InternationalJournalofScientificResearchandEngineeringDevelopment-– Volume5Issue4,July-August2022

RESEARCHARTICLE OPENACCESS

AN INTEGRATED DEEP LEARNING MODEL OF RoBERTa AND LSTM WITH TRANSFORMER FOR SENTIMENT ANALYSIS

Ansu K Raju[1], Athira B[2]

1. P G scholar, Department of Computer Applications, Musaliar College of Engineering and Technology, Pathanamthitta
2. Assistant Professor, Department of Computer Science and Engineering, Musaliar College of Engineering and Technology, Pathanamthitta

Abstract:

Sentiments are powerful tool and a driving force to make decisions. The significance of sentiment analysis is expanding more every passing second. Because of the very same reason, the studies related to sentiment analysis are evolving itself. Major platforms like social media, product review panels and lots of other forums where people express their emotions or feelings are also on the rise. When this occurs, the emotions that arise form the foundation for strategic planning in various spheres in life. However, these studies or researches seek to extract the actual emotions from a spoken word or a given or shown remark. Factors such as lexical diversity, an unbalanced data set and the long-distance dependency of the inputs pose hurdles to the accuracy of the emotions analysed or extracted. There is the possibility of combining different model types in the same model structure to exploit the advantages of each model. Hence, a combination of a sequence model and a transformer model finds its way in. Prior to providing data input to the model, data augmentation and preprocessing needs to be performed. Doing so solves the problem of lexical diversity and balances the dataset. RoBERTaTokenizerFast word embedding's role in data augmentation is to synthesis the samples by substituting similar word vectors and to oversample the minority classes in the dataset. Comparatively, the problem with the sequence model such as short-term memory requires a longer computation time than the RoBERTa approach that uses parallel processing. With the hybrid combination of RoBERTa and LSTM, the execution time problem and the long-distance dependency problem can be solved effectively. For the study, a custom dataset is created based on the datasets employed include IMDB, Sentiment140, Twitter US Airline Sentiment Dataset, Sentiment Analysis Data and Covid 19 Sentiment Analysis on Complete Data featuring positive, negative and neutral emotions. Results from the experiment demonstrate the improvement in performance with progression in training.

Keywords: - Sentiment Analysis, Robustly Optimized Bidirectional Encoder Representations from Transformers approach, Long Short-Term Memory, Hybrid Model, Transformer Model, Recurrent Neural Network, Word Embedding**.**

I. INTRODUCTION

Emotion is a very basic ingredient of all human interaction. Rather than being a pure thought, view or attitude, emotion relies primarily on sentiment rather than reason. Analysing emotions offers in-depth

interpretation and context-based insight into opinions, enabling subjective information to be identified and abstracted by closely analysing conversations and forums in order to comprehend the underlying sentiment. Doing so is no easy task. Indeed, even before being able to assess what sentiment a particular sentence reveals, one must first understand what "sentiment" really is to begin with. Might it be generic, and can sentiment be grouped into clear categories such as happy, sad, angry or bored? Alternatively, would it be a dimensional term, and sentiment to be evaluated on more of a two-way kind of spectrum? On top with the question of definition, in any sentence uttered by people there are several shades of interpretation. There are complex ways in which people express their opinions; use of rhetorical way such as sarcasm, irony and implicit meaning can throw the sentiment analysis off track. Therefore, one of the best approaches to understand these means is to understand the context: when you know how a paragraph begins and also the vocabulary mannerisms, can strongly influence the sentiment of subsequent phrases (Grinvald, 2021; Raj, 2021; Gupta, 2018; MonkeyLearn Inc., 2022).

Harnessing the strengths of these two methods, the possibility of a combined model emerges which is more efficient compared to its parent models. More to the point, the proposed method integrate the Robustly optimized Bidirectional Encoder Representations from Transformers approach (RoBERTa) from the Transformer family (Liu et al., 2019) and the Long Short-Term Memory (LSTM) from the Sequence family (Hochreiter and Schmidhuber, 1997). The RoBERTa models excel at sequence-to-sequence modelling in efficiently yielding representative word embeddings for the texts. Moreover, the LSTM stands out in temporal modelling to encode the long-distance dependency nature from the given input. It also augments the lexical diversity of the vocabulary through data augmentation with synonym substitutions. Data augmentation also mitigates the problem of the lopsided dataset.

