2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377043

Sentiment Analysis Using Language Models: A Study

Spraha Kumawat
Department of Computer Science and Engineering
Amity University, Noida, India
spraha.kumawat@s.amity.com

Inna Yadav
Department of Computer Science and Engineering
Amity University, Noida, India
inna.yadav1@s.amity.edu

Nisha Pahal
Department of Computer Science and Engineering
Amity University, Noida, India
npahal@amity.edu

Deepti Goel
RIL, Mumbai, India
deeptigoyal2003@gmail.com

***Abstract*—Sentiment analysis is concerned with extracting sentiment to ascertain the attitudes, and emotions associated with the text. It is broadly applied to voice of the customer as they convey their experience and feelings more blatantly which is why comprehending customer's emotions is a must. To identify and classify these unstructured emotions natural language processing (NLP) and machine learning approaches have been adopted in recent times. The main issue with the existing techniques is the inability to deal with the correct interpretation of context owing to lack of labeled data. In this paper, we studied deep neural network based language models to interpret and classify textual sequences into positive, negative or neutral emotions which remove the bottleneck of explicit human labeling. These models were analyzed and evaluations were performed on the Twitter US Airline Sentiment dataset. We have observed a considerable amount of improvements with respect to prior state-of-the-art approaches which closes the gap with supervised feature learning.**

***Index Terms*—Deep Learning, Sentiment Analysis, Language Models, Transformers**

I. Introduction

In today's environment sentiment analysis is widely used to mine subjective information to recognize polarity within content. This information can be used to predict public response in order to discover how people feel about a particular product, or topic. For instance, by sentiment analysis, we can detect if buyers are happy about costing and product service. On social media, the responses from customers are rapid and quick which can be helpful to the brands if they want to display their products. Since reviews are of various varieties, sentiment analysis can be divided into 3 three levels to simplify the analysis.

In general, *Document level* being foremost one, has the principal task to organize the document in order to conclude that every document expels point of view on a single unit. Second comes, *Sentence level* that has the principal task to identify whether a sentence shows a positive or negative opinion. This level of analysis has firmly allied with subjectivity classification which differentiates between objective statements that reveal fact-based information and subjective statements that reveal point of views and judgements. The third level is *Aspect-based* level which carries out fine-grained analysis based on a rating system. This level straight away look into the opinion itself. For instance: the 5-star rating system could possibly look like – 5-star is excellent, 4 star is moderate, 3 is ordinary, 2 is unsatisfactory and 1 star is terrible.

In sentence level which determines whether a text is positive or negative, multi-class sentiment analysis uses different clusters or categories such as excellent, moderate, ordinary etc. to better understand the emotions entailed in the text. However, multi-class classification proves out to be a challenging task due to its mathematically quantifying and complexity in understanding the context behind human feelings. For instance, recognizing polarity of the sentence *The beginning of movie was great, but climax wasn't justified* is difficult to analyze due to inherent subjectivity which leads to skewed results. Also, if sentences are short there might not be enough context to generate a reliable sentiment analysis. This motivated us to review the latest studies that have employed deep learning based approaches to capture a general understanding of a domain problem and achieve the goal of building a dependable sentiment classifier.

II. Related Works

Before the recent advancement in deep neural network models, reduction operation such as lemmatization, rule-based systems, probabilistic language models [1] ruled NLP. The authors in [2] used a Naive Bayes classifier solution to interpret and perform sentiment classification task.

The authors in [3] stated that it was prologue of word embedding such as word2vec [4] by Google, fastText by Facebook [5], with the most significant pre-trained examples which led to the attainment of deep learning in NLP. Rane et al. [6] proposed sentiment analysis of US airline service tweets using word embeddings with different classifiers such as KNN (K- Nearest Neighbour), Support Vector Machine(SVM), Decision Tree, Random Forest etc. Word2vec basically uses a shallow neural network to learn

word embeddings wherein pre-trained word embedding only transmit formerly gained knowledge to the first layer on neural network, whereas remaining is abstracted from scratch. Subsequently, the pre-trained word embedding are context free. For these reasons, deep neural networks were explored as the main standard despite their complexity to produce better contextual word representation.

