Twitter Sentimental Analysis Using Rule Based and Machine Learning Method

Thejaswini N

PG Scholar
Artificial Intelligence and Machine Learning
Visvesvaraya Technological University
For Post Graduate Studies
Bangalore, India

Mr. Yadhu Naik B H

Professor
Artificial Intelligence and Machine Learning
Visvesvaraya Technological University
For Post Graduate Studies
Bangalore, India

Abstract: *In social media like twitter ,Information is present in large amount ,extracting and analyzing sentiment of tweets from twitter gives several usage in various fields. We can analyze sentiment of tweets by using two methods , rule – based method and machine learning method.so here creating application called tweet analyzer which automatically extracts tweets from twitter using API keys and analysis the sentiment of the live tweets using rule- based method (TextBlob,VANDER) and also analyzing already extracted sentiment140 dataset using machine learning algorithms. the sentiment140 dataset which contains 16,00,000 tweets. In order to automate classification of tweets here employed naive Bayes algorithm ,support vector machine algorithm, logistic regression and evaluating each model using accuracy score and f1 score. the results of each algorithm are tabulated and the best algorithm to fit is logistic regression than SVM, naive Bayes and deployed application using streamlit.*

Keywords*: twitter sentiment analysis, API Keys, naïve bayes, SVM, Logistic regression Classifier, streamlit*

Date of Submission: 02-02-2022 Date of acceptance: 16-02-2022

I. INTRODUCTION

The Internet has been really helpful in allowing people all over the world to express themselves. This is accomplished through blog posts, online discussion forums, item audit sites, and other means. This client's material is heavily relied upon by people all around the world. As an example before making a purchase choice, someone who wish to buy a product will study reviews and comments about it. However,It is big deal for a single person to read all of the available reviews in one sitting. It would be completely pointless. As a result, this procedure can be simplified by automating it.

Opinion mining, often well-known as sentiment analysis, is a natural language processing method.to figure out whether a paragraph is favourable, bad, or neutral. When big amount of consumer feedback is gathered through social media, it is frequently utilised by marketing sectors to monitor client satisfaction with a service, product, or brand and identify customer demands.

Machine Learning (ML) plays a critical role in this process[1]. The ML technique of Sentiment Analysis (SA) assists the system in determining the feeling of a given statement. system is based on multiple machine learning set of rules that be able to recognizes nature of feeling or a group of sentiments.Duringstudies, Machine Learning approaches In determining polarity, it beat knowledge and vocabulary based techniques.

II. LITERATURE SURVEY

DenikurniantoNugroho et al. (2021) worked on Prediction of the US presidential election in 2020 In accordance with Twitter data extraction and sentiment analysis using lexiconinformation used here was gathered from Twitter one week before the US presidential election.VADER sentiment analysis model was employed in this study. The data cleaning technique in this study is based on a text mining method. VADER sentiment analysis model can produce predictions the actual results of the US presidential election.As a consequence of this research, the Democratic Party is predicted to gain 22 votes versus the Republican Party's 19 votes. According to the BBC, the Democratic Party received 24 votes, while the Republican Party received only 20 votes.

Nikhil Yadav et al. (2021) Product Evaluation Using Twitter Sentiment Analysis and Machine Learning.Twitter, a microblogging service, is a vast storehouse of public opinions directed against a variety of people, products, companies, and other items. The system of analysing one's public ratings is known as sentiment evaluation. When sentiment analysis is integrated with Twitter, it provides useful insights into what people are saying.Sentiment analysis is thus critical for identifying how the general public feels about certain services or items.This study focuses on the various ways for categorising product criticisms (which can take the form of tweets) and analysing regardless of whether or not massive behaviouris it positive, bad, or indifferent?(neutral), as well as the application of that analysis to product market appraisal.

III. IMPLEMENTATION

A. DATABASE

'Twitter' is a large psychological database that is regarded as a treasure mine of data. Complex searches, such as retrieving every tweet about a specific topic or pulling a specific user's non-retweeted tweets, are possible with Twitter's API. Tweepy, a Python library, is used in conjunction with this. Developers can get public Tweet data for requested available Tweets using the GET /tweets endpoint. To begin, import all of the necessary packages and set up the token and key variables.

To analyze the sentiment of tweets from the Sentiment140 datasetWe plan to employ a machine learning pipeline that includes the classifier (Bernoulli Naive Bayes , logistic regression and support vector machine) as well as Term Frequency-Inverse Document Frequency.

B. SOFTWARE (TOOLS AND TECHNIQUES)

For this project PyCharm or spyder were the platform used to write the python program.PyCharm is the best integrated development environment I've ever used. PyCharm allows you to access the command line, connect to a database, build a virtual environment, and manage your version control system all from a single window, saving you time by eliminating the need to switch between windows. For deployment here streamlit is employed.

