Comparative Study on Classification Algorithms for Sentiment Analysis

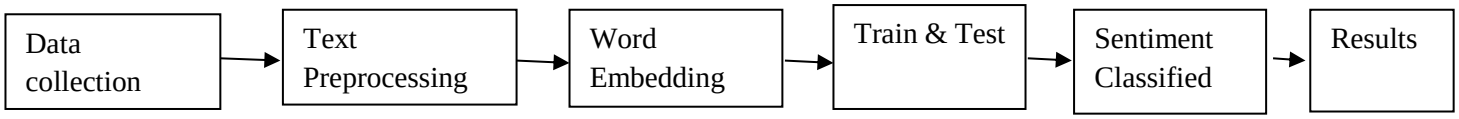
[1]T. Devishree, [2] K.S Jeen Marseline
[1] Research Scholar, [2]Head, Department of Information Technology
[1]Department of Computer Science
[1,2] Sri Krishna Arts and Science College, Coimbatore, India

Abstract - The paper is based on comparative study with the classifiers named Support Vector Machine (SVM) and Recurrent Neural Network Long-Short Term Memory (RNN LSTM). The process has taken place by two classification algorithms to measure the sentiment from customer reviews. Sentiment analysis may also know to be opinion mining, which is the process of determining whether the text reflects positive, negative or neutral sentiment. Using this analysis, business managers can acquire deep perception into customer opinions about their product. Customer opinion can bring any changes to a brand's success and the decision to monitor it can be the difference between a well-produced product and a missed opportunity. It can also inform marketing and product strategy by revealing chances to reframe the customer experience. By applying few metrics measures to produce the accuracy and finally, concluded by the comparative measures of the reviews from both the classifiers and finally among those classifier RNN LSTM shows the better results.

Keywords - sentiment analysis, SVM, LSTM RNN

I. INTRODUCTION

Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language [2]. The best business organization understands sentiment of their customers, what people are saying, how they are saying it, and what they actually mean. People express opinions in complex ways; rhetorical devices like sarcasm, irony, and implied meaning can mislead sentiment analysis. Sentiment Analysis is the domain of understanding these emotions with software. The object of sentiment analysis has typically been a product or a service whose review has been made public on the Internet. This might explain why sentiment analysis and opinion mining are often used as synonyms, although, we think it is more accurate to view sentiments as emotionally loaded opinions [3]. Comparing with many other fields, advances in Deep Learning have brought Sentiment Analysis into the foreground of cutting-edge algorithms. The work has been processed in few steps: Data collection, Text preprocessing, Word Embedding, Training and testing, Sentiment Detection, Sentiment Classified and Results.



The Social media possess enormous data which can be used for many purposes like analysis, classification. There are many classification algorithms in data mining especially for this sentiment or trend analysis can be done by Linear classifiers like logistic regression, Naïve Bayes classifier and Fisher's linear discriminant, support vector machines, quadratic classifiers, k-nearest neighbor, neural networks. And So these classification algorithm plays a great role in the fields like Email spam classification, Bank customer loan willingness prediction, sentiment analysis and drug classification, were the list gets extended. The classification Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering. Coming to the Decision Tree, given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. As like this, each classification algorithm varies by its method, processing way and they get differ in their accuracy perception. In addition, this comparative work is on two classifications are Support Vector Machine (SVM) and Recurrent Neural Network Long-Short Term Memory (RNN LSTM).

(i) SUPPORT VECTOR MACHINE (SVM)

$$K(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} \mathbf{X}_i \cdot \mathbf{X}_j & \text{Linear} \\ (\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C)^d & \text{Polynomial} \\ \exp(-\gamma |\mathbf{X}_i - \mathbf{X}_j|^2) & \text{RBF} \\ \tanh(\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C) & \text{Sigmoid} \end{cases}$$

SVM (support vector machine) is a supervised machine learning algorithm which can be used for classification or regression problems. It is implemented practically by using kernel, the technique to transform the data. After training the new data to same space to predict which category they belong and categorize the new data into different partitions and achieve it by training data.

SVM applies an iterative training algorithm, which is used to minimize an error function. In this, for categorical variables a dummy variable is created with case values as either 0 or 1. The main aim of the SVM is to train dataset that assigns new unseen objects into a particular category. It achieves this by creating a linear partition of the feature space into two categories. The kernel function has as follows:
From this, the $\mathbf{X}_{i*}\mathbf{X}_j$ Linear function has taken for the evaluation.