Deep Neural Architectures such as seq2seq based Recurrent Neural Networks (RNN), Long Short Term Memory(LSTM) were trained in an unsupervised manner to refine the representation of the words according to their context. The authors in [7] proposed Sentiment Analysis Model using LSTM on weather and mood related tweets. In [8], the authors used LSTM and CNN based methodologies for sentiment analysis in financial domain to predict change in price based on the polarity. These strategies handle the challenge of missing context. Nevertheless, the necessity of training the actual task model from beginning is not eradicated.

To capture a general understanding of language of specific downstream NLP tasks, the idea of training a separate language model came into existence. The authors in [9] developed Transformer based architecture that uses attention mechanism with alignment in seq2seq architectures for translation and to capture relationships between the words of a sentence. Transformers facilitate parallelization during training and has proved very successful in improving the SOTA in many NLP tasks as compared to RNN's which tend to be very slow to train. This has led to the development of pre-trained systems such as BERT (Bidirectional Encoder Representations from Transformers) [10], GPT (Generative Pre-trained Transformer) [11], Roberta [12] which have been trained with huge general language datasets, and can be fine-tuned to specific language tasks. The authors in [13] investigated the modeling power of contextualized embeddings from pre-trained language models. The work in [14] paper reviews the latest studies employed in the field of deep learning to solve sentiment analysis problems. It shows comparative study for the different models and input features.

The discussed deep learning based models provides a solution to automatically extract complex data representations from large volumes of unsupervised data. To realize multi-classification for sentiment analysis[1], we have utilized these approaches in NLP space.

III. Sentiment Classification techniques

The task of multi-class classification has gained attention for the reason that most of the work related to sentiment analysis of contents centralizes binary classification of data while in real life it can be categorized in more than two classes.

[1]https://www.kaggle.com/crowdflower/twitter-airline-sentiment

A. LSTM

Long short-term memory (LSTM) is based on artificial recurrent neural network (RNN) and have the capablity to learn long-term dependencies. In LSTM, the information flows through a mechanism known as cell states. To control the cell state, LSTM exploits three gates namely, input gate, output gate and forget gate. The input gate decides which information should enter the cell state and the output gate determines which information should be going to the hidden state of the next cell. Finally, the forget gate decides what information to forget or thrown away and what is relevant to be kept by multiplying 0 to a position in the matrix. LSTM preserves the past information as the only inputs it has seen are from the past. Bi-directional LSTM on the other hand consists of two RNNs to take care of both past and future. Since it conserves both forward and backward information about the sequence at every time step, understanding the context is better. They offer greater performance and this is the reason they are widely used.

B. BERT

BERT stands for Bidirectional Encoder Representations from Transformers. It is a transformer based machine learning technique that pre-trains deep bidirectional representations from unlabeled text to create contextualized word embeddings. The bidirectional nature allows to acquire the context of a word depending upon its surrounding that is, left and right of the word. Figure 1 depicts that the top layer is the contextualized representation of each input word [2]. BERT follows a 2-step process to solve NLP-tasks:

- train a language model on a large unlabelled text corpus (unsupervised or semi-supervised)
- fine-tune the trained model to specific NLP tasks (supervised manner)

BERT leverage attention mechanism to learn complex patterns in the data. The attention mechanism enhances the fine-tuning performance by letting each token from input sequence focus on some other token. BERT offers wide variety of practical applications that ranges from sentiment analysis to Question Answering tasks to Named Entity Recognition.

C. Roberta

A Robustly Optimized BERT Pre-training Approach (Roberta) introduces dynamic masking pattern and allows training on longer sequences. Roberta was implemented in PyTorch and modifies the key hyper-parameters to predict the sections of text that are intentionally hidden.

