International Journal For Technological Research In Engineering
Volume 10, Issue 9, May-2023 ISSN (Online): 2347 - 4718

A RESEARCH PAPER ON SENTIMENT ANALYSIS OF TWITTER POSTS USING MACHINE LEARNING APPROACHES

[1]Amit Kumar, [2]Prof. Ritesh kumar
[1]Research Scholar, [2]Assistant Professor
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
PKG Group of Institutions
KURUKSHETRA UNIVERSITY, KURUKSHETRA

Abstract— ***This study deals with the issue of sentiment analysis on Twitter, which involves categorising tweets based on whether they exhibit positive or negative emotions. As all know, 'Twitter' is a platform that connects news and social networking issues to online social media micro blogging website that enables users to post 140-character limit to an issue update. It has more than 200 million users as a registered end user [24], of whom 100 million are active users and half of them log in at least once in a day, it is a free online service that is growing in a strictly increasing manner. Each day, it generates about 250 million tweets [20]. In this work, we represent a public opinion by analyzing sentiments stated in the tweets in light of this significant use. Numerous applications need to justify public mood, including businesses environment and deals to gauge the market reaction to their goods, the prediction of political outcomes, and the analysis of socioeconomic phenomena like the stock exchange, etc. The goal of this project is to create a practical classifier that can accurately and automatically identify the sentiment of an unidentified Twitter stream.***

Keywords: Machine Learning, Twitter

1. INTRODUCTION

This Twitter sentiment analysis project falls under the categories of "Pattern Classification" and "Data Mining." Both of these concepts are intimately connected and entangled, and they may be properly described as the automated (unsupervised) or semiautomatic process of catching "useful" arrangements in vast sets of data (supervised). In order to effectively identify individual unlabeled data samples (tweets) which come out to be the best version of the pattern model with the best characteristics accordingly, the project will mainly depend on "Natural Language Processing" methods for withdrawing considerable patterns and features from the enormous dataset of tweets.

The separation of formal language-based appearance and informal blogging-based characteristics may be characterized into two broad groupings that can be utilized for modeling patterns and categorization. Language-based characteristics deal with formal linguistics that carries the parts of speech in which each sentence is specified with the preceding sentiment polarity of certain words and phrases. Prior sentiment polarity distinguishes the inherent inbuilt inclination of certain words and phrases to communicate particular and specific feelings in general. For instance, the term "great" has a strong connotation of positivity, while the word "evil" carries a strong connotation of negativity. Therefore, anytime a word with a meaningful sense is employed in a phrase, there is a great chance that the statement as a whole will be fruitful. On the other hand, a syntactical approach to the issue is Parts of Speech labeling. It indicates to the ability to recognize the part of speech in a phrase corresponding to each word, including a noun, pronoun, adverb, adjective, verb, interjection, etc., belongs to. By examining the frequency distribution of these parts of speech in a particular category of labeled tweets, either alone or in the composition of other parts of speech, patterns may be derived. The elements of Twitter are more infrequent, attach to how individuals express themselves on social networking online platforms, and compress their ideas into the constraint of 140 characters that Twitter provides. Twitter hashtags, retweets, word capitalization, word lengthening, question marks, the use of URLs in tweets, exclamation points, online emoticons, and internet shorthand/slang are a few examples of these. The two kinds of classification methods are supervised vs. unsupervised, and non-adaptive vs. adaptive/reinforcement methods. When we have pre-labeled data samples available, we may train our classifier using an instructed technique.

2. OBJECTIVES

The public's perceptions of political leaders or the concepts they have about the laws and regulations that are in existence, etc.

Analyze and categorize data.
Analyze the sentiment of the tweet (review basis).
Classification of the sentiment of public opinion into-
-Positive
-Negative

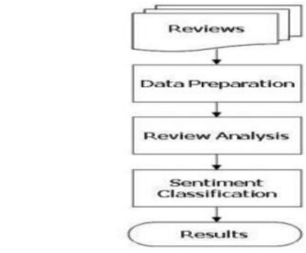


Figure 1: A typical sentiment analysis model.

