Extending Bidirectional Language Model for Enhancing the Performance of Sentiment Analysis

Eashan Arora, Sakshi Mishra, K. Vimal Kumar and Pawan Upadhyay

Abstract Sentiment analysis provides an analysis about the writer's emotion conveyed in the text. It uses natural language analysis to predict the sentiment of the writer. The proposed work describes an effective and efficient natural language approach to realize multi-classification for textual data according to their respective sentiments. Proposed system employs a RNN language model based on Long-Short term memory (LSTM) over pre-trained word vectors that was generated from different language model for sentiment classification. In order to improve the performance of sentiment analyser, we have used a bidirectional language model—embeddings from language model layer (ELMo). We have merged the Natural language processing and Deep learning techniques to analyse and achieve an important emotion from the long sentences fed to the system. The proposed system is compared with state-of-art approaches such as—Simple Multilayer Perceptron, recurrent neural network (RNN) and LSTM. These techniques are thus being applied to classify sentiments on the dataset of imdb-movie review (Mass et al. in Learning word vectors for sentiment analysis. Association for Computational Linguistics, pp 142–150, 2011 [1]). These techniques have improved the system to capture syntactic and semantic relationship that can help in identifying the sentiments.

Keywords Bidirectional language model · Sentiment analysis · ELMo · Multilayer perceptron · RNN · LSTM · Stacked LSTM

E. Arora (✉) · S. Mishra · K. V. Kumar · P. Upadhyay
Jaypee Institute of Information Technology, Noida, India
e-mail: eashanarora2798@gmail.com

S. Mishra
e-mail: mishrasakshi1208@gmail.com

K. V. Kumar
e-mail: vimalkumar.k@gmail.com

P. Upadhyay
e-mail: pawanupadhyay@gmail.com

V. K. Gunjan et al. (eds.), *Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies*, Lecture Notes in Electrical Engineering 643,
https://doi.org/10.1007/978-981-15-3125-5_15

1 Introduction

In various areas, the feedback provided by the customers regarding some product/service has an incomparable relevance according to the context. Opinions which are expressed in the form of feedbacks provide an opportunity to get an idea about the likes and dislikes of community on the whole. Thus, for any data centric company and their data scientists who look into extracting some meaning out of an unstructured textual data, sentiment analysis will be one of the initial steps towards such a task with a very little investment. Using natural language processing (NLP), extracting textual features that can be used for sentiment analysis preserves its popularity according to their usage. The two main focuses in NLP are language processing and language generation. The ability to recognize the text and understand the meaning is performed in language processing whereas language generation involves generating natural language from linguistic data. The emphasis of this paper will be on the processing of textual data and mapping it features to perform efficient sentiment analysis. This paper also addresses the core problem of sentiment analysis on a particular context.

The main contribution of this paper is the improvement of sentiment analysis by making use of new word embeddings from a language model that can capture the language specific features. This language specific feature contributes in learning the parameters of network based their corresponding labels. The main idea is to automatically train model without pre-processing of data. The literature study shows [2] that a deep contextualized language model such as ELMo will be a very powerful language embedding model for sentiment analysis because it makes use of a bidirectional analysis using the network. Also, it generates the word vectors based on the word usage in terms of syntactic and semantics in that sentence. The network should hold a memory in order to deal with sequential data such as a text. The meaning of a word has dependency not only on the context of the previous word but on the whole sentence. Thus, the proposed system uses a bidirectional language model which has been trained on 1 Billion words from benchmark dataset and accepts strings in the form of raw or tokenized form. The model output is mean pooled vector representation of the input fed to it.

The paper is organized as follows: in Sect. 2, we present a brief description on some related works on sentiment analysis and our approach towards it. The dataset used and its description presents in Sect. 3. Section 4 describes the software packages and models we used in our proposed study. Experimental results are presented and analysed in Sect. 5. We also present upcoming work and approach towards it in last section of the paper after concluding section that summarizes the findings of this proposed work.

2 Related Work

Review given by people gives the user's thought which plays an essential role in gathering up any important information. Resources such as online reviews and personal blogs provide an opportunity to understand and seek other's opinion. This makes sentiment analysis an important task.

Sentiment analysis has grown up remarkably during the early 2000s. However, first approach towards such analysis has been done by Hatzivas-Silogou and Mckneown [3] in 1997. In this research, they used wall street journal data to perform research on identifying the semantic orientation of adjectives. Since dictionaries do not show sentiments, they made a basic assumption that sentence itself would serve that information. They demonstrated that the conjunction between the adjectives provide an indirect information regarding the sentiments of the sentences.

