Research Article

Bidirectional Encoder Representations from Transformers (BERT) Language Model for Sentiment Analysis task: Review

Ms. D.Deepa[1], Dr.A.Tamilarasi[2]

[1],Department of Computer Science and Engineering, Kongu Engineering Collge, Perundurai, Tamilnadu,

[1]deepa@kongu.ac.in

[2] Department of Computer Applications, Kongu Engineering Collge, Perundurai, Tamilnadu
drtamil@kongu.ac.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract. The latest trend in the direction of sentiment analysis has brought up new demand for understanding the contextual representation of the language. Among the various conventional machine learning and deep learning models, learning the context is the promising candidate for the sentiment classification task. BERT is a new pre-trained language model for context embedding and attracted more attention due to its deep analyzing capability, valuable linguistic knowledge in the intermediate layer, trained with larger corpus, and fine-tuned for any NLP task. Many researchers adapted the BERT model for sentiment analysis tasks by influencing the original architecture to get better classification accuracy. This article summarizes and reviews BERT architecture and its performance observed from fine-tuning different layers and attention heads.

Keywords: contextual representation, language models, linguistic knowledge, sentiment analysis, BERT, fine-tuning, attention heads

1. Introduction

At the moment, the Sentiment Analysis task is being viewed as an important strategy, which influences most of the business to improve their profit by perceiving customer's opinions (Saura, Palos-Sanchez et al. 2019). This task can be accomplished by Natural Language Processing (NLP) and Machine Learning (ML) approaches which use the statistical frequency-based embeddings like TF-IDF, Countvectorizer and Bag of words. All these approaches fail to capture high-level semantic meanings behind text data. (Zhao and Mao 2017).

NLP makes the machines to understand the language. The complication in NLP is to symbolize the knowledge from the plentiful text from online (Chowdhary 2020). Even though the most of the machine learning research works on sentiment classification proved outperforming accuracy (Gamal, Alfonse et al. 2019, Patel and Chhinkaniwala 2019, Rietzler, Stabinger et al. 2019, Singh and Goel 2019, Wu, Weld et al. 2019, Kanakaraddi, Chikaraddi et al. 2020, Wang and Lin 2020), all it needs well-built training data (Roh, Heo et al. 2019), massive human efforts to represent feature's contextual information (BİRİCİK, Diri et al. 2012, Dzisevič and Šešok 2019). The feature engineering process was simplified by the neural network word embedding build using Pre-trained algorithms, which better represents the shallow semantic but fails to represent the different contexts of the word (Horn 2017). The conceptual representation of text may well facilitate to feed the implicit sentiment conveyed with specific other explicit concepts (Cambria 2013).

Sentiment Analysis requires syntactic and semantic rich information for polarity classification (Rezaeinia, Rahmani et al. 2019). Each word in the sentence sequence may have different contexts depends on the surrounding words (Tang, Wei et al. 2014). The various Deep learning architectures like CNN, RNN, LSTM are attempted to learn the context of the word sequence with long term dependency (Rojas-Barahona and Compass 2016). Investigations have been done to bidirectionally (Wang, Jiang et al. 2016) learn the context of the word observing the long term dependency. Finally, concluded that the sentiment analysis needs a model to learn the context of words by holding the whole sentence context without any limitation in the length and bidirectionally looking at the context of the surrounding words parallelly.

(Tenney, Xia et al. 2019) Pre-trained Language models are different from traditional pre-trained embeddings in the way it learns word embedding for the given context. The pre-trained language model builds feature representations from unsupervised larger corpus combined with finetuning, which solves labeled data deficiency for training and domain-specific context. BERT (Devlin, Chang et al. 2018), Roberta (Cui, Che et al. 2019), DistilBERT (Sanh, Debut et al. 2019), XLNet(Cui, Che et al. 2019, Sun, Wang et al. 2019), ENNIE Sun, Wang et al. 2019), ERNIE 2.0 (Sun, Wang et al. 2020) are the most popularly used bidirectionally Pre-trained language models. (Mao, Wang et al. 2020). BERT is encompassing 12 attention layers. Embedding is 512 in size.it can be used for downstream tasks like Question answering, NER, Sentiment Analysis, etc., BERT is fine-tuned to improve the classification accuracy by introducing one more output layer depends on the downstream application (Devlin, Chang et al. 2018) and has achieved amazing results in many language understanding tasks.

To apply pre-trained representations to NLP tasks the two main strategies are obtainable. The feature-based approach, which uses the pre-trained representations as a additional features to the downstream task. Or the fine-tuning-based approach, which trains the downstream tasks by fine-tuning pre-trained parameters. The contributions of this survey are to investigate the use of BERT's feature-based approach and fine-tuning-based approach for the sentiment analysis task discuss the further research work on the sentiment analysis section by excluding other sections specified in the review from paper, Pre-trained Models for Natural Language Processing: A Survey (Qiu, Sun et al. 2020, Zhao, Lin et al. 2020)

2. Text Feature Representation

Sentiment classification accuracy highly relays on feature extraction and representation methods (Rahman, Biplob et al. 2020). Also, all the contextualized information about the word is to be inferred and utilized to give meaningful classification. So the recent sentiment classification task is done using a pre-trained language model rather than statistical and Pre-trained embedding.

2.1 Classical Feature Representation

The recent work on statistical feature embeddings like TF-IDF (Alomari, ElSherif et al. 2017, Dey, Jenamani et al. 2017, Das and Chakraborty 2018), (Kim, Seo et al. 2019) (Rahman, Biplob et al. 2020), Countvectorizer (Tripathy, Agrawal et al. 2016, Vijayaraghavan and Basu 2020) and Bag of words(El-Din and Applications 2016, Cummins, Amiriparian et al. 2018, Karimi, Rossi et al. 2020, Raju and Sridhar 2020) .fails to capture high-level semantic meanings behind text data (Zhao and Mao 2017). The introduction of Deep Learning (DL) uses neural network architecture(Erhan, Courville et al. 2010), which shaped feature embedding automatic with pre-trained model word2vec (Mikolov, Sutskever et al. 2013) can use either CBOW or skip-gram architecture to build distributed feature representation from the unsupervised corpus (Tenney, Xia et al. 2019). The other pre-trained models are Glove (Pennington, Socher et al. 2014) and fasttext (Bojanowski, Grave et al. 2017). All these models differ in its performance depends on the vocabulary size, dealing with Outofvocabulary (OOV) and mainly representing the context of the text. These models give only one embedding vector for a word, even though the word has participated in different contexts (Tenney, Xia et al. 2019).

