Project Title:

Sentiment Analysis of Tweets

Project Members:

19BCE0249 (Kaustubh Dwivedi)

19BCE0276 (Harine A)

19BCE0253 (Nishma Avalon Rebello)

19BCE0200 (Soumyaraj Roy)

Report submitted for the

First Project Review of

Course Code: CSE3013 – AI

Slot: F1

Professor: Dr. W.B. Vasantha

Submitted on: 26 March 2021

1. ABSTRACT

Sentiment analysis is a significant tool in social media monitoring and is often performed on textual data, as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools, for example, Brand watch Analytics make the process quicker and easier than ever before, because of Realtime monitoring capabilities. The applications of sentiment analysis are broad and powerful as it helps big companies to decide the nature of their customers and implement changes. The ability to extract insights from social data has been a practice which is being widely adopted by organizations across the world. Shift in the sentiments on social media have been shown to correlate with shifts in the stock market. Being able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategize and plan for the future. It can also be an essential part of one's market research and customer service approach. Not only a company can see what people think of their products or services, they can see what people think about the competitors too. The overall customer experience of a company's users can be revealed quickly with sentiment analysis, also it can get far more granular too. This project targets to enact on such a tool for all purpose utility. There are various social media platforms where people express their views and opinions. Twitter is one such social media platform, where public expresses their opinion freely and in a vast quantity. We have chosen twitter in this project for data collection, experimentation and analysis. Analyzing the reactions on twitter, can give a true opinion analysis of majority of the people. We will check the success rate of different algorithms and implement an algorithm to our own understanding and incorporate an aspect that will allow us to judge sentiments in the tweets. Sentiment analysis helps us to get a detailed review of this huge information in an organized way.

Keywords: Sentiment analysis, Brand watch analytics, twitter, social media, opinion analysis.

2. INTRODUCTION

This project deals with Sentiment Analysis, also called opinion mining, which is a Natural Language Processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

We will use the concepts of Natural Language Processing (NLP), which is a field of Artificial Intelligence, in which computers are programmed how to process, analyze and understand large amounts of natural language data and hence derive meaning from human language in a smart and useful way.

Need of sentiment analysis

i) In Business:
In marketing field, the companies use it to develop their strategies or to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches; and why do the consumers not buy some products.

ii) In Politics:
In the political field, it is used to keep track of political view and to detect consistency and inconsistency between statements and actions at the government level. It can also be used to predict election results.

iii) Public Actions:
Sentiment analysis can also be used to analyze and monitor social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

By automatically sorting the sentiment behind reviews and social media conversations in open platform like Twitter, a company such as an E-Commerce company, can make faster and more accurate decisions for its market.

In this project we will perform the following tasks:

1. Firstly, Collection of Datasets from Kaggle.

2. Pre-process the data (transformation of data).

3. Train the data using python code.

4. Test the data.

5. Analyze the data and show result.

6. Analyze and show the level of accuracy of the prediction (analyze file).

3. **Literature Survey**

S.NO	Authors and Year (Reference)	Title (Study)	Concept / Theoretical model/ Framework	Methodology used/ Implementation	Dataset details/ Analysis	Relevant Finding	Limitations/ Future Research/ Gaps identified
1.	Erik Cambria and Soujanya Poria, Alexander Gelbukh, Mike Thelwall, 2017	Sentiment Analysis Is a Big Suitcase	This paper focuses more about deep learning and its implementation using NLP(natural language program)	This paper basically explains about the syntactic layer, and further breaks down to Microtext Normalization, Sentence Boundary Disambiguation, Part-of-Speech Tagging and their detailed explanation . further more it gives us a more deep understanding of the semantics and the pragmatics layer.	We can understand about the how sentimental analysis can be implemented through NLP, its structure and the break down and use of each process. Also the paper provides the problems to deal with in each process and alternatives.	Sentiment analysis enormous bag of natural language processing (NLP) issues. sentimental analysis has for some time been confused with the undertaking of polarity detection. This, notwithstanding, is only one of the numerous NLP issues that should be addressed to accomplish human-like execution in sentiment analysis.	Some NLP undertakings, in any case, require more than a simple data driven way to deal with accomplish human-like execution more pros and cons are to be discussed regarding the sentiment approach towards NLP.
2.	Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Imad Zaza, 2017	Twitter Vigilance: a Multi-User platform for Cross-Domain Twitter Data Analytics, NLP and Sentiment Analysis	This paper proposes the twitter vigilance architecture, which is a cross-space, multi-client apparatus for gathering and dissecting Twitter information, giving aggregated measurements, dependent on the volume of tweets and retweets, clients' impact organization, Natural Language Processing and sentiment Analysis of textual content.	This journal first gives us an insight of what a social media analytics platform is, and also explains how sentiment analysis of social media can be influenced such as surveying customer sentiments, anticipating monetary and market results , anticipating political race results, giving early recognition and cautioning for unfavourable technical issues as well as for disaster response surveillance systems.	In order, to build this architecture, the important aspects concerned are data, NLP and Sentiment Analyses based metrics, API availability, User network analysis, real time analysis and full faceted search. The paper also gives a detailed analysis about the architecture.	The extraction of Part-of-Speech (POS) labelled keywords and calculation of catchphrase event at various time goal. sentiment polarity extraction for each single tweet; this sort of data can be valuable to evaluate and appraise the overall notion of the Twitter local area in regards to a particular channel or search. lower-level measurements are utilized to weight keyword events (registered as NLP-based measurements, as recently depicted) to assess the most powerful	This paper explains the architecture very well, with help of case study and understand the implementation of the sentiment analysis through NLP and social media analytics.

						keywords for conclusion examination, just as distinguishing conceivable sources and motivations to clarify or decipher explicit supposition patterns.	
3.	S.Muthukumaran, Dr.P.Suresh 2018	Text Analysis for Product Reviews for Sentiment Analysis using NLP Methods	This paper clarifies various strategies for sentiment analysis and displays a productive methodology. It likewise features the significance the item surveys are of most extreme significance for the purchasers to choose depending on their interests with respect to item's different angles for instance a monitor, processor speed, memory.	The proposed method they have analyzed various types of algorithms, for predicting semantic orientation. They utilized four-stage supervised learning algorithm to derive the semantic direction of descriptive words from constraints on conjunctions. The texts are at that point tokenized into tokens and the stop-words are recognized and taken out. The audits for a couple of mainstream telephones have been gotten by building a web crawler. The web crawler has been written in Python utilizing a scraping library called Beautiful Soup. Alongside the survey text, some extra information bunches and the lexical semantic highlights are appeared to have higher exactness than.	This paper distinguishes solid pieces of information of subjectivity utilizing the consequences of a technique for bunching words as indicated by distributional similarity. Basically this paper uses Flip kart Reviews Database as dataset for this project . The main source of data used is the product reviews from Amazon. They utilize Dirichlet distribution and Bayesian Classification , that represents a supervised learning method as well as a statistical method for classifications of various words and their meaning.	The paper mainly focuses on the implementation of the sentiment analysis. We can also understand more about opinion mining as well as sentiment analysis. The architecture proposed utilizes a non-supervised sentiment order, approach for sentiment classification and it is assessed utilizing a dataset of online client surveys of cell phones. This paper shows that, the framework performs very well in opinion arrangement of client surveys with high exactness. implemented fuzzy functions to emulate the effect of various linguistic hedges such as dilators, concentrator and negation on opinionated phrases help the system to achieve more accuracy in sentiment classifications.	As future work of this journal, we can refine rule set to extricate more reliance relations from datasets and that will assist with improving the precision and review estimations of the framework by characterizing algorithms. In the event that the framework ready to right all the spelling and syntactic blunders present in the survey reports in the pre-processing step itself that will improve the review estimation of the System execution.
4.	Alex Mordkovich, Kelly Veit, Daniel Zilber 2011	Detecting Emotion in Human Speech	This paper mainly focuses on detection of emotion in speech, they use a free software(praat) which is used to process the audio data and extract various statistics which includes voice report. The	In this paper in order to analyse various emotions of the speaker, the sound accounts and record information are pre-processed in a custom four-stage pipeline to produce an information document in which every expression is	In this paper for preparation, cross-approval, and testing, they utilized the Emotional Prosody Speech and Transcripts acquired from the Linguistic Data Consortium. This information comprises	Through this paper we could analyse the detection of emotion in the human speech by analysing various aspects and they concluded that we could take the sample inputs and classify them under 14 emotions, when trained on the complete training	The feature selection was executed in two ways. The main route was to discover the mix of 1-3 features that limits training error. The subsequent methodology was to utilize a forward or in

