International Journal of Advanced Research and Review

www.ijarr.org

Analyzing Twitter Data for Text Classification

[1] Thirluka Balanandini, [2] V. Sravani,

[1] Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract

Predicting the polarity of words and subsequently classifying them into positive or negative sentiment is the major emphasis of sentiment analysis, a classification issue. The two most common kinds of classifiers are those that rely on a vocabulary and those that employ machine learning. Word Sense Disambiguation and SentiWordNet are examples of the former, whereas RNN Classifier, Multinomial Naive Bayes (MNB), Logistic Regression (LR), Support Vector Machine (SVM), and others are examples of the latter. In this paper, we make use of two preexisting datasets: one from Stanford University's "Sentiment140" with 1.6 million tweets, and another from CrowdFlower's Data for Everyone library with 1,837 entries; both datasets have already been classified according to the sentiments conveyed within them. We compare the results achieved by the following sentiment classifiers—Textblob, Sentiwordnet, MNB, LR, SVM, and RNN—using the aforementioned dataset to categorize tweets as positive or negative. In addition to the aforementioned machine learning methods, the datasets have also been subjected to an ensemble version of MNB, LR, and SVM. In addition, you may utilize the learned models mentioned earlier to forecast the sentiment of fresh data.

Index Terms

Text Analysis on Twitter, Sentiment Analysis, RNN, Multinomial Naive Bayes (MNB),

I. INTRODUCTION

The term "sentiment analysis" refers to the process of deducing an article's emotional tone from a huge body of text by using methods like Natural Language Processing (NLP). It delves into the author's feelings and thoughts about the subject matter discussed in the work. You may find this content in a variety of places, including documents, social media posts, and databases. There are many ways to categorize people's feelings: positive, negative, neutral, or objective. Both lexicon-based and machine learning-based approaches may be used for this classification. When it comes to classifying sentences, there are two main approaches. One uses an existing dictionary with pre-assigned scores to each word; the other uses machine learning algorithms to train a model with labelled data; and finally, it uses the model to predict a class for new text. With millions of people regularly posting public comments, Twitter has become an invaluable resource for sentiment research on any subject. As a result, it is currently among the most popular microblogging sites. To uncover the dominant sentiments in tweets, this article uses lexical and machine learning based methods. For Lexicon-based Sentiment analysis, Textblob, SentiWordNet, and Word Sense Disambiguation provide the right meaning of a word in a certain context. As for machine learning-based methods, RNN Classifier, MNB, and LR have been used. This study presents a comparison based on how well each method

predicts the emotion of a certain tweet. We are also comparing the outcomes of an ensemble strategy that uses majority voting of MNB, LR, and SVM on the datasets with the rest of the approaches.

II. RELATED WORK

A vast and rapidly expanding area, sentiment analysis is managed as a natural language processing activity. To address the issue at hand, several technological techniques and libraries have been developed, ranging from document level classification to phrase level classification. A lexicon-based technique, which uses pre-built sentiment dictionaries, and a machine learning-based approach, which trains the computer using the available data, are two ways to do opinion mining that address the issue. Data extraction from Twitter, language processing, and sentiment analysis on text using machine learning algorithms are all part of the full method. Deep neural networks have been more effective than older methods of machine learning. In his paper, Ibrahim[15] lays out a method for election forecasting that involves finding buzzers in campaign-era Twitter data, doing sentiment analysis by assigning a polarity to each subtweet, and finally determining the overall sentiment of all tweets using this polarity. To enhance accuracy compared to employing separate machine learning algorithms, Rincy Jose[7] suggested an ensemble strategy of majority voting. The data gathered is assessed using a lexicon-based methodology in the study by Khin Zezawar [18] on the teacher assessment technique or student feedback. In a study by Dan Li, an upgraded model LSTM is employed for text classification instead of regular RNN to enhance the performance of regular RNN [17]. This article provides an overview of the various algorithms and an analysis of their outcomes by drawing on all the algorithms and methods utilized in the preceding studies. When predicting the likelihood of a collection of characteristics belonging to a certain class label, Multinomial Naive Bayes Classifier makes use of Bayes Theorem. It relies on the premise that various occurrences' probabilities are unrelated to one another. [1] Binary logistic regression classifier that divides an input value into two categories according to the possible values for the output. It approximates the result using the sigmoid function. It utilizes the sigmoid function for data modeling. The sigmoid function has an output between 0 and 1, and a classification threshold of 0.5 is used. [12]

