

Predicting Polarity Using Sentimental Analysis

V. Mani^{1*}, Deepika S.², Harishkumar R. S.², Ranisnekha M.² and Mukesh P. S.²

¹Assistant Professor, Department of CSE, M. Kumarasamy College of Engineering,
 Karur, Tamil Nadu, India. Email: maniv.cse@mkce.ac.in

²UG Student, Department of CSE, M. Kumarasamy College of Engineering, Karur,
 Tamil Nadu, India.

*Corresponding Author

Abstract: Sentiment analysis is a study of people's sentiments, opinions, attitudes, emotions in a written language or text. Rapid Growth in the field of sentiment analysis and explore to find the sentiment on various social media platforms using techniques of machine learning with analyzing sentiment, and also helps in the analysis of polarity or subjectivity. The most widely used social media site is Twitter, where people share their thoughts in the form of tweets and hence it becomes the major data sources of sentimental analysis. Recently the more used social media platform such as Twitter. Their people express their thoughts and opinion as a tweet representation. It is the major data resources of analyzing the sentiments. Sentiments are classified to different group like positive, negative or neutral. Such analysis process helps to differentiate the sentiments also classifying them into different groups comes under prediction of sentiment. Very first we pre-process the dataset, feature extraction in which meaningful insights are extracted from the dataset, then extracted features are applied for classification model using machine learning Random Forest algorithm, Regression algorithm, etc. But with the advancement of the python language and to reduce the code complexity we have analyzed the polarity using the python packages, API and Algorithm which are available. This model proved to be highly effective and accurate on the analysis of feelings. At last the trained classification model are tested in order to check the performance it is measured by accuracy.

Keywords: Classification, Random forest, Regression, Sentimental analysis.

I. INTRODUCTION

Basically, Sentiment analysis refers to the analysis of an idea or feeling or opinion using data such as text or images. Mostly it helps in the decision-making process of companies. For example, if public perception about a product is not that good, a company may try to change the product or stop production to avoid losses and raise the business value in future using this prediction.

There are various resources for example, reviews, public interviews, opinion polls, surveys and so on. As more and more people join social media sites, like Facebook and Twitter can be parsed for public sentiment.

Text understanding is an important concern. Some machine learning techniques are used, including various supervised and unsupervised methods. In natural language processing, algorithms such as the Random Forest algorithm and logistic regression are used to predict the polarity in a text. The Twitter analysis depends may vary on different levels [1].

Analyzing thoughts sharing, conversation, opinion and form of tweets is fundamental in evaluating the business strategy, political analysis, and public actions.

Data collected from the resources like reviews, blogs, tweets are processed and classified into various groups called text classification.

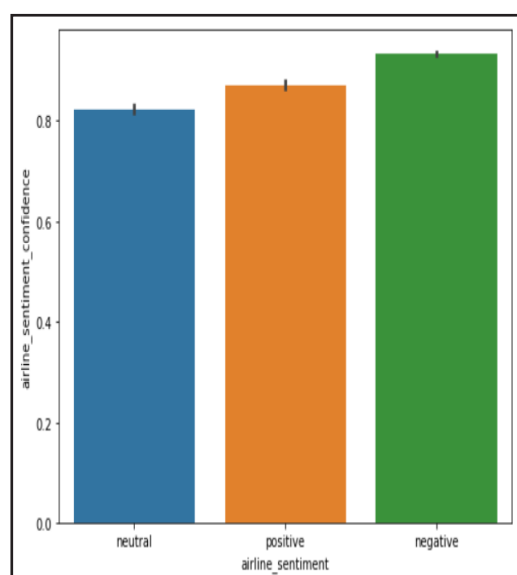


Fig. 1: Classified Groups

Taking tweets from six US airlines, the task is to determine whether there is a positive, negative or neutral people's

perception about the airline. This uses a common supervised learning task where, given a text string, you must classify the text string into predefined types.

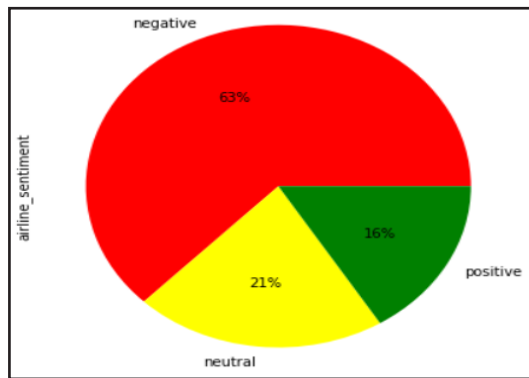


Fig. 2: Percentage of Groups

Initially, importing the essential libraries and the relevant dataset. We will perform exploratory data analysis and text preprocessing to convert textual data to numeric data. Then we will train and test models using machine learning algorithms. Fig. 1 and 2 shows the classification and percentage of groups.

