Sentiment Analysis of Twitter using Machine Learning

Hruthik Mekala
*[1]Computer Science and Engineering, VIT-AP University, Amaravathi, Andhra Pradesh, India.
hruthikmekala2003@gmail.com

ABSTRACT

Sentimental analysis of Twitter data is an area that has experiencing significant growth in recent times after the twitter changes into X. The ability to identify opinions from tweets using machine learning techniques has helped the researchers to simplify the process. This paper represents the use of machine learning algorithms and Scikit -learn in sentimental analysis of Twitter data. We perform analysis of Twitter datasets which are publicly available by Natural language Tool Kit. We train and test the data using machine learning classifiers such as Logistic Regression. After analysis of the datasets by using the features and machine learning algorithm we can get accuracy maximum of 77.8 %.

Keywords: Sentimental Analysis, Machine Learning, NLTK, Scikit -learn, Twitter data

I. INTRODUCTION

As we look into today the world is moving with new innovations in every field, people also habituated to the technology often. The Social media become one of enterntainment purpose. The Social media sites like Facebook, Instagram, Twitter, WhatsApp etc.., are turned to everyone's daily routine. Not only Social Media sites there are few online shopping sites like Amazon, Flipkart, Myntra, etc., which are mostly used now. The companies need the feedback of the customers to improve their business strategies. They need to know about the satisfactory level of the product. The customers send emails or comments, make reviews on these. It become difficult for the companies to gather all these and analyse separately. It takes large amount of time and work for the company.

Sentimental analysis is also known as opinion mining in which the data is a mixture of emotions. To identify these emotions from the customers feedback is the main thing for the companies to improve their strategies in the market. There are various kinds of emotions are filled in them. Let us consider the Twitter platform, in which a large number of users tweets daily on the site on different things. As we know there are positive and negative responses, we need to consider the both and sometimes it may be neutral. To filter out the responses and to take measures on them we use technologies like Machine Learning which helps us to analyse the data and make it easier to identify the opinions.

By using the NLTK we can train and test the machine learning classifiers which are used for the features we have provided. We can classify whether it's a positive tweet or negative tweet.

Knowledge-based techniques classify the text-based which is unambiguous like happy, sad, afraid, and bored. Some context bases are not likely to not involving any words influenced by related to assign available to use the words likely to understand to explicit emotions.

Logistic regression is a machine learning-based classification approach that uses training data to learn how to classify the sentiment of a text. In sentiment analysis, the two categories are typically "positive" or "negative", and the neutral class is ignored because logistic regression works with binary classification logic.

We use python for the coding of these algorithms in which we can import various libraries.

Python is both Application-oriented and Problem-oriented language, it provides a dynamic and automatic management for systems which supports different programming models incorporated with imperative, object-oriented, functional and procedural with a comprehensive standard library. Here we have used four python packages.

NumPy: NumPy is a Python library which adds reinforce for multi-dimensional arrays. It speedily integrates with a wide variety of databases and integrates with C/C++ also.

Pandas: The Pandas library is generally used for data science, This is because the Pandas library is used in conjunction with other libraries that are used for data science. It is built on top of the NumPy library which means that a lot of the structures of NumPy are used or replicated in Pandas.

The data produced by Pandas is often used as input for plotting functions in Matplotlib, statistical analysis in SciPy, and machine learning algorithms in Scikit-learn.

Scikit: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Natural Language Toolkit: Natural Language Toolkit (NLTK) is a package of libraries. It helps in writing programs for representative token and to process common language using Python which is written in English. It includes graphical demonstration and sample data which accompanied by concepts of processing tasks, plus a cookbook.

II. Keywords

Sentiment Analysis: Sentiment analysis, also known as opinion mining, is the process of determining the sentiment expressed in a piece of text, whether it's positive, negative, or neutral. It involves analyzing text data to understand the subjective opinions, emotions, or attitudes conveyed within it. Sentiment analysis is commonly used in social media monitoring, customer feedback analysis, and market research to gain insights into public opinion and sentiment trends.