The various Deep learning architectures like CNN, RNN, LSTM(Tang, Qin et al. 2016, Xu, Meng et al. 2019) (Agarwal, Mittal et al. 2015, Rojas-Barahona and Compass 2016, Jianqiang, Xiaolin et al. 2018, Alharbi and de Doncker 2019, Ethayarajh 2019, Feng, Wang et al. 2019, Han, Bai et al. 2019, Hung 2020, Kumar, Srinivasan et al. 2020, Ombabi, Ouarda et al. 2020, Tran and Phan 2020) are attempted to learn the context of the word sequence with long term dependency. Investigations have been done to bidirectionally learn the context of the word observing the long term dependency and BI-LSTM (Melamud, Goldberger et al. 2016) better learns the context of the words (Kiperwasser and Goldberg 2016, Mousa and Schuller 2017, Li, Yang et al. 2019, Xu, Meng et al. 2019, Kumar, Verma et al. 2020, Li, Qi et al. 2020, Wang, Zhu et al. 2020) because of the left and right direction but not simultaneously (Devlin, Chang et al. 2018).

2.2 Feature representation using Pre-trained Language models

The language model better learns the context by predicting the next word in the sequence. Deep conceptual feature representation is given by the pre-trained language model ELMo (Peters, Neumann et al. 2018) by combining the weight of all the internal layers and uses two BiLSTM to provide shallow word embedding. ULMFiT (Howard and Ruder 2018) is a unidirectional language model to provide word embedding. GPT2 (Radford, Wu et al. 2019), BERT (Devlin, Chang et al. 2018), XLM all use transformers and support subword embedding. Among these transformer models, BERT pre-trained language model learns the context bidirectionally at the same instant. The common idea about all the Pre-trained models is the unsupervised Pre-training and transformers eradicate the limitation in learning the context with longer-term dependencies and restrict the language model fixed-length sentences (Mathew and Bindu 2020).

2.3 Language Models

Language modeling is the task of assigning a probability to a given sequence of words occurring in a sentence. Language models determine the probability of the word by analyzing text corpus (Jelinek 1990). An algorithm interprets the text corpus essentially the features and characteristics of the language and uses those features to understand new phrases. Recently, the use of neural networks in the development of language models has become very popular. Domain adaptation is a necessary procedure to cast the general domain sentiment lexicon or sentiment classifier for practical use. Neural network approaches are achieving better results than statistical language models in most of the challenging Natural language tasks to allow for better modeling of longer contexts (Arisoy, Sainath et al. 2012).

3. BERT Architecture

BERT is a deep bidirectional transformer architecture introduced by Google. It supports multilingual universal language representation of 104 languages. BERT is pre-trained from unlabelled Wikipedia (2,500M words), and Book corpus (800M words) to obtain contextual embeddings. The BERT model can be used in two steps: Pre-training and Fine-tuning. The two pre-trained models of BERT are BERTBASE (Layers=12, Hidden states`=768, Attention Heads=12, Total Parameters=110M) and BERTLARGE (Layers=24, Hidden states =1024, Attention Heads =16, Total Parameters=340M). The maximum sequence length of the input sentence is 512 tokens and BERT uses a wordpiece tokenizer strategy for Tokenization. The 'cased' model of the BERT preserves the original text cases for tokenization whereas the 'uncased' BERT model changes the text into lowercase before tokenization. The vocabulary size of BERT-Base, uncased is of 30,522 words.

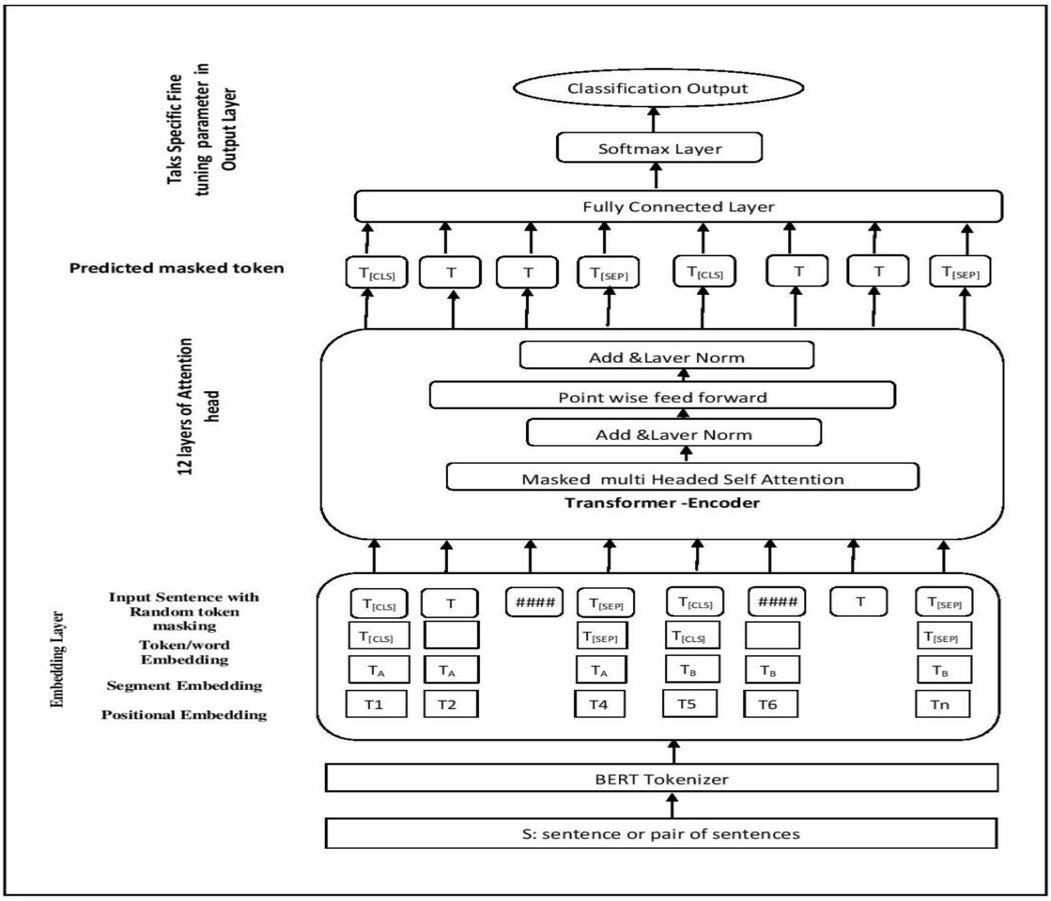

Figure 1. BERT Architecture

3.1 BERT Character/sub-word Embedding

The tokenization process involves splitting the input text into a list of tokens that are available in the BERT dictionary also called vocabulary. To deal with out of vocabulary, BERT uses a technique called Byte Pair Encoding (BPE) (Gage 1994) based WordPiece tokenization. BPE merges frequently occurring subword pairs which are composed of multiple sentences and character-level /sub-word embeddings. Padding is done with these sub-word sequences to make them into a fixed length of BERT vocabulary. To generate an approximate vector for the original word, the three acceptable alternative strategies used are averaging the embeddings, summation of subword embeddings and taking the last token. Average sub-word embedding vectors are the most adapted strategy in BERT. Tokenized text is checked against part of token or padding using attention masks.

