentiment analysis. Sarcasm detection is considered as the primary task in text categorization problem. It requires a deeper understanding of natural language. Poria Soujanya et al. [11] developed models for extracting sentiment, emotion and personality features based on a pre-trained convolutional neural network for sarcasm detection.

The dominant sequence transduction models include an encoder and decoder which consist of either convolutional neural or complex recurrent networks [12]. An attention mechanism has been added to this model. Vaswani Ashish et al. propose a new architecture named, Transformer, based on all the above. The experiments show that this model on two machine translation tasks takes less training time. The model achieves 28.4 BLEU (Bilingual Evaluation Understudy) scores — a standard metric for evaluating a generated sentence to a reference sentence) on the WMT 2014, an English to-German translation dataset. This improves the existing results by over 2 BLEU and achieves a score of 41.0. They take 3.5 days on eight GPUs for training.

The above discussed works include word embedding pre-trained on large amounts of unlabelled data by using the algorithms such as word2vec and Glove for initializing the initial layer of a neural network. And the rest of the network is then trained on data of a particular task. Many of the current models for supervised NLP tasks are models pre-trained on language modelling *(which is an unsupervised task)*, and then fine turned in a supervised manner to a specific task. The remaining section of this paper describes such pre-trained models.

III. TYPES OF PRE-TRAINED MODELS

Pre-trained transformer models are mainly used in transfer learning. Fig. 1 shows the process of transfer learning which uses a large generic model pre-trained on lots of text and can be later trained on smaller datasets for specific tasks.

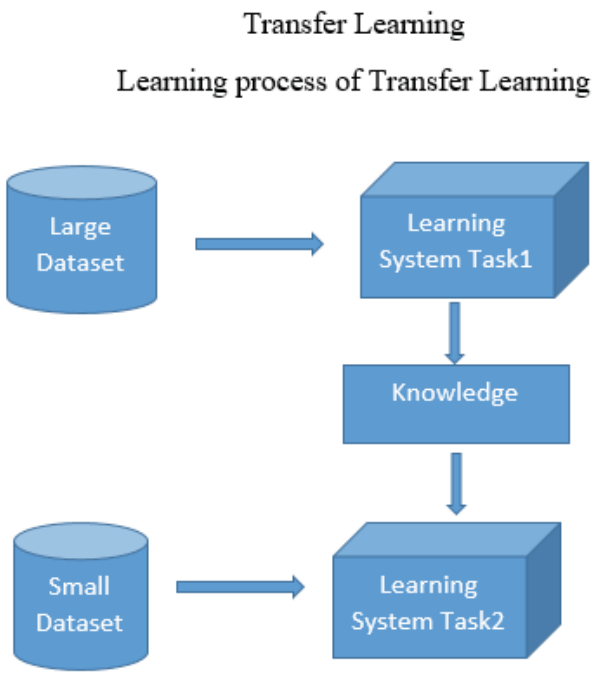

Fig 1. Transfer Learning

The commonly used multi-purpose model in NLP applications include machine translation, question answering systems, chatbots, sentiment analysis etc. The main component of these multi-purpose NLP models are language modelling which will predict the next word or character in a sequence using transformer model. The various transformer models include ULMFiT, Transformer, OpenAI's GPT-2, BiGRU, Google's BERT, Transformer-XL and XLNet.

1. ULMFiT

ULMFiT(Universal Language Model Fine-Tuning) released in the year 2018 achieves good results using new NLP techniques. In this method, after training on WIkitext 103 dataset, a fine-tuning using this model to a new dataset is applied.

Anand Sarthak et al. [13] present an approach and the system description for SubTask A of SemEval 2019 Task 9 related to Suggestion Mining from Forums and Online Reviews. The authors evaluate whether a sentence consists of a suggestion or not. They suggested a model based on Universal Language Model fine tuned for Text Classification. Various pre-processing techniques are applied before training the language and the classification model. The results obtained are analysed using the trained model by achieving F1 score of 0.7011.

2. Transformer

The Transformer model plays a vital role in the recent developments in NLP. Earlier, Usage of Recurrent Neural Networks (RNN) were being implemented in machine translation as well as question answering systems. The new architecture achieve better performance than RNNs and CNNs by reducing the resources for training computation. According to Google, Transformer "applies a self-attention mechanism which directly models relationships between all words in a sentence, regardless of their respective position". For example, consider the sentence "She found the shells on the bank of the sea."[3] The model needs to know that the "bank" represents the sea shore and not a money transaction place. Transformer model can easily understand this in a single step.

