STM JOURNALS

Journal of
Software Engineering Tools & Technology Trends

ISSN: 2394-7292
Volume 11, Issue 1, 2024
January—April
DOI (Journal): 10.37591/JoSETTT

https://journals.stmjournals.com/josettt

Review JoSETTT

Twitter Emoticon Interpretation Using Machine Learning Algorithms in Sentiment Analysis

R. Pushpa*, P. Priyadarshani, S. Santhosh Kumar

Abstract

In the current era, thousands of people share their opinions every day on the well-known microblogging platform Twitter in the form of tweets. A tweet must be brief and straightforward in order to be effective, though sentiment analysis of Twitter data will be the main emphasis of this study. Sentiment analysis study encompasses NLP and text data mining. We will conduct sentiment analysis on Twitter data using several logistic machine learning approaches. Nonetheless, our attention will be directed toward methods and varieties of sentiment analysis in which we will learn how to retrieve tweets from Twitter. In addition, we will uncover some common metrics and compare various machine learning approaches on the same dataset.

Keywords: Twitter, sentiment analysis (SA), machine learning, logistic regression, positive, negative, natural language processing, TfidfVectorizer, stemming

INTRODUCTION

Sentiment analysis is the study of people's opinions, attitudes, and behavioral reactions to particular situations or events based on the written language that is accessible. Due to the growth of subjects like machine learning and the mixing of these with the previously employed statistical techniques of natural language processing, it has emerged as one of the most active study topics. Sentiment analysis (SA) has suddenly become very popular as a result of people's increased interest in and use of social media, which includes reviews of various works of art and topics, forum discussions about hot-button issues, blogs and microblogs about opinions and information sharing, Twitter, and social networks, discussing topics that are trending locally or internationally. As more people turn to digital media for communication and self-expression, businesses and individuals are able to glean the information they want from user's thoughts and feelings.

This study discusses strategies and tactics that may be used to directly support systems for opinion-driven information searching. The Python module "scikit-learn" and a few additional libraries were utilized for this work in order to assess user sentiment and examine the current situation for any given topic or issue.

***Author for Correspondence**
R. Pushpa
E-mail: pushpaadhinathan19@gmail.com

Student, Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry, India

Received Date: February 29, 2024
Accepted Date: March 28, 2024
Published Date: April 05, 2024

Citation: R. Pushpa, P. Priyadarshani, S. Santhosh Kumar. Twitter Emoticon Interpretation Using Machine Learning Algorithms in Sentiment Analysis. Journal of Software Engineering Tools & Technology Trends. 2024; 11(1): 1–6p.

In this research, we suggest efficient methods for leveraging user tweets for sentiment analysis or opinion mining. The remainder of the document is structured as follows: The relevant works that have been completed are presented; the Section after that deals on the literature survey. The datasets, data preparation, data cleaning, and other methods covered in this study are discussed; the experiments and the outcomes are discussed; the study is concluded by what and how work might be further expanded.

RELATED WORK

Sentiment Analysis

Understanding people's feelings, perspectives, attitudes, and opinions is known as sentiment analysis. Another name for it is opinion mining. Sentiment analysis locates and examines the sentiment expressed in a text. Sentiment analysis is therefore used to identify viewpoints, convey sentiment, and categorize polarity. Opinion mining is a technique used to examine the opinions expressed by users in online reviews. One use of natural language processing (NLP) is opinion mining, which monitors consumer sentiment on a product. For instance, movie evaluations assist new viewers in determining whether or not to watch a film. However, the vast number of reviews leads to an information overload because to the lack of automated techniques for determining their sentiment polarity. The three layers of sentiment analysis are as follows: aspect, sentence, and document (Figure 1).