Context embeddings constructed can be word-level similarity embeddings or sentence-level similarity embeddings. The word-level similarity is not meaningful with BERT embeddings because these are contextually dependent, and these context changes depend on the many sentences it appears. So sentence embeddings similarity comparison is acceptable a single sentence queried against a dataset of other sentences to find the most similar sentence. BERT construct different context vector of the word for its different context.

3.2 BERT pre-training task

BERT pre-training task includes two new unsupervised predictive tasks: Masked Language Model (MLM), hides certain words during training and tries to determine the missing word and Next Sentence Prediction (NSP) whether the second sentence came after the first sentence. The transformer encoder reads the entire sequence of words at once and learns the context of a word based on all of its surrounding words. The contextual representation is a high-dimensional vector space of word and sentence. The input representation of each sentence is given by the summation of positional embedding specify the position of the token within the sequence; segment embedding distinguishes one sentence from another and word embedding. A special classification token ([CLS]) is added at the stating of each sentence and in the final hidden state aggregate sequence representation of this token is used as for classification tasks. Another special token, [SEP] is added, to mark the end of a sentence or the separation between two sentences. The summation of all these embeddings forms the input layers to the BERT.

3.3 BERT fine-tuning

With the pre-trained BERT model, an untrained layer is added at the top of the output layer. Then train the new model for our classification task. First, the Bottom layers of the pre-trained BERT model were already encoded with extensive information about the language, and this encoded information is used a feature for the classification task. The fine-tuning tasks take much less time to train a new set of training data with these features and no need to learn the language from scratch. A major drawback of the NLP task is that these models are requested to build, which in turn to get improved accuracy, a massive amount of time and energy to train the model. But the outsized pre-trained features embeddings obtained from BERT allow us to fine-tune our task with a much smaller dataset and minimal task-specific adjustments to give a good performance.

3.4 Transfer Learning and BERT domain adaptation

In sentiment analysis and its applications domain adaptation has been identified as a key issue, this is because that many topics are characterized by their sublanguages, such as special terminologies and jargons. Being unfamiliar with these topics or domains will lead to misunderstanding of the sentiment conveyed. For the same reason, it has been observed that the performances of sentiment analysis usually drop when using sentiment lexicons of the general domain or other irrelevant domains. Therefore, the direct use of lexical resources is often unfeasible. All the time constructing the model with domain-specific data from the scratch will consume not only the time and effort but also suffer from domain-specific labeled training data. So the choice of using a generalized model feature set with a small amount of domain-specific data will make this job easier. Domain adaptation is a necessary procedure to cast the general domain sentiment lexicon to domain-specific lexicon.

Domain adaptation is part of transfer learning, which transfer the knowledge from the source domain to the target domain to bridge the distribution gap between the training data and the test data. Models that are purely trained from training data are applied directly to the test data. The problem arises when the model trained with a large amount of manually labeled data of one domain, are asked to make predictions on examples from a new domain with little or no labeled data. Supervised learning algorithms fix the same distribution drawn from the training data and the test data, Traditional supervised learning cannot achieve high performance on this new domain. BERT resolves these issues with a two-step procedure: self-supervised domain-specific BERT language model finetuning, followed by supervised task-specific finetuning. With the BERT pre-training, select a small amount of annotated domain-specific source data is taken for the fine-tuning. The performance of the model is validated with the new set of test data.

4. Related Work

4.1 Pre-train BERT with the domain-specific corpus

BERT pre-trained model can be used for different 11 downstream NLP tasks and these tasks commonly use all the layer information except the output classification layer. The BERT model is pre-trained on an unsupervised general-domain corpus. BERT can be further pre-trained from the scratch on a new domain-specific corpus with a masked language model and next sentence prediction tasks (Mathew and Bindu 2020).

(Rezaeinia, Ghodsi et al. 2017, Rezaeinia, Rahmani et al. 2019) improved accuracy of sentiment analysis by pre-training the BERT on Google News with about 100 billion words. (Reimers and Gurevych 2019) presented computationally efficient Sentence-BERT (SBERT), which pre-train the BERT with siamese and triplet network structures to build semantically meaningful embeddings. COVID-Twitter-BERT proposes by (Müller, Salathé et al. 2020). They pre-trained BERT with a large corpus of Twitter messages on the topic of COVID-19 collected from January 12 to April 16, 2020, including input sequences size of 30000-word vocabulary. A hierarchical

BERT context representation from different corpus Twitter and Reddit was invented by (Srivastava, Varshney et al. 2020). These context representations are jointly used to find the sarcasm of the utterance.

BERT-based text classification model BERT4TC is proposed by (Yu, Su et al. 2019) to improve the performance of the classification with task-specific knowledge. They constructed three different auxiliary sentences and feed it into the input layer for pre-training and given the classification as a binary sentence-pair task. The model further post-trained to solve the domain challenge. A BERT optimized model RoBERTa (Liu, Ott et al. 2019) with pre-trained with Dutch section of the OSCAR corpus was developed by (Delobelle, Winters et al. 2020) for Dutch-specific language task and showed state of art results on most of the NLP tasks, specifically Dutch sentiment analysis task.

4.2 Taking advantage of the Masked Langauge Model

In BERT basically, 15% tokens of every sentence are made to and predicted during language understanding. The contextual information is generated for these masked tokens only. BERT predicts the masked word with the probability of,

- 80% of the time masked tokens are replaced
- 10% of the time masked tokens are replaced with a random token
- 10% of the time masked tokens are left unchanged and these unchanged tokens are ignored during prediction

When 15% of the tokens are masked in a sentence, the context embedding generated for one masked word in a sentence looks it's surrounding words in the same sentence and don't get knowledge on the other masked tokens. Similarly, the context embeddings are generated for all the other masked tokens in the sentence. Researchers experimented Masking task in different ways to get deeper contextualization information about the tokens. To solve this problem, the RoBERTa model was proposed by (Liu, Ott et al. 2019) with a dynamic masking task and removed the next sentence prediction task from the BERT.To get more contextualization the set of sentences used for pre-training is duplicated 10 times and all the sentences are involved with dynamic masking tasks. They showed state-of-the-art results on standard datasets. (Tian, Gao et al. 2020) proposed Sentiment Knowledge Enhanced Pre-training (SKEP) to build the embedding for the sentiment word by masking automatically extracted sentiment words. (Garg and Ramakrishnan 2020), (Li, Ma et al. 2020) generated adversarial examples from the sentence by Masked Language modeling. The randomly masked tokens are predicated all possible tokens that are grammatically correct and semantically coherent with the context. The newly predicted all context representations of each sentence are used for classification. The examples are generated to resolve the shortage of domain data for classification tasks and thus reduces overfitting.