Natural language understanding involves a wide range of diverse tasks such as semantic similarity assessment, document classification, textual entailment and question answering. Since unlabelled structured set of texts are plenty and labelled dataset are in shortage, it is a challenge for differently trained models to act perfectly. Radford Alec et al. [14] illustrated the results of these jobs as generating pre-trained model on different structured unlabelled texts and consequently well trained (refined) on each particular job. They made use of job oriented conversions in better training to get useful changes by reducing the sample structure. To show the importance of the results they proved their method on variety of benchmarks. The task-agnostic model they proposed showed better results for 12 tasks. In Stories Cloze Test for commonsense reasoning, RACE for question answering and MultiNLI for text entailment they made an increase of 8.9, 5.7 and 1.5 percentage respectively.

3. OpenAI's GPT-2

GPT-2, a transformer based model was trained to predict the next word in Internet text about 40GB[3]. GPT-2 displays a broad set of capabilities like the ability to generate conditional synthetic text samples of exceptional quality. GPT-2 also outperforms on Wikipedia, news or books compared to other models. GPT-2 can achieve great score for learning on language tasks like question answering, reading comprehension, summarization and translation from the raw text.

Radford Alec et al. [15] demonstrate language models on WebText consisting of millions of webpages

without any explicit supervision. In question answering, this model can achieve 55 F1 score based on the CoQA dataset. And the performance is improved by log-linear fashion. The largest model of GPT-2 contains 1.5B parameter Transformer that attains 7 tested language modelling datasets. These findings lead a promising path for building language processing systems.

4. BiGRU

The short-term memory problem of RNN(Recurrent Neural Networks) is solved by Bidirectional Gated Recurrent Unit(BiGRU) pre-trained model. Predicting a paragraph of text can efficiently be done in this model. This model has internal mechanisms called gates that can regulate the flow of information. It is used in state-of-the-art deep learning applications like speech synthesis, speech recognition, natural language understanding etc.

Sentiment analysis based on text is essentially part of NLP. Sentiment classification process extracts features through models. Comment text is used for classification. No fixed format for such text and sentiment feature information is dispersed in this text. Learning of sentiment classification is more complex in this model. The established model contains fine-grained feature extraction on BiGRU and attention. Using skip-gram model, the vocabulary is vectorised. Noise is filtered by Naive Bayes algorithm. Finally, feature extraction is done by BiGRU and fine-grained attention. Feng Xuanzhen and Xiaohong Liu[16] propose fine-grained attention model based on a long review. Here attention layer focuses on features in different levels - word, sentence and paragraph. JD review and IMDB datasets are used for validation.

5. Google's BERT

In Bidirectional Encoder Representations (BERT) , context of a word in text considers from left and right sides of a word[3]. It differs from other models which is unidirectional and gains better result. Designing of this model is for doing multi-task learning which performs different NLP tasks simultaneously. This pre-training model is the first unsupervised, bidirectional and deeply system for NLP tasks. BERT outperforms other models on 11 Natural Language Processing (NLP) tasks.

The rapid growing of biomedical documents leads to biomedical text mining. Extraction of biomedical literature is popular among researches due to its valuable. The deep learning play a vital rule for the improvementt of effective biomedical text mining models. Lee Jinhyuk et al. [17] analyse how BERT can be adapted for biomedical corpora and named new model as BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining. BioBERT achieves better results compared to BERT and previous models, which exhibited in this manner : 0.62% and 2.80% F1 score improvement in biomedical named entity recognition and relation extraction respectively. In biomedical question answering, they get 12.24% MRR improvement.

6. Transformer-XL

Transformer-XL is about 1800 times faster than Transformer. It is a new architecture for NLP that helps machines understand and no fixed-length limitations [3]. One of the key ideas in the Transformer XL is relative positional encodings. Other than single embedding, the Transformer XL computes an embedding between any two tokens which leads to the attention between the two words.

Transformers have the advantage of studying longer-term dependencies and are restricted by static length scenario of putting up language modelling. Dai Zihang et al. [18] propose a model that includes segment level repetition technique and new positional encoding method that makes learning dependency based on static length and also not changing the temporal coherence. Their method gave an answer to the context of fragmentation problem as well as it empowers longer term dependency. It outperforms 80% and 450% longer than RNN and vanilla Transformers. The following results show their achievement-increase on enwiki8 by 0.99, text8 by 1.08, WikiText-103 by 18.3 and Penn Treebank on 54.5.