Consider the example sentence—*movie was good but bit slow before interval*, in which good and slow are the two dissimilar words that are connected by the conjunction but their main focus were on identifying the sentiment orientation of adjectives and not the whole sentence. When analysing a word's semantics it is being found that each word have an isolated meaning of itself and also having some meaning in context of a sentence and this concept is called as the "Principle of Compositionality". According to this principle [4], the meaning of word depends on the meaning of its predecessor in the whole sentence. Thus describes its meaning in respect of whole text.

In recent years, artificial neural network gains more importance in the field of sentiment analysis. Some techniques that are designed for the representation of sentences through semantic composition are convolution neural network (CNN), recurrent neural network (RNN), Long-short-term memory (LSTM), etc.

Recurrent neural network came in across during 1980s but gained an importance since last few years because of increase in computational power and also helps in dealing up with large amounts of data. Because of its internal memory, RNN's are able to store important information regarding the input and then predicting an output through it. But RNNs has problem in dealing with the long-term dependencies [5]. This problem leads to the invention of LSTM during 1997 (discovered by Hochreiter and Schmidhuber). It is an extension for recurrent neural network that makes the usage of larger internal memory thus to deal with long-term dependencies and also enables RNNs to remember their inputs over a long period of time [6].

In recent years, sentiment analysis was explored a lot. In a research on sentiment classification using product reviews [7], the researchers identified the class of sentiment polarities both at the sentence-level and review-level. Later on, the prediction of sentiments using texts on social media was explored and to improve its accuracy new features were introduced [8].

To deal with the problem of polysemy i.e. the concept that word has different meanings according to usage; proposed system uses ELMO since other embedding models focuses on one vector output representation, combining of different meanings of single word all together. ELMo [9] is the embedding technique proposed by

Allen NLP. Unlike other embedding models it is dynamic as it changes every time depending on the context even if the word is same, totally depending on the nature of the sentence.

Dataset

To illustrate and evaluate the need for bidirectional language model for sentiment classification, data for training and testing was collected from an online source platform-Kaggle. The IMDBmaster dataset is a set of 7998 reviews from different people having their particular views on a particular movie. Each of these movie reviews is classified as either being "positive" or "negative". The data is divided into a train and validation set, trained on 5358 sentences and validation on 2640 reviews. Among these set, 50% are positive reviews and rest 50% are negative. The training database contains the imdb reviews as input which accepts it in the form of raw strings or tokenized strings.

3 Proposed Work

The proposed system aims to provide a deep learning model that can capture the language specific features in an appropriate manner such that it can improve the accuracy of sentiment analysis. In order to capture the language specific features, a bidirectional language model is being used. The features captured are represented in the form of vectors which are further used by sentiment analysis models to map the relation with their respective sentiments. The sentiment analysis model is developed using different models such as multi layer perceptron model, long short term memory (LSTM) network and stacked LSTM network.

3.1 Word Embedding Using Bidirectional Language Model

Recent developments in improving computational power of computers have taken the development in NLP research to a new dimension. The language model takes a new shift in how words are vectorized or encoded. Word-embeddings have been a major reason how different NLP model deal with language. Previously models like continuous bag-of-words (CBOW), Skip gram model and glove have been used for embeddings task. But, each of them has certain disadvantages. Due to these disadvantages a bidirectional sentence level analysis is being employed in this system instead of textual feature extraction using a directional (left-to-right) sentence level analysis. One such model is embeddings from language model (ELMO) that uses bidirectional analysis (biLM-bidirectional language model) to learn both word (e.g., syntax and semantics) and linguistic context. Currently the bidirectional language model shown in Fig. 1 has been used for 6 Downstream NLP tasks namely: question answering, textual entailment, semantic role labeling and named entity extraction.

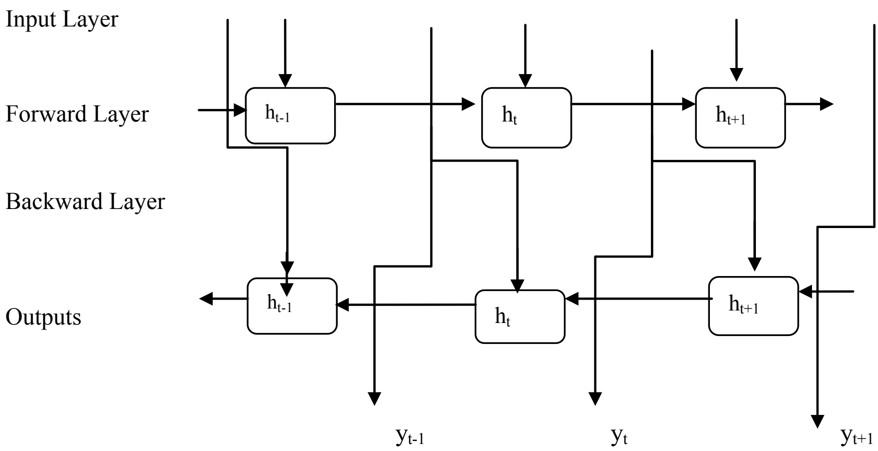

Fig. 1 Bidirectional language model

ELMO gives an embedding according to the context of the sentence it's used in, so that it can capture both word-level meaning and contextual information. The language model gains the understanding by getting trained to predict the next word based on the current word in the current sentence. Since ELMO employs bidirectional approach, it actually goes a step further and trains the model such that it can extract the sense of the next word and also the previous word.