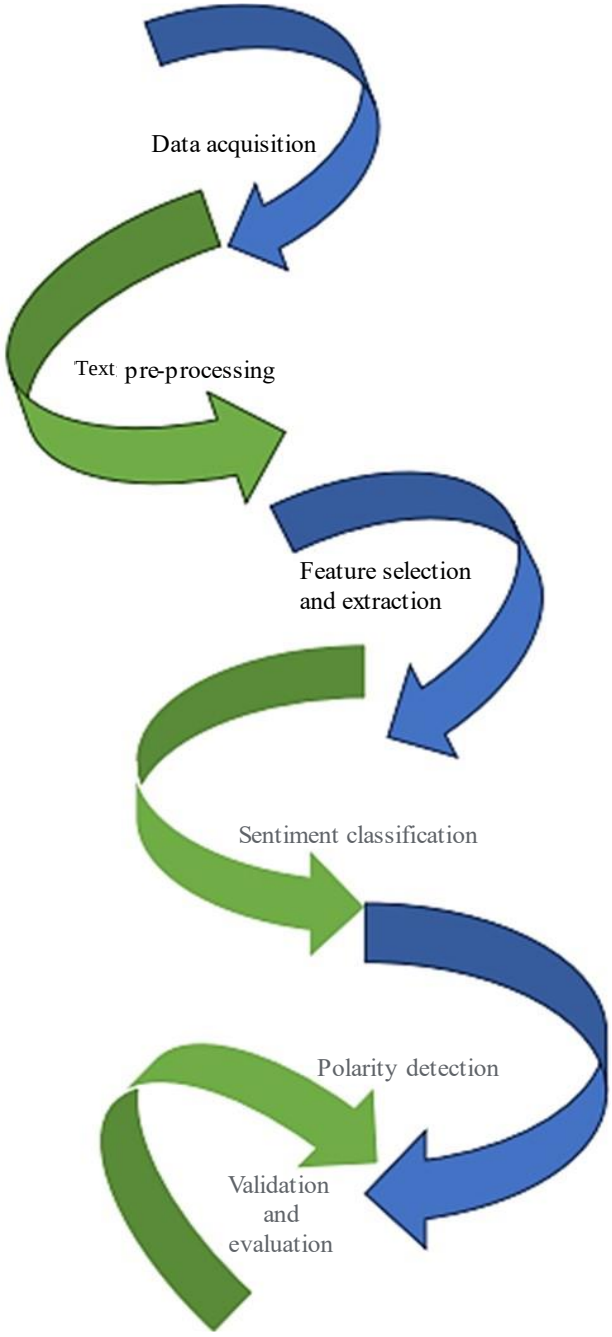

Figure 1. An approach for sentiment analysis.

Document-level: It assigns a favorable, negative, or neutral classification to the document. Sentiment categorization at the document level is the term for it.

Sentence-level: Categorizes the sentences as neutral, negative, or positive. Sentence-level sentiment categorization is the term for it.

Aspect-level: This categorizes the sentiment according to the particular features of things. We refer to it as sentiment categorization at the aspect level.

DIFFERENT APPROACHES FOR SENTIMENT ANALYSIS

There are several methods for doing sentiment analysis on linguistic data; the best method to utilize will depend on the platform you are using and the type of data you are working with. The majority of sentiment analysis research uses machine learning or lexicon-based analytic methods. Machine learning approaches employ machine learning algorithms to regulate the processing of data. They categorize linguistic data by expressing it in vector form. Conversely, the dictionary-based, or lexicon-based, technique uses a dictionary lookup database to classify the linguistic input. Using lexicon databases like WordNet, SentiWordNet, and treeks to interpret linguistic data, it computes sentiment polarity at the sentence or document level throughout this classification process.

Lexicon-based Approach

Using lexical databases such as SentiWordNet and WordNet, the lexicon-based technique makes sentiment predictions. Every word in the phrase or text is given a score, and the feature from the lexicon database that is available is used for annotation. It determines the polarity of the text by looking at a list of terms, each of which has a weight assigned to it. It then extracts information that helps to determine the text's overall attitude.

Opinionated words that are categorized as positive or negative word types, together with a definition of the term as it appears in the present context, can be found in lexicon dictionaries or databases. Sentiment polarity is ascribed to the document and a numerical score is assigned to each word. The average score is calculated by adding together all of the numerical values.

Machine Learning Approach

The sentiment analysis literature often uses a machine learning technique. With this method, the words in the phrase are seen as vectors and subjected to analysis using a variety of machine learning methods, including Maximum Entropy, SVM, and Naïve Bayes. In light of this, the data is trained, which may be used using machine learning algorithms.

LITERATURE SURVEY

A number of researchers have been working on Twitter and they occasionally publish their findings [1]. They have improved the categorization results by utilizing a variety of sentiment analysis approaches. Their work is also beneficial to this study as it addresses the sentiment analysis strategies, feature selection strategies, and other pre-processing procedures they have employed. For better clarity and comprehension of the selected issue, this study has assessed researches for both supervised and lexicon-based techniques, as well as for Twitter and non-Twitter data. The primary emphasis of this research is the supervised approach for sentiment analysis tasks.