A. Relevance of the project

This software allows to identify the sentiment or emotion of the given text that is posted on social media.

B. Scope of the Project

Of many analytical scopes, at which people excel all the rest, is the skill of sensing sentiments. The influence of open judgment is an area of major relevance during this age of high mass media. Finding individuals with the right skills to peruse and provide bias- free assessment can be easily solved utilizing a matching algorithm or its derivatives. Given that the field is an optimisation domain that makes extensive use of AI, creating an algorithm of this sort would greatly benefit the domain's learning capabilities and performance.

II. EXISTING SYSTEM

Of the most popular methods for solving sentiment analysis approaches using ML algorithms (Wongkar and Angdresey, 2019; Rahat, Kahir, and Masum, 2019) and approaches using DL as explained in section

2.1, the existing system is centred on the RoBERTa transformer model and the LSTM-RNN model. The transformer models (Vaswani et al., 2017) adhere to parallelised processing for execution, thereby facilitating improved computation and execution time. Given that RoBERTa is an optimised version of the highly promising transformer model BERT, it features a stable version of the pre-trained model with data from a gigabyte range. Based on RoBERTa, it contains 12 layers, 768 hidden state vectors and 125 million parameters, and undergoes self-supervised training using a large corpus of English data. Clearly, the RoBERTa models are capable of effectively generating representative word embeddings for the texts, allowing the model to outshine other sequence-to-sequence models.

Indeed, the model offers the possibility to create meaningful word embeddings as feature representations. Therefore, the layers that follow the RoBERTa layer can easily capture the useful information. The stack of sequence models in the RNN family contains encoder-decoder pairs to handle temporal modelling and solve the long-distance dependency of the given input. While the problem of longer execution time exists with sequence models, it can effectively address the problem of long-distance dependency. With LSTM, it is capable of remembering the previous information. By using this feature, LSTM can predict the successive occurrence in the sequence more easily.

III. PROPOSED SYSTEM

Combining two models, one from the transformer family and the second from the rnn sequence model family, into a single model to capitalise the merits of the respective models and assess on how this model performs against a custom dataset compiled from publicly available datasets with all the required sentiment polarities. The proposed approach adopts user-defined data balanced across all three label classes. It relies on the RoBERTaTokenizerFast for data embedding. Out of the different versions of RoBERTa, Roberta-base has been opted to fill the first layers in the stack of the proposed model where the input text will eventually be filled. Preprocessing stages include removal of tags, punctuation and symbols, conversion of raw text to lower case, lemmatisation and vectorisation. The prediction phase of the proposed method is integrated into a Graphical User Interface (GUI) to allow better user interaction with the system.

IV. LITERATURE REVIEW

C. "Multilingual Bidirectional Encoder Representations from Transformers (MBERT) and XLM-RoBERTa (XLM-R)," [Younas et al],2020

The above are the two Deep Learning (DL) methods for sentiment analysis. The authors, Younas et al. (2020), adopted these methods for analysing the sentiment of

multilingual Twitter tweets. Specifically, the dataset comprises of tweets during the 2018 general elections in Pakistan. Tweets are in two languages, English and Roman Urdu, and the dataset has a total of 20,375 tweets categorized under three labels, namely positive, negative and neutral sentiments. The dataset has been split into training and testing samples (80% and 20% respectively). The authors train and assess the performance of the model separately through the paper. In the hyperparameter tuning, these authors chose a learning rate of 2×10−5 for MBERT and 2×10−6 for XLM-R. For MBERT and XLM-R, the experimental results are 69% and 71%, respectively, indicating that XLM-R performs significantly higher than its competitors.

D. "Deep Bidirectional Long Short-Term Memory (DBLSTM),"[Anbukkarasi and Varadhaganapathy], 2020 .

For this study, Anbukkarasi and Varadhaganapathy (2020) presented a character-based sentiment analysis of self-collected Tamil tweets. There are 1500 positively, neutrally and negatively marked tweets in the dataset. They included preprocessing steps removing unnecessary symbols, special characters and numbers that do not play any role in the meaning of the tweets. The cleaned model gets exposed to the model trained with Word2Vec to achieve word embedding. 80% of the data is utilised for training and the rest of the data is categorised as test samples. During the experiments, the proposed method achieved 86.2% accuracy in predicting the sentiments of the inputs.