$$g(z) = \frac{1}{1 + e^{-z}} \qquad (1)$$

Support vector machines (SVMs) are a kind of classification that search for the best possible hyperplane that physically separates the classes from one another. It uses the hyperplane to categorize any newly added points. With x being the input vector and w being the hyperplane's normal,

$$\vec{w}.\vec{x} + b \geq 0 \qquad (2)$$

then x belongs to the positive class and

$$\vec{w}.\vec{x} + b < 0 \qquad (3)$$

it is considered to be part of the negative category. the eleventh In a recurrent neural network (RNN), each layer is linked end-to-end such that the output of one layer may be used as an input in the layer below it. Additionally, it reduces parameter complexity by using the same parameters in each cycle/layer. Learning the past using RNN is possible, but with each partial extraction, the older inputs' contribution to the output decreases and eventually disappears, causing us to lose potentially crucial knowledge. One name for this is the "vanishing gradient" issue. Using the network's memory—the cell state—and the gates that control the flow of information via individual cells, Long Short-Term Memory networks solve the problem of coping with long-term dependence. An input gate it, a forget gate ft, an output gate ot, and a memory cell Ct are all part of the set of vectors that exist at each time step t. The hidden layer ht's output is calculated using all of them together. The LSTM network consists of four stages, numbered 1, 2, 3, and 4. 1) The model starts by cleaning up the current cell of any extraneous data. The "forget gate" (ft) describes this phenomenon. 2) The amount of additional information that will be added to cell state is determined in the following phase. The term for this is the input gate (it). Step three involves updating the current memory cell (Ct) with a combination of data supplied via earlier stages. 4) The output ht is calculated in the last phase. This modified form of the cell state determines the amount of new memory to send to the next LSTM unit,

and the output is dependent on that. We will print this result and send it on to the next network phase. [14] Here are the symbols and formulae used in Long Short-Term Memory (LSTM):

$$f_t = \sigma(P_f * x_t + V_f * h_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma(P_i * x_t + V_i * h_{t-1} + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(P_c * x_t + V_c * h_{t-1} + b_c) \quad (6)$$

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (7)$$

$$o_t = \sigma(P_o * x_t + V_o * h_{t-1} + b_o) \quad (8)$$

$$h_t = o_t * \tanh(C_t) \quad (9)$$

$$\sigma - Sigmoid\ activation\ function$$

$$\tanh - Hyperbolic\ tangent\ activation function$$

$$x_t - Input\ at\ time\ t$$

$$C_{t-1} - Memory\ \ from\ previous\ block$$

$$h_{t-1} - Output\ of\ previous\ block$$

$$C_t - Memory\ from\ current\ block$$

$$h_t - Output\ of\ current\ block$$

$$P_i, P_c, P_f, P_o, V_i, V_c, V_f, V_o - Weight\ matrices$$

$$b_i, b_c, b_f, b_o - Bias\ vectors$$

Paper	Dataset	Method	Result
paper1	Twitter data during campaign period	Sentiment aggregation method for each subtweet.	MAE-0.6%
paper2	political online news from folha	Sentiment lexicon-SentiLex	Highest accuracycross validation-83.24%
paper3	Live data from Twitter Streaming API	Ensemble Approach	Accuracy-71.48%
paper4	Live data from Twitter	Textblob,W-WSD	W-WSD Accuracy-62%

TABLE I RESULT-SUMMARY OF PREVIOUS PAPERS

paper1	Buzzer Detection and Sentiment Analysis for predicting presidential Election Result in a Twitter Nation.
paper2	Sentiment based Features for Prediction Election Polls:Case Study on the Brazilian Scenario
paper3	Prediction result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach
paper4	Machine learning-Based Sentiment Analysis for Twitter Accounts

TABLE II PAPER-DESCRIPTION

Table 1 displays an overview of the outcomes from the preceding study. The articles that were described are included in Table 2.

