Twitter Sentiment Analysis

Submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Engineering (SEM VI)

Mini Project Synopsis Report - Stage-I

by

Prathamesh Chikane

Roll No. 05

Subodh Halpatrao

Roll No. 15

Sutej Kulkarni

Roll No. 28

Under the Supervision of

Prof. J.P.Patil

DEPARTMENT OF INFORMATION TECHNOLOGY

KONKAN GYANPEETH COLLEGE OF ENGINEERING

KARJAT-410201

April 2020

Certificate

This is to certify that the project entitled **Twitter Sentiment Analysis** is a bonafide work of Prathamesh Chikane **(Roll No. 05), Subodh Halpatrao (Roll No. 15), Sutej Kulkarni (Roll No. 28)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Undergraduate** in **DEPARTMENT OF INFORMATION TECHNOLOGY.**

Supervisor/Guide

Prof. J.P.Patil

Department of Information Technology

Head of Department
Prof. J.P.Patil
Department of Information Technology

Principal Dr.
M.J. Lengare
Konkan Gyanpeeth College of Engineering

i

Project Report Approval

This thesis / dissertation/mini project report entitled **Twitter Sentiment Analysis**
by **Prathamesh Chikane (Roll No. 05), Subodh Halpatrao (Roll No. 15), Sutej Kulkarni
(Roll No. 28)** is approved for the degree of **DEPARTMENT OF INFORMATION TECHNOLOGY.**

Examiner.

……………………………

………………………………

DATE.

PLACE.

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature

Prathamesh Chikane (Roll No. 05)

Signature

Subodh Halpatrao (Roll No. 15)

Signature

Sutej Kulkarni (Roll No. 28)

Date.

Abstract

Nowadays, people from all around the world use social media sites to share information. Twitter for example is a platform in which users send, read posts known as `tweets' and interact with different communities. Users share their daily lives, post their opinions on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them. The aim of this paper is to present a model that can perform sentiment analysis of real data collected from Twitter. Data in Twitter is highly unstructured which makes it difficult to analyze. However, our proposed model is different from prior work in this field because it combined the use of supervised and unsupervised machine learning algorithms. The process of performing sentiment analysis as follows: Tweet extracted directly from Twitter API, then cleaning and discovery of data performed. After that the data were fed into several models for the purpose of training. Each tweet extracted classified based on its sentiment whether it is a positive, negative or neutral.

Acknowledgements

Contents

List of Figures

List of Tables

Abbreviations

NLP:- **Natural Language Processing**

ML:- **Machine Learning**

NN:- **Neural Network**

RNN:- Recurrent Neural Network

Chapter 1

INTRODUCTION

1.1 Introduction

Sentiment Analysis is a technique widely used in text mining. Twitter Sentiment Analysis, therefore means, using advanced text mining techniques to analyze the sentiment of the text in the form of positive, negative and neutral. It is also known as Opinion Mining, is primarily for analyzing conversations, opinions, and sharing of views (all in the form of tweets) for deciding business strategy, political analysis, and also for assessing public actions.

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract, quantity and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction or event. The attitude may be judgement or evaluation,

affective state (emotional state of the author or speaker), or the intended emotional communication. As a part of Natural Language Processing, algorithms like Naive Bayes and Neural Networks like RNN are used in predicting the polarity of the sentence. sentiment analysis of Twitter data may also depend upon sentence level and document level.

1.2   Objectives

In this section we mention objectives of our project:

•   To build a system/software using machine learning techniques that analysis the sentiments of users on the twitter platform.

•   The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

1.3   Motivation

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Morever the response on twitter is more prompt and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis). Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analysing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm's stock market value.

1.4   Purpose, Scope, and Applicability

1.4.1   Purpose

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral.

Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences.

1.4.2   Scope

Business:

Companies use Twitter Sentiment Analysis to develop their business strategies, to assess customer's feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.

Politics:

In politics Sentiment Analysis Dataset Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset Twitter is also used for analyzing election results.

Public Actions:

Twitter Sentiment Analysis also is used for monitoring and analyzing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.