3. METHODOLOGY

SENTIMENT SENTENCE EXTRACTION & POS TAGGING:

The fundamental need for POS tagging is to tokenize reviews after removing STOP words that have no relation to sentiment. The remaining phrases are transformed to tokens after STOP words like "am, is, are, the, but," and similar words have been properly removed. The POS tagging process involves these tokens.

The taggers of part-of-speech (POS) have been identified in natural language processing (NLP) in order to categorise words into a form of machine algorithm to understand their parts of speech. A POS tagger is highly recommended to the users for sentiment analysis for two particular reasons defined as: 1) Pronouns and nouns of the speech often have no feeling attached to the taggers/users (there is no strong AI there).

NEGETIVE PHRASE IDENTIFICATION:

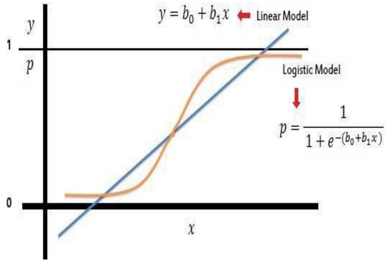
The fundamental need for POS tagging is tokenization of learner after the elimination of STOP phrases that have no relation to sentiment. The remaining phrases are transformed to tokens after the appropriate elimination of STOP words like "am, is, are, the, but," and so forth. These tokens are used for POS tagging.

In order to categorise words according to their parts of speech, part-of-speech (POS) taggers have been identified. A POS tagger is very important for sentiment analysis for the two reasons listed below: 1) Pronouns and words like nouns often don't convey feeling. With the use of a POS tagger, it is possible to filter out such terms; 2) A POS tagger may also be used to separate words that can be used in various parts of speech.

SENTIMENT CLASSIFICATION ALGORITHMS:

Bernoulli Naïve Bayes classifier:

The data taken which is assumed to be multivariate Bernoulli distributed, Bernoulli (NB) implemented the naive Bayes training and several classification methods to the nodes; On the other side, even if there may be several parameters that are going to take existence, each parameter is considered to be a labelled as binary-value (Bernoulli, boolean). Since this class only accepts binary-valued feature vectors as sample representations, BernoulliNB 10.



Implementation Rule
The following two steps are discussed to train the dataset are given as Unpacking of data: The enormous collection of Amazon.com reviews is available as.json files. To understand the given dataset from those files and then dump it into a pickle file for quicker access and object serialisation, a tiny piece of Python code has been written.
Therefore, the first data retrieve is carried out in this step utilising Python File Handlers.

The Description of Preparing Data for Sentiment Analysis:

i) Consequently, the pickle file is loaded in upcoming state, and all data try to fulfil the exceptions needed for sentiment analysis is going to be eliminated. There are many columns in the data, as shown in our example dataset on Page 11, but just the rating and text the review are the upcoming resources we needed. The "review Summary" column is therefore removed from the existance.

4. RESULT

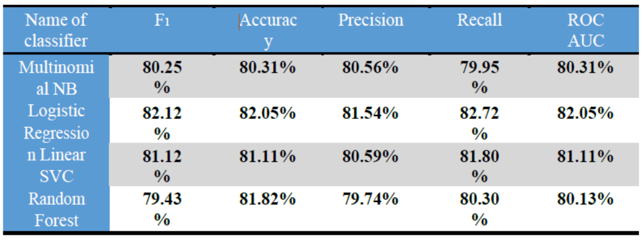
Results and Sample Output
The computer is now able to determine if a statement that has been input will get a favourable or negative response as a result of this training dataset of public assessments.

The considerable amount of relevant examples among the recovered instance is defined as precision (also marked as positive predictive value), while the amount of relevant instances that have been retrieved relative to the total number of relevant instances is marked as recall (also known as sensitivity). Therefore, a comprehension of and a measurement of relevance are the foundations of both precision and recall.

The F1 score, as we all aware it is commonly known as the F-score or F-measure, is a gauge of a test's precision. The approaching value is calculated by the fraction of total number of accurate positive outcomes with the total positive outcomes retrieved by the classifier (p) and the counting of accurate positive outcomes by the total number of relevant samples (r), respectively (all samples that should have been identified as positive). In the process, F1 score is here the harmonic average value of accuracy and recall, having the upper bound of 1 (perfect precision and recall) and a lower bound of 0.