Machine Learning: Machine learning is a subset of artificial intelligence (AI) that enables systems to automatically learn and improve from experience without being explicitly programmed. It involves the development of algorithms and statistical models that allow computers to perform specific tasks by learning patterns and relationships from data. Machine learning algorithms are trained on labeled datasets to recognize patterns and make predictions or decisions on new, unseen data.

NLTK (Natural Language Toolkit):NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources, such as WordNet, along with a suite of text processing libraries for tasks such as tokenization, stemming, tagging, parsing, and classification. NLTK is widely used for natural language processing (NLP) tasks, including text analysis, information retrieval, machine translation, and sentiment analysis.

Scikit-learn: Scikit-learn is a popular open-source machine learning library for Python. It provides simple and efficient tools for data mining and data analysis, built on top of NumPy, SciPy, and Matplotlib. Scikit-learn offers a wide range of supervised and unsupervised learning algorithms, including classification, regression, clustering, dimensionality reduction, and model selection. It also provides utilities for data preprocessing, model evaluation, and cross-validation, making it an essential toolkit for machine learning practitioners and researchers.

Table 1: Table comparison with existing models

S.NO	**Prior Art**	**Limitations / Demerits of Prior Art**	**Merits of the proposed idea to overcome the limitations of the prior art**
1	Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn	1)Data bias: Considering the different datasets which may inherits the training data effects the accuracy of the model. 2)The effectives of the model depends on the features extraction.	Considering a balanced dataset to understand the effectiveness of training model we can achieve better accuracy.
2	Twitter Sentiment Analysis Using Supervised Machine Learning	Limitations include potential data quality issues due to Twitter's noisy nature, biases and generalization challenges stemming	Employ robust data preprocessing techniques to mitigate noise, address biases through diverse dataset

		from training data, and difficulties in capturing nuanced sentiments and context within Twitter's character limit.	curation, and leverage advanced NLP methods like contextual embeddings to capture nuanced sentiments within Twitter's constraints.

III. Methodology

General Sentimental Analysis: A basic system that performs with a level of sentiment analysis on Twitter based on movies with following steps: a. Fetch data from Twitter. b. With Feature selection and Word Features, get both positive and negative tweets. c. With the help of Training set both Positive and Negative tweets and using Classifier we can analyze tweets who are positive and negative.

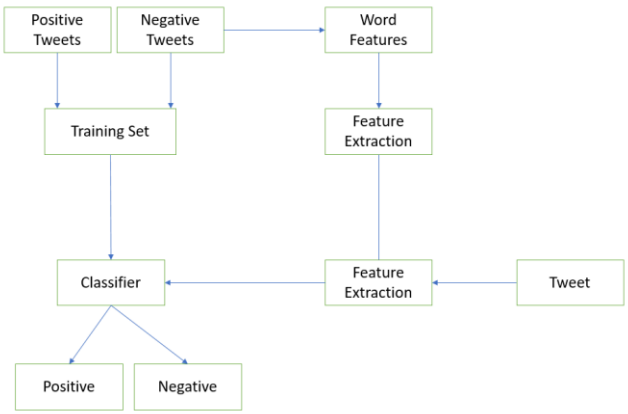

Figure 1: General Sentimental Flowchart

IV. Proposed work

Data Collection and Preprocessing:
You installed the Kaggle API and downloaded the Twitter sentiment dataset using it.
Preprocessed the data by cleaning and converting it into a format suitable for analysis.

Data Processing:
Loaded the data into a pandas DataFrames.
Renamed the columns for better understanding.
Handled missing values. Converted the target variable to binary, where 0 represents negative tweets and 1 represents positive tweets. Performed stemming on the text data to reduce words to their root forms.

Feature Engineering:
Utilized TF-IDF vectorization to convert text data into numerical format, which is suitable for machine learning algorithms.

Model Training:
Utilized logistic regression as the machine learning algorithm for sentiment analysis. Split the data into training and testing sets. Trained the logistic regression model on the training data.

