Sentiment Analysis of Twitter Tweet Using Machine Learning

Alok Ranjan Tripathy, Amit Kumar Moharana, Alakananda Tripathy, and Sipra Sahoo

Abstract To express emotions, the users are using emoticons. Previously, for the classification of text, emoticons, or image machine learning techniques, are being used, but emoticons to text being ignored, resulting in the misrepresentation of many emotions. Sentiment analysis is the evaluation of people's attitudes, thoughts, and feelings in order to determine whether they are optimistic, negative, or neutral. In recent years, the use of such emoticons on social media has nearly doubled. This study proposed an algorithm and pattern for sentiment analysis that makes use of both text and emoticons, this study shows that when emoticons are used, the emotion associated with them outweighs the sentiment expressed by textual data processing. This paper defines a social media sentiment analysis scheme that categorizes posts as positive or negative depending on the general polarity of the message. These classifiers often give a judgment score very similar to the decision boundary for a significant number of posts, implying that they are clearly confused rather than absolutely incorrect about these tweets. This paper uses different techniques like Naïve Bayes, SVM, LSTM, logistic regression model to analyze the sentiment.

Keywords Sentiment analysis · Naive Bayes · SVM · Twitter tweet dataset · LSTM

A. R. Tripathy · A. K. Moharana (✉)
Department of Computer Science, Ravenshaw University, Cuttack, Odisha, India

A. Tripathy · S. Sahoo
Department of Computer Science and Engineering, S'O'A Deemed to be University, Bhubaneswar, Odisha, India
e-mail: alakanandatripathy@soa.ac.in

S. Sahoo
e-mail: siprasahoo@soa.ac.in

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 860, https://doi.org/10.1007/978-981-16-9488-2_39

1 Introduction

The popularity of social networks and society dependent on mobile media has led the young scientists to continue to work on feeling study. Organizations from those days have been keen to evaluate consumers or the public view of their social media products [1]. Online services are linked to the site, internet forum, remarks, tweets, and product data evaluation of social communication [2].

There is an increase in the popularity of the social and electronic media networks. Societies also encouraged to carry out research on feeling analysis via the extreme use of the internet by companies all over the world [3]. Web texts have shaped the market and socio-economic processes. To analyze the sentiments, different methods drive the computer control based on four text classifiers, namely, Naïve Bayes, SVM, LSTM, and Random Forest.

Sentimental Analysis

It is also known as emotion AI or opinion polling. It mostly focuses on identifying subjective data. It is useful to determine how pleased customers are with goods and services. Sentiment analysis examines the polarity of language, determining whether it is positive or negative. Change the notion, boost production, and advertise with the aid of these polarities to help eliminate some negative. The various steps to analyze sentiment data are given in Fig. 1.

Section 2 discusses the related work, the proposed model is being discussed in Sects. 3, and 4 highlights the analysis of results, and Sect. 5 concludes the paper.

The paper highlights the different machine learning techniques to analyze the sentiment of the Twitter tweet.

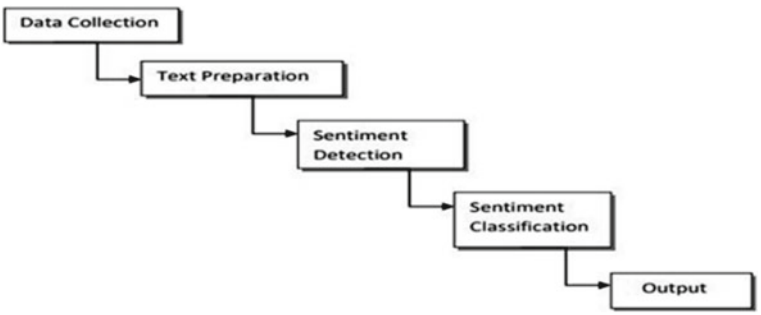

Fig. 1 Steps to analyze the sentiment data

2 Related Work

Badr et al. [4] used SVM and NB algorithms to analyze social media sentiment and used ant colony and particle Swarm optimization methods to achieve 73.62%, 77.30%, and 80.54% accuracy for Naïve Bayes and 76.71%, 80.54% for SVM, respectively. Two approaches were discussed by the author Kawade, [5], sentiment score and polarity count, to analyze the social network Uri assault tweets, achieving 94.3% accuracy for negative results and 5.7% accuracy for positive results.