III. PROPOSED APPROACH

A.Dataset Description • A dataset of tweets labeled as positive or negative has been selected for the sentiment analysis model. "Sentiment140" is a dataset that originates at Stanford University and is used to train the algorithm. A few of the fields included in the collection include the tweet's polarity (0=negative, 4=positive), its id, its date, its user, and its content. A total of 1.6 million samples were used, with 0.8 million indicating a positive attitude and the remaining 0.8 million a negative trait. Another dataset that was used was "Crowdflower's Data for Everyone library." This dataset had 13871 samples, out of which 8493 were negative, 3142 were neutral, and 2236 were positive. The training set size was 70% and the testing set size was 30% for both datasets. Section B. Approach The tweets need to be NLTK preprocessed before they can be used directly for analysis. This will make them readily usable. Tokenizing the tweets, converting all uppercase letters to lowercase, using a language translator to convert to a common language, removing stopwords and punctuation, classifying as Part Of Speech (POS), extracting data from HTML and XML files, using a spellchecker, and converting to a common language were all part of the preprocessing. Following this preprocessing, several machine learning algorithms are given the cleaned-up text. The first is MNB, which uses probability and the idea that word occurrences in distinct classes are independent to operate. The next step is to apply LR on the provided datasets. For binary classification, it is a popular linear classifier. Support Vector Machines (SVMs) are the next algorithms employed; they choose the best hyperplane to divide the classes. A RNN with LSTM, a powerful tool for sequential learning, was subsequently implemented. In our example, it learns from the whole sequence, which it receives as input. The RNN model used in the strategy under discussion is shown in Figure 1. Additionally, algorithms for sentiment analysis that are based on corpora or dictionaries are used and evaluated alongside these machine learning methods. Using majority voting as an ensemble strategy on MNB, LR, and SVM, as well as for classification and result comparison, has helped enhance the outcomes a little. As the data flows through the processes, it is shown in Figure 2.

IV. RESULTS & COMPARISON

Table 2 shows the results obtained by various classifiers on the two datasets. Graph 1 is the graphical representation of the results in table 1 for Dataset 1 and Graph 2 is a similar representation for the other Dataset.

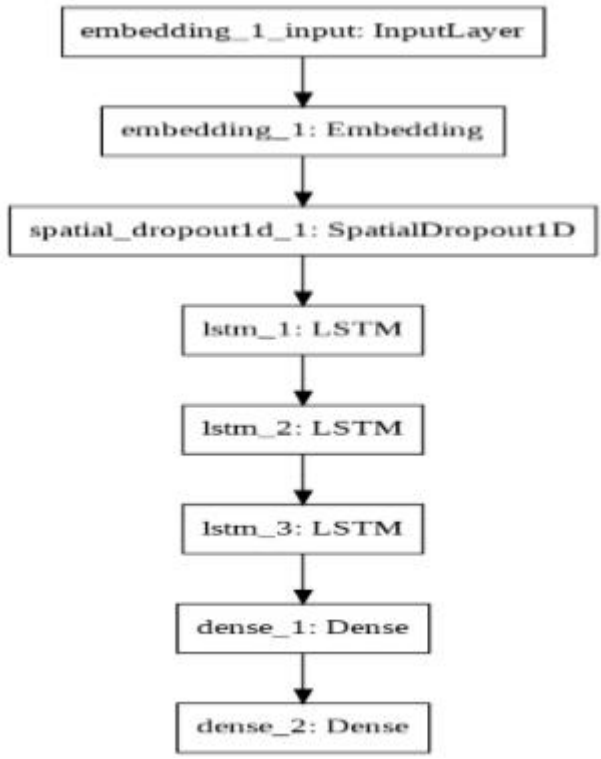

Fig. 1. LSTM-NETWORK

Approach	Dataset1(From Stanford)	Dataset2(From Crowdflower)
SentiWordNet	45%	40 %
Multinomial Naive Bayes	76.52%	63.88%
Logistic Regression	76.44%	65.42%
Support Vector Machine	77.98%	68.90%
Recurrent Neural Network With LSTM	82%	66%
Ensemble	77.67%	68.50%

TABLE III COMPARISON OF ALL ALGORITHMS

Approach	Precision	Recall	F-Score	accuracy
Multinomial Naive Bayes	77%	77%	76%	76.52%
Logistic Regression	76%	76%	76%	76.44%
Support Vector Machine	78%	78%	78%	77.98%
Recurrent Neural Network With LSTM	81%	84%	82%	82%
Ensemble	78%	78%	78%	77.67%

