Twitter Dataset Based Sentimental Analysis with Machine Learning Approach

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

BE CSE - AIML

Submitted by:

ENDLURU VIGNESH

20BCS6891

VUPUUTURI BHARATH

20BCS6586

KETHURI AJAY

20BCS6585

KOYYADA AKSHAY KUMAR

20BCS6369

Under the Supervision of:

DR RANJAN WALIA

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

MONTH & YEAR
MARCH - 2023

Table of Contents

ABSTRACT

A Twitter sentiment analysis is the process of determining the emotional tone behind a series of words, specifically on Twitter. A sentiment analysis tool is an automated technique that extracts meaningful customer information related to their attitudes, emotions, and opinions. Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service, or idea. Sentiment analysis or opinion mining refers to identifying as well as classifying the sentiments that are expressed in the text source.

Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of people on social media for a variety of topics. Therefore, we need to develop an Automated Machine Learning Sentiment Analysis Model in order to compute customer perception. Due to the presence of non-useful characters (collectively termed as noise) along with useful data, it becomes difficult to implement machine learning models on them.

The goal of tweet sentiment analysis is to find the positive, negative, or neutral sentiment part in the tweeter data. Sentiment analysis can help any organization to find people's opinions of their company and products. We have applied sentiment analysis on twitter data set. Our model takes input tweet, sentiment, and output selected text starting and ending in input tweet. We are using Natural Language Processing model with one more layer to find sentiment positions in tweets. The extra layer is used to find the sentiment part of a tweet. Our model can achieve 80 percent accuracy on the validation data set.

Timeline/Gantt Chart

S.N	Strategies	1st week	2nd week	3rd week	4th week	5th week	6th week
1)	Problem Identification						
2)	Research & Analysis						
3)	Design						
4)	Coding						
5)	Implementation & testing						
6)	Project finalisation						
7)	Documentation						

1. **INTRODUCTION**

Nowadays, the age of Internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. Nowadays, millions of people are using social

network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising.

People mostly depend upon user generated content over online to a great extent for decision making. For e.g., if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision. The amount of content generated by users is too vast for a normal user to analyze. So, there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis (SA)tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements. Textual Information retrieval techniques mainly focus on processing, searching, or analyzing the factual data present. Facts have an objective component but, there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks.For example, recommendations of items proposed by a recommendation system can be predicted by considering considerations such as positive or negative opinions about those items by making use of SA.

1. Problem Definition:

In this project, we try to implement an NLP Twitter sentiment analysis model that helps to overcome the challenges of sentiment classification of tweets. We will be classifying the tweets into positive or negative sentiments. The necessary details regarding the dataset involving the Twitter sentiment analysis project are:
The dataset provided is the MP Twitter Dataset which consists of 5000 tweets that have been extracted using the Twitter API. The various columns present in this Twitter data are:
Target: the polarity of the tweet (positive or negative)
Ids: Unique id of the tweet
Label: Type of the tweet
Tweet: Text of the tweet

2. Project overview:

In this project, a dataset is taken for reference of tweets for instance racist tweets. After training and testing, the model is ready to classify whether the tweet is positive or negative accordingly. The only drawback/ limitation is that dataset is small. It leads to low accuracy.

1.3 Hardware Specifications:

- Processor – 64-bit eight-core, 2.5GHz per core.
- RAM – Minimum 4GB required.
- Hard Disk – SSD or HDD minimum 40GB free space required.

1.4 Software Specifications:

- Edition - Windows 10/11
- OS build 19043.1526
- Python installed – version 3.7 to 3.10
- nltk installed on windows
- Any python compiler or system terminal

2. LITERATURE REVIEW

Literature review has been selected as the research method. The literature review is selected in order to gain knowledge and deep understanding about various Natural Language Processing models and their efficiency so that the most suitable and efficient method can be selected from the identified models.
Sentiment Analysis is the procedure of determining the instance of the class to which the text belongs and estimating the emotion of the text by outputting the label given for the particular mood behind it.

2.1 Existing System:

The paper [1] (Mittal, 2016) describe the requirement and impact of the sentiment analysis on on-line platform. They need additionally bestowed a listing of sentiments of emotions, interjections and comments that are extracted from posts and standing updates. They need got result to knowing whether or not {the on-line the web the net} reviews and posts are being useful to client or not and that on-line websites being most popular by the purchasers.

