International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

A MACHINE LEARNING-SENTIMENT ANALYSIS ON MONKEYPOX OUTBREAK; AN EXTENSIVE DATASET TO SHOW THE POLARITY OF PUBLIC OPINION FROM TWITTER TWEETS

Dr. H Joseph Williams, Professor, Department Of ECE, SICET, Hyderabad
M.Srimukhi, Asma, N.Kavya, P.Sri Chandana
Department Of ECE, SICET, Hyderabad

Abstract:
On Twitter, one of the most popular social media sites, there was an increase in tweets containing positive, negative and neutral messages about the virus in a short time. Because of the diversity of tweets, researchers continued to conduct public opinion polls and measure the extent of public opinion regarding the virus. Then, the method we proposed in this study examined the smallpox virus by following Twitter users who shared their thoughts on the social media site. The proposed method performs sentiment classification using various criteria and classifiers to evaluate the sentiment of collected tweets. Early detection of viral behavior in tweets could lead to better understanding and control of the virus. Tweets are divided into three sentiment categories: positive, negative, and neutral. Our proposed method is based on three different machine learning algorithms, including Nave Bayes classifier, random forest classifier, and support vector machine classifier. Since these three classifications have different advantages, tweets from Twitter can be classified during the application process. Results are plotted using the matplotlib package provided with Python.

Keywords - tweets, Monkey flower, emotion, classifier, learning model.

I. Introduction

Critical analysis is a type of text analysis that uses techniques to identify the emotions of a text, whether good, bad, or neutral. Using data for emotional intelligence can help businesses better understand what people are generally talking about. Twitter has nearly 30 million active users who send an average of 500 million tweets every day, making it one of the most important social media platforms for news, information and interaction with international brands and prominent images. So it's no surprise that businesses see this Weibo platform as an important tool for marketing and customer support. Twitter allows businesses to reach a wide audience and connect directly with customers. Twitter analytics allows businesses to understand their target audience, track what is being said about their business and competitors, and identify emerging issues. When analyzing Twitter data, various metrics such as number of comments or number of retweets seem insufficient to get the full picture. It is really important to understand the importance of this information. Do they or do they not talk about specific products or topics? Assessing emotions leads to correct decisions. It ensures a good understanding of the subject or the whole being discussed.NeedofsentimentalAnalysis

1) Economic Evolution: It is not important compared to all information, there is only money in the economy. Ho

wever, the analysis helps eliminate important features of businessspecific profiles. Sentiment analysis will provide businesses in the industry with a great opportunity to improve their brand and audience. This will benefit any businesstoconsumer business, whether it is restaurants, entertainment, hospitality, mobile consumer, retail or travel. Page layout.

2) Research needs: Evaluation, analysis, theory, classification, etc. Research needs on the topics are other important factors leading to the expansion of SA. Additionally, his research topics include text mining, machine learning, natural language processing, artificial intelligence and polling, audio, content analysis, etc. It will be based on computer science disciplines, including

2. literaturesurvey

G. P. Zhang's theory (2000) - Classification according to GP is one of the most active tasks in the research and use of neural networks. Zhang's (2000) theory. Rich and everexpanding knowledge. This section provides an overview of the most important developments in the field of neural network classification. In particular, the relationship between neural and traditional classifiers, the balance between learning and generalization, the selection of different features, and the effect of error rates are examined. The purpose of this review is to provide an overview of the literature in this field and to encourage readers interested in the research area. Machine learning, dictionary-based, and hybrid methods are three categories of classification theory. When evaluating volume generated, more complex metrics such as number of followers/friends, number of likes/shares/retweets per post and number of shares, reaction rate, and other combination metrics are often included. However, measuring user sentiment is not an easy task. After receiving the necessary information, the system performs a logical analysis. He said there are two main methods of sentiment analysis: machine learning and dictionary methods.

D.Can, S.Narayana (2012) - It was the researchers who wanted a real opinion: Analysis of public reaction time. They collect responses from the microblogging site Twitter. Twitter is one of the social media platforms where users can express their thoughts, feelings and opinions on all kinds of topics. Twitter messages from American voters constitute a large amount of data used to gauge public opinion of each candidate and predict who will win. It was thought that the attacks people shared on Twitter would affect the entire election process. They also investigated the effects of emotional surveillance on these public events. They also showed how fast this overtheair emotional analysis is compared to content analysis, which can take days or weeks to complete. The system for them has proven to analyze all Twitter data for opinion polls, candidates, support and more and generate profits. They also investigated the effects of emotional surveillance on these public events. They also show how quick the evaluation of this hypothesis is compared to traditional content, which can take days or weeks to complete. Six of them found a technology that can analyze opinions about elections, candidates, support and more from all Twitter data and return good results.