4.3 BERT with Co-trained knowledge

The pre-trained embeddings of BERT can be used for 11 downstream tasks like sentiment analysis, question answering, named entity recognition, intent detection, etc., But the BERT is lacking of task-specific knowledge and domain-related knowledge. To further improve the performance of the BERT model with additional semantic knowledge with the help of dictionaries and other deep learning models. Bi-LSTM deep learning architecture and Gated Recurrent Unit is mostly adapted with BERT to get more contextualization.

(Rezaeinia, Ghodsi et al. 2017, Rezaeinia, Rahmani et al. 2019) used Part-of-Speech (POS) tagging techniques, lexicon-based approaches, and statistical embeddings as semantic information. The bidirectional GRU model is used to learn the semantic information of the left and right contexts of words (Sun, Gao et al. 2019). (Arase and Tsujii 2019) inject semantic relations through phrases and sentential paraphrases between a sentence pair into a pre-trained BERT model through classification. (Ke, Ji et al. 2019) proposed SentiLR, which acquires context dependency as well as word-level linguistic knowledge from sentiment dictionary SentiWordNet for the part-of-speech tag. (Araci 2019) utilized the benefits of BERT pre-trained models in the finance domain to overcome the overfitting due to the shortcomings of training data. Ngram and long term dependencies information collected through CNN and LSTM models parallelly with BERT pre-training. (Zhang, Wu et al. 2019) developed SemBERT to improve sentiment prediction through modified BERT pre-training using CNN (Convolution Neural Network) to capture semantic information from semantic role labeling. Also tried subword- level masking for Pre-training. (Wei, Liao et al. 2020) proposed a model for implicit sentiment analysis. After BERT pre-training, a Bi-LSTM layer is introduced to capture more contextual information and then the attention layer collects the weights of the keyword during sentence encoding and finds the difference between the multi-polarity attentions with the knowledge of sentiment dictionary.

4.4 Fine-tuning BERT with the Output layer

The contextualized embedding of BERT is used as features and available at the output layer. The [CLS] at this layer collects the classification information of a sentence. During Post-training the additional domain-specific training data is understood in the same way as the pre-training process of BERT. The information collected in [CLS] token of all the sentences can be used for any specific NLP task just by fine-tuning the output layer with all task-specific hyperparameters values. Researchers have fine-tuned the BERT for their sentiment analysis task by adding an output layer and tuning the parameters.

FinBERT, a language model for finance Sentiment classification is conducted by adding a dense layer after the last hidden state of the [CLS] token and followed as a recommended practice for using BERT for any classification task. Then, the classifier network is trained on the labeled sentiment dataset (Araci 2019). Recently fine-tuning the BERT model for the finance domain has gained the attention of the researchers. (Hiew, Huang et al. 2019) used fine-tuned BERT to classify the Chinese stock market unlabelled reviews into three levels as positive, negative and neutral. Numbers of positive, negative and neutral reviews are used to calculate sentiment index and results in enhanced results compared to other deep learning methods. Analyzing the stock market reviews to decide on further investment is simplified by concentrating on negative sentiments. (Zahera, Elgendy et al. 2019) fine-tuned BERT model by adding 10 BERT stacked layers on top of the outputs too for multi-label classification of tweets. Then, we add an extra dense layer with a sigmoid activation function. Further, we used two loss functions (binary cross-entropy, focal loss) to minimize the errors during training our models. (Kocoń, Zaśko-Zielińska et al. 2019) fine-tune the BERT output layer for multiclass classification. (Zhang, Wu et al. 2019) have tried subword- level masking for Pre-training in his SemBERT model.

(Alaparthi and Mishra 2020) fine-tuned the parameters of BERT at the output layer to classify IMDB movie review and solved the overfitting problem. Also, he compared the results of the dictionary approach, machine learning approach and deep learning approach with BERT. The BERT pre-trained model outperformed the other methods was reported. (Reimers and Gurevych 2019) experimented three pooling strategies (CLS, MEAN and MAX) in the SBERT model by adding a pooling layer to the output of BERT / RoBERTa to derive a fixed-sized sentence embedding. (Zhao, Li et al. 2020) modeled RoBERTa to perform sentence pair matching process to match all the entities in the review with the negative sentiments entity list. Considering the number of entities in the list the more models are used for entity detection in the way each model is for each entity. The model identifies more entities will be viewed for decision making. (Sousa, Sakiyama et al. 2019) fine-tuned the BERT model for manually annotated stock new articles and achieved a better F1 Score. (Yang, Xu et al. 2019) BERT is influenced by 30 search terms of Financial and Economic Attitudes Revealed by Search (FEARS) index to get word embeddings. Stock investors are interested in these search terms to look at the review to decide for investment. The embeddings of these search terms are further used by the attention layers to classify the reviews.

(Lin, Madotto et al. 2019) added a linear layer at the top of the BERT for sales prediction of the target product. (Pruthi, Dhingra et al. 2019) proposed word recognition model. (Yin, Meng et al. 2020) developed SentiBERT with the basic BERT pre-training additional phrase-level compositional semantic information collected using an attention mechanism. The labeled phrase-level sentiment learned left out for other tasks. (Wu, Lv et al. 2019) proposed a new conditional contextual masked augmentation on labeled sentences by substituting words in the sentence sequence to overfitting problems. (Azzouza, Akli-Astouati et al. 2019) Concatenate pre-trained representations of word embeddings with the BERT representation method called TwitterBERT to enhance sentiment classification. (Lei, Zhang et al. 2019) improved BERT with Hierarchical Sequence Classification (Silla, Freitas et al. 2011) and Conditional Random Field (CRF) (Wu, Lv et al. 2019) layer to consider the interrelationship between classes and dependency between the tokens. This improved the contextual representation and consent to use for adjacent microblogs. (Ling 2020) used BERT pre-trained on massive Chinese corpus and all the parameters are fine-tuned in the output layer to predict the sentiment of Coronavirus blog posts with greater accuracy.

Sentence-level sentiment analysis was proposed by (Shen, Liao et al. 2019) using BERT to get contextual vector and BiGRU (Bidirectional Gated Recurrent Unit) to perform sentiment analysis. A similar combination is used by (Lu, Zhu et al. 2020) to resolve the ambiguity in Chinese vocabulary. With BERT Embedding, BiGRU is used to integrate the Chinese grammar rules to standardize the output of adjacent positions of the dictionary words. He used four categories of sentiment dictionary as sentiment words, non-sentiment words, degree words, negative words to formulate the regulator. The benefit of BERT pre-trained contextual information was extended to narrative text by (Lyu, Ji et al. 2020). The traditional document classification performs a summation of the polarity of each sentence without considering the context of relevant sentences. In this research, an approach to sentiment analysis of a narrative text is done by combining the contextual representation of the contextually relevant sentences. We use BERT as an encoder to generate sentiment labels for a single sentence. The user reviews may not hold sentiment additionally it includes sarcasm, humor, hate speech. The sentiment label of the user reviews is changed due to these extra elements. To solve this issue (Badlani, Asnani et al. 2019) proposed a two-step model. In the first step, the elements interleaving the label of the sentence are identified and categorized. This information is feed into the sentiment classification task in the second stage.