II. SHORT SURVEY

Sentiment analysis through the vocabulary of a person, language or the knowledge they acquire by the similarity between the word embedding and sentiment lexicon [1]. To find the user reviews with the sentiments and to promote them to the emotion findings which they can use the lingual feeding to the support of the reviews acquired. This method uses the corpus-based methods and knowledge-based methods. The sentiment lexicons and the word embedding. They split the datasets and analyzed by using the two different methods one is the word or the document embedding and the other is the similarity between review and emotions.

By classifying the emotions or the sentiments in the text from the user [2]. Count the number of negative and positive words in the text. When the number of a positive word is greater than the number of negative word, it returns a positive sentiment and vice versa. Otherwise, it will return a neutral sentiment. Firstly the text preprocessing is done and after that the text transformations takes place then the feature selection is done to analyze the best results for the views where their emotions or their opinion is detected.

The sentiments and emotions for the recognition for the healthcare [3]. Analysis for the healthcare which is used to detect the sentiments that include intensity analysis, emotional analysis, polarity analysis, subjective analysis. To detect the patients' mind and their sentiments and emotions which is used to save the time of the doctors and also they are used to make the medications, treatments. This is used to give treatment by analyzing the records of the patients in a short span of time. This detection is used to switch over the analysis of the user

records and this is helpful for the treatment as soon as possible. This can be the one to give fast medications so as to enable to provide them with the usage of this analysis machines which is able to save them as the source to detect the user problems easily and quickly.

Social media blogs, discussion forums and the data by which the comments and reviews from the net sources are done [4]. The unstructured data also had a great effort in the working of the client reviews. From this, the author has found the analysis of sentiments through the unstructured data also they gain emotion of the reviewers. we have a tendency to various approaches which enables sentiment analysis on Twitter website. Performing on linguistic communication process (NLP) and connected fields [7-8].

Stemming and Lemmatization are two essential processes of preprocessing module during feature extraction but also there is a need for additional dictionary support for indexing and searching, which enhances accuracy of feature extraction [5]. We will divide the dataset into training and testing. The training set is nearly 25,000 reviews which induce word vectors in the model. We will evaluate the performance of classifier after cross-validating parameters on the training set, then use linear SVM. Bag-of-Words is a technique for natural language processing where features are extracted from a sentence, document, website, etc. Perform classification to evaluate the test data [6].

III. IMPLEMENTATION

Initially, importing the essential libraries and the relevant dataset. We will perform exploratory data analysis and text preprocessing to convert textual data to numeric data. Then we will train and test models using machine learning algorithms.

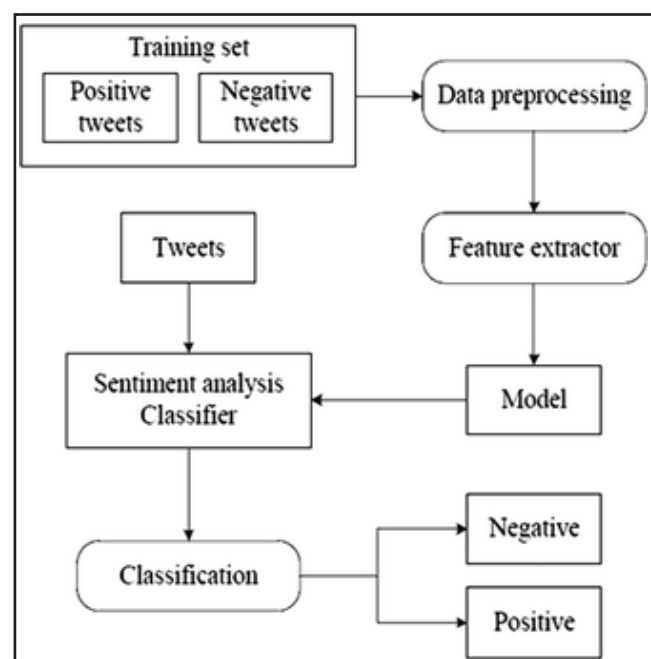


Fig. 3: Architecture

A. Installing and Importing Required Libraries

The first step is always to import the required libraries like pandas, numpy, nltk etc. which are needed to create sentimental analysis model. Steps used in this implementation is shown Fig. 3.