The paper [2] (Anto, 2016) describe the merchandise rating mistreatment sentiment analysis. In promoting of any product the producer can get the proper result from the client feedback. After got feedback they'll changes to his product in step with the feedback. Some users continually fail to convey their feedbacks. Objective of this paper is to avoid the problem of providing feedbacks and supply the technique which might provide automatic feedback on the premise of information collected from twitter. They used the technique SVM and got result eightieth accuracy. This system offers quick and valuable feedback.

The paper [3] (Saragih, 2017) describe regarding the client engagement by analysis the comments on social media in transport on-line. They used technique TF-IDF. The result shows that the class “Feedback system by driver” and “Feedback system by user” have the foremost comments for 3 means that of transports on-line, whereas class “service quality for driver “has the littlest comments. This feedback of social media is accustomed evaluate the performance of this business transport on-line.

This paper [4] (Mamgain, 2016) describe regarding the sentiment analysis of people's opinions relating to high faculties in India. They need represented comparison between the result obtained by the subsequent machine learning algorithms: Naive Bayes and SVM and Artificial Neural Network model: Multilayer Perception. Naive Bayes {Thomas Bayes mathematician} outperforms SVM for the aim of matter polarity classification that is fascinating as a result of the model utilized by Naive Bayes is easy (use of freelance probabilities) and therefore the likelihood estimates made by such a model are of caliber. Yet, the classification selections created by the Naive Bayes model portray a decent accuracy as a result of whenever a call with the upper likelihood is being created.

2.2 Proposed System:

In this project, the model is created by machine learning algorithms and Natural Language Processing. Model has access for a dataset called Mp twitter that includes text of some various types of tweets. With the help of the dataset, the model can differentiate between the emotion behind the text. In this model group of tweets are trained with the help of labels and its unique id. The model precise that input text is given then classifies it and shows the type of sentiment.

3.PROBLEM FORMULATION

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

4.OBJECTIVES

The proposed work is aimed to carry out work leading to the development of an approach for sentiment analysis with Natural Language Processing. The proposed aim will be achieved by dividing the work into following objectives:

1. Create a model for classifying the tweet into respective sentiment.
2. Creating a new dataset of latest tweets consisting 5000 of them.
3. After training and testing, find its accuracy accordingly.
4. Make sure the accuracy is above 65%.

5. Data visualize on the model.

5.METHODOLOGY

5.1 Pre-processing of the datasets

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,

1. Remove all URLs (e.g www.xyz.com), hash tags (e.g. #topic), targets (@username)
2. Correct the spellings; sequence of repeated characters is to be handled
3. Replace all the emoticons with their sentiment.
4. Remove all punctuations, symbols, numbers
5. Remove Stop Words
6. Expand Acronyms (we can use a acronym dictionary)
7. Remove Non-English Tweets

5.2 Feature Extraction

The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram [18]. Machine learning techniques require representing the key features of text or documents for processing.

These key features are considered as feature vectors which are used for the classification task Some examples features that have been reported in literature are:

1. Words And Their Frequencies: Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. showed better results by using presence instead of frequencies.

2. Parts Of Speech Tags Parts of speech like adjectives, adverbs and somegroups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3. Opinion Words And Phrases Apart from specific words, some phrases and idioms which convey sentiments can be used as features. e.g cost someone an arm and leg.

4. Position of Terms The position of a term with in a text can effect on how much the term makes difference in overall sentiment of the text.

5. Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.

6.REFERENCES

[1] Mittal (2016) Research on impact of sentiment analysis.

[2] Anto (2016) rating mistreatment sentiment analysis.

[3] Saragih (2017) client engagement by analysis the comments on social media in transport on-line.

[4] Mamgain (2016) sentiment analysis of people's opinions relating to high faculties in India.

[5]. Zahratu Sabrina explained what are the different approaches for sentiment analysis, 2023.

[6]. 2023, Shamsuddeen Hassan Muhammad implemented sentiment analysis for various languages in Africa.

[7]. 2021, Vedurumudi Priyanka used ensembling technique, which combines various classifiers, in order to improve the accuracy for the dataset taken from Kaggle.

[8]. 2023 Imane Lasri used sentiment analysis for Moroccan public universities from twitter using big data technologies.

[9]. 2022, Yili Wang used a few other approaches for sentiment analysis which includes probabilistic classifier, Linear classifier, rule-based classifier, hybrid approach and other approaches.

[10]. 2022, Masoud AminiMotlagh collected data, pre- processed it, detected and classified accordingly. Thisresearch paper included some of the machine learning techniques like KNN, SVM, NAÏVE BAYES and Bagging.

[11]. 2022, Astha Modi analyzed how different techniques like LSTM, Naïve bayes, decision tree and SVM are working. Accuracies are also recorded and future scopes of every technique also mentioned in it.