			statistic in the report (pitch, pulses, voicing, jitter, and harmonicity) are determined through this report. It also emphasises about Mel-frequency cepstral coefficients (MFCCs) which are a common set of features used in voice processing algorithms.	an information test addressed as a single line. This subsequent information record is stacked into MATLAB as a stylized design matrix. The chose K-Means clustering as their classification algorithm for simplicity purpose. Also they experimented with the SVM implementations in Liblinear, LibSVM, and the MATLAB-builtin SVMClassify for our classification tasks.	accounts of expert actors discussing dates and numbers with different emotional intonations. The semantic content of the expressions is proposed to be sincerely unbiased, as a type of mental control in the examples.	data set.	backward search heuristic. Neither one of the approaches improved outcomes fundamentally. These search algorithms for new features end up being slow. Basically, the main aim of this proposal was to analyse various emotions through the voice. Now they could actually classify under 14 emotions but it can be still improved to expand the classifications to improve the accuracy.
5.	Milad Sharif, Soheil Norouzi 2011	Sentiment based model for Reputation system in Amazon	This paper mainly proposes that create tools to evaluate the semantics of item audits and determining the polarity of opinions. Also to assess the strength of an opinion is utilizing audits with numeric ratings and preparing (semi-)supervised learning calculations to arrange surveys as certain or negative	In this paper we could analyse that the semantic orientation and strength of a survey is anticipated by following the adjustments in the related financial factors of a dealer. the technique utilizes two diverse parallel classifiers (for example Innocent Bayes and semi-directed recursive auto-encoder) to foresee the exceptional cost of an item. The notion investigation calculation (for example RAE) was conveyed to acquire the semantics of the item surveys and gave a model to the exceptional costs	The data set incorporated in this paper details of the different transactions that occurred on Amazon.com for a wide range of software items. The data set gathered from freely accessible data at Amazon.com by utilizing Amazon Web Services. The data set incorporates two sections, transaction history and reputation data. The initial segment comprises of transaction IDs at every product and the cost at which the products were sold. The second piece of data set incorporates the reputation history of every trader that had a product available to be purchased during the time frame which the data set was gathered.	Basically this journal helps us to analyse the semantic orientation and strength of a review incorporated in the amazon service d by tracing the changes in the associated economic variables of a merchant. Various algorithms and methods such as multivariate Bernoulli event model and Semi-supervised Auto-Encoder (RAE) architecture are well explained and utilized in their model. Basically, they conducted a survey in which they used different binary classification models to accurately predict the polarity of the premium price that a merchant gets based on the costumer reviews	This paper provides us a very simple survey of analysing different classification model and determined the accurate one which can predict these distributions more accurately than other models. Further more research to be done on how its it better than other models and how the accuracy can be improved.

4. PROPOSED WORK AND IMPLEMENTATION

Methodology adapted:

1. Use data sets from Kaggle (An online platform capable of providing users with datasets)

2. Pre-processing the data

Cleaning, normalization, transformation, feature extraction and selection, and so on are all part of the pre-processing. The result of pre-processing would be coherent and uniform data that can be used to improve the performance of the classifier.

3. Use python code to train the data set

The machine learning modules which we plan to use include SVM, Logistic regression, random forest, Naïve Bayes and text blob

One of the most basic text classification algorithms is the Naive Bayes classifier. It's a simple classifier based on the Bayes theorem that makes naive assumptions about the feature variables' independence.

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model

Support vector machines so called as SVM is a supervised learning algorithm. The Ideology behind SVM is to find a hyperplane that best separates the features into different domains.

