A Synopsis on

Sentiment Analysis Framework for Social Media

Submitted in partial fulfillment of the requirements
of the degree of

Bachelor of Engineering

in

Information Technology

by

Riddhi Prajapati (15104015)
Pratiksha patil (16204011)
Aafreen shaikh (16204009)

Name of Guide:Prof.Sunil A.Sushir

Department of Information Technology
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI
2018-2019

CERTIFICATE

This is to certify that the project Synopsis entitled ***"Sentiment Analysis Framework for Social Media"*** Submitted by ***"Riddhi Prajapati (15104015)"*** **"Pratiksha Patil (16204011)" "Aafreen Shaikh (16204009)"** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Information Technology.***to the University of Mumbai,is a bonafide work carried out during academic year 2018-2019

(Prof.Sunil A.Sushir)
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane

Date:

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)

_____

Riddhi Prajapati (15104015)
Pratiksha patil (16204011)
Aafreen shaikh (16204009)

Date:

Abstract

In today's world, Social Media is one such platform where people openly express their emotions, reviews, feedback or even personal experiences publicly.This feedback is regardingproducts, posts, services which they find proper or not. This is quiet useful as it lets everyone know about the quality and other measures which is good enough. Currently Microblogging websites like twitter, facebook are filled with data and opinions. Twitter is one amongst most frquently used microblogging website where users tweet about a particular topic. Hence an individual might require hours to figure out whether the users are giving a positive or neg-ative response for that particular topic. Sentiment Analysis is an automated process which identies and classies opinions or sentiments from a piece of text. Sentiments are of two categories Positive, Negative and Nuetral. Sentiment Analysis in Twitter is dicult as com-pared to the general ones due to presence of large number of misspellings, hashtags, etc.Various Machine Learning algorithms are adopted for classification of data set into Positive or Negative classes based on their sentiments such as Naive Bayes, Decision Tree, etc. This paper contains implementation of Naive Bayes, Random Forest and Logistic Regression.The output of which are displayed in terms of Wordcloud and confusion matrix. The main emphasis of this research is on the classification of emotions of tweets' data gathered from Kaggle of For actual implementation of this system Python with Anaconda Navigator is been used.

Introduction

Currently, social networking sites are at an all-time high, so from there, a large amount of data is generated. Social Networking websites and micro blogging websites in todays world has become the biggest web destinations for people to communicate with each other, to express their thoughts about products or movies, share their daily experience and communicate their opinion about real time and upcoming events, such as sports or political elections etc. To achieve a large, diverse data set of current public opinions or sentiments, Twitter could be used as a valuable resource which allows users to send and read small messages called Tweets. Twitter is a reservoir for a large amount of data. So this data is extremely useful for predicting results of political activities, new initiatives led by government, or research and deciding on what content to share with the audience. Input to our model is the raw data extracted from tweets. For the same, we automate the process of tweet extraction and categorizing it into two categories i.e. positive or negative. The content in twitter generated by the user is with the advent of social media over the last decade. The efforts to determine peoples attitudes with respect to a specific topic or event have garnered a wide research interest in natural language processing and introduced Sentiment Analysis.

Sentiment Analysis is systematic method of gaining knowledge from opinions or emotions. It's application has shown significance in business and marketing field. As social media gained its importance in recent days, Sentiment analysis turned out to be one of the best era for research. The sentiment analysis of customer's social media data is very important in the present day business scenarios. Customers share their reviews and their comfort towards the products on social media. This information can be used for various application such as market

research, product feedback and analysing customer service effectiveness. The analysis of the sentiment could lead to many interesting results.

Objectives

The objectives of this project are as follows :

- To Preprocess the data set and classification of the same into positive and negative.
- Sentiment analysis to determine the attitude of the mass. Whether it is positive or negative towards its subject of interest.
- Graphical representation of the sentiments in form of Bar graph and Word cloud.
- To implement algorithms for classification of data set and display the same using Confusion matrix
- To determine the algorithm with highest accuracy on data set.

Literature Review

Ms. Farha Nausheen and Ms. Sayyada Hajera Begum in their paper "Sentiment Analysis to Predict Election Results Using Python" proposed lexicon based sentiment analyzer which classifies the tweets based on its sentiment value. They have considered the tweets from US presidential elections 2016. They have performed classification based on polarity and subjectivity measures. The proposed system retrieves tweets and performs political sentiment analysis. Tools like Twython, NLTK, TextBlob are employed. They present comparison between the top candidates for presidential elections 2016.

Anukur Goel, Jyoti Gautam and Sitesh Kumar in their paper "Real Time Sentiment Analysis of Tweets Using Naive Bayes" proposed the use of SentiWordNet2 for Opinion Mining providing sentiment score to each synset of WordNet. For implementation of the same they have used python with NLTK along with python-twitter APIs. The scores are of three types: positivity score, negativity score and objectivity score. So instead of using Naive Bayes in traditional way, they have used scores so that more accuracy can be attained. Also they have implemented Naive Bayes using sentiment140 training data using twitter database.