Model Evaluation:
Evaluated the model's performance using accuracy score on both training and test data.

Model Saving and Future Predictions:
Saved the trained model using pickle for future use.
Demonstrated how to load the saved model and make predictions on new data.

V. Algorithm

Logistic Regression:

logistic regression is used in this code to classify sentiment in the Twitter dataset. It learns a decision boundary between positive and negative sentiments based on the TF-IDF features extracted from the text data during training, and then predicts the sentiment of unseen tweets in the test set.

Model evaluation

Accuracy Score

```
[ ] # 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 = 77.8%

```
X_new = X_test[200]
print(Y_test[200])

prediction = model.predict(X_new)
print(prediction)

if (prediction[0] == 0):
  print('Negative Tweet')

else:
  print('Positive Tweet')
```

```
1
[1]
Positive Tweet
```

```
X_new = X_test[132]
print(Y_test[132])

prediction = model.predict(X_new)
print(prediction)

if (prediction[0] == 0):
  print('Negative Tweet')

else:
  print('Positive Tweet')
```

```
1
[1]
```

VI. RESULTS AND DISCUSSION

We successfully executed the sentimental analysis of twitter data using machine learning algorithm logistic regression. Here we can represent the work flow of the results by breaking down them into points

1. **Data Collection:** We obtained a Twitter sentiment dataset from Kaggle containing 1.6 million tweets labeled with sentiment scores.
2. **Data Preprocessing**: We cleaned the text data by removing special characters, converting to lowercase, tokenizing, removing stopwords, and stemming.
3. **Data Splitting:** We split the dataset into training and testing sets for model evaluation.
4. **Feature Engineering:** We converted the textual data into numerical data using TF-IDF vectorization.
5. **Model Training**: We trained a logistic regression model on the training data.
6. **Model Evaluation:** We evaluated the model's performance on both the training and testing datasets, achieving an accuracy of around 81% on the training data and 77.8% on the test data.
7. **Model Saving**: We saved the trained model using pickle for future use.
8. **Model Inference:** We demonstrated how to load the saved model and make predictions on new data, correctly classifying tweets as positive or negative based on the model's predictions.

The final results of the trained model are satisfactory which is giving better accuracy then previous models. This helps us to improve the analysis of customers feedback data which results in development of business stratagies.

VII. References

[1] Shihab Elbagir and Jing Yang, “Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn”, The Hong Kong Polytechnic University and City University of Hong Kong, Publisher: Association for Computing Machinery New York, NY, United States, 21 December 2018.

[2] George B. Aliman, Tanya Faye S. Nivera, Jensine Charmille A. Olazo, Daisy Jane P. Ramos, Chris Danielle B. Sanchez, Timothy M. Amado, Nilo M. Arago, Romeo L. Jorda Jr., Glenn C. Virrey, Ira C. Valenzuela, Title of the book: “Sentiment Analysis using Logistic Regression”, Journal of Computational Innovations and Engineering Applications JULY 2022: 35-40.

[3] Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth international conference on computing, communications and networking technologies (ICCCNT), Tiruchengode, pp 1–5

[4] Yadav, N., Kudale, O., Rao, A., Gupta, S., Shitole, A. (2021). Twitter Sentiment Analysis Using Supervised Machine Learning. In: Hemanth, J., Bestak, R., Chen, J.IZ. (eds) Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies, vol 57. Springer, Singapore.

[5] Majumder, Sayan and Aich, Anuran and Das, Satrajit, Sentiment Analysis of People During Lockdown Period of COVID-19 Using SVM and Logistic Regression Analysis (March 5, 2021). Available at SSRN: https://ssrn.com/abstract=3801039 or http://dx.doi.org/10.2139/ssrn.3801039

[6] Srinivasan, R., Subalalitha, C.N. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. *Distrib Parallel Databases* **41**, 37–52 (2023).

[7] A. L. Yadav, K. Soni and S. Khare, "Heart Diseases Prediction using Machine Learning," *2023 14th*

International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023,pp.1-7,doi:10.1109/ICCCNT56998.2023.10306469.