LSTM Based Text Classification

Dirash AR*[1], SK Manju Bargavi*[2]

[1,2]*Department of MCA, Jain University, Bangalore, Karnataka, India.*
kavyapitchai@gmail.com

Abstract—*This paper will review the literature on various methods and algorithms for analysing text sentiment. This paper will go through different machine learning algorithms for detecting sentiment in a text. Many of the algorithms used here had drawbacks, such as taking longer to train the model and using small datasets to train, resulting in lower performance. The performance of the model was improved by using LSTM, and it took less time to train the model. When compared to LSTM, many other approaches, such as RNN and CNN, are inefficient. Various companies use Sentimental Analysis to better understand their customers' reactions to their goods.*

Keywords—*Machine Learning, Natural Language Processing, Sentiment analysis, Text Analytics*

I. INTRODUCTION

Since the dawn of the Internet, social media has been a commonly used and important networking method. It is a powerful tool for disseminating knowledge and expressing opinions. Due to the large number of people who use social media on a daily basis, a large number of reviews, feedbacks, and articles have been produced. Many businesses connect with their customers through social media.. It is important for businesses to immediately determine whether a consumer review is positive or negative; this is known as "sentiment analysis."

Deep learning is a deep machine learning architecture that is inspired by our brain and consists of several layers of perceptron. Deep learning has been used in sentiment analysis in a number of successful studies. Methods that learn from a series of words include Long Short Term Memory (LSTM) and Dynamic Convolutional Neural Network (DCNN). Both strategies outperform traditional approaches. Long Short Term Memory networks, or LSTMs, are a type of short-term memory network. By designing weight coefficients between connections, the LSTM accumulates long-term relationships between distant nodes. Speech recognition, Natural Language Processing, and image captioning are only a few of the applications that these networks.

Sentiment analysis is the process of interpreting and categorizing emotions in text data using text analysis techniques. In the early stages, methods such as Naive Bayes, Support Vector Machines, and others are used to classify sentiment. Deep learning techniques (CNN, RNN, ANN) such as neural networks have recently gained popularity due to their impressive results. Businesses may use sentiment analysis software to determine how customers feel about products, brands, and services based on online reviews. Sentiment analysis is extremely useful in social media monitoring because it allows us to see how the public feels about a subject. Because of real-time tracking capabilities, social media monitoring tools like Brand watch Analytics make this process faster and simpler than ever.

Identifying the sentiments of customer about products and services is a game changing strategy if a company wants to take an edge on its competitors. Through this sentiment analysis companies can classify the customer conversation about a brand by key features are utility over brand products and service that customer care about or customers underlying reactions and interactions about these features or utilities and with the recent advances in the field of deep learning ability of RNN based algorithm to analyse text as improved considerably.

Recently, LSTM has become the most common method for sentiment classification. Hoch Reiter and Schmid Huber suggested LSTM in 1997, and it was refined and popularised by many people in subsequent work. They are now commonly used and perform exceptionally well on a wide range of problems. The long-term dependence problem is expressly ignored by LSTMs. They don't have to work hard to remember details for a long time; it's almost second nature to them. All recurrent neural networks take the form of a chain of neural network modules that replicate. This repeating module has a very basic structure at the level of RNNs, such as a single tanh layer.