This language model is basically concatenations of the activations on several layers of the biLMs (bi-directional Language Models). Different layers of a language model encode different kind of information on a word (e.g. Part-Of-Speech tagging is well predicted by the lower level layers of a biLSTM while word-sense disambiguation is better encoded in higher-levels).

3.2 Sentiment Analysis Approach

Sentiment analysis is being used in many other NLP applications, so that, it can improve the overall accuracy of those applications. In general, sentiment analysis can be used for predicting customer satisfaction based on their reviews. Based on these review, it can also be used for finding the popularity of a particular product. Since it has wide scope for integration, the need for improving its accuracy is predominant. In this proposed system, the deep learning models are used for mapping the textual features with its corresponding sentiment labels. The efficiency of bidirectional model is evaluated using different sentiment analysis models such as multi-layer perceptron, long short term memory (LSTM) and a stacked LSTM.

Multi-layer perceptron based sentiment analyzer

The multilayer perceptron network used for sentiment analysis has one input layer and an output layer. But, there is flexibility in the number of hidden layers to be used. In this proposed system, the number of hidden layers was kept in varying manner and the accuracy of network is found to be good when there were three hidden layers being used in it. This is due to improvement in the learning of network parameters.

LSTM based sentiment analyzer

The recurrent neural network (RNN) was introduced in this proposed system to capture the features in an appropriate manner. Since the recurrent neural network has a vanishing gradient issue, the long short term memory network (as shown in Fig. 2) was later introduced in place of simple RNN. As in recurrent neural network, when the error is propagated it goes through the layers of neurons which are connected to themselves. The hidden layers which are connected to other hidden layer have network weight parameter called the recurrent weight. The weight is applied many times on top of itself which causes the gradient error to decline rapidly. That is, the weight of the layers on very far left are a bit updated much slower than the weight on the other layers on the far right and this creates a domino effect.

$$
\begin{aligned}
&\text{IF:}\\
&\quad \text{WREC} = \text{SMALL} \rightarrow \text{Vanishing gradient}\\
&\quad \text{WREC} = \text{Large} \rightarrow \text{Exploding gradient}
\end{aligned}
$$

As LSTM are capable of handling long term dependencies. For LSTMS WREC = 1, there is only 2 point-wise operations as well understand for the down and there's

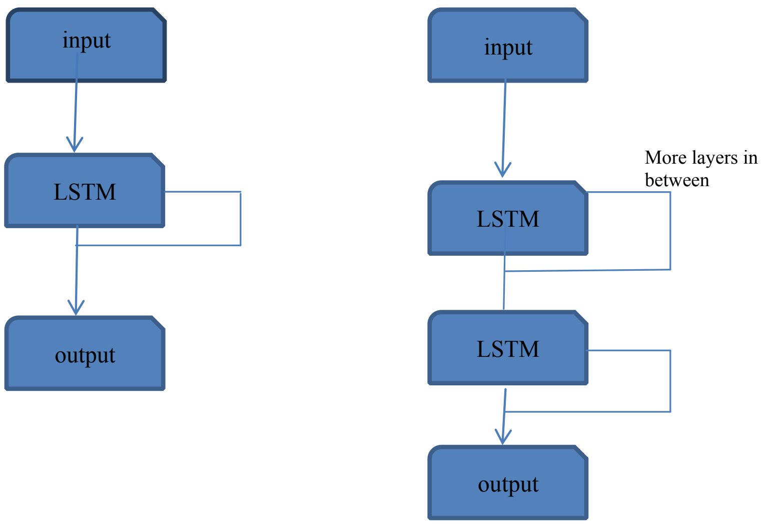

Fig. 2 LSTM model versus stacked LSTM model [10]

no complex neural network layered operations. LSTM'S have a memory cell which just goes through time, sometimes it might be removed or sometimes things might be added to the memory cell. Otherwise it flows through times freely and therefore when you back propagate through these LSTMS there is not much of a problem of vanishing gradient.

Stacked LSTM

Since LSTM model consist of single hidden layer followed by feedforward output layer, the textual features are learnt in a coarse manner. In order to fine tune the learning of features there is need for multiple LSTM layers. So, the proposed system uses stacked LSTM layers which is the extension to this model that has multiple hidden LSTM layers as shown in Fig. 2. Stacked LSTMs were first introduced by Graves et al. in their application of LSTMs to speech recognition. Stacked LSTM helps to increase the depth of network which in turn helps to increase the accuracy and efficiency of the model. A single large hidden layer can be used to approximate functions but increasing the depth of the network is another solution that only requires few neuron and trains faster. In stacked LSTM, one LSTM layer provides a sequential output which then serves as an input for the next LSTM layer. It has been argued that stacked LSTM layer allows network to learn at a different time scales over an input to make a better use of parameters [10].