Using a variety of characteristics, including unigrams, bigrams, pos-tagging, hash-tags, ngrams, and so on Agarwal *et al.* have employed a variety of features and feature selection techniques, such as chi-square, information gain, and semantic features and ideas [2]. Supervised machine learning is implemented after thorough data pre-processing in the technique by Khan *et al.* [3]. In order to ensure that machine learning is not restricted to any one area, they gathered labeled datasets from a variety of disciplines. They employed several training sets to teach the SVM classifier with two distinct feature

sets: 1) Information gain (IG) with feature presence, and 2) feature frequency. 3) Feature frequency; and 4) Cosine similarity with feature existence. They discovered that the existence of features is preferable to their frequency.

Various methods and classifiers, including maximum entropy (MaxEnt), naive bayes (NB), lexicon-based approach, support vector machines (SVM), and others, have been employed sometimes. The results have been assessed using a variety of criteria, including f-measure, accuracy, precision, and recall [4]. A mixed language NB classifier on unigrams achieved 71.5% accuracy [5]. The NB classifier's usage of semantic characteristics raises the f1-measure against unigram by 6.47% and the pos + unigram by 4.78%. Our study primarily focuses on merging machine learning classifiers and demonstrates that the combined approach outperforms independent classifiers in terms of outcomes. With a limited dataset and few features, this research also provides comparison findings to the feature hashing + lexicon- based features employed by Da Silva *et al*. [6].

In their discussion on the importance of thorough preprocessing in raising the evaluation measure, Jianqiang and Xiaolin provided six distinct preprocessing techniques [7]. In light of this, our method filters the tweets using a strong preprocessing. A technique for determining the opinions on every facet of a product was put out by Khan and Byeong [8], and further research into this area might be beneficial [9, 10].

PROPOSED SYSTEM

Our goal, is to do sentiment analysis on Twitter data. Our plan is to construct a classifier that is composed of various machine learning classifiers, since achieving sentiment analysis for data from Twitter is our main objective, as shown in Figure 2.

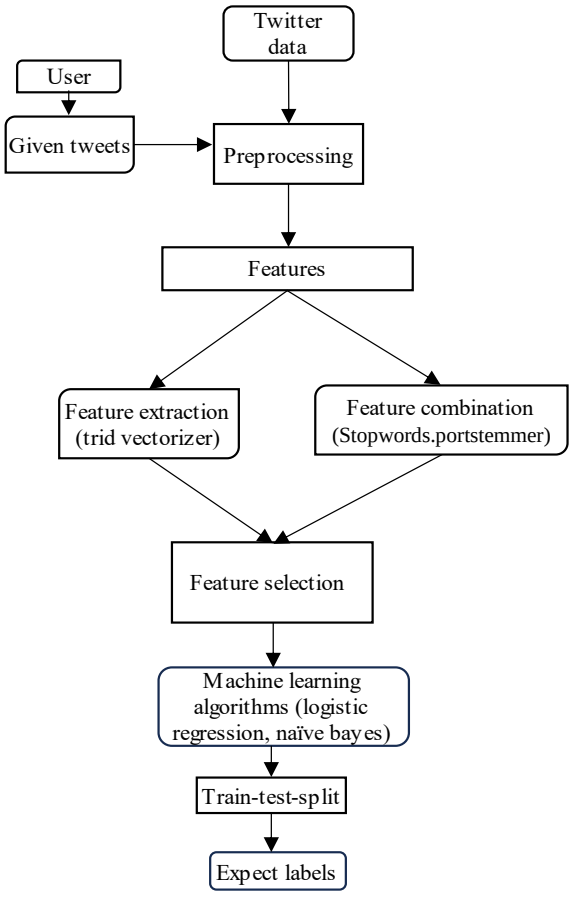

Figure 2. Block diagram of twitter sentiment analysis.

Step 1: To begin, we will download tweets from the Kaggle website and import them into Python.

Step 2: After that, we preprocess these tweets to make them suitable for feature extraction and mining.

Step 3: Following pre-processing, we feed the data into our trained classifier, which categorizes the data into positive or negative classes according to the outcomes of the training process.

Algoritms

Logistic Regression

This probabilistic classifier is used for classification but not for regression. This classifier is used to predict the probability of the variable. Also called a linear regression model. Its functions are somehow more complex. The predicted value of a variable will be converted into binary numbers like 0 and 1.