Nguyen et al. [6] attained an F1-score of 90.2% accuracy utilizing the Vietnamese student feedback corpus and the LSTM support vector machine technique. Taking Twitter tweets, reviews and applying different machine learning algorithm and deep learning algorithm such as NB, SVM, LSTM, and Random Forest. Amazon and IMDB movie reviews were taken by Bansal and Kaur [7] and consider NB, J48, BFtree, and oneR classifier among these the faster one in learning is NB and the more promising one is oneR, for generating the accuracy in classified instances. In [8], Shreyas R Labhsetwar et al. use a dataset of 1000 labeled sentences, 500 +ve and 500 −ve, to conduct sentence-level analysis. It appears that using conjunct analysis with sentence-level analysis will improve accuracy. They discovered that the ML method is inefficient, therefore he used WordNet to improve the accuracy by around 80%. The POSICTCLAS tool for Chinese text is being suggested by Bhargav et al. [9] where the average review size is about 600 words in education review, average reviews of length are about 460 terms in stock review, and 120 words average length in computer review.

Singh et al. are taking sentiment analysis on movie and product review dataset in Turkish and English language using SVM classifier to obtain 91.33% accuracy [1]. Khan et al. [10] worked on Urdu dataset for sentimental analysis and polarity detection.

3 Proposed Model

The Twitter dataset using machine learning and deep learning classifiers such as Naive Bayes, LSTM, SVM, and Logistic Regression is being discussed in the proposed model. The procedure is discussed in Fig. 2.

The different steps of the process include:

- Data collection is very essential. In order to discover or analyze sentiment, a collection of Twitter tweets from the database is used in the proposed model.
- Data Preprocessing is important in data mining. When there is irrelevant, redundant information, noisy or inaccurate data, knowledge discovery become more difficult. It takes a long time to prepare data, and after this step is over, it is time to create the training set. The different steps of preprocessing of data consist of:
  - **Remove Punctuation**: It removes the special characters.

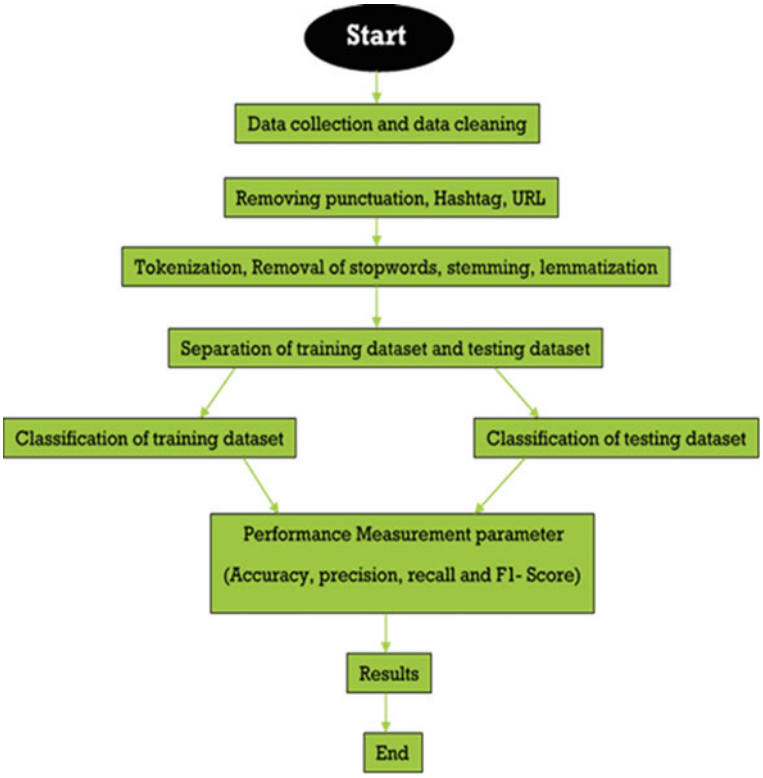

Fig. 2 Flowchart of the proposed model

- **Tokenization**: It breaks a long text into little segments or words, which are referred to as tokens. The phrase is broken up and converted into a token format.
- **Remove Stop words**