4 Result Analysis

Firstly, the proposed bidirectional model performance is evaluated over the previous built models such as, multi-layer perceptron and LSTM. So that importance of bidirectionality in sentiment analysis is detected properly. Initially for the multilayer perceptron model, a dense network layer with default 256 neurons was used and for input vector the training data samples are fed to the ELMO embedding layer which is of a 2D shape vector with 1024 as its dimension.

For the LSTM based model, the analysis on this layer was performed and we found that LSTM does not support masking and lengths of different variable size. To overcome this, a masking function in the ELMO model is employed. Further to train the ELMO embedding layer, input training sentences are used. Afterwards embeddings from the Language Model are fed to the LSTM layer with a 3D shape of the same dimensions as the dense network.

The further validation is done on the testing data samples; the model is trained for 10 epochs with each of network layers and each network yields a particular accuracy as shown in Table 1 (Fig. 3).

Table 1 Accuracy of bidirectional model over different sentiment classification models

ELMo + Baseline model	RUN1	RUN2	RUN3	RUN4	RUN5
MLP	86.62	87.7	87.4	87.35	86.51
LSTM	87.25	86.43	87.23	87.23	87.42
Stacked LSTM	87.12	85.98	86.82	87.16	86.79

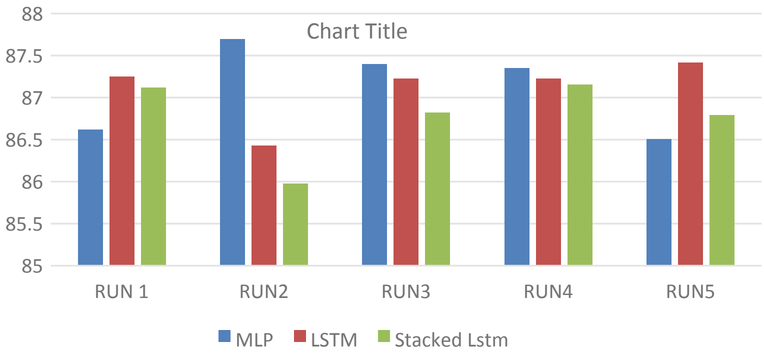

Fig. 3 Results of proposed sentiment analyzer in different runs

5 Conclusion and Future Work

The proposed sentiment analyzer was developed and was found to improve the accuracy of sentiment analyzer with a bidirectional language model-ELMO. While analyzing a particular text, the semantic information of a specific word cannot be extracted by considering it as an individual unit. But, the semantic information can be extracted using the context usage in the whole sentence. That is why we used a model that not only focuses on the context of the word but also on the context of the whole sentence—both forward and backward. We tried implementing the Bidirectional model in different layers of the network such as multilayer perceptron network. LSTM and Stacked LSTM. A network is built to classify the sentiment of the IMDB movie review dataset. It is found that the performance of proposed sentiment analyzer using stacked LSTM network and bidirectional language model was so good. It is found to be 87% accurate compared to other variants developed. The use of bidirectional language model has significant improvement in sentiment analyzer. But, still there is scope for further improvement which can be developed using bidirectional model to predict the sentiment analysis as well.

References

1. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. Association for Computational Linguistics, pp. 142–150
2. Peters, ME Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations, NAACL
3. Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the association for computational linguistics (EACL '97), Association for Computational Linguistics, Stroudsburg, PA, USA, pp 174–181. https://doi.org/10.3115/979617.979640
4. Mehler A (2007) Compositionality in quantitative semantics. A theoretical perspective on text mining. In: Aspects of automatic text analysis. Studies in fuzziness and soft computing, vol 209. Springer, Berlin, Heidelberg
5. Hochreiter S (1991) Untersuchungen zu dynamischen neu-ronalen netzen.Master's thesis, Institut fur Informatik, Technische Universitat, Munchen
6. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
7. Fang X, Zhan J (2015) Sentiment analysis using product review data. J Big Data 2(1). https://doi.org/10.1186/s40537-015-0015-2
8. Alaoui IEI, Gahi Y, Messoussi R, Chaabi Y, Todoskoff A, Kobi A (2018) A novel adaptable approach for sentiment analysis on big social data. J Big Data 5(1). https://doi.org/10.1186/s40537-018-0120-0
9. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long Papers), vol 1, pp 2227–2237
10. Sak H, Senior A, Beaufays F (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proceedings of the annual conference of the international speech communication association, (INTERSPEECH), pp 338–342