1.4.3   Applicability

Sentiment Analysis in business, also known as opinion mining is a process of identifying and cataloging a piece of text according to the tone conveyed by it. This text can be tweets, comments, feedback, and even random rants with positive, negative and neutral sentiments associated with them. Every business needs to implement automated sentiment analysis. If you doubt it, here's a little perspective. The accuracy can never be 100%. And of course, a machine does not understand sarcasm. However, according to a research, people do not agree 80% of the time. It means that even if the machine accuracy does not score a perfect 10, it will still be more accurate than human analysis. Also, when the corpus is huge manually analyzing is not an option. Hence, sentiment analysis in business is more than just a trend.

The applications of sentiment analysis in business cannot be overlooked. Sentiment analysis in business can prove a major breakthrough for the complete brand revitalization. The key to running a successful business with the sentiments data is the ability to exploit the unstructured data for actionable insights. Machine learning models, which largely depend on the manually created features before classification, have served this purpose fine for
the past few years.

At ParallelDots, we have powerful sentiment analysis API that uses deep learning which provides an accurate analysis of the overall sentiment of the given text.

Chapter 2

LITERATURE SURVEY AND

PAPER REVIEW

Literature Survey: Is the process of analyzing, summarizing, organizing, and presenting novel conclusions from the results of technical review of large number of recently published scholarly articles. In this chapter we survey previous research done on automatic image annotation, we have studied about following papers published by some experts

2.1   Literature Survey

1. Twitter Sentimental Analysis

Author: Aliza Sarlan, Chayanit Nadam, Shuib Basri 2016

In this paper, author suggested an method of analyzing tweets using Natural Language Processing (NLP) for classifying tweets as positive, negative or neutral. The central idea of NLP techniques is based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules. Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task , it has been handled at the sentence level and more recently at the phrase level. NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.

2. A Study of Sentiment Analysis of Twitter Data of tweets

Author: Abdullah Alsaeedi, Mohamad Zubair Khan

This paper briefly describes commonly used sentimental analysis methods including machine learning based approach and global and local feature extraction. This paper is organized as follows. Section 2 briefly describes commonly used twitter sentimental analysis methods including global and local feature extraction. Section 3 covers the low-level features which are generally extracted from tweets. Section 4 overviews the mostly used supervised learning methods for the classification or analyzing the sentences or tweets

These methods often rely on supervised classification approaches where sentiment detection is framed as a binary which are positive and negative. This approach requires labeled data to train classifiers. This approach, it becomes apparent that aspects of the local context of a word need to be taken into account such as negative (e.g. Not beautiful) and intensification (e.g. Very beautiful)

3.REAL TIME SENTIMENT ANALYSIS OF TWITTER POSTS

Author:Prakruthi V, Sindhu D, Dr S Anupama Kumar 2018

In paper the author discusses about the systems are designed to retrieve information using twitter data and then classify them based on the semantics of knowledge contained. Authors Lokmanyathilak Govindan Sankar Selvan and Teng-Sheng Moh have developed the framework in paper which makes use of real-time Twitter data stream, that are cleaned and analyzed and then fast feedback is acquired through opinion mining. Paper deals with the opinion extracted and collected from the popular social media platform named Twitter. For Comparison of market status of two enterprises they copy the two dictionary files one for positive and other for negative words from the repository in the backend as those files are used for analyzing and scoring terms from tweets. The paper mainly conveys about the sentimental analysis of tweets using R language which is helpful for collecting the sentimental information in the form of either positive score, negative score or someplace in between them. Then they execute the analysis of tweets that are in size of TBs which means big data using R language and Rhadoop Connector. In paper author has considered a way of advancing the present sarcasm detection algorithms by including improved pre-processing and text mining techniques like emoji and slang detection. To analyze bulky live data streams, three considered key features are density distribution, negativity and influence. To do this in paper authors Ming Hao, Christian Rohrdantz and others have considered Pixel cell-based sentiment

calendar where every opinion is represented in form of cell. The cell color is the sentiment value, i.e, green for positive, gray for natural and red for negative. In paper Web-based tool named SWAB (Social Web 30 Analysis Buddy) which is integrating qualitative analysis and large-scale data mining techniques together is proposed here. Prototype of this tool is demonstrated by analyzing matter posted by student on Twitter is demonstrated. In paper authors Sonia Anastasia and Indra Budi have used the Sentiment analysis tools: R and Rapidminer. Data is collected by crawling through Twitter API with R. Then, Rapidminer is used in data pre-processing, classification, for classifiers performance evaluation and Net Sentiment Score is calculated which associates with customer satisfaction using classification results. In paper tweets are collected and all unimportant words are deleted from the tweet collection to support the classification process, then tweets are filtered by means of a Bayes Naive classifier, which was earlier trained and whose intention is to select messages that represent news about fresh cyber-attacks and malware. In this paper exemplary malware activities presently circulating on the network have been identified.