Text Blob is a Python library meant for processing textual data.

4. Testing phase

5. Analyze the data and display results

Perform Sentiment Analysis on Tweets After gathering and cleaning our data set, we are ready to execute the sentiment analysis algorithm on each tweet. Then, we will calculate an average score for all the tweets combined.

6. Visualization of results using graphs and charts

We plan to use Pyplot to display figures. matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc

7. Analyze as well as show the accuracy-level of the prediction (analyses file)

HARDWARE AND SOFTWARE REQUIREMENTS:

SOFTWARE REQUIREMENT

Operating System: Windows 7, Windows XP, Windows Vista or higher versions

Programming Language: Python

Coding Platform: Any Python Platform such as Anaconda, Spyder or Jupyter Notebook

Modern Web Browser: Preferably Chrome or Firefox

APIs: Twitter

Kaggle Data Set

HARDWARE REQUIREMENT

RAM: 1GB or more

Processor: Any Intel Processor

Hard Disk: 6GB or more

Speed: Min 1 GHz

No additional hardware components are required.

5. Dataset used / Tools used

a. We are taking this dataset from Kaggle.com. The data embodies the relationship mapping tweets to their author's sentiments: positive or negative. The tweets have been extracted using the twitter api.

b. Reference paper we are taking in consideration is

Sailunaz, K. (2018). Emotion and Sentiment Analysis from Twitter Text (Unpublished master's thesis), University of Calgary, Calgary.

The link for the following project is mentioned below:

https://prism.ucalgary.ca/handle/1880/107533

c. Our project differs the above research paper that we are analyzing the tweets and using Machine Learning models and Natural Language Processing concepts to predict whether a tweet tweeted by the user is positive negative or neutral and analyze the other trends with highest accuracy. Our is a practical hands-on approach by using NLP and ML concepts.

6. Expected result

The main expectation from this project is to create a Machine Learning Classifier that can be used to predict the sentiment of a tweet to a very higher level of accuracy. We will test our dataset on several different Machine Learning Algorithms. As our model's task is to predict the sentiment of tweet, we will select the model with the highest accuracy score. We will not only check the success rate of different algorithms but also implemented an algorithm to our own understanding and incorporated an aspect that allows us to judge Sentiment in the tweets. In order to calculate the overall polarity of a post, we referred to the previous researches and added some value by introducing new metrics. This sentiment Analysis Model which we will create can be used for many purposes for e.g.: - Prediction of movie by the reviews, product reviews, to know about the opinion on certain topic by analyzing comments on social media, Video rating on YouTube etc.

7. References:

[1] E. Cambria, S. Poria, A. Gelbukh and M. Thelwall, "Sentiment Analysis Is a Big Suitcase," in *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74-80, November/December 2017, doi: 10.1109/MIS.2017.4531228.

[2] D. Cenni, P. Nesi, G. Pantaleo and I. Zaza, "Twitter vigilance: A multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis," 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, USA, 2017, pp. 1-8, doi: 10.1109/UIC-ATC.2017.8397589.

[3] maran, S.Muthuku & esh, P.Sur. (2017). Text Analysis for Product Reviews for Sentiment Analysis using NLP Methods. International Journal of Engineering Trends and Technology. 47. 474-480. 10.14445/22315381/IJETT-V47P278.

[4] Detecting Emotion in Human Speech (2011), Alex Mordkovich , Kelly Veit , Daniel Zilber ,stanford university

[5] SENTIMENT' BASED MODEL FOR REPUTATION SYSTEMS N AMAZON (2011), Milad Sharif ,Soheil Norouzi, stanford university

[6] Victoria Ikoro, Maria Sharmina, Khaleel Malik, and Riza Batista-Navarro : Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers. 2018

[7] A Machine Learning based Framework for Sentiment Classification: Indian Railways Case Study (IJITEE ISSN: 2278- 3075, Volume-8 Issue-4, February 2019)

[8] A Survey: Sentiment Analysis Using Machine Learning Techniques for Social Media Analytics (IJPAM InternationalJournal of Pure and Applied Mathematics)