7. XLNet

XLNet, a generalized autoregressive pre-training model, enables learning bidirectional contexts. Maximizing the expected likelihood over all permutations of the factorization order can be possible in this model. Its autoregressive formulation outperforms BERT. Furthermore, XLNet integrates ideas from Transformer-XL and BERT. Experimently, XLNet outperforms BERT and other models on 20 tasks. XLNet proposed a new objective called Permutation Language Modeling in the pre-training phase.

Yang Zhilin et al. [19] conclude that XLNet is a generalized Auto Regressive(AR) pre-training method. Its objectives include permutation combination in language modeling. NLNet architecture integrates Transformer-XL and two-stream attention mechanism. This model achieves better results for various tasks with better improvement. XLNet can be widely expanded for the tasks in vision and reinforcement learning. The authors evaluate on IMDB, Yelp-2, Yelp-5, DBpedia, AG, Amazon-2, and

Amazon-5 datasets and can achieve better results by decreasing error rate by 16%, 18%, 5%, 9% and 5% respectively compared to BERT.

IV. COMPARATIVE STUDY OF PRE-TRAINED MODELS

We analysed movie review dataset with 25000 samples in the latest pre-trained models named BERT and BiGRU. The result obtained is shown in Table 1. BERT pre-trained model got more accuracy with 94.08%. Keras ktrain library is used for implementation. One cycle policy is used for finding learning rate. GPU based execution improves computational speed. Epoch based accuracy and loss are also displayed in Fig. 2 and Fig. 3 respectively, for BERT and Fig. 4 and Fig. 5 illustrates the same for BiGRU model.

TABLE1. COMPARISON OF MODELS

Dataset	*Model Name*	*Loss*	*Accuracy*
IMDb	*BERT*	*.2301*	*.9408*
IMDb	*BiGRU*	*.5681*	*.7206*