(Wang, Lu et al. 2020) used BERT to classify COVID related Weibo posts collected from 1 January 2020 to 18 February 2020. After the classification, the classes are summarized on topics. As (Hiew, Huang et al. 2019) did stock prediction with negative sentiments, in this analysis negative sentiments are given more concentration. (Srivastava, Varshney et al. 2020) utilized BERT pre-trained information along with CNN to capture local information, and together this information was used as whole text representation in the attention mechanism and further classification was done.

4.5 Taking advantage of intermediate Layers

The BERT model is composed of 12 or 24 layers of transformer encoder. All these layers are having 12 or 16 attention heads and internal states depend on the BERT model we choose for our classification. Multihead attention is essential to different sections of input in parallel. The attention head in one layer looks for necessary information from the previous selective heads and knowledge of this current head might be looking by selective more than one head in the top layers. All these attention heads and internal states will carry linguistic knowledge like token relevance Pen Tree bank - syntactic dependencies and tagging information where the classification task probe all this knowledge to get benefitted (Clark, Khandelwal et al. 2019).

A language-independent hierarchical multi-layer classifier model was introduced by (Biesialska, Biesialska et al. 2020). This model was adapted to the architecture of the Transformer encoder (Vaswani, Shazeer et al. 2017) with multi-head attention applied parallelly and the order of the words is represented using relative position representations. On top of the self-attention layer, a bi-attention layer is used to bring together the representations from all the heads. Finally, the combined representations were used for sentiment classification. This model was evaluated on sentiment analysis tasks in English, Polish and German language. (Liu, Shen et al. 2019), (Xu, Zhu et al. 2020) also attempted to capture local and global sentiment similarity of words by taking advantage of the attention mechanism.

To inherit the correlation between the aspects terms in the sentence (Hou, Huang et al. 2019) experimented with an ensemble method by integrating Graph Convolution Networks (GCN) and selective self-attention mechanism. Due to the distance between the aspect terms and opinion words are far in GCN and is vulnerability. To overcome this problem in GCN, the aspect term select the opinion word directly, which has top attention weights. Similar work was carried out by (Song, Wang et al. 2019) to capture the dependency between context and target using the Attentional Encoder Network (AEN). (Wang, Shen et al. 2020) experimented with another model R-GAT+BERT (Relational Graph Attention with BERT) to capture the long term dependency between the aspect and target.

Restriction in accessing the local information and long-term dependency in CNN and RNN model is solved by (Lv, Hu et al. 2019) extract-attend-predict model by utilizing the BERT deep self-attention encoder to capturing long-distance words. This model encompasses two additional layers including the BERT-Based Encoding Layer and the sentiment prediction layer. BERT-Based Encoding Layer which performs actual embedding of the sentence and gives the contextual vector representations. In the early Aspect based sentiment analysis task, the aspect terms extracted are not embedded with the sentence and neighbor aspect information. So Aspect Term Extraction Layer was introduced to extract aspect information from the self-attention encoder. Multi-Granularity Attending Layer performs two kinds of attending operations. In coarse-grained attending, find the most relevant sentence with aspect representation. Followed by fine-grained attending, all the similar pair of aspect words and sentence words were found. Finally, the Sentiment Prediction layer predicts the sentiment of all the aspects.

(Sanh, Debut et al. 2019) developed smaller and faster than the original BERT model by reducing 40% of layers called DistilBERT and can be used for sentiment classification tasks and gives 97% accuracy. (Sajjad, Dalvi et al. 2020) experimented 40% layer dropping on BERT, RoBERTa, and XLNet. The modified model gives 98.2% of accuracy compared with the results of DistilBERT, DistilRoBERTa and GLUE. They have experimented with five Layer-dropping Strategies. Dropping some K layers, like k= 4 to 6 top-Layers then fine-tuning the last layer for specific tasks gives a better result. The neighboring layers are usually having similar semantic information. Hence dropping alternative layers will not affect the performance. The layers having below the threshold values won't be taking part in the next level attention. Such layers can be dropped. But dropping any two adjacent layers, some K numbers of the middle and bottom layer represents the relationship between the word pieces degrades the performance and not recommended.

(Zhang and Lu 2019) introduced parallel location encoding on Multi-Attention Network (MAN) with Location-Point-Wise Feed-Forward Networks (LPFFN), which is pre-trained on BERT. In the given sentence there will be more aspects and context words. The semantic relevance including the hidden relevance between all the target words and context words obtained from MAM using the location encoding information. The semantic information passed to Context-Target-Interaction Layer. Bi-GRU architecture is used to get target word representation where the selective interaction between target and context interaction is selected dynamically based on the degree of influence. Coefficient Loss Forwarding Mechanism retains the context-target pair information in various Contexts - Target Interaction Layer. (Kou, Yang et al. 2020) explored auxiliary talks and Co-training framework to integrate Self-Supervised Attention into the BERT model. First in the auxiliary task, using random masking of the tokens in original sentence S, generate S1 which doesn't change the label and all the masked tokens are scored as '0'. From this scoring, the tokens which are influencing the correct prediction are identified. This task-oriented weighting score is learned by self-supervised Attention

(SSA), and more generalize the model. Second, the co-training framework optimizes target specific tasks with a loss function. It uses two representations. They are the sentence representation for the target-specific task and the token representation influenced by Self –supervised attention. Additional SSA is added to externally score the irrelevant tokens which are identified from the auxiliary and co-training task to get more token importance for target task prediction.

Author	Year	Corpus used for Pre-training	Language	Name of the Model
Mathew et al	2020	Book corpus	English	Movie Reviews(BERT)
Rezaeinia, S. M., et al	2019	Google News	English	Movie Reviews, Customers reviews, SST, Rotten Tomatoes, SST-1(Word Embeddings)
Reimers., et al	2019	Siamese and triplet network structures	English	Sentence Bert(SBERT)
Müller, M., et al.	2020	Twitter messages on the topic of COVID-19	English	COVID TWITTER BERT
Srivastava, H., et al	2020	Book corpus	English	Twitter and Reddit(Sarcasm Detection)
Yu, Su et al	2019	post-training of BERT	English	BERT4TC models
Liu, Ott et al	2019	Book corpus,CC-News,OpenWebText,Stories	English	RoBERTa
Delobelle, Winters et al	2020	OSCAR Corpus	Dutch	RoBERTa