O. Almatrafi, S. Pala, B. Chavan et al. (2014) - The authors emphasized the locationbased approach. They say sentiment analysis involves extracting sentiments from text on specific sites using natural language processing (NLP) and machine learning. They explored many applications of locatiobased sentiment analysis using data that facilitates data collection from multiple sources. The article easily accesses the site tweet feature on Twitter, allowin g information (tweets) to be collected from specific sites to identify patterns and trends. As part of their research, they are examining the 2014 Indian general elections. They extracted 600,000 tweets from both political parties over seven days. They use machine learning algorithms, such as the Naive Bayes algorithm, to create a classifier that can classify tweets as positive or negative. They used a Python module to tap into the perceptions and behavior of users across different spheres of both political parties and plotted their findings on a map of India.

Dr. Ratnadeep R. Deshmukh (2014).different words. The purpose of the feature selection method is to select important content from the text for sentiment analysis. Machine learning, dictionarybased, and hybrid methods are three categories of classification theory. The system developed by the authors for the general analysis of Twitter data faces new problems due to the heterogeneity of data distribution due to the abundance of daily posts using different languages. The purpose of the feature selection method is to select important content from the text for sentiment analysis. Machine learning, dictionarybased, and hybrid methods are three categories of classification theory. When measuring volume generated, more complex metrics such as number of followers/friends, number of likes/shares/retweets of a post, and engagement, price response, and other composite metrics are often included. On the other hand, analyzing user opinions requires effort. It is the process of identifying and classifying the ideas or thoughts expressed in a particular passage. This system requires a key and access to the Twitter API.

B. Sun, V. Ng, et al. (2016) - Panda Library contains data structures and tools for working with data systems used in many different fields, including statistics, finance, social, and more. This library provides an integrated, easy touse framework for working with data. Analyze and manage large data sets. It is intended to be the future standard for computing in Python. It works well as an addition to existing Python work, as well as using and extending existing data management functions in other programming languages such as R. Explain the layout and features of pandas. The library called Natural Language Toolkit (NLTK) contins many software modules, many data structures, tutorials, problems, some statistics, machine learning classes including speech, etc. is available. The main purpose of .NLTK is word processing or analyzing information in human language. The corpus is provided by NLTK and is used to train the classifier.

III. Methodology

A. Overview

Our experiment is illustrated in Figure 1.0, starting with data collection, interpretation, and planning. We pre-remove retweets, symbols, hashtags, user tags, stop words, numbers, and equal words and replace emojis with text.

B. Data collection

Create a data set to be completed and the data provided by this set should be accessed for comments on social media platforms. It is recommended to use the Python library Tweepy, which has an API to extract the via's relationship to the server. Remove retweets when tweeting.

C. dataPre-processing

Preprocessing is a part of Natural Language Processing (NLP) that helps convert raw text into an understandable format. In this research, we use many tools and methods for prioritization. Since our files are in multiple languages, we need to make sure that all files are in the same language. To do this, we leveraged Python's built-in cleantext package to translate all non-English tweets into English. Then come all the retweets and tags.

Hashtags, stop words, tokenization, stop words and duplicate words are removed. We will cover each of the following tasks:

1) Removing user tags: When tweets are shared, duplicates are created and this can greatly affect educational standards, are reported and corrected, so retweets should be removed. "RT" means retweet and "@Someone" means user tag. They are also excluded.

2) Emoji and text replacement: Emojis are small digital images and symbols that people use to convey their thoughts and feelings. We convert these images into comparison text to improve our training model. All text has been converted to lowercase to avoid double-checking.

3) Hashtags, numbers and symbols removed: Hashtags are words used to find and store similar information on social media. The pound sign (#), often seen before periods, is a powerful tool in relationships. However, it was removed from the database as it was not required for education standards. Regular expressions are used to remove numbers, repeated expressions, and punctuation marks (RegEx). This speeds up the learning process while also reducing memory usage.

4) Elimination of stop words: Stop words are words that contain little important information for the sentence, such as

' toâ, –
me, "my", "we". To avoid noise in the dataset, we remove them using the Python package library stop message.

5) Tokenization: Tokenization is done by dividing the text into smaller parts of a word using natural language tools. Tokenization is required so that the video in the review view can be easily removed.