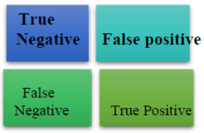
The main characteristic curve of receiver of operating system, or ROC curve, is nothing but a graphical representation used in statistics to show how much fluctuation is noticed to show the discriminating threshold which affects a binary classifier system's ability to diagnose problems. By displaying the whole data in a two-by-two contingency table for each threshold level, the Total Operating Characteristic (TOC) develops on the concept of ROC only. The TOC strictly provides more information than the ROC since the ROC only provides two bits of relative information for each threshold. The likelihood that a classifier would rate a randomly occured positive instance higher than a randomly selected negative one when using normalised units is equal to the area under the curve (often abbreviated as the AUC). This is assuming that "positive" ranks higher than "negative".

The area which can be calculated with in the desirable range is given as follows, which is acceptable that actually makes sense (the boundary value for the integration are choosen as large T has a lower bound on the x-axis).



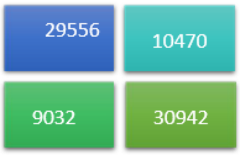
Name of classifier	F1	Accuracy	Precision	Recall	ROC AUC
Multinomial NB	80.25 %	80.31%	80.56%	79.95 %	80.31%
Logistic Regression	82.12 %	82.05%	81.54%	82.72 %	82.05%
Linear SVC	81.12 %	81.11%	80.59%	81.80 %	81.11%
Random Forest	79.43 %	81.82%	79.74%	80.30 %	80.13%

The Confusion Matrix Format is as follows:



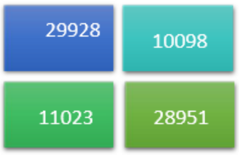
True Negative	False positive
False Negative	True Positive

The Confusion Matrix of Each Classifier are as follows:



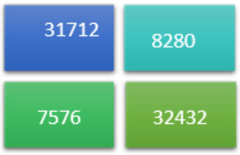
29556	10470
9032	30942

Classifier 1: Multinomial NB



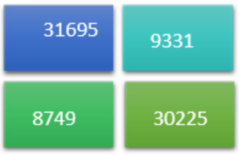
29928	10098
11023	28951

Classifier 3: Liner SVC



31712	8280
7576	32432

Classifier 4: Bernoulli Naïve Bayes



31695	9331
8749	30225

Classifier 4: Random Forest

The following are the images of such sample output after successful dataset training using the classifiers:

5. CONCLUSIONS

In this report, we used ‘Sentiment analysis’ which is the procedure of categorising writings respect to the emotions they express. This article discusses representative approaches used in the three basic processes of a typical sentiment analysis model, which are data preparation, analysis, and sentiment categorization.

In recent research works, it have been clearly seen that the sentiment analysis, a developing field in text mining and computational linguistics, etc became the interest of numerous researchers and the increment of their significance is getting high. In-depth of the methodologies where looking for extracting opinions and features of the actions it can be establish in the future research. We can incorporate the novelty in the classification of the models which can adapt the ordered labels characteristics in rating inference into the account. Many real time applications can be marked on a usage of sentiment analysis outcomes are also anticipated to appear in future scope.

REFERENCES

[1] Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. Discovery Science, Lectures Notes in Computer Science, 2010, Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1_1

[2] Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[3] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceeding of international conference on language Resources and Evaluation(LREC), 2010.

[4] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceeding of AAAI conference on Weblogs and Social Media (ICWSM),2010.

[5] Google App Engine https://developers.google.com/appengine/.

[6] Google Chart API <https://developers.google.com/chart/>

[7] Tweet Stream: Simple Twitter Streaming API Access http://pypi.python.org/pypi/tweetstream

[8] Twitter REST API <https://dev.twitter.com/docs/api>

[9] Twitter Sentiment, an online application performing sentiment classification of Twitter. <http://twittersentiment.appspot.com/>

[10] Ian H. Witten, Eibe Frank & Mark A. Hall. Data Mining – Practical Machine Learning Tools and Techniques.