D. Google/Electra

ELECTRA stands for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately" — is a novel discriminator technique proposed by Google researchers to

[2]https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

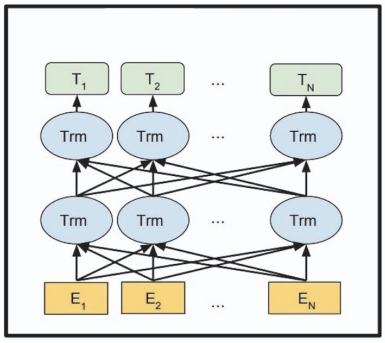

Fig. 1. BERT Language Model

provide the benefits of BERT. It is released as an open-source model and includes a ready-to-use pre-trained language representation models using relatively little compute. These models are trained to differentiate between "real" input tokens vs "fake" input tokens generated by another neural network, similar to the discriminator of a GAN. [15].

IV. Experiments

This section will describe the methodologies used to train, test and fine-tune the models namely Bidirectional LSTM, transformer based BERT, Roberta and Google/Electra, as well as the results obtained.

A. Dataset

For experimentation, we have utilized 3MB "Twitter US Airline Sentiment" annotated data-set released by Crowdflower to perform supervised training of a model. This dataset contains information of travelers in February 2015 who expressed their feelings on twitter. The dataset comprises of 14,640 tweets of six different US airlines such as US Airways, Virgin America, Delta, United, and Southwest classified into 3 emotion classes namely positive, negative and neutral. Figure 3 shows number of sentences belonging to classes positive, negative and neutral. We apportion the data into training, and test sets. A typical split of 80 percent for the training data and 10 percent each for validation set and test set was made.

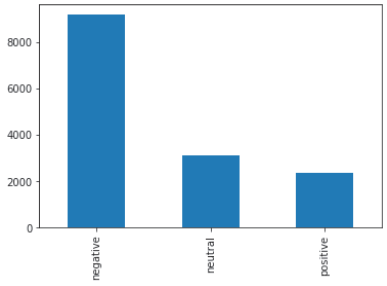

Fig. 2. Number of sentences belonging to each class

Fig. 3. Word cloud view of the dataset

1) Experimental Results: To improve model performance on multi-classification semantic problem we applied Bidirectional LSTM's. Bi-directional LSTM was trained on following parameters: batch size = 512, epochs = 10, and using categorical_crossentropy. We used ADAM optimizer to update the weights and the accuracy metric was calculated and reported for each epoch. As figure 3 reflects an unequal distribution of classes within a dataset (imbalanced dataset) parameter classweights was assigned to handle the issue. Given following parameters we achieved an accuracy of approx. 73 percent on test dataset. Figure 4 shows confusion matrix to visually observe how well the network made predictions.

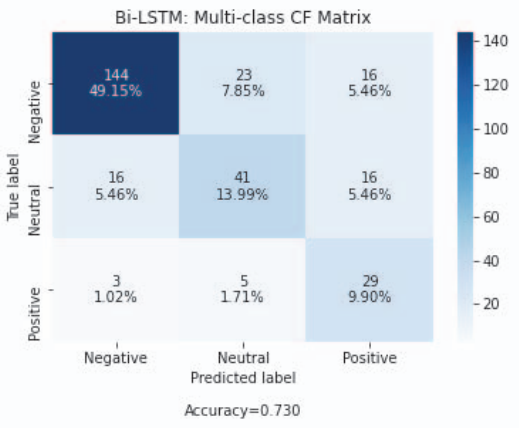

Fig. 4. Confusion Matrix

Further, we adapted state-of-the-art transformer models to perform sentiment analysis in a fast and easier way using the open-source framework *SimpleTransformers* [16]. We basically trained three models: BERT, Roberta and Electra models with a batch size = 2, and epochs = 3 (due to computational power constraints). The other essential parameters were obtained as default setting such as max_seqlength=128,

Model Name	Test Accuracy
BERT	0.812
Roberta	0.808
Electra	0.798

TABLE I
TEST ACCURACY RESULTS

learning_rate= 4e-5 and adam_epsilon= 1e-8. To overcome the class imbalance issue, matthews correlation coefficient (mcc) ranging between -1 and +1 has been used. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction [3].

The test accuracy obtained using BERT, Roberta and Electra is depicted in table I and figure5. The result clearly shows the importance of pretrained language models in the toolbox of every NLP practitioner [17].