4.REAL TIME TWITTER SENTIMENT ANALYSIS USING 3-WAY CLASSIFIER

Author: Alaa S, Al Shammari 2018

In, authors used twitter to collect a corpus to preform linguistic analysis of the corpus and building a sentiment classifier using nave Bayes to determine the polarity of the documents. The experimental evaluation shows improvement of the proposed techniques than the previous methods. In, authors collect 3 different datasets (HASH, EMOT and iSieve) and do a pre-processing of it. Then, evaluating the training data which found that some features may not be good to be combined with other features. While, proposed a unigram model as baseline which gain 4 % at both binary classification and 3-way classification. Then, they investigated of both tree kernel model and feature based models which exceeded the unigram model. The combination of parts-of-speech and prior polarity features outperforms other features based on features analysis. In, authors focused at political sentiment analysis problem at real time using nave Bayes to classify tweets under 4 categories (positive, negative, neutral and unsure) where the system applicable to analyze tweet at election events. The method that used is generic which can be used in other domains such as movies events. While, studied the sentiment analysis problem for electronic products domain where tweet posts related to that field was analyzed. The machine learning techniques outperforms symbolic techniques based on sentiments identification. The evaluation of enhanced vector feature was tested using several classifiers such as nave Bayes, Maximum Entropy, SVM and Ensemble

classifiers where the accuracy of these classifiers was almost similar. The performance of the proposed feature vector improved at the electronic products opinion domain. Another insignificant effort has been directed under the development of sentiment analysis systems to support the research community. Microsoft azure developed a real-time twitter sentiment analysis demo that helps to represent the polarity of tweets of Bing maps where it is expressed as opinion as positive, negative or neutral.

There are several classifiers used for opinion mining and sentiment analysis of microblogging platforms. I used Simple Voter algorithm and Naïve Bayes algorithm which classify tweets as positive, negative, or neutral opinion. The positive and negative dictionaries have been downloaded from the internet with total of 2014 positive words and 4783 negative words. In the proposed system, it works based on sentence level. So, each tweet decomposed into number of separated words. At this level, we score each tweet based on this equation:

Score = Number of positive words - Number of negative words

If Score > 0, then the sentence is positive.

If Score < 0, then the sentence is negative.

If Score = 0, then the sentence is neutral.

As shown above, each tweet will be tokenized into separated words where each word will be compared based on matching of positive dictionary term or even negative dictionary term. Then, the score will be assigned for each tweet based on the probability using Naïve Bayes algorithm or the high majority using Simple Voter algorithm. Therefore, the tweets will be classified into 3 categories (positive, negative, or neutral) based on the score.

5. TWITTER SENTIMENTAL ANALYSIS WITH DEEP CONVOLUTION NEURAL NETWORK

Author: Aliaksei Severyn, Alessandro Moschitti.

This paper describes our deep learning system for sentiment analysis of tweets. The main contribution of this work is a new model for initializing the parameter weights of the convolutional neural network, which is crucial to train an accurate model while avoiding the need to inject any additional features.

In this paper the author has described deep convolutional neural network for sentiment analysis of tweets. Its architecture is most similar to the deep learning systems presented in that have recently established new state-of-the-art results on various NLP sentence classification tasks also including sentiment analysis. Convolutional neural networks have been also successfully applied in various IR

applications, e.g. While already demonstrating excellent results, training a convolutional neural network that would beat hand-engineered approaches that also rely on multiple manual and automatically constructed lexicons, e.g. requires careful attention. This becomes an even harder problem especially in cases when the amount of labelled data is relatively small, e.g., thousands of examples. It turns out that providing the network with good initialisation parameters can have a significant impact on the accuracy of the trained model

The architecture of the convolutional neural network for sentiment classification in this paper is mainly inspired by the architectures used in for performing various sentence classification tasks. Given that the training process of this model requires to run the network on a rather large corpus, design choices that the authors made are mainly driven by the computational efficiency of our network. Hence, different from which presents an architecture with several layers of convolutional feature maps, we adopt a single level architecture. Nevertheless, single-layer architectures have been shown in to perform equally well. The network is composed of a single convolutional layer followed by a non-linearity, max pooling and a soft-max classification layer