TABLE IV DETAILED RESULT FOR DATASET1

Approach	Precision	Recall	F-Score	accuracy
Multinomial Naive Bayes	68%	64%	53%	63.88%
Logistic Regression	64%	65%	65%	65.42%
Support Vector Machine	67%	69%	68%	68.90%
Recurrent Neural Network With LSTM	71%	69%	68%	66%
Ensemble	67%	69%	67%	68.50%

TABLE V DETAILED RESULT FOR DATASET2

As has been observed from the results in table 3, RNN with LSTM gives the best accuracy. This is also in accordance with

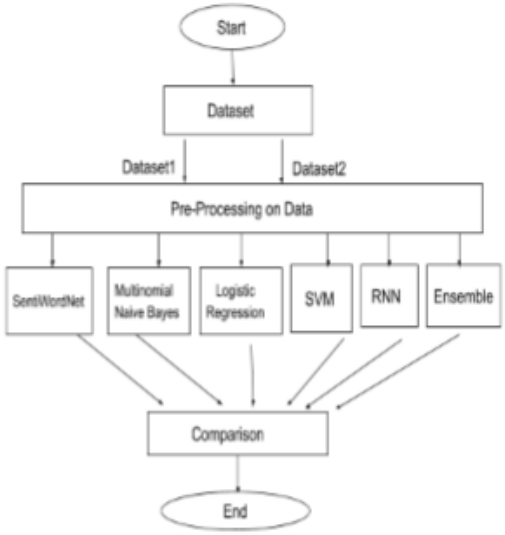

Fig. 2. Flow-Chart

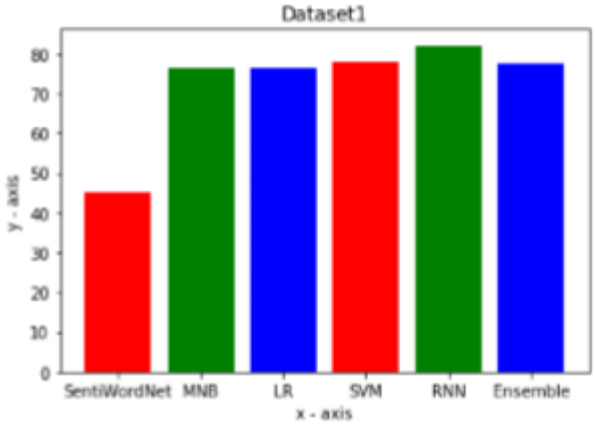

Fig. 3. Comparison for Dataset1

Label	Positive	Negative
Positive	193297	46064
Negative	66629	174010

TABLE VI CONFUSION MATRIX OF NAIVE BAYES FOR DATASET1

Label	Positive	Negative
Positive	182531	56830
Negative	56243	184396

CONFUSION MATRIX OF LOGISTIC REGRESSION FOR DATASET1

RNN is generally understood to learn from the whole sequence, which means that it will, in this instance, acquire knowledge about the entire phrase rather than individual words, in contrast to other algorithms.

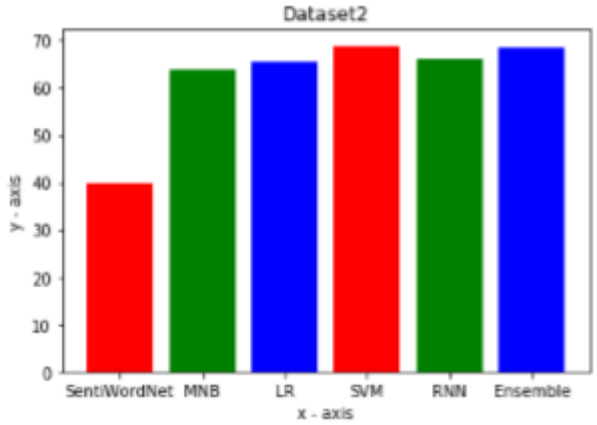

Fig. 4. Comparison for Dataset2

Label	Positive	Negative
Positive	184801	54560
Negative	53332	189527

TABLE VIII CONFUSION MATRIX OF SVM FOR DATASET1

Label	Positive	Negative
Positive	185603	53758
Negative	53397	187242

TABLE IX CONFUSION MATRIX OF ENSEMBLE APPROACH FOR DATASET1

Label	Negative	Neutral	Positive
Negative	2505	18	0
Neutral	902	72	22
Positive	550	10	82

TABLE X CONFUSION MATRIX OF NAIVE BAYES FOR DATASET2

Label	Negative	Neutral	Positive
Negative	1989	344	190
Neutral	476	388	132
Positive	197	99	346