Ms.K.Saranya and Dr.S.Jayanthy in their paper "Onto-based sentiment classification using Machine Learning Techniques" proposed the use of semantics and ontology for text classification combining the same with machine learning techniques for getting better results. They employed SVM (Support Vector Machine), NB (Naive Bayes), kNN (k nearest neighbors) classifiers. They identified affective class hierarchy in WordNet with extracting it. Later assigning emotions to semantic roles and creating an emotion hierarchy.

Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta in their paper "Sentiment Analysis of Tweets Using Machine Learning Approach" have proposed Support Vector Machine, Adaboosted Decision Tree and Decision Tree based hybrid sentiment classification model for improving the overall accuracy of the classifier in the classification of tweets. For analytical evaluation of the proposed classifier they have employed accuracy and f-measure. They have provided comparative results which prove that hybrid model improved the overall classification accuracy and f-measure of sentiment prediction as compared to traditional existing techniques for classification. Their proposed approach classified the tweets as Positive and Negative.

Problem Definition

The problem in Sentiment analysis is classifying the polarity or accuracy of given text at the document, sentence, or feature/aspect level. Whether the expressed opinion is a document, a sentence or an entity feature/aspect is positive, negative or neutral.

Proposed System Architecture/Working

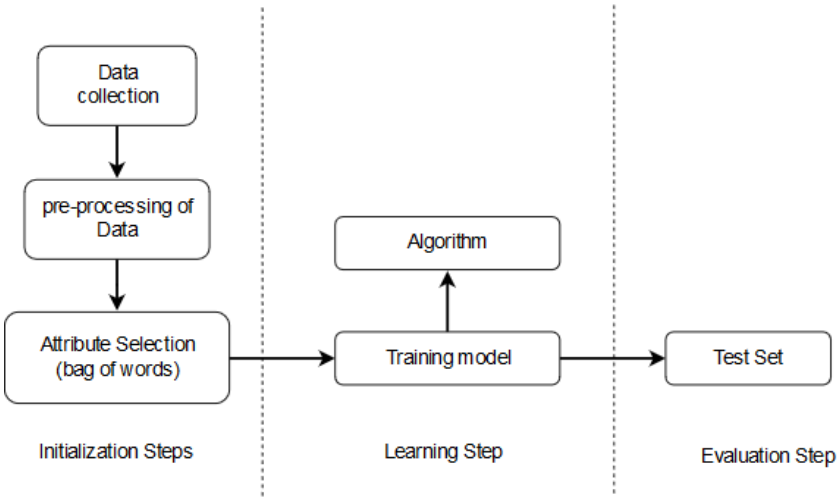

Figure 1: Workflow of Sentiment Analysis

In the very first stage of the proposed model, all the essential python libraries have to be imported for data set manipulation, mathematical functions, visualization, etc. Also libraries for sentiment like Natural Language Toolkit, Word cloud, etc are imported. The libraries for different models also needs to be imported.Then data set is saved into new dataframe and displayed. There are tweets which are labelled 0 are considered as positive and tweets with label 1 are considered as negative randomly.

Conclusion

We have completed our project using python as language. We have used Html and Css for output representation. Although there was a problem with merging of python with html, through number of tutorial we were able to merge it. In this paper, Naive Bayes, Random Forest and Logistic Regression based sentiment are represented for improving the overall accuracy of classifiers in the classification of tweets. We were able to determine the Positivity and Negativity of each tweet. Hence for this we apply pre-processing techniques so that proper classification is performed based on unbiased user tweets. Those were represented in the form of diagrams like Bar graph, Word cloud. Although many classifiers, data is fed as an input to the training data set, and the process we have proposed classifies the tweets into positive and negative. We have took comparative observations against the Naive Bayes, Random Forest and Logistic Regression. Here, Logistic Regression when trained 32000 tweets of data set and tested, the system shows an accuracy of 96.19 percent. Also for each algorithm, we displayed an Confusion Matrix.

References

[1] Abror Abduvaliyev, Al-Sakib Khan Pathan, Jianying Zhou, Rodrigo Roman and Wai-Choong Wong ,"On the vital Areas of Intrusion Detection Systems in Wireless Sensor networks",IEEE Communications Surveys & Tutorials, Accepted For Publications, 2013-in press.

[2] H.H. Soliman, et al,"A comparative performance evaluation of intrusion detection techniques for hierarchical wireless sensor networks", Egyptian Informatics Journal (2012) 13, 225238.

[3] Ms. Farha Nausheen,Ms. Sayyada Hajera Begum,"Sentiment Analysis to Predict Election Results Using Python",2018

[4] Ankur Goel,Jyoti Gautam,Sitesh Kumar,"Real Time Sentiment Analysis of Tweets Using Naive Bayes",2016

[5] Ms.k.Saranya,Dr.s.Jayanathy,"Onto based Sentiment Classification using Machine learning",2017

[6] Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta,"Sentiment Analysis of Tweets using Machine Learning Approach",2018

1 Publication