2.2 Paper Comparison

Sr. No.	Paper Title	Author (Pub. Year)	Description
1	Twitter Sentiment Analysis	Aliza Sarlan, Chayanit Nadam, Shuib Basri (2016)	In this paper, author suggested an NLP based technique for classification of the sentences from the dataset also known as tweets used In Twitter Natural Language Processing (NLP) for classifying tweets as positive, negative or neutral
2	A Study of Sentiment Analysis of Twitter Data of tweets	Abdullah Alsaeedi, Mohamad Zubair Khan	This paper briefly describes commonly used sentimental analysis methods including machine learning based approach and global and local feature extraction. This paper is organized as follows. Section 2 briefly describes commonly used twitter sentimental analysis methods including global and local feature extraction. Section 3 covers the low-level features which are generally extracted from tweets. Section 4 overviews the mostly used supervised learning methods for the classification or analyzing the sentences or tweets

3	real time sentiment analysis of twitter posts	Prakruti V, Sindhu D, Dr. Anupama K (2018)	In paper the author discusses about the systems are designed to retrieve information using twitter data and then classify them based on the semantics of knowledge contained. The paper mainly conveys about the sentimental analysis of tweets using R language which is helpful for collecting the sentimental information in the form of either positive score, negative score or someplace in between them
4	Real time Twitter Sentiment Analysis using 3-way classifier/Lexicon based approach	Alaa S, Al Shammari (2018)	In this the author proposed system works based on sentence level. So, each tweet decomposed into number of separated words. At this level, we score each tweet based on this equation: Score = Number of positive words - Number of negative words If Score > 0, then the sentence is positive. If Score < 0, then the sentence is negative. If Score = 0, then the sentence is neutral.
5	Twitter Sentimental Analysis with neural networks	Aliaksei Severyn, Alessandro Moschitti.	This paper talks about the training process to achieve the sentiments of the sentences/tweets (positive, negative & neutral) using neural networks.

Figure 2.1: Comparison Table

Chapter 3

SURVEY OF METHODOLOGY

3.0.1 Naïve Bayes Classification

A classifier is a machine learning model that is used to discriminate different objects based on certain features. A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem, Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem
Types of Naive Bayes Classifier:

Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

Bernoulli Naive Bayes:

This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

Gaussian Naive Bayes:

When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

Natural Language Processing (NLP)

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.Most NLP techniques rely on machine learning to derive meaning from human languages.

NLP entails applying algorithms to identify and extract the natural language rules such that the unstructured language data is converted into a form that computers can understand. When the text has been provided, the computer will utilize algorithms to extract meaning associated with every sentence and collect the essential data from them. Sometimes, the computer may fail to understand the meaning of a sentence well, leading to obscure results.

3.0.3   Neural Networks

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process.

Surprisingly, one model that performs particularly well on sentiment analysis tasks is the convolutional neural network, which is more commonly used in computer vision models. The idea is that instead of performing convolutions on image pixels, the model can instead perform those convolutions in the embedded feature space of the words in a sentence. Since convolutions occur on adjacent words, the model can pick up on negations or n-grams that carry novel sentiment information.

RNNs are probably the most commonly used deep learning models for NLP and with good reason. Because these networks are recurrent, they are ideal for working with sequential data such as text. In sentiment analysis, they can be used to repeatedly predict the sentiment as each token in a piece of text is ingested. Once the model is fully trained, the sentiment prediction is just the model's output after seeing all n tokens in a sentence. RNNs can also be greatly improved by the incorporation of an attention mechanism, which is a separately trained component of the model. Attention helps a model to determine on which tokens in a sequence of text to apply its focus, thus allowing the model to consolidate more information over more timesteps.

Chapter 4

REQUIREMENTS AND

ANALYSIS

4.1   Problem Definition

Twitter Sentiment Analysis is the automated process of analyzing text data and sorting it into sentiments positive, negative or neutral. Performing Sentiment Analysis on data from Twitter using machine learning can help companies understand how people are talking about their brand. With more than 321 million active users, sending a daily average of 500 million Tweets, Twitter allows businesses to reach a broad audience and connect with customers without intermediaries. Monitoring Twitter allows companies to understand their audience, keep on top of what's being said about their brand and their competitors, and discover new trends in the industry. Are users talking positively or negatively about a product? Well, that's exactly what sentiment analysis determines.