  It is a Natural Language Toolkit (NLTK) that includes a collection of terms. It just functions as a filter and used to filter out natural language data after or before processing.
- **Stemming**

  The practice of reducing a term to its underlying word by removing superfluous characters. It reduces the term to its simplest form.
- **Lemmatizing**

  This converts a word's result into its dictionary or canonical form. The resulting lexical form is known as lemma.
- **Vectorization**

This converts the word into a number, which is executed faster. It aids word embedding or word vectorization approach by employing vectorization technology.

– **Bag of word (BOW)**

As the machine learning algorithm is unable to operate with text, it must be transformed into numbers. Because the algorithm requires a vector input, first the materials are transformed into fixed length vectors.

- **Classification of Training Data**

It is the next step, four different classifiers are used for the proposed model.

– **Naive Bayes**:

This is a strong algorithm. It is a probabilistic model that looks like Bayes theorem of the algorithm. It is used to determine the assumption of predictor independence. It is the most basic model that is straightforward to construct.

– **Logistic Regression**:

It is a supervised classification method that uses statically learning techniques. It is a regression model, after all it predicts the probability in a given dataset. It makes use of the sigmoid function. Because it detects defaulters, this approach is employed in the financial industry. It may also be used to forecast binary classes. Logical regression's output, or goal value, is binary in nature. It is used to detect spam, diagnose cancer, and forecast diabetes. Logistic regression may be classified into three types based on number categories: binary, multinomial, and ordinal. Binary and multinomial models are the two types of models.

– **Support Vector Machine (SVM)**:

The enhanced feature sets were utilized for sentiment classification after the outliers were removed using clustering. SVM is mostly used to classify sentiments. It categorizes good and negative feedbacks. Precision, Recall, F-Measure, and Accuracy are all factors that influence the algorithm's performance.

– **Bullet Long Short-Term Memory (LSTM)**:

The inputs are multiplied by weight then the bias is added, and so on, until the output from the last layer is obtained in the feed-forward networks. However, because these networks do not retain memory, they cannot be utilized to process sequential data. This sort of network's input and output are also fixed. Long Short-Term Memory (LSTM) networks can learn long-term dependencies. These networks perform admirably in a wide range of situations. Long-term dependence is not a concern with LSTMs because they are expressly intended to avoid it. It is achieved by LSTM memorizing information over lengthy periods of time, as this is their inherent behavior. The chain structure also exists on LSTM networks, but the repeating module has a different structure.

It contains four interacting layers instead of a single layer of a single neural network.

4 Result Analysis

The proposed model is classified using Naïve Bayes classifiers. In this, the Twitter Tweets are taken as dataset, around 60,000 pieces of data are taken of which 30,000 pieces are used for training and 30,000 pieces are used for testing. The word vector of positive and negative reviews is kept separated, and also it keeps track of positive and bad ratings. Then, using conditional probability, best words are determined. After training the data using Navies Bayes Classifier, the result is obtained as shown in Fig. 3.

There is a comparison graph between SVM and logistic regression based on accuracy and prediction, and it is shown in Fig. 4.

Precision = Correct true predictions/Total number of positive prediction

$$Precision = \frac{TPP}{TPP + FPP} \tag{1}$$

Recall is to figure out the percentage of item actually present in the input.

Recall = Correct true predictions/total of true positive prediction and false-negative prediction

$$Recall = \frac{TPP}{TPP + FNP} \tag{2}$$

F-Measure is the ratio of the harmonic mean of accuracy and memory, which combines precision and recall.