C. ALGORITHMS

Naïve Bayes:

Naive Bayes is one of straightforward supervised machine learning algorithm which means input dataset to algorithm should contain label of target output[1] . naive bayes algorithm works based on Bayes' theorem. There are strong assumptions about feature independence, i.e algorithm basically take responsibility that every input features/variable is unrelated to the others such a basic assumption about real-world data .It is mostly utilised in text classification tasks that require a large training dataset. Here ,we also using naïve bayes classifier to build machine learning model to automate classification of tweets. It is a probabilistic classifier, which means it makes predictions based on an object's likelihood.Bayes theorem states that probility of an event occurring given the probability of another event that has already occurred.every pair of features being classified is independent of each other and having equal contribution.

Support Vector Machine:

SVM/support vector machine is also common Supervised Learning technique that may be utilized to resolve both classification and regression issues[1].Mostly importantly utilised in Machine Learning for Classification difficulties.SVM algorithm's purpose is to discover finest line or decision borderline for dividing n-dimensional space into classes. to facilitate in future, so easily place fresh data point in the appropriate category. hyperplane is a best-case decision boundary.There are a lot of hyperplanes can be created upon which to divide the two types of data points.Aim is to discover plane by way of the largest margin, or distance among data points from both classes.extreme vectors that assist create the hyperplane are chosen via SVM.These vectors also called support vectors.If support vectors are detached, position of the hyperplane will changed.

Logistic Regression:

Machine learning and data mining applications rely heavily on classification algorithms. Classification challenges account for roughly 70% of Data Science issues. There are many different classification problems to choose from, but logistics regression is a popular and effective regression method for addressing binary classification issues[2][5]. Various classification issues, such as spam detection, can be solved using logistic regression. Diabetes forecasting, determining whether buyer will purchase a definite product or switch to aopposing, Just a few instances include assessing if a person will click on a given advertisement link and a variety of other scenarios.

For binary classification, Logistic Regression is one of the best simple and extensively used Machine Learning approaches. It's simple to set upand can be used to begin any binary classification task.core foundational principles of deep learning can help it. Logistic regression is used to define and guess the association among one dependent binary object and one independent feature. Instead of modelling a regression line, use a "S" shaped logistic task to calculate two maximum values in logistic regression (0 or 1).

IV. RESULTS AND DISCUSSION

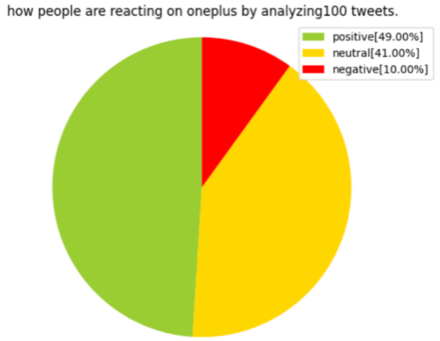


Fig1: Pie chart visualization for oneplus product analysis

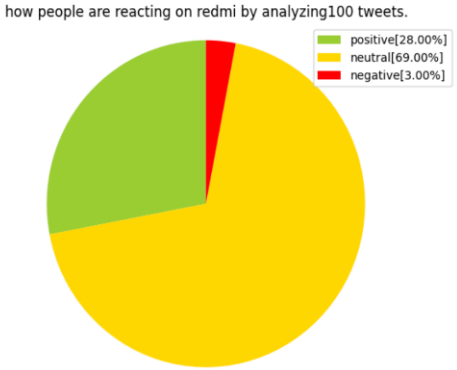


Fig2: Pie chart visualization for redmiproduct analysis

By visualizing above two pie chart we can easily analyze how people are reacting sentiment towards on redmi and oneplusproduct respectively by analyzing 100 latest live tweets by using rule based method and also we can conclude easily which is best among two products. also we can improve our business/product/service as a owner by analyzing sentiment of tweets for brand monitoring. similarly we can also use this model anywhere to automate product evaluation.