4.2   Requirements Specification

For implementation, in software we will require the following software and hardware specifications:
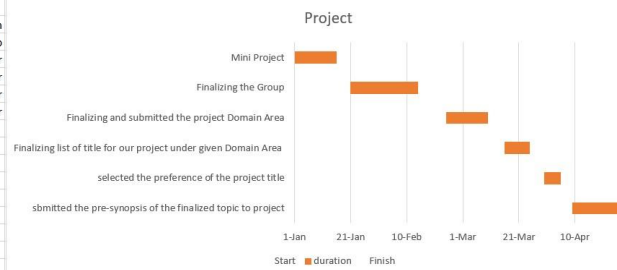
- **Software Specification**

For implementation we will require applications such as MATlab, Jupiter Notebook,Python IDE,Django,Flask etc. That can be used to build an environment and to train machine based on it.

- **Hardware Specification**

To implement the project we will require a computer with specification such as multicore CPU, graphics card, hard disk upto 500GB, upto 8GB RAM. Input devices such as keyboard, optical mouse.

4.3   Planning and Scheduling

Task Name	Start	duration	Finish
Mini Project	1-Jan	15	17-Jan
Finalizing the Group	21-Jan	24	14-Feb
Finalizing and submitted the project Domain Area	24-Feb	15	13-Mar
Finalizing list of title for our project under given Domain Area	16-Mar	9	27-Mar
selected the preference of the project title	30-Mar	6	6-Apr
sbmitted the pre-synopsis of the finalized topic to project	9-Apr	16	27-Apr

Project

Chapter 5

CONCLUSIONS

5.1   Conclusion

After researching through various papers related to Twitter Sentiment Analysis, We have concluded that, a system can be developed that can analyze twitter sentiments. Various experiments had being conducted using different methodologies, the best results are seen in the methods that are based on neural network ,Naïve Bayes. Hence looking at the results, we have decided to take the same approach for developing our system. We will test our system against twitter datasets and compare our results.

Bibliography

Journals and Conferences:

[1] M. Rambocas , and J. Gama, "Marketing Research :The Role of Sentiment Analysis". The 5th SNA-KDD Workshop'11. University of Porto, 2013

[2] A. K. Jose, N. Bhatia, and S. Krishna, "Twitter Sentiment Analysis". National Institute of Technology Calicut,2010.

[3] P. Lai, "Extracting Strong Sentiment Trendfrom Twitter". Stanford University, 2012.

[4] M. Comesaña, A. P. Soares , M. Perea, A.P. Piñeiro, I. Fraga, and A. Pinheiro, " Author ' s personal copy Computers in Human Behavior ERP correlates of masked affective priming with emoticons," Computers in Human Behavior, 29, 588–595, 2013. [6] A.H.Huang, D.C. Yen, & X. Zhang.

[5] A. Pak, and P. Paroubek , "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Special Issue of International Journal of Computer Application, France:Universitede Paris-Sud, 2010.

[6] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R.Passonneau, "Sentiment Analysis of   Twitter Data," Annual International Conferences. New York:Columbia University, 2012.

[7]H. Saif, Y. He and H. Alani, "Alleviating Data Scarcity for Twitter Sentiment Analysis". Association for Computational Linguistics, 2012.

[8] R. Prabowo, and M. Thelwall, "Sentiment Analysis:A Combined Approach," International World Wide Web Conference Committee (IW3C2), 2009. United Kingdom: University of Wolverhamption.

[9] H. Saif, Y.He, and H. Alani, "Semantic Sentiment Analysis of Twitter," Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. United Kingdom: Knowledge Media Institute, 2011.

[10]J. Spencer and G. Uchyigit, "Sentiment or: Sentiment Analysis of Twitter Data," Second Joint Conference on Lexicon and Computational Semantics. Brighton:University of Brighton, 2008

26

Websites:

https://ieeexplore.ieee.org/Xplore/home.jsp

(Accessed on 25/01/2020)

https://en.wikipedia.org/

(Accessed on 04/02/2020)

https://sci-hub.tw/

(Accessed on 22/02/2020)

https://skymind.ai/wiki/neural-network

(Accessed on 08/03/2020)

https://www.sciencedirect.com/science/article/pii/S1389041718300482

(Accessed on 25/03/2020)