Table 1. Research work on BERT Pre-trained with Domain-Specific Corpus

Author	Year	Language	Application Domain	Masking type
Yiming Cui.,et al	2019	Chinese	BERT	Whole word masking
Junru Zhou.,et al	2019	English	LIMIT-BERT(Linguistic Informed Multi-Task BERT)	Syntactic Phrase Masking, Semantic Phrase Masking and Whole Word Masking
Zhiyong Wu.,et al	2020	English	BERT	Perturbed Masking
Liu, Ott et al	2019	English	RoBERTa	Dynamic masking
Tian, Gao et al	2019	English	BERT and RoBerta (Sentiment Knowledge Enhanced Pre-training)	Sentiment masking
Li, Ma et al	2020	English	BERT	contextualized perturbation generator

Table. 3. Research work on BERT Pre-trained model fine-tunning

Author	Year	Language	Application Domain	The architecture proposed in the output layer
Araci et al	2019	English	FinBERT	fine-tuning only a small subset of model layers for decreasing training time without a significant drop in performance
Hiew,Huang et al	2019	Chinese	Financial Sentiment Index	fine-tuned BERT to all unlabelled posts
Zahera, Elgendy et al	2019	English	Tweets	build 10 BERT stacked layers on top of the BERT outputs

Kocoń, Zaśko-Zielińska et al	2019	Polish	Polish text reviews	adding one additional output layer
Zhang, Wu et al	2019	English	SemBERT Model	sub-word level masking
Alaparthi and Mishra	2020	English	Movie reviews	fine-tuned the parameters at the output layer
Zhao, Li et al	2020	English	Finance text	fine-tuned RoBERTa to do sentiment analysis and use cross-entropy loss as a loss function
Lin, Madotto et al	2019	English	Sales Prediction	stack another linear layer with Softmax function on top of BERT
Ling	2020	Chinese	BERT	all the parameters are fine-tuned in the output layer
Wu, Lv et al	2019	English	BERT	a conditional masked language model

Table. 3. Research work on BERT Pre-trained model fine-tunning

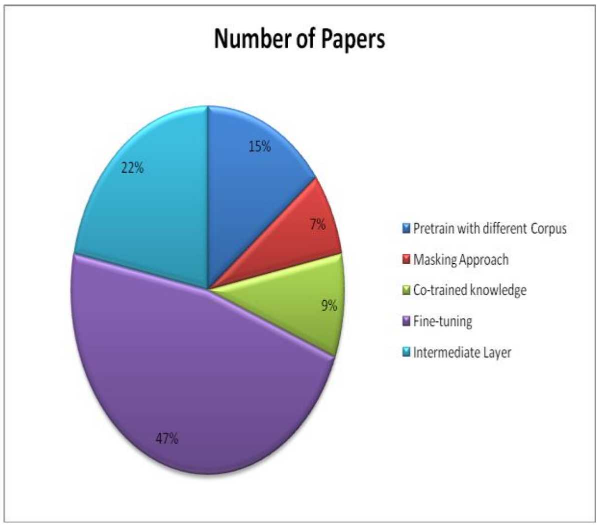

Figure 2. Contributions of the author in influencing the Layers of BERT model

4. Summary and Conclusion

Recent research works on sentiment analysis tasks utilized the benefits of pre-trained language model embedding rather than pre-trained embeddings. Since the sentiment analysis task requires more contextual information learned at a particular instance by considering the whole text. Every time training the model from the scratch with limited training data which always suffers from the amount of labeled data is a costlier process and the accuracy of the classification depends on the hypotheses generated from these training data. So get the benefits of pre-trained language models with domain-specific corpus attracted the researchers to get outperform results on the sentiment classification task. To get more language understanding, researches leverage the original architecture of the BERT model to improve classification accuracy. Since the sentiment analysis task is different from other NLP tasks in the sense it needs more semantic and contextual information of the tokens. Also, it is necessary to give attention to the important tokens that influence the context of the whole sentence and select these tokens are challenging ones. The suitable other deep learning architectures to infer this semantic and context information from the language were adapted with the model with the necessary fine-tuning of the output layer of the BERT model. Some times the research work is prohibited due to the requirement of high computing. To overcome these issues, the experiments are done by distilling the different parts of the original architecture, and removing the alternate layers of BERT gives better results even with a minimized model. Also, the sentiment classification polarity levels can be further increased from binary class to fine-grain classes b in most of the research work and improved the accuracy by finetuning the output layer with hyperparameter values drop out, batch size, learning rate, epoch. The Application domains which is shortage of domain data like finance, medicine get benefitted through BERT pre-trained language models. Similarly, many languages like

Dutch, Chinese, Polish, Persian, Italian, German, French, Arabic, Japanese and Russian experimented BERT model for the sentiment analysis task and got outperforming results.

References

1. Agarwal, B., et al. (2015). "Sentiment analysis using common-sense and context information." 2015.
2. Alaparthi, S. and M. Mishra (2020). "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey." arXiv preprint arXiv:2007.01127.
3. Alharbi, A. S. M. and E. J. C. S. R. de Doncker (2019). "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information." 54: 50-61.
4. Alomari, K. M., et al. (2017). Arabic tweets sentimental analysis using machine learning. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer.
5. Araci, D. (2019). "Finbert: Financial sentiment analysis with pre-trained language models." arXiv preprint arXiv:1908.10063.
6. Arase, Y. and J. Tsujii (2019). "Transfer Fine-Tuning: A BERT Case Study." arXiv preprint arXiv:1909.00931.
7. Arisoy, E., et al. (2012). Deep neural network language models. Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT.
8. Azzouza, N., et al. (2019). TwitterBERT: Framework for Twitter Sentiment Analysis Based on Pre-trained Language Model Representations. International Conference of Reliable Information and Communication Technology, Springer.
9. Badlani, R., et al. (2019). "Disambiguating Sentiment: An Ensemble of Humour, Sarcasm, and Hate Speech Features for Sentiment Classification." 337.
10. Biesialska, K., et al. (2020). "Sentiment Analysis with Contextual Embeddings and Self-Attention." arXiv preprint arXiv:2003.05574.
11. BİRİCİK, G., et al. (2012). "Abstract feature extraction for text classification." 20(Sup. 1): 1137-1159.
12. Bojanowski, P., et al. (2017). "Enriching word vectors with subword information." 5: 135-146.
13. Cambria, E. (2013). An introduction to concept-level sentiment analysis. Mexican international conference on artificial intelligence, Springer.
14. Chowdhary, K. (2020). Natural language processing. Fundamentals of Artificial Intelligence, Springer: 603-649.
15. Clark, K., et al. (2019). "What does BERT look at? an analysis of BERT's attention."
16. Cui, Y., et al. (2019). "Pre-training with whole word masking for Chinese BERT."
17. Gopalakrishnan, R., Mohan, A., Sankar, L. P., & Vijayan, D. S. (2020). Characterisation On Toughness Property Of Self-Compacting Fibre Reinforced Concrete. In Journal of Environmental Protection and Ecology (Vol. 21, Issue 6, pp. 2153–2163)..
18. Das, B. and S. J. a. p. a. Chakraborty (2018). "An improved text sentiment classification model using TF-IDF and next word negation."
19. Delobelle, P., et al. (2020). "RobBERT: a dutch RoBERTa-based language model." arXiv preprint arXiv:2001.06286.
20. Devlin, J., et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding."
21. Dey, A., et al. (2017). Lexical TF-IDF: An n-gram feature space for cross-domain classification of sentiment reviews. International Conference on Pattern Recognition and Machine Intelligence, Springer.
22. Dzisevič, R. and D. Šešok (2019). Text Classification using Different Feature Extraction Approaches. 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), IEEE.
23. El-Din, D. M. J. I. J. o. A. C. S. and Applications (2016). "Enhancement bag-of-words model for solving the challenges of sentiment analysis." 7(1).
24. Erhan, D., et al. (2010). Why does unsupervised pre-training help deep learning? Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.
25. Ethayarajh, K. J. a. p. a. (2019). "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings."
26. Feng, S., et al. (2019). "Attention-based hierarchical LSTM network for context-aware microblog sentiment classiGage, P. J. C. U. J. (1994). "A new algorithm for data compression." 12(2): 23-38.
27. Gamal, D., et al. (2019). "Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains." 1(1): 224-234.
28. Garg, S. and G. Ramakrishnan (2020). "BAE: BERT-based Adversarial Examples for Text Classification." arXiv preprint arXiv:2004.01970.
29. Han, H., et al. (2019). "Augmented sentiment representation by learning context information." 31(12): 8475-8482.