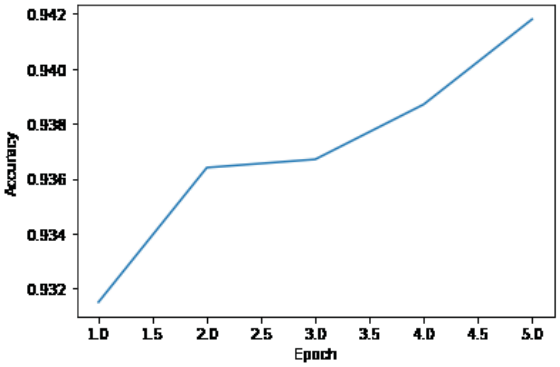

Fig 2. BERT – Accuracy

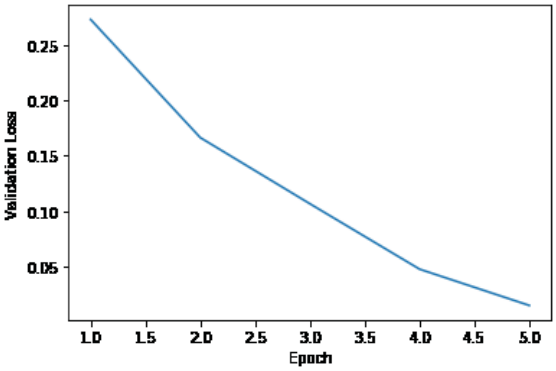

Fig 3. BERT - Loss

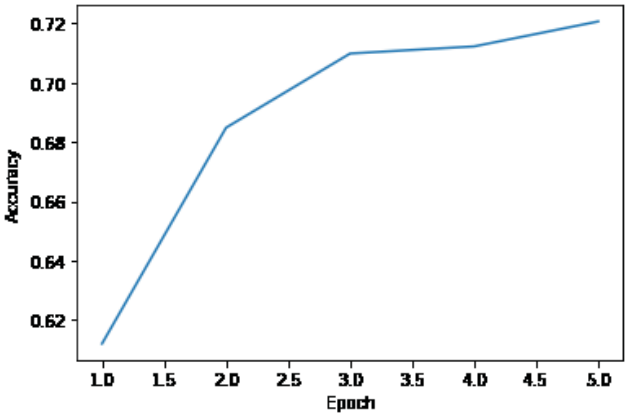

Fig 4. BiGRU – Accuracy

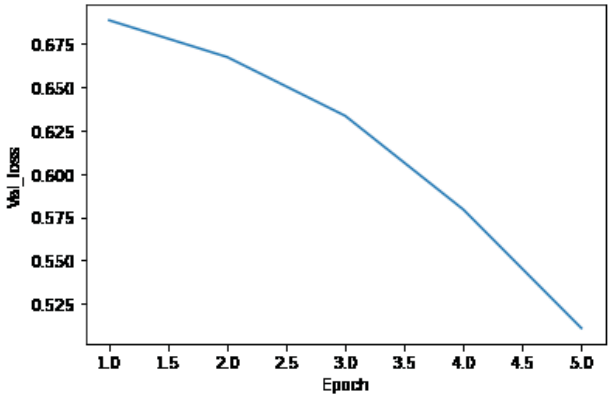

Fig 5. BiGRU – Loss

V. CONCLUSION

Extracting polarity from dataset in Sentiment Analysis is a challenging phase. Pre-trained models can achieve this task easily without the need to build a model from scratch. Pre-trained models serve as standards for either improving an existing model or testing our own models against it. Different pre-trained models like ULMFiT, Transformer, OpenAI's GPT-2, BiGRU, Google's BERT, Transformer-XL and XLNet are presented in this paper. Even though XLNet outperforms in more tasks compared to other models, its computational complexity is very high. It takes more training time due to its autoregressive bidirectional nature and also needs better hardware. Performance evaluation with the two latest models, BiGRU and BERT using IMDB dataset shows an improvement in accuracy with number of epochs with BERT exhibiting superior performance. We can further extend the study by evaluating and comparing the performance of various models with more datasets.

References

[1] SunShiliang Chen Luo and Junyu Chen, "A review of natural language processing techniques for opinion mining systems.", Information fusion 36, pp. 10-25, 2017

[2] Houlsby Neil et al. "Parameter-efficient transfer learning for NLP." arXiv preprint arXiv:1902.00751, 2019.

[3] https://www.analyticsvidhya.com

[4] RezaeiniaSeyed, Mahdi Ali Ghodsi and RouhollahRahmani. "Improving the accuracy of pre-trained word embeddings for sentiment analysis." ,arXiv preprint arXiv:1711.08609, 2017.

[5] Yu Liang-Chih et al. "Refining word embeddings for sentiment analysis.", Proceedings of the 2017 conference on empirical methods in natural language processing. 2017.

[6] BhoirSnehal, TusharGhorpade and Vanita Mane., "Comparative analysis of different word embedding models.", 2017 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2017.

[7] Camacho-ColladosJose and Mohammad TaherPilehvar., "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis." ,arXiv preprint arXiv:1707.01780 , 2017.

[8] Iacobacci Ignacio, Mohammad TaherPilehvar and Roberto Navigli., "Embeddings for word sense disambiguation: An evaluation study.", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)., 2016.

[9] Jianqiang Zhao, GuiXiaolin and Zhang Xuejun., "Deep convolution neural networks for twitter sentiment analysis.", IEEE Access 6, 2018, pp. 23253-23260.

[10] GaoYazhiet al., "Convolutional neural network based sentiment analysis using Adaboost combination." ,2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016.

[11] PoriaSoujanyaet al., "A deeper look into sarcastic tweets using deep convolutional neural networks.", arXiv preprint arXiv:1610.08815, 2016.

[12] Vaswani Ashish et al., "Attention is all you need.", Advances in neural information processing systems, 2017.

[13] Anand Sarthak et al., "Suggestion Mining from Online Reviews using ULMFiT.", arXiv preprint arXiv:1904.09076, 2019.

[14] Radford Alec et al., "Improving language understanding by generative pre-training." ,URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.

[15] Radford Alec et al. ,"Language models are unsupervised multitask learners." ,OpenAI Blog 1.8, 2019.

[16] Feng Xuanzhen and Xiaohong Liu., "Sentiment Classification of Reviews Based on BiGRU Neural Network and Fine-grained Attention.", Journal of Physics: Conference Series. Vol. 1229. No. 1. IOP Publishing, 2019.

[17] Lee Jinhyuk et al. ,"Biobert: pre-trained biomedical language representation model for biomedical text mining.", arXiv preprint arXiv:1901.08746, 2019.

[18] DaiZihang et al., "Transformer-xl: Attentive language models beyond a fixed-length context." ,arXiv preprint arXiv:1901.02860, 2019.

[19] Yang Zhilin et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding.", arXiv preprint arXiv:1906.08237, 2019.