(ii) **RECURRENT NEURAL NETWORK LONG-SHORT TERM MEMORY (RNN LSTM)**
A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit dynamic temporal behavior for a time sequence. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs [1].The storage can also be replaced by another network or graph, if that incorporates time delays or has feedback loops. Such controlled states are referred to as gated state, which is part of long short-term memory (LSTMs) and gated recurrent units. The LSTM, it's a special kind of recurrent neural network which works for many tasks much better than the standard version. Almost all exciting results based on recurrent neural networks are achieved with them. In recurrent neural network, during the gradient back-propagation phase, the gradient signals can end up being multiplied a large number of times (as many as the number of time steps) by the weight matrix associated with the connections between the neurons of the recurrent hidden layer. A memory cell is composed of four main elements: an input gate, a forget gate and an output gate. The gates serve to modulate the interactions between the memory cell itself and its environment. The input gate can allow incoming signal to alter the state of the memory cell or block it. The output gate can allow the state of the memory cell to have an effect on other neurons or prevent it. Finally, the forget gate which work with the unwanted data and throws the over fitting data.

II. LITERATURE SURVEY

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities [6]. Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level [7]. And in the, With the advent of Web 2.0, people became more eager to express and share their opinions on web regarding day-to-day activities and global issues as well. Evolution of social media has also contributed immensely to these activities, thereby providing us a transparent platform to share views across the world [8]. Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories [10]. In some reports on classification or object detection competitions, deep networks obtain better results than those using SVM or other methods. We remark that both deep learning and SVM can be used to improve classification performance. The outputs of the neurons in the earlier layers feed into the neurons in the later layers [11]. The training process is more complex because the errors need to be back-propagated over different layers. Based on the progresses in machine learning communities, deep learning both as a method and a conception has already been applied to remote sensing filed such as classification by combination of spectral-spatial feature Zhao and Du (2016). But there is not so much research on how to decide the parameters when applying deep learning to classification tasks. In this paper, comparisons between the kernel linear method (SVM) and LSTM's efficient 'adam' optimization algorithm is used to perform the sentiment analysis with some metrics towards the customer's reviews and produce the accuracy score. The score acquired by both classifiers may show different in their performance, that going to be the further procedure.

III. PERFORMANCE METRICS

Most of the research studies on the analysis field has experimented by the metrics measures like precision, recall, f1 score either by combined or separately. The metrics may be included or excluded according to the problem and their purpose of the work. Each performance shows different accuracy and support scores.
The experiments have been performed on anaconda platform using python script and backend as tensorflow. The performance evolution is the method which is used to measure the uses of segmentation in SVM classifier and is carried out using the measures like Precision, Recall, f1 Score and Accuracy. In LSTM have used the softmax matrix, optimizer which taken place here is adam and accuracy metrics. The performance of the classifier can be analyzed using the following performance measures [15]. SVM perform well with both linear and non-linear data sets, here linear have been applied. SVM requires extensive training time [7]. Initially, the preprocessing takes place and the word embedding that has mentioned the dimension. The data set have been used here is product reviews of customers that is publicly available. In both classifiers the data set have been trained and tested. Then performed measures using some metrics individually, they are as follows:

(i) **Precision:** Precision is the ratio of correctly predicted positive to the total predictive positive, that which return only relevant instances.

$$\text{Precision} = \frac{TP}{TP+FP}$$

(ii) **Recall:** Recall is the ratio of correctly predicted positive observations to the all observations in actual class, which identify all relevant instances.

$$\text{Recall} = \frac{TP}{TP+FN}$$

(iii) **F1 score:** The F1 score metric, single metric that combines recall and precision with the weighted average. Therefore, this score takes both false positive and false negatives.

F1 Score = 2 *(Recall*Precision) / (Recall + Precision)

(iv) **Accuracy:** It is simply a ratio of correctly predicted observation to the total observations. Accuracy is a great measure but only while having symmetric datasets where values of false positive and false negatives are almost similar.

Accuracy = TP+TN / FP+FN+TN

Here,

- True positives(TP): the points labeled as positive that are actually positive
- False positives(FP): the points labeled as positive that are actually negative
- True negatives(TN): the points labeled as negative that are actually negative
- False negatives(FN): the points labeled as negative that are actually positive

On applying these metrics on the classification algorithms, have attained the accuracy measures which are to be compared, to know the classifier that shows better results.