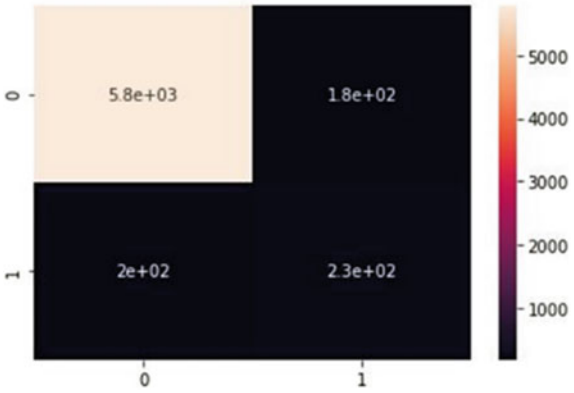

Fig. 3 Naive Bayes classifier

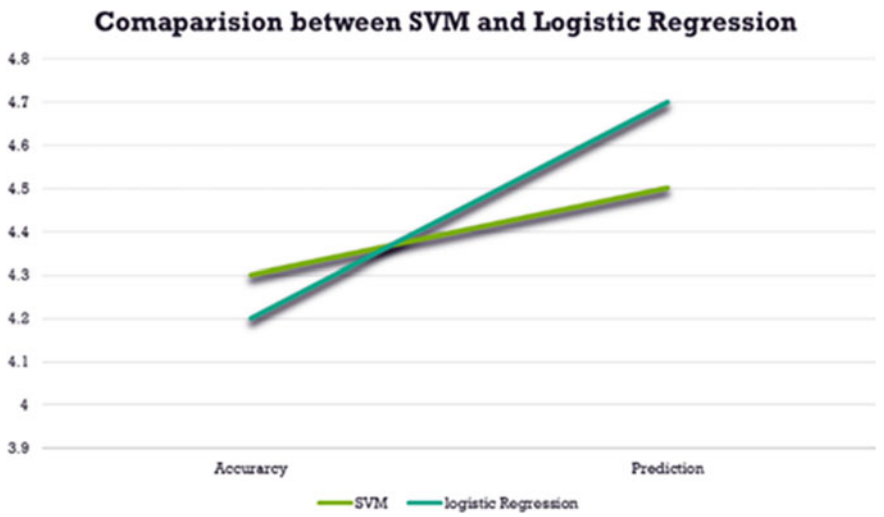

Fig. 4 Comparison between SVM and logistic regression

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Accuracy is the statistical metric for determining how effectively a classification test properly detects or eliminates a condition. The proportion of genuine findings (including true positive and true negative) among the total number of cases is investigated in the accuracy:

$$Accuracy = \frac{TPP + TNP}{TPP + TNP + FPP + FNP} \tag{4}$$

In Eqs. (1), (2) and (4), TP signifies the True Positive Prediction, False Positive Prediction is denoted as FPP, True Negative Prediction is denoted as TNP, and FNP is denoted as False Negative Prediction.

Figure 5 shows the count of words using LSTM. The accuracy and loss factor of LSTM using bidirectional NN are shown in Fig. 6.

The results for the different parameters for the different classifiers are shown the Table 1.

5 Conclusion

In the proposed work, different techniques of deep learning and machine learning are applied to sentiment analysis for review of the tweets in Twitter. It is the most popular and famous topic through which can know the sentiment of any post or any type of