B. Load the Dataset

Load the dataset to train your model. Developing a sentiment analysis model to predict public perceptions.

Theread_csvused to loading dataset. Either can go with download and save the dataset in your computer for further use or directly link to any URL.

C. Exploratory Data Analysis

Before creating the model, perform analysis on data to get some insights. Let's examine dataset a little bit to see if there are any trends. But before that, let's change the default plot size to get a better view of the layers [9].

Here you can visualize your dataset as maps, visualizing the polarity using bar plot and pie chart, and run the scripts that Python Seaborn library used to draw the Countplot method. You can see

- The percentage of public tweets for each airline.
- The distribution of sentiments of all the tweets.
- The distribution of sentiment for each individual airline.

D. Data Preprocessing

The tweets contain many slang words and punctuation marks. We need to clean up our tweets before applying the machine learning model. Before cleaning up the tweets, however, the dataset is split into a feature. Our feature set contains only tweets. The corresponding label is the impression of the twitter dataset. Using the Pandas DataFrame's iloc method [10].

Once the data is divided into training and testing, you can preprocess the data to clean it up, as our dataset contains many irrelevant or noisy data to eliminate such kind of data if present. To do so, you will use regular expressions for easy removal.

E. Representing Text in Numeric Form

Machine and deep learning deals with numeric data using many approaches to convert to number format. Input contains many text, those text have to be converted to number before model building. To do so, there are three main approaches, namely bag of Words, TF-IDF, and Word2Vec.

F. Split the Data Set

The dataset will be divided into two main sets 80% and 20%. We will use the

- 80% dataset for training the model.
- 20% dataset for testing the trained model.

G. Training and Evaluating the Model

Used the Random Forest algorithm, logistic regression to its ability to act on non-normalized data.

H. Making Predictions from the Model

Once the model has been trained we need to call the predict method in the class. Classification metrics namely confusion matrix, accuracy, F1 measure used to evaluate the classification model by machine learning.

IV. RESULT

A. Random Forest

```
[[1719 108 43]
 [325 243 46]
 [134 62 248]]
```

	precision	recall	f1-score	support
negative	0.79	0.92	0.85	1870
neutral	0.59	0.40	0.47	614
positive	0.74	0.56	0.64	444
micro avg			0.75	2928
macro avg	0.70	0.62	0.65	2928
weighted avg	0.74	0.75	0.74	2928

Accuracy 0.75

B. Logistic Regression

```
[[2635 124 29]
 [462 399 58]
 [212 83 390]]
```

	precision	recall	f1-score	support
negative	0.80	0.95	0.86	2788
neutral	0.66	0.43	0.52	910
positive	0.82	0.57	0.67	685
micro avg			0.78	4392
macro avg	0.76	0.65	0.69	4392
weighted avg	0.77	0.78	0.76	4392

Accuracy 0.78

V. CONCLUSION

In Sentiment analysis, NLP tasks are the most commonly used, which helps to determine the public perception of a particular topic. We learned how Python libraries contribute to the analysis of emotions and sentiments on social media platforms. We carried out an analysis of public tweets about six US airlines, using Random Forest for approximately 75% accuracy and 78% excellent results using logistic regression.

REFERENCES

- [1] O. Araque, G. Zhu, and C. A. Iglesias, "A semantic similarity-based perspective of affect lexicons for sentiment analysis," *Knowledge-Based Systems*, vol. 165, pp. 346-359, 2019.
- [2] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, IEEE, 2013.
- [3] A. Vij, and J. Pruthi, "An automated Psychometric Analyzer based on sentiment analysis and emotion recognition for healthcare," *Procedia Computer Science*, vol. 132, pp. 1184-1191, 2018.
- [4] B. Xiang, and L. Zhou, "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 434-439, 2014.
- [5] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. Ch. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214-224, 2017.
- [6] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment - Topic detection from text," *Computing Science*, 2014.
- [7] Yao metal et al., "Exploring the sentiment strength of user reviews," *International Conference on Web-Age Information Management*, pp. 471-482, 2016.
- [8] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle," *Proceedings of the ACL 2012 System Demonstrations*, pp. 115-120, July 2012.
- [9] Pankaj, P. Pandey, Muskan, and N. Soni, "Sentiment analysis on customer feedback data: Amazon product reviews," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019.
- [10] H. Aoyama, W. Souma, and I. Vodenska, "Enhanced news sentiment analysis using deep learning methods," *Journal of Computational Social Science*, vol. 2, pp. 33-46, 2019.