The training and evaluation for bidirectional LSTM took 1 hour while transformer based models took approximately 4 hours each, when run on CPU machine.

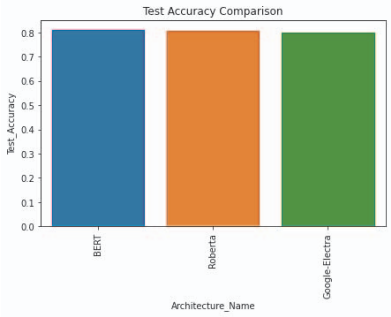

Fig. 5. Visualizing Test Accuracy

Figure 6-8 shows training learning curve obtained from the training dataset to know how well the model is learning, validation learning curve calculated from a hold-out validation dataset to know how well the model is generalizing and mcc curve for all the three models.

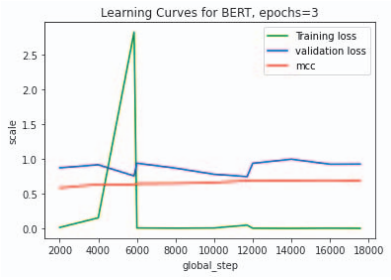

Fig. 6. Learning Curve for BERT

[3]https://en.wikipedia.org/wiki/Matthews$_correlation_coefficient$

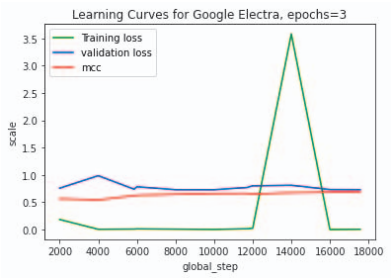

Fig. 7. Learning Curve for Electra

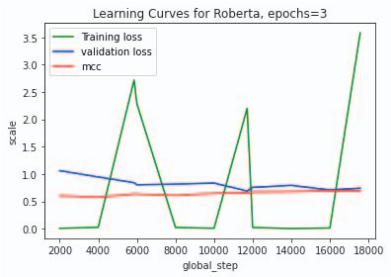

Fig. 8. Learning Curve for Roberta

V. CONCLUSION

Sentiment analysis refers to automatic assortment, accumulation, and classification of data into various opinions such as positive, negative or neutral. In this paper, we have discussed the core of deep learning models and related techniques such as transformer based architecture to address sentiment analysis in social network data. The approaches referred here automatically extract complicated data representations from large collection of unsupervised data to cover adversity of textual characteristics. We conducted experiments to evaluate bi-LSTM, BERT, Roberta and Electra models on tweets dataset. The experiments give us a broad perspective on applying deep learning models with an observation that language modeling is well suited for capturing facets of a natural language.

REFERENCES

[1] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[2] D. D. Das, S. Sharma, S. Natani, N. Khare, and B. Singh, "Sentimental analysis for airline twitter data," in *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 263, 2017.

[3] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[4] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[5] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," *arXiv preprint arXiv:1712.09405*, 2017.

[6] A. Rane and A. Kumar, "Sentiment classification system of twitter data for us airline service analysis," in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1. IEEE, 2018, pp. 769–773.
[7] J. Qian, Z. Niu, and C. Shi, "Sentiment analysis model on weather related tweets with deep neural network," in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 2018, pp. 31–35.
[8] W. Souma, I. Vodenska, and H. Aoyama, "Enhanced news sentiment analysis using deep learning methods," *Journal of Computational Social Science*, vol. 2, no. 1, pp. 33–46, 2019.
[9] Y. Guo, Y. Zheng, M. Tan, Q. Chen, J. Chen, P. Zhao, and J. Huang, "Nat: Neural architecture transformer for accurate and compact architectures," in *Advances in Neural Information Processing Systems*, 2019, pp. 737–748.
[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[13] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting bert for end-to-end aspect-based sentiment analysis," *arXiv preprint arXiv:1910.00883*, 2019.
[14] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, 2020.
[15] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.
[16] T. Rajapakse, "Simple transformers," 2020.
[17] S. Ruder, "Nlp's imagenet moment has arrived," *Gradient, July*, vol. 8, 2018.