D. Data Labels

At TextBlob we use polarity scores to determine labels. There are three labels: good, bad and neutral. TextBlob is another dictionarybased sentiment analysis (rulebased sentiment analysis) we used in our research. We create a Python loop that iterates over all lines in the file and retrieves the polarity and content using the textblob() function. The polarity score is a floating number between 0 and 1, and its state is still between 0 and 1. In this study, we are interested in the transcribed polarity score as shown in the equation below.

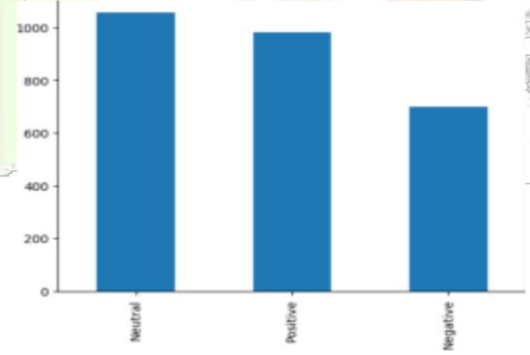

Fig 2.0 : graphical representation of positive, negative and neutral tweets

E.VectorizationWord embedding, also known as word vectorization, is a language processing method that converts a word or phrase in a sentence into number vectors that can be used to analyze the semantics, analogues, and approximations of a word. It is easier to train and extract features in machine learning when using the term embedding. TF-IDF is adopted to perform hypothesis testing using vectorization.

1) TFIDF: Time frequency (TF) and inverse document frequency (IDF). While IDF focuses on the frequency count of words in a word, TF focuses on all words in a document. The TF formula is found in Equation 1.
TF (t, d) = of term (t) in file Frequency (d) (1)
All terms in file (d)

$$\text{TF (t, d)} = \frac{\text{Frequency of term (t) in the document (d)}}{\text{Total word in the document (d)}} \quad (1)$$

Purpose The purpose of IDF is to determine how much text there will be for each word in the document. We need IDF because it maximizes the value of infrequently occurring expressions while reducing the weight of frequently occurring expressions. Equation 2 can be used to calculate the IDF.

$$IDF\ (t) = \log2\ \left(\frac{\text{Total Documents (N)}}{\text{1+Total Documents with term (df(t))}} \right)\ (2)$$

IDF (t) = log2 (Total Documents (N)) (2)
1+Term Total Documents (df(t))
TF - IDF expression in Equation 3 are Results formula 1 and union of 2
TF ☈ IDF = tf .idf (t, d, N) = tf (t, f).idf (t, N) (3)

TF − IDF = tf .idf (t, d,N) = tf (t, f).idf (t,N)

F. Learning Models -Various types of machine learning

are used in this project to design, build and evaluate many models. The remaining 80% of the dataset is used for training, while only 20% is used for validation. Here, the accuracy, precision, recall, and F1 score of each model are used to evaluate its performance. Each learning algorithm uses sklearn's preset hyperparameter values. The actual algorithm is discussed in more detail below.

1) Random Forest: We use the random forest model. It is an algorithm in ensemble machine learning classification. This algorithm produces a large number of decision trees, which means most of the trees will choose one class. This scheme collects the results of the prediction tree and randomizes them. Node size, number of trees, and number of samples are the three minimum hyperparameters that must be met for the random forest to work properly. Use the bagging technique (also known as bootstrap aggregation), which uses training data to create a unique set of training data. Results are determined by the preferred price.

2) Support Vector Machine: classify data into a set of objects using a powerful SVM model that provides the boundaries of the cluster. Hyperplanes are an important feature of SVM for determining boundaries (separations) between classes. There are three hyperplanes. Positive hyperplane, negative hyperplane and ideal hyperplane are three types of hyperplane.

$\vec{w}.\vec{x} + b = 1$ for Positive hyperplane (1)

$\vec{w}.\vec{x} + b = -1$ for Negative hyperplane (2)

$\vec{w}.\vec{x} + b = 0$ for optimal hyperplane (3)

Equation 1-3 represents this number of hyperplanes:
w-.x-+ b = 1 (for positive hyperplane (1))
w-.x-+ b =-
1 (for negative hyperplane (2)). is x. Ensure that the model has its best hyperplane, where the large edges should be maximized. For nonlinear problems, the method is sufficient to solve them using longer dimension appropriate kernels, allowing to separate them. SVMs include Polynomials, Gaussians and kernels. Such as Gaussian Radial Basis Function (RBF).

1) Naive Bayes: Another method used for classification is Naive Bayes. It is a probabilistic classifier that uses relevant events to determine the relevant class of its input. The probability calculation for each category is defined in Equation 4.