II. LITERATURE SURVEY

[1] Here author conducted researches on sentiment classification for Chinese document. Here they used five learning methods (centroid classifier, K-nearest neighbour, winnow classifier, Naive Bayes and SVM) and Four feature selection methods (MI, IG, CHI and DF) and are investigated on a Chinese sentiment review dataset with a size of 1021 docs. The experimental results show that IG is the best at selecting sentimental words and SVM is the best for sentiment classification. [2] They demonstrate a device that gathers Tweets from social media sites. They used supervised learning algorithms and machine learning algorithms (Naive Bayes, maximum entropy classification). This method was able to identify emotions with a high degree of accuracy. [3] They proposed an advanced model that combined Naive Bayes, SVM, and Maximum Entropy with WordNet-based Semantic Orientation to extract synonyms and similarity for the content function. By using semantic analysis, this model produces a better performance, and the accuracy has also improved. [4] They developed a supervised statistical sentiment analysis framework that can detect the sentiment of short informal textual messages like SMS as well as the sentiment of a word within a message (term-level task). We introduced a number of features based on lexical categories and surface form. [5] This paper addresses sentiment polarity categorization, which is a basic issue in sentiment analysis. Experiments on sentence-level categorization as well as review-level categorization were conducted. [6] They use an LSTM-based feature extraction method to identify human behaviours in this paper; however, since they use a smaller dataset, the accuracy is not up to par. Dataset used here is by WISDM. [7] In this paper they used machine learning method to performed for sentiment classification of reviews, the use SVM algorithm. The problem is that they perform only Document level sentiment but not word level sentiment analysis. [8] They used machine learning algorithms to analyse the Twitter dataset and found that SVM and naive Bayes had the highest accuracy as compared to the other models. Their precision isn't up to par. [9] We use deep learning techniques to evaluate the emotions of Thai tweets. Deep learning techniques substantially outperform many traditional techniques, according to the findings: Except for Maximum Entropy, we perform an experiment to find the best LSTM and DCNN parameters using Nave Bayes and SVM. The best classifier, we prove, is DCNN, followed by LSTM. [10] The experimental results show that the C-LSTM outperforms both the CNN and the LSTM in these tasks and can produce excellent results. For sentence representation and text classification, C-LSTM is used. (SST) benchmark is the algorithm used. The drawback is that Tensor-based operations or tree-structured convolution can be used instead of regular convolution. [11] They are Using deep learning algorithm (RNN, CNN, DNN) for sentimental analysis. Using deep learning methods, sentiment analysis can be accomplished in more efficient and accurate way. The dataset used here is huge so it will take more time to train the model. [12] This paper will use the Deep learning algorithm CNN and LSTM to Generates caption for an image by using the flicker dataset. The model CNN to analyse the image uploaded and LSTM technique is used to generate captions for that image. [13] They present a model for sentiment classification in social media short texts based on LSTM. For sentimental classification, the LSTM technique was used. They used English movie reviews from IMDB. [14] The LSTM model was used in this paper to suggest a sentiment analysis using deep learning techniques. The model was not up to par, but it can be improved with the help of the Bidirectional LSTM network. [15] On the basis of long-term historical data, the authors proposed an LSTM-based model for predicting share price. They did so with the aid of the Raw Stock Price Database. They used LSTM to determine the best time period for forecasting the share's future price. The disadvantage is that it can only forecast future growth and not current growth.

III. ARCHITECTURE

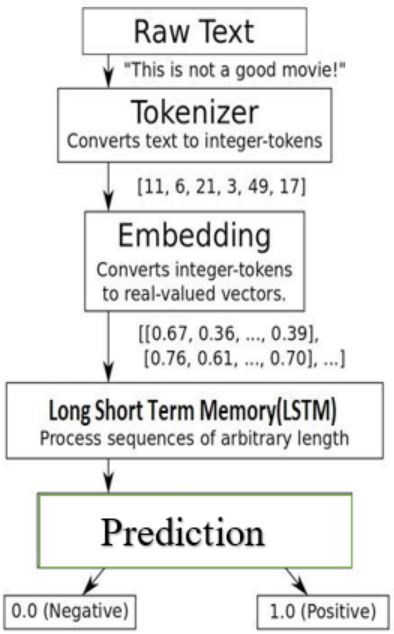

Fig. 1: Architecture of the model

3.1 Raw Text

The model is based on the IMDB dataset, which is owned by Amazon and stands for "Internet Movie Database." It contains knowledge about movies, video games, web series, and television series, among other things, which can be downloaded in a format suitable for neural networks using the keras.dataset. This dataset includes 25000 IMDB movie reviews, all of which have been reprocessed and labelled as either positive or negative.

3.2 Tokenizer

Tokenization is the method of separating or tokenizing a string of text into a list of tokens. A token can be thought of as a component of a sentence, and a sentence can be thought of as a component of a paragraph. Tokenization (splitting a string into its desired constituent parts) is essential to all NLP tasks. Since sentiment information is often sparsely and unusually expressed by a single cluster of punctuation, I believe tokenization is even more critical in sentiment analysis than it is in other areas of NLP**.**

3.3 Embedding

The field of Natural Language Processing (NLP), which combines computer science, artificial intelligence, machine learning, and computational linguistics, gave birth to the concept of word embedding. The text mining technique of word embedding establishes a relationship between words in textual data (Corpus). The context in which words are used determines their syntactic and semantic meanings. The distributional hypothesis proposes that words that appear in similar contexts have semantically similar meanings. The two main approaches to word embedding are count-based embeddings and prediction-based embeddings. Language relationships are captured by embeddings. Character embeddings are dense vector representations.