Performance Metrics of Sentiment Classification

Typically, predetermined metrics like accuracy, precision, and recall are used to assess sentiment categorization performance.

Accuracy

In order to find which model gives better result, then it is necessary to find the accuracy (Eq. (1)). Accuracy for any model can be given as:

$$Accuracy = \frac{tp+tn}{tp+fn+tn+fp} \quad (1)$$

Where,

Tp= true positive, case was positive and it predicted positive;
Tn= true negative, case was negative and it predicted negative;
Fn= false negative, case was positive and it predicted negative; and
Fp= false positive, case was negative and it predicted positive.

RESULT AND DISCUSSION

For twitter sentiment analysis to be implemented successfully and assessing the effectiveness of the sentiment analysis method, the most dependable statistic is accuracy, which is determined by the classifier as shown in Figure 3.

```
[ ]  # accuracy score on the training data
     X_train_prediction = model. predict(X_train)
     training_data_accuracy = accuracy_score(Y_train, X_train_prediction)

[ ]  print('Accuracy score on the training data :', training_data_accuracy)

     Accuracy score on the training data : 0.81018984375

[ ]  # accuracy score on the test data
     X_test_prediction = model. predict(X_test)
     test_data_accuracy = accuracy_score(Y_test, X_test_prediction)

[ ]  print('Accuracy score on the test data :', test_data_accuracy)

     Accuracy score on the test data : 0.7780375

Model Accuracy on Training Data = 81%

Model Accuracy on Test Data = 77.8%
```

Figure 3. Result analysis.

CONCLUSION

This research concludes that machine learning techniques, particularly logistic regression, may more accurately predict attitudes. These have been used to the study of natural language processing, particularly in the area of sentiment analysis and the identification of subjective data such as opinion and emotion in written texts. Still, each of them can forecast the sentiment to varying degrees of accuracy. In this study, finding public opinion and doing opinion mining are our main objectives. Generally, through tweets, people express their ideas, feelings, etc. However, we may not always be able to accurately gauge their thoughts and emotions. Therefore, by using sentiment analysis on such tweets, we are able to determine the number of people that support and oppose the mission. Furthermore, our methodology has a restricted number of features for learning the classifiers; in our future research, we will use better feature selection methods such as Information Gain and Chi-Square, to find best the solutions.

Acknowledgment

We owe a debt of gratitude to our dear advisor, Mr. K. Muthukumaran, whose invaluable counsel, recommendations, and immense assistance made it possible for us to finish our project. We have found a lot of inspiration in him.

REFERENCES

1. Pak Alexander, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the International Conference on Language Resources and Evaluation, LREc. 2010; 10: 1320–1326.
2. Agarwal A, Xie B, Vovsha I, Rambow O. Sentiment analysis of twitter data. Proceedings of the workshop on languages in social media. Association for Computational Linguistics. LSM'11: Proceedings of the Workshop on Languages in Social Media. 2011 Jun; 30–38.
3. Khan Farhan Hassan, Usman Qamar, Saba Bashir. A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. Knowl Inf Syst. 2017; 51(3): 851–872.
4. Narr Sascha, Michael Hulfenhaus. Language-independent twitter sentiment analysis. Knowledge Discovery and Machine Learning (KDML), LWA. 2012.
5. Saif Hassan, Yulan He. Semantic sentiment analysis of twitter. International Semantic Web Conference. Berlin Heidelberg: Springer; 2019.
6. Da Silva, Nadia FF, Eduardo R. Tweet sentiment analysis with classifier ensembles. Decis Support Syst. 2020; 66: 170–179.
7. Jianqiang Z, Xiaolin G. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. IEEE Access. 2021; (5): 2870–2879.
8. Khan Jawad, Byeong Soo Jeong. Summarizing customer review based on product feature and opinion. Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. IEEE. 2016; 158–165.
9. Kavya Suppala, Narasinga Rao. Sentiment Analysis Using Naïve Bayes Classifier. Int J Innov Technol Explor Eng (IJITEE). 2019 Jun; 8(8): 264–269. ISSN: 2278–3075.
10. Durgesh Patel, Sakshi Saxena, Toran Verma. Sentiment Analysis using Maximum Entropy Algorithm in Big Data. Int J Innov Res Sci Eng Technol. 2016; 5(5): 8355–8361. (http://www.ijirset.com/upload/2016/may/246_49_Sentim ent.pdf)