Conditional probability* Prior probability (4)
Proof
Mathematically,

$$\frac{\text{Conditional probability * prior probability}}{\text{Evidence}} \qquad (4)$$

$$P\left({}^{y}/_{X}\right) = \frac{P\left({}^{X}/_{y}\right) P(y)}{P(X)}$$

Among these:

P(y): previous outcome

P (X/ y): probability
P (y/X):

Probability after br>P (X): Marginal probability (evidence)

There are many variations of Naive Bayes; However, in this research we use Multinomial Naive Bayes, a model that focuses on classifying problems in data processing.

IV. Results and Discussion

This section presents the experimental procedure used to evaluate the effectiveness of the proposed model. In our research, we developed, refined and analyzed 3 models based on registration, vectorization and normalization methods. All models are trained and evaluated using three machine learning algorithms: random forest, support vector machine (SVM), and naive Bayes classifier.

TABLE I

Comparisons of Model's Performance

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.8946	0.8851	0.8958	0.8694
Random Forest	0.8778	0.8806	0.8648	0.87752
Naïve Bayes	0.7421	0.7928	0.7632	0.7594

Table 1 shows the results of the model created using textblob and the TFIDF vectorizer for registration. Based on our tests of these three algorithms. Support vector machine has the best performance with an accuracy of 0.8946.

5. Conclusion

Analysis and inference are hot topics in machine learning. The purpose of theory analysis is to analyze texts according to the theory they contain. Sentiment analysis is a new research field that has recently received great attention in the field of computational language and text mining. In this project, we discuss a simple technique to segment tweets into good and bad groups using machine learning and Python. We can further improve our distribution by continuing to extract more features from tweets. The Twitter API is very useful for processing data requests and can provide valuable insight into what the public is thinking.

VI. Future scope

More word embedding methods and approaches (eg: doc2Vec) and text tagging (eg: Azure Machine Learning) will be added in future work to tune the performance of the model. We also hope to use deep learning and Transformer algorithms to improve thinking and predictive thinking.

References

1) G.P Zhang (2000). “Identifying Consensus,” Proceedings of the 20th International Conference on Computational Linguistics, page 1367 Association for Computational Linguistics, Stroudsburg, PA, USA.

2) H. Wang, D. Can, F. Bar, S. Narayana et al (2012). “Twitter Sentiment Analysis and Opinion Mining,” A Report on Sentiment Analysis and Opinion Mining from Social Media Data. United Kingdom: Knowledge Media Institute, 2011

3). deep. Almatrafi, S. Parack, B. Chavan, et al. (2014). - Twitter as an institution for sentiment analysis and sentiment mining. –,

7. Proceedings of the International Conference on Language Resources and Assessment. European Association for Linguistic Resources, Valletta, Malta

4) Manisha Mishra and Monika Srivastav (2014). - Proceedings of the 42nd Annual Meeting of the Artificial Intelligence - Communication Association, ACL - 04.

5) Dr. Ratnadeep R Deshmukh (2014). — Spam Theory and Analysis — Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM — ACM, New York, NY, USA, 219–230.

6) Sumneet Kaur, Aman Puri, Yashi Jain (2019). - Analysis of sentence-level sentiment polarity classification using discourse analysis - , Proceedings of the 21st International World Wide Web Conference, WWW - 12. ACM, New York, NY, USA, 191–200.

7) P. Pang, L. Lee (2000). - Information-theoretic approach to emotional polarity classification - . Good Web, Proceedings of the February WICOW/AIRWeb Collaborative Workshop on Good Web - 12th ACM, New York, NY, USA, 35–40.

8) Wilson T, Wiebe J, Hoffmann P (2018). - Identify topic polarities in sentence-level thinking -.
Proceedings of the Human Language Technology and Natural Language Processing Standards Conference. Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA. 347-354: I.

9) Staphord bengesi, Timothy Oladunni, Ruth Olusegun and Halima Audu (2023), — Machine Learning — Sentiment Analysis of the Monkey Pox Epidemic: A Report on the Polarity of Public Opinion from Twitter

Tweetsâ , https://ieeexplore.ieee.org/stamp/stamp.jsp?arnu mbe r=10036414, Volume 11.
10) Babacar Gaye, Azuli Wulamu (2019), "Sentiment Analysis of Online Reviews Using Machine Learning Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering ICPT, 34(33), 46-55.

11) C. Sitaula, A. Basnet, A. Mainali and T.B. Shahi, “Deep learning-based sentiment analysis for COVID-19 related tweets in Nepal,” Comput. Intel. Neuroscience, vol. 2021, p. 1-11 November 2021.