E. ''Convolutional Neural Network (CNN) Method,'' [Dholpuria, Rana, and Agrawal], 2018

Based on the CNN model, the paper (Dholpuria, Rana, and Agrawal, 2018) explicates the sentiment analysis method for the imdb movie reviews dataset. With 3000 reviews, the dataset contains positive and negative labels. Preprocessing steps were performed by the authors to remove characters irrelevant to sentence meaning, symbols, duplicate words and stop words. The dataset is 75% distributed among the training set, while the rest of the data is contained in the test samples. The authors compared the CNN model with Naive Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbour (KNN) and Ensemble model with the observation that CNN predicts with a comparatively high accuracy of 99.33%. Thus, the authors, Thinh et al. (2019), propose CNN and RNN models for residual learning based sentiment analysis through this paper. An Internet Movie Database (IMDB) dataset with 50,000 data samples of positive and negative classes are used. Two sets of data are equally divided for training and test categories. In the CNN model, the CNN layers have 128 and 256 filters, while the RNN model has 128 units of LSTM, Bi-LSTM and Gated Recurrent Unit (GRU). The feature extractor based on the CNN model has achieved 90.02% accuracy compared to the RNN model.

F. ‘‘Bidirectional Encoder Representations from Transformers (BERT),” [Dhola and Saradva], 2021.

For the sentiment analysis, Dhola and Saradva (2021) have used both Machine Learning (ML) and DL techniques. While the ML techniques are SVM and Multinomial Naive Bayes (MNB), BERT and LSTM were selected from the pool of DL algorithms. The dataset contains 1.6 million tweets with only positive and negative labels. They have undertaken the preprocessing steps including tokenisation, stemming, lemmatisation and removal of stop words and punctuation. The data is apportioned into 80% and 20% for training and testing data respectively. The BERT method outperformed other techniques with an accuracy of 85.4%.

V. CONCLUSION

The power of sentiment analysis has already demonstrated its imperative in the realm of data science. Decision-making in business — in any kind of business — as well as in politics strongly relies on how people feel about the approach at hand or how they react to it. The quality of a product or method can definitely be gauged from the ratings or comments of the beneficiaries. Hence, the use of sentiment analysis will be increasingly essential to identify the most relevant attitudes or sentiments. Just a minor tweak can lead to a major spike in real-world outcomes. Conducting studies to develop a universal approach to sentiment analysis is the watchword. Given that, this project has aimed at proposing a model which is a combination of two different models, RoBERTa and LSTM, to handle the tasks at which each is suited. Finding the dataset to train the model posed a significant challenge. In the existing approach, prediction of positive and negative polarity is accomplished, whereas the proposed model succeeds in providing an additional neutral polarity. Enhanced performance on a custom dataset containing reviews from diverse domains demonstrates that the model is applicable more broadly. For now, the model is designed to support only one language and its modality. Taking it further, the key extension of the project will be to explore whether it can deliver equivalent levels of performance across multilingual or multimodal data environments.

REFERENCE

[1] Boaz Grinvald. Sentiment Analysis A Step By Step Guide (2021). [accessed on:21.07.2022]. 2021. URL: https://www.revuze.it/blog/sentiment-analysis-a-step-by-step-guide-2021/.

[2] Shashank Gupta. Sentiment Analysis: Concept, Analysis and Applications. [accessed on: 21.07.2022]. 2018. URL: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f.

[3] MonkeyLearn Inc. Sentiment Analysis: A Definitive Guide. [accessed on: 21.07.2022]. 2022. URL: https://monkeylearn.com/sentiment-analysis/#:~:text=Sentiment\ %20analysis\%20(or\%20opinion\%20mining,feedback\%2C\%20and\%20understand\ %20customer\%20needs.

[4] DataRobot. Introduction to Sentiment Analysis: What is Sentiment Analysis? [accessed on: 23.07.2022]. 2018. URL: https://www.datarobot.com/blog/introduction-to-sentiment-analysis-what-is-sentiment-analysis/.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: ArXiv abs/1907.11692 (2019)