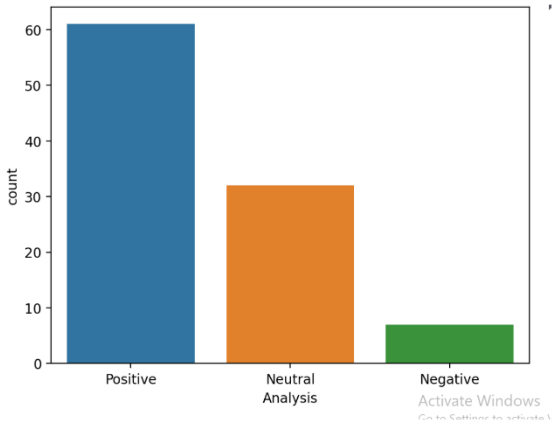
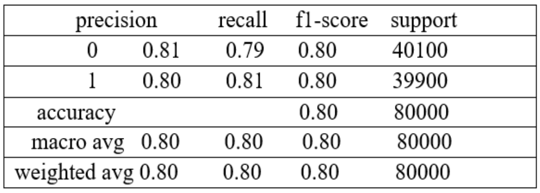


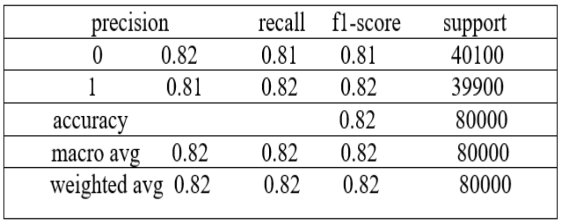
Fig3: Bar graph visualization for sentimental analysis of a particular person

By visualizing above bar graph we can easily analyze sentiment of the particular person by using rule based method.which helps to celebrity analytics so that we can able to choose best celebrity for endorsement.
Classication report of naïve bayes algorithm:



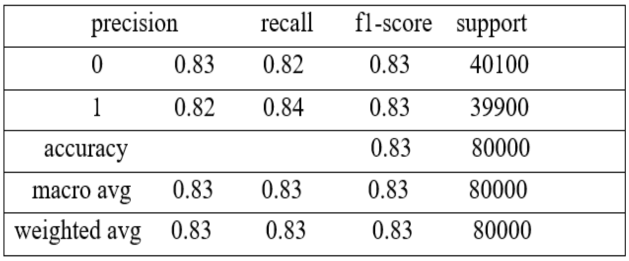
	precision	recall	f1-score	support
0	0.81	0.79	0.80	40100
1	0.80	0.81	0.80	39900
accuracy			0.80	80000
macro avg	0.80	0.80	0.80	80000
weighted avg	0.80	0.80	0.80	80000

Classication report of Support vector machine algorithm :



	precision	recall	f1-score	support
0	0.82	0.81	0.81	40100
1	0.81	0.82	0.82	39900
accuracy			0.82	80000
macro avg	0.82	0.82	0.82	80000
weighted avg	0.82	0.82	0.82	80000

Classication report of logistic regression algorithm:



	precision	recall	f1-score	support
0	0.83	0.82	0.83	40100
1	0.82	0.84	0.83	39900
accuracy			0.83	80000
macro avg	0.83	0.83	0.83	80000
weighted avg	0.83	0.83	0.83	80000

We may deduce the following details after examining all of the models:
In terms of model accuracy, Logistic Regression outperforms SVM, which in turn outperforms Bernoulli Naive Bayes.
F1-score: For classes 0 and 1, the F1 Scores are:
(a) Bernoulli Naive Bayes (accuracy = 0.80) SVM (accuracy = 0.81) Logistic Regression (accuracy = 0.83) for class 0
(b) Bernoulli Naive Bayes (accuracy = 0.80), SVM (accuracy = 0.82), and Logistic Regression (accuracy = 0.83) for class 1.
In our issue statement, Logistic Regression is using Occam's Razor, which states that if the data has no assumptions, the simplest model works best. Because our dataset has no assumptions and Logistic Regression is a straightforward model, the notion applies to the above-mentioned dataset. As a result, we conclude that the Logistic Regression model is the best fit for the dataset.

Reference

[1]. NikhilYadav,TwitterSentimentAnalysisUsing MachineLearningForProductEvaluation,isbn 9784-1-7281-4689-0

[2]. Sahar A. El_Rahman, FeddahAlhumaidiAlOtaib, and Wejdan Abdullah AlShehri, “ Sentiment Analysis of Twitter Data”, 2019 International Conference on Computer and Information Sciences (ICCIS), ISBN: 978-1-5386-8125-1, 2019.

[3]. HetuBhavsar, RichaManglani" Sentiment Analysis of Twitter Data using Python"International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN: 2395-0056 p-ISSN: 2395-0072.

[4]. Mishra, Prerna, RanjanaRajnish, and Pankaj Kumar. (2016) "Sentiment analysis of Twitter data: Case study on digital India." International Conference on Information Technology (InCITe)-The Next Generation IT Summit on the Theme-Internet of Things: Connect your Worlds. IEEE, 148 – 153.

[5]. Pavel, Andrei, et al. (2017) "Using Short URLs in Tweets to Improve Twitter Opinion Mining." 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 965 – 970.

[6]. sara Rosenthal, NouraFarra and PreslavNakov, "SemEval-2017 task 4: Sentiment analysis in Twitter", Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017.