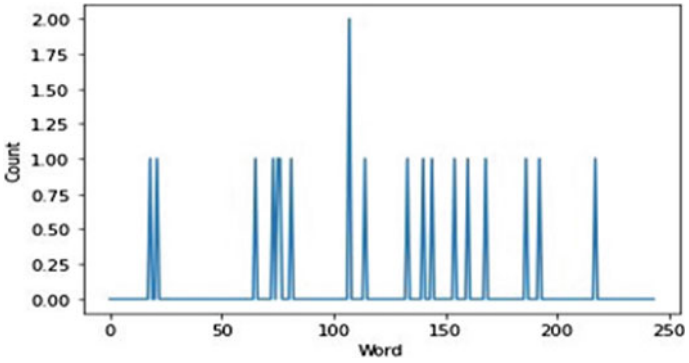

Fig. 5 Count of words using LSTM

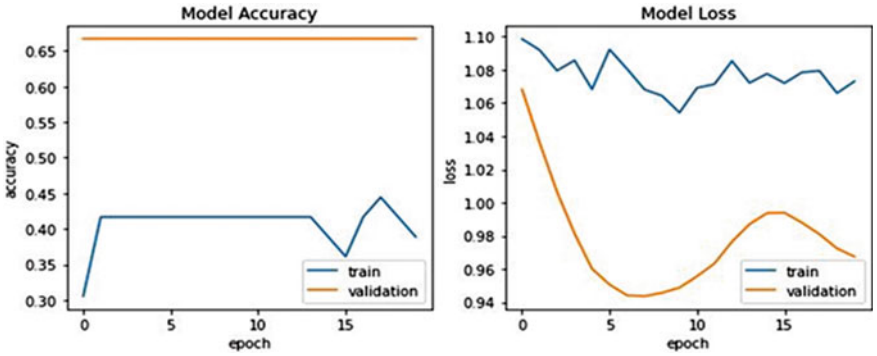

Fig. 6 Accuracy and loss factor of bidirectional LSTM using NN

Table 1 Comparison based on different parameters for different techniques

	SVM	Logistic regression	LSTM	Naive Bayes
Accuracy	0.831	0.74	0.88	0.843
Precision	0.54	0.63	0.75	0.71
Recall	0.55	0.563	0.63	0.59
F1-Score	0.55	0.581	0.78	0.64

text document. The intension of people about other things or people can be known through the sentiment analysis. In this model, the sentiment analysis classification is performed, and their prediction is done through Naïve Bayes, SVM, LSTM, and logistic regression based on different parameters like accuracy, precision, recall, and F1-score. It is observed that for LSTM has better accuracy, precision, recall and F1-score is more as compared to the other model.

References

1. Singh J, Singh G, Singh R (2017) Optimization of sentiment analysis using machine learning classifiers. HCIS 7(1):1–12
2. Mishra P, Danda P, Dhakras P (2018) Code-mixed sentiment analysis using machine learning and neural network approaches. arXiv preprint arXiv:1808.03299
3. Ullah MA, Marium SM, Begum SA, Dipa NS (2020) An algorithm and method for sentiment analysis using the text and emoticon. ICT Exp 6(4):357–360
4. Badr EM, Salam MA, Ali M, Ahmed H (2019) Social media sentiment analysis using machine learning and optimization techniques. Int J Comput Appl 975:8887
5. Kawade DR, Oza KS (2017) Sentiment analysis: machine learning approach. Int J Eng Technol 9(3):2183–2186
6. Nguyen VD, Van Nguyen K, Nguyen NLT (2018) Variants of long short-term memory for sentiment analysis on Vietnamese students' feedback corpus. In: 2018 10th international conference on knowledge and systems engineering (KSE). IEEE, pp 306–311
7. Bansal P, Kaur R (2018) Twitter sentiment analysis using machine learning and optimization techniques. Int J Comput Appl 179(19):5–8
8. Labhsetwar SR (2019) Sentiment analysis of customer satisfaction using deep learning. Res J Comput Sci (IRJCS) 6:709–715
9. Bhargav PS, Reddy GN, Chand RR, Pujitha K, Mathur A (2019) Sentiment analysis for hotel rating using machine learning algorithms. Int J Innov Technol Expl Eng (IJITEE) 8(6):1225–1228
10. Khan MY, Nizami MS (2020) Urdu sentiment corpus (v1. 0): linguistic exploration and visualization of labeled dataset for Urdu sentiment analysis. In: 2020 international conference on information science and communication technology (ICISCT). IEEE, pp 1–15