IV. EXPERIMENTAL RESULTS:

The Support Vector Machine(SVM) which applied the metrics precision, recall, f1 score and resulted with average values of sentiment as positive, negative and neutral for the reviews. The table (Fig.1) below has shown the values acquired on SVM:

	precision	recall	f1-score	support
Negative	0.72	0.89	0.80	1700
Neutral	0.53	0.31	0.39	641
Positive	0.66	0.46	0.54	434
Average /total	0.67	0.69	0.66	2775

Fig.1

And the performance metrics of Long-Short Term Memory (LSTM) applied and the values acquired are shown below table (Fig.2).

Label	**Value**
Positive	64.40
Negative	89.84
Score	0.40
Accuracy	0.83

Fig.2

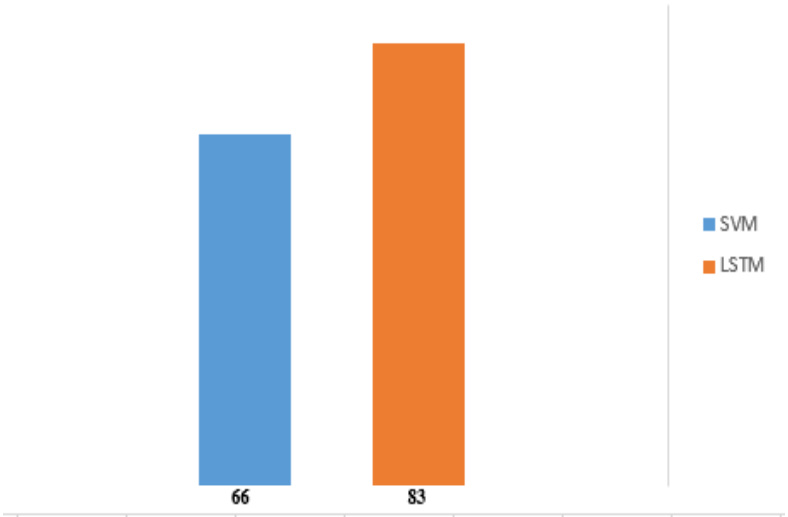


Fig.3

V. CONCLUSION

The performance metrics values of both Support vector machine (SVM) and Long-Short Term Memory Recurrent Neural Network (LSTM RNN) have been compared to see the better results from one of the both classifiers. The values acquired through LSTM the scores and accuracy such as score:0.40 and accuracy:0.83. Therefore, have to look at other parameters to evaluate the performance of the model. In this model, have got 0.83 which means model is approximate 80% accurate. Whereas the sentiment analysis of given dataset validated and achieved the separate accuracy results for positive and negative. From the graph, Fig.3 by comparing the both results, clearly know that LSTM shows the better results than SVM. Since, the data sets are not symmetrical in the quantity cannot show the accurate results but acquired the maximum level of approximate of the accuracy.

REFERENCE

[1] https:// en.wikipedia.org
[2] Liu B.Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Handbook of Natural Language Processing, Marcel Dekker, Inc. New York, NY, USA (2009).
[3] The evolution of sentiment analysis—A review of research topics, venues, and top cited papers Author links open overlay panel, Mika V.Mäntylä[a]DanielGraziotin[b]MiikkaKuutila[a], ELSEVIER ,Computer Science Review, Volume 27, February 2018, Pages 16-32
[4] S. Mulay, P. Devale and G. Garje," Intrusion Detection System Using Support Vector Machine and Decision Tree", *in International Journal of Computer Applications*, vol. 3, no. 3, pp. 40-43, 2010.
[5] J.M. Sokolova and G. Lapalme, A systematic analysis of performance Measures
[6] Sentiment analysis using product review data**,** Xing Fang, Justin Zhan.
[7] Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Google Scholar
[8] A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, kumar ravi and VadlamaniRavi
[9] Analyzing social media remark using sentimental analysis, T.Devishree, K.S.Jeen Marseline, 2018 IJRTI, Volume 3, Issue 6.
[10] Joachims T.Probablistic analysis of the rocchio algorithm with TFIDF for text categorization, presented at the ICML conference 1997
[11] Ruiz M, Srinivasan P. Hierarchical neural networks for text categorization, presented at the ACM SIGIR conference 1999