30. Hiew, J. Z. G., et al. (2019). "BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability." arXiv preprint arXiv:1906.09024.
31. Horn, F. J. a. p. a. (2017). "Context encoders as a simple but powerful extension of word2vec."
32. u, X., et al. (2019). "Selective attention based graph convolutional networks for aspect-level sentiment classification." arXiv preprint arXiv:1910.10857.
33. Howard, J. and S. J. a. p. a. Ruder (2018). "Universal language model fine-tuning for text classification."
34. Hung, B. T. (2020). Domain-Specific Versus General-Purpose Word Representations in Sentiment Analysis for Deep Learning Models. Frontiers in Intelligent Computing: Theory and Applications, Springer: 252-264.
35. Jelinek, F. J. R. i. s. r. (1990). "Self-organized language modeling for speech recognition." 450-506.
36. Jianqiang, Z., et al. (2018). "Deep convolution neural networks for twitter sentiment analysis." 6: 23253-23260.
37. Kanakaraddi, S. G., et al. (2020). Comparison Study of Sentiment Analysis of Tweets using Various Machine Learning Algorithms. 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE.
38. Tholkapiyan, A.Mohan, Vijayan.D.S, A survey of recent studies on chlorophyll variation in Indian coastal waters, IOP Conf. Series: Materials Science and Engineering 993 (2020) 012041, 1-6.
39. Ke, P., et al. (2019). "Sentilr: Linguistic knowledge enhanced language representation for sentiment analysis."
40. Kim, D., et al. (2019). "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec." 477: 15-29.
41. Kiperwasser, E. and Y. J. T. o. t. A. f. C. L. Goldberg (2016). "Simple and accurate dependency parsing using bidirectional LSTM feature representations." 4: 313-327.
42. Kocoń, J., et al. (2019). Multi-level analysis and recognition of the text sentiment on the example of consumer opinions. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019).
43. Kou, X., et al. (2020). "Improving BERT with Self-Supervised Attention."
44. Kumar, A., et al. (2020). "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data." 57(1): 102141.
45. Kumar, A., et al. (2020). "ATE-SPD: simultaneous extraction of aspect-term and aspect sentiment polarity using Bi-LSTM-CRF neural network." 1-22.
46. Lei, J., et al. (2019). BERT Based Hierarchical Sequence Classification for Context-Aware Microblog Sentiment Analysis. International Conference on Neural Information Processing, Springer.
47. Li, L., et al. (2020). "Bert-attack: Adversarial attack against BERT using BERT " arXiv preprint arXiv:2004.09984.
48. Li, W., et al. (2020). "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification."
49. Li, Z., et al. (2019). "Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation." 7: 72928-72935.
50. Lin, Z., et al. (2019). Learning to learn sales predictions with social media sentiment. Proceedings of the First Workshop on Financial Technology and Natural Language Processing.
51. Ling, J. (2020). Coronavirus public sentiment analysis with BERT deep learning.
52. Liu, N., et al. (2019). "Attention-based Sentiment Reasoner for aspect-based sentiment analysis." Human-centric Computing and Information Sciences 9(1): 35.
53. Liu, Y., et al. (2019). "Roberta: A robustly optimized BERT pre-training approach."
54. Lu, Q., et al. (2020). "Bi-GRU Sentiment Classification for Chinese Based on Grammar Rules and BERT." 13(1): 538-548.
55. Lv, Y., et al. (2019). Extract, Attend, Predict: Aspect-Based Sentiment Analysis with Deep Self-Attention Network. 2019 IEEE 21st International Conference on High-Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), IEEE.
56. Lyu, C., et al. (2020). Incorporating Context and Knowledge for Better Sentiment Analysis of Narrative Text. Text2Story@ ECIR.
57. Mao, Y., et al. (2020). "LadaBERT: Lightweight Adaptation of BERT through Hybrid Model Compression."
58. Mathew, L. and V. Bindu (2020). A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), IEEE.
59. Melamud, O., et al. (2016). context2vec: Learning generic context embedding with bidirectional lSTM. Proceedings of the 20th SIGNLL conference on computational natural language learning.

60. Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems.
61. Mousa, A. and B. Schuller (2017). Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.
62. Müller, M., et al. (2020). "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter." arXiv preprint arXiv:2005.07503.
63. Ombabi, A. H., et al. (2020). "Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks." 10(1): 1-13.
64. Patel, N. V. and H. J. I. J. o. D. S. S. T. Chhinkaniwala (2019). "Investigating machine learning techniques for user sentiment analysis." 11(3): 1-12.
65. Pennington, J., et al. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).
66. Peters, M. E., et al. (2018). "Deep contextualized word representations."
67. Pruthi, D., et al. (2019). "Combating adversarial misspellings with robust word recognition." arXiv preprint arXiv:1905.11268.
68. Qiu, X., et al. (2020). "Pre-trained models for natural language processing: A survey."
69. Radford, A., et al. (2019). "Language models are unsupervised multitask learners." 1(8): 9.
70. Rahman, S. S. M. M., et al. (2020). An Investigation and Evaluation of N-Gram, TF-IDF and Ensemble Methods in Sentiment Classification. International Conference on Cyber Security and Computer Science, Springer.
71. Raju, K. V. and M. Sridhar (2020). Based Sentiment Prediction of Rating Using Natural Language Processing sentence-level Sentiment Analysis with Bag-of-Words Approach. First International Conference on Sustainable Technologies for Computational Intelligence, Springer.
72. Reimers, N. and I. Gurevych (2019). "Sentence-Bert: Sentence embeddings using siamese BERT-networks." arXiv preprint arXiv:1908.10084.
73. Rezaeinia, S. M., et al. (2017). "Improving the accuracy of pre-trained word embeddings for sentiment analysis."
74. Rezaeinia, S. M., et al. (2019). "Sentiment analysis based on improved pre-trained word embeddings." 117: 139-147.
75. Rietzler, A., et al. (2019). "Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification."
76. Roh, Y., et al. (2019). "A survey on data collection for machine learning: a big data-ai integration perspective."
77. Rojas-Barahona, L. M. J. L. and L. Compass (2016). "Deep learning for sentiment analysis." 10(12): 701-719.
78. Sajjad, H., et al. (2020). "Poor Man's BERT: Smaller and Faster Transformer Models."
79. Sanh, V., et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter."
80. Saura, J. R., et al. (2019). "Detecting indicators for startup business success: Sentiment analysis using text data mining." 11(3): 917.
81. Shen, J., et al. (2019). Sentence-level sentiment analysis via BERT and BiGRU. 2019 International Conference on Image and Video Processing, and Artificial Intelligence, International Society for Optics and Photonics.
82. Silla, C. N., et al. (2011). "A survey of hierarchical classification across different application domains." 22(1-2): 31-72.
83. Singh, R. and V. Goel (2019). Various machine learning algorithms for twitter sentiment analysis. Information and Communication Technology for Competitive Strategies, Springer: 763-772.
84. Song, Y., et al. (2019). Targeted Sentiment Classification with Attentional Encoder Network. International Conference on Artificial Neural Networks, Springer.
85. Sousa, M. G., et al. (2019). BERT for Stock Market Sentiment Analysis. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE.
86. Srivastava, H., et al. (2020). A novel hierarchical BERT architecture for sarcasm detection. Proceedings of the Second Workshop on Figurative Language Processing.
87. Sun, X., et al. (2019). "Word representation learning based on bidirectional grus with drop loss for sentiment classification." IEEE Transactions on Systems, Man, and Cybernetics: Systems.
88. Sun, Y., et al. (2020). ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. AAAI.
89. Sun, Y., et al. (2019). "Ernie: Enhanced representation through knowledge integration."
90. Tang, D., et al. (2016). "Aspect level sentiment classification with deep memory network."

91. Tang, D., et al. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
92. Tenney, I., et al. (2019). "What do you learn from context? probing for sentence structure in contextualized word representations."
93. Tian, H., et al. (2020). "SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis." arXiv preprint arXiv:2005.05635.
94. Tran, T. K. and T. T. J. I. A. Phan (2020). "Capturing Contextual Factors in Sentiment Classification: An Ensemble Approach." 8: 116856-116865.
95. Tripathy, A., et al. (2016). "Classification of sentiment reviews using n-gram machine learning approach." 57: 117-126.
96. Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems.
97. Vijayaraghavan, S. and D. J. a. p. a. Basu (2020). "Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms."
98. Wang, K., et al. (2020). "Relational Graph Attention Network for Aspect-based Sentiment Analysis." arXiv preprint arXiv:2004.12362.
99. Wang, S., et al. (2020). "Emotion-Semantic-Enhanced Bidirectional LSTM with Multi-Head Attention Mechanism for Microblog Sentiment Analysis." 11(5): 280.
100. Wang, T., et al. (2020). "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model."
101. Wang, X., et al. (2016). Combination of convolutional and recurrent neural networks for sentiment analysis of short texts. Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers.
102. Wang, Z. and Z. J. C. C. Lin (2020). "Optimal feature selection for learning-based algorithms for sentiment classification." 12(1): 238-248.
103. Wei, J., et al. (2020). "BiLSTM with Multi-Polarity Orthogonal Attention for Implicit Sentiment Analysis." 383: 165-173.
104. Wu, T., et al. (2019). "Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis." 26(4): 1-27.
105. Wu, X., et al. (2019). Conditional BERT contextual augmentation. International Conference on Computational Science, Springer.
106. Xu, G., et al. (2019). "Sentiment analysis of comment texts based on BiLSTM." 7: 51522-51532.
107. Xu, Q., et al. (2020). "Aspect-based sentiment classification with a multi-attention network." Neurocomputing 388: 135-143.
108. Yang, L., et al. (2019). Leveraging BERT to improve the FEARS index for stock forecasting. The First Workshop on Financial Technology and Natural Language ProcYin, D., et al. (2020). "SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics." arXiv preprint arXiv:2005.04114.
109. Yu, S., et al. (2019). "Improving BERT-based text classification with an auxiliary sentence and domain knowledge." IEEE Access 7: 176600-176612.
110. Zahera, H. M., et al. (2019). Fine-tuned BERT Model for Multi-Label Tweets Classification. TREC.
111. Zhang, Q. and R. J. F. I. Lu (2019). "A MultiAttention Network for Aspect-Level Sentiment Analysis." 11(7): 157.
112. Zhang, Z., et al. (2019). "Semantics-aware BERT for language understanding." arXiv preprint arXiv:1909.02209.
113. Zhao, L., et al. (2020). "A BERT based Sentiment Analysis and Key EntityDetection Approach for Online Financial Texts."
114. Zhao, M., et al. (2020). "Masking as an Efficient Alternative to Finetuning for Pre-trained Language Models."
115. Zhao, R. and K. J. I. t. o. f. s. Mao (2017). "Fuzzy bag-of-words model for document representation." 26(2): 794-804.

Biography

D.Deepa is currently working as a Assistant professor in Department of Computer Science and Engineering, Kongu Engineering College, Tamilnadu, India and also she is a Ph.D. candidate in information and communication engineering under Anna University Chennai. She completed his master degree in Computer Science and Engineering at Kongu Engineering College, Anna University, in 2011. Her major research interests are Natural Language Processing, Sentiment Analysis and Deep Learning Models.

Dr. A. Tamilarasi received the B.Sc. degree from the University of Madras, Chennai in 1983, the M.Sc. degree from Bharathiar University, Coimbatore in 1985, the M.Phil degree from University of Madras, Chennai in 1987 and the Ph.D. degree from the University of Madras, Chennai in 1994. She is currently a Professor and Head in the Department of Computer Applications, Kongu Engineering College, Perundurai. Her research interests are in Data Mining, Soft Computing, Theory o f Computation and Theoretical Computer Science. She has authored 12 books, more than 100 International Journals & National Journals and Presented several papers in International and National Conferences. She has undertaken several research projects sponsored by various organizations. She is a member in Computer Society of India (CSI), ISTE (Indian Society for Technical Education), IEEE and various other International Computer Societies and organization for knowledge exchange and enhancement