Word embeddings are dense vectors with a much lower dimensionality than other types of vectors. Second, the distance and direction of the vectors represent the semantic relationships between terms. It's a text representation in which words of the same meaning have equivalent representations. In other words, it represents words in an extremely complex structure in which similar words are clustered together based on a corpus of relationships. This part is usually done by an embedding layer in deep learning frameworks like TensorFlow and Keras, which stores a lookup table to map the words represented by numeric indexes to their dense vector representations.

3.4 Long Short Term Memory

The dataset is split into two parts: a training set and a test set. To solve a simple sentiment analysis problem, we'll build a neural network model. The LSTM algorithm is used to build a sentiment analysis classification model. Long short term memory is abbreviated as LSTM. They're a kind of RNN (recurrent neural network) that's great for sequence prediction. We may categorise feedback based on the following criteria. We can identify feedback based on emotion in a variety of ways, but we're using the most current technique, LSTM networks. The model can predict sentiment analysis on text using this approach, and it is very accurate.

3.5 Prediction

This text classification method looks at the input text and determines whether the underlined emotion is positive or negative, as well as the probability of such positive or negative statements. Probability represents the strength of a positive or negative claim. Using the Long Short-Term Memory networks algorithm, the model's accuracy is 86.68 percent.

VI. CONCLUSION

We suggest a sentiment classification method for text data based on LSTM in this paper. Users from all over the world express and share their views on a variety of topics. Since manual analysis of large volumes of such data is difficult, there is a legitimate need for computer processing. People's views and perceptions about goods, services, politics, social activities, and organisation strategies are analysed using sentiment analysis. The most interesting types of textual documents for sentiment analysis are reviews (from places like TripAdvisor, Amazon, and IMDB) and social network updates (mostly from Twitter and Facebook). When there is more training data, DL methods such as LSTM perform better with 85 percent accuracy in sentiment classification. A number of sentiment mining and classification schemes are investigated and referred to. Our findings revealed that the Long Short Term Memory Networks algorithmic standard outperformed others in terms of precision. We plan to expand this research in the future so that various embedding models can be considered on a wider range of datasets.

V. REFERENCES

[1] Tan, Songbo and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with applications,* 2008.

[2] Hemalatha, I. Varma, G. Saradhi and G. A, "Sentiment analysis tool using machine learning algorithms}{,," *International Journal of Emerging Trends \& Technology in Computer Science (IJETTCS,* 2013.

[3] Gautam, G. Yadav and Divakar, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis,"

in *2014 Seventh International Conference on Contemporary Computing (IC3)}*,, 2014.

[4] Kiritchenko, Svetlana, Z. Xiaodan and M. S. M, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research,* 2014.

[5] Fang, X. Zhan and Justin, "Sentiment analysis using product review data," *Journal of Big Data,* 2015.

[6] Chen, Y. Zhong, K. Zhang, J. Sun, Q. Zhao and Xueliang, "LSTM networks for mobile human activity recognition," in *2016 International conference on artificial intelligence: technologies and application*s, 2016.

[7] Luo, F. Li, C. Cao and Zehui, "Affective-feature-based sentiment analysis using SVM classifier," in *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD*), 2016.

[8] Kharde, Vishal and Sonawane, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971,* 2016.

[9] Vateekul, P. Koomsubha and Thanabhat, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE*, 2016.

[10] Zhou, C. Sun, C. Liu, Z. Lau and Francis, "A C-LSTM neural network for text classification," *arXiv preprint arXiv:1511.08630,* 2016.

[11] A. Qurat, T. Ali, M. Riaz, A. Noureen, A. Kamran, M. Hayat and B. Rehman, "Sentiment analysis using deep learning techniques: a review," *Int J Adv Comput Sci Appl,* 2017.

[12] Tan, Y. H. Chan and C. Seng, "Phrase-based image caption generator with hierarchical LSTM network," *Neurocomputing,* 2018.

[13] Wang, J.-H. Liu, Ting-Wei, L. Xiong and W. Long, "An LSTM approach to short text sentiment classification with word embeddings," in *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 201*8, 2018.

[14] M. R, D. S and J. B, "Sentiment Analysis of US Airlines Tweets Using LSTM/RNN," in *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 2019.

[15] Ghosh, A. Bose, S. Maji, G. Debnath, N. Sen and Soumya, "Stock price prediction using LSTM on Indian share market," in *Proceedings of 32nd international conference on*, 2019.