TABLE XI CONFUSION MATRIX OF LOGISTIC REGRESSION FOR DATASET2

Label	Negative	Neutral	Positive
Negative	2135	257	131
Neutral	493	388	115
Positive	215	82	345

TABLE XII CONFUSION MATRIX OF SVM FOR DATASET2

Label	Negative	Neutral	Positive
Negative	2180	225	118
Neutral	550	345	101
Positive	254	62	326

TABLE XIII CONFUSION MATRIX OF ENSEMBLE-APPROACH FOR DATASET2

in its own right before tackling a vocabulary list. Two probabilistic classifiers are logistic regression and Naive Bayes. Because of this, there is little to no variation in their accuracies.of paramount importance. Alternatively, Support Vector Machines (SVM) outperform MNB and LR because they are geometric classifiers that take independencies into account. While applying the same algorithms to a smaller dataset, it was found that the larger dataset produced superior results in terms of accuracy. Since the data being observed is highly unstructured and does not adhere to any rules regarding grammar, spelling, or symbols, it can be concluded that supervised machine learning algorithms perform better than unsupervised lexicon-based algorithms. However, in order to achieve better results, supervised algorithms require a larger dataset. The comprehensive results of the performance metrics for Dataset 2 are shown in Table 5, whereas those for Dataset 1 are shown in Table 4. A weighted average of the accuracy, recall, and F-measure values are taken into account.

V. CONCLUSION

This project applies and compares a number of methods, some of which are lexicon-based and others of which are machine-learning based. Classification using traditional dictionaries is much less accurate when applied to novel data/text than machine learning based models trained on related data. This is due to the fact that the observational text, in this case tweets, is very informal and does not adhere to traditional grammar and spelling conventions, resulting in extremely unstructured data. The findings of the comparison are also readily apparent when looking at various machine learning techniques. So far, RNN stands out as the most accurate algorithm in use.

REFERENCES

[1] Harpreet Kaur, Veenu Mangat, Nidhi, “A survey of sentiment analysis techniques”, February 2017,2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

[2] Ali Hasan, Sana Moin, Ahmad Karim And Shahaboddin Shamshirband, “Machine Learning-Based Sentiment Analysis For Twitter Accounts”, February 2018. 3.

[3] Biswarup Nandi, Mousumi Ghanti, Souvik Paul, “Text Based Sentiment Analysis”, November 2017, 2017 International Conference on Inventive Computing and Informatics (ICICI). 4.

[4] Bhagyashri Wagh, Prof. J. V. Shinde, Prof. P. A. Kale, “A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques”, December 2017,International Journal of Emerging Research in Management & Technology. 5.

[5] Hamid Bagheri, Md Johirul Islam, “Sentiment analysis of twitter data”, Iowa State University . 6.

[6] Apoorv Agarwal, Vivek Sharma, Geeta Sikka, Renu Dhir, “ Opinion mining of news headlines using SentiWordNet”, March 2016, 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[7] Rincy Jose, Varghese S Chooralil, “Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation”, November 2015, 2015 International Conference on Control Communication Computing India (ICCC).

[8] Umar Farooq, Tej Prasad Dhamala, Antoine Nongaillard, Yacine Ouzrout, Muhammad Abdul Qadir, ”A word sense disambiguation method for feature level sentiment analysis”, December 2015, 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA).

[9] Lesk Algorithm-Wikipedia

[10] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, “Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python”, May 2017,International Journal of Computer Applications.

[11]Support Vector Machines(SVM) — An Overview

[12]UnderstandingLogisticRegression,(https://www.geeksforgeeks.org/understandinglogistic-regression/)

[13]Shadi Diab, Al-Quds Open University, Ramallah, Palestine, ”Optimizing Stochastic Gradient Descent in Text ClassificationBased on Fine-Tuning Hyper-Parameters Approach.”, December 2018,International Journal of Computer Science and Information Security (IJCSIS).

[14]Fenna Miedema, Prof. dr. Sandjai Bhulai, “Sentiment Analysis with Long Short-Term Memory networks” , 2018, Vrije Universiteit Amsterdam.

[15]Mochamad Ibrahim, Omar Abdillah, Alfan F. Wicaksono and Mirna Adriani, “Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation”, November 2015, IEEE International Conference on data mining workshop(ICDMW).