Analyzing the Pulse of Twitter: Sentiment Analysis using Natural Language Processing Techniques

Author Info
Venkata Ravi Kiran Kolla
Sr. Software Engineer
USA

Abstract—Social Media have gained more attention nowadays. Public and confidential opinions about a huge variety of matters are expressed via countless social media. Twitter is one of the most popular social media that is growing exponentially. Twitter gives organizations a fast and efficient way to analyze customers' opinions toward critical success in the market. Constructing a program for sentiment analysis is an approach to be used to computationally estimate customers' opinions. This paper describes the design of sentiment analysis, extracting a large size of tweets. Prototyping has been used in this design. Results classify customers' opinions with the help of tweets into positive and negative, represented in the form of a pie chart and html page. However, the program has planned to construct on a web application system, but due to the restraint of Django which can be worked on a Linux server or LAMP, for further, this approach need to be done.

INTRODUCTION

Many people are using social network sites to express their sentiments, and opinion and disclose their daily lives. However, people write anything such as social activities or any comments on products. Through the online communities gives an interactive forum where clients brief and influence others. Moreover, social media comes up with a chance for business that giving a platform to build relations with their clients such as social media to advertise or speak directly to client for connecting with their perspective of products and services.

In contrast, clients have all the power regarding what they want to see and how they respond. With this, the company's success & failure is publicly shared and comes up with word of mouth. However, the social network can change the behavior and decision making of consumers, for example, mentions that 87% of internet customers have an effect on their purchase and decision by customer review. So that, if company can catch up quickly on what their customer's think, it would be more beneficial to organize to react on time and come up with a good strategy to compete their competitors.

A. Problem Statement

In spite of the availability of software to extract data regarding an individual's sentiment on a specific product companies and other data workers still face issues related to the data extraction.

- Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the fast development of the Web in the entire world, people are using social media like Twitter that generates big amount of opinion texts in the form of tweets which is available for the sentiment analysis. This translates to a huge volume of information from a human point of view which make it hard to extract a sentence, go through them, analyze them, summarize them and organize them into an understandable format.

- Difficulties of Sentiment Analysis with unsuitable English

Informal language defines as the use of colloquialisms and slang words in the communication, using the conventions of the language like 'would not' or 'wouldn't'. Not all systems are able to check sentiment with the use of informal words and this could destroy the analysis and decision- making process.

Emoticons, are a pictorial form of human facial expressions, which in the absence of body language used to draw a receiver's attention to the temper of a sender's verbal communication, improving and changing its understanding. For instance, Represents a happy state of mind. Systems do not have sufficient data to allow them to find feelings out of the emoticons. As humans frequently turn to emoticons to correctly express what they cannot say with words. Not being able to analyze this takes the company to a loss. Short-form is often

used even in case of short message service (SMS). The use of short-form will be used more often on Twitter so as to help to minimize the characters used. This is because Twitter has put a limit to 1 4 0 in case of characters. F o r e x a m p l e, 'Tba' refers to be announced**.**

B. Objective

The purpose of the study are first, to study the sentiment analysis which in view to analyze opinions of a customer of a company's product; and second, is to create a program for client' review on a product that allows a company or a client to sentiment and analyzes a huge amount of tweets into a useful format.

METHODOLOGY

This project has been divided into two phases. First, literature study is conducted, then the system development. Literature study includes conducting studies on wide range of sentiment analysis techniques and mechanism that currently in used. In phase 2, application needs and functionalities are defined prior to its development. Also, architecture and interface design and how the program will interact are also identified. In developing the application, several tools were utilized, such as Python Shell 2.7.2 and Notepad.

LITERATURE REVIEW

Opining Mining

Opinion mining can be defined as the broad area of natural language processing or text mining that involves the computational study of sentiments, reviews and emotions expressed in text. Although, view or attitude based on emotion rather than reason is often colloquially referred to as a sentiment. Hence, going towards an equivalent for opinion mining or sentiment analysis.

As said that opinion mining has various application niches including accounting, law, entertainment, technology, politics etc. In older days many social media have provided web users avenue for opening up to share their thoughts and opinions.

Twitter

It is a popular real time service that allows user to express short information defined as tweets that are limited to 140 characters. Users write tweets to share their opinion about various aspects regarding their daily lives. Twitter is an ideal platform for the extraction of opinion of public on specific issues. A collection of them is used as the first corpus for sentiment analysis refers the importance of opinion mining or natural language processing.

Twitter, with 500 million users and million messages per day, has become a valuable asset for companies to keep an eye on their reputation and brands by extracting and analyzing the sentiment of the tweets by the public about their products, services market and even about competitors. Provided that, from the social media generated opinions growth of the web, large quantity of opinion texts such as tweets, reviews, opinions or any discussion groups and forums are available for analysis, thus making the world wide web the quick, most promising and easily achievable medium for sentiment analysis.

Microblogging with E-commerce

A microblogging platform like Twitter is similar to a conventional blogging platform the only difference is single posts are shorter. Twitter has put a limit for a small number of words for the quick transmission of information or a quick exchange of thoughts. However, small or large companies are starting to the potential of microblogging

as an e-commerce marketing tool. Though, microblogging platform is developed in a short time for promoting foreign trade website by the use of a foreign microblogging platform.

The moment of sharing, interactive, community-oriented features 87uhjn are opening an e-commerce, launched a new bright spot which it can be shown that microblogging platform has enabled companies do brand image, product important sales channel, improve product sales, talk to consumer for a good interaction and other business activities involved.

said, in fact, the companies creating such products have started to poll them to get a knowledge of sentiment for a product. Many times, these companies consider user reactions and reply to users.

Social Media

It is a group of applications that have been created on the ideological and technological foundations of Web and that is allowed to construct and exchange of user generated contents. In the important discussion of Internet World Start, we determined that a trend of internet users is growing and continuing to expand more time with social media by the total time they spend on mobile devices and social media. On the contrary, businesses are using social networking sites to identify and communicate with customers, business can be visualized as damage to productivity caused by social network. As social media is easy to post to the public, it can harm confidential information to spread out.

On the other hand, discussed that the advantages of engaging in social media have gone beyond simply sharing to build company's honor and bring in career growth opportunities and income. Moreover, mentioned that the social media is also have a good application in advertisement by organizations for growth, professionals for recruiting, learning online and electronic commerce. Electronic commerce or E-commerce refers to the purchase and sale of goods or services online which can via social media, such has Twitter which is convenient due to its 24-hours availability, ease of customer service and global reach.

Reasons of why business tends to use more social media is to get insight into client behavior, market and present a chance to learn about customer opinions and reviews.

Twitter Sentiment Analysis

The sentiment can be identified in the comments or tweet to get useful indicators for many different purposes. Also, and stated that a review can be classified into two groups, which is negative and positive words.

Lexicon-based Approach

Lexicon-based methods uses predefined list of words where every word is associated with a particular sentiment. The lexicon methods vary in accordance to the context in which they are constructed and include calculation for a document from the semantic orientation of phrases in the documents. Moreover, It states that a lexicon sentiment is used to detect word-carrying opinion and then to predict opinion expressed in the phrase. It has been shown that the lexicon methods that have a basic paradigm which are:

i. Preprocess tweets, post by removing punctuation
ii. Initialize a total polarity score equal 0
iii. Check Whether token is present in a dictionary, If token is positive, s will be positive (+)

Else, s will be negative (-)

However, It is mentioned that one advantage of leaning-based method, is that it has the capacity to grasp and construct trained models for particular purposes. On the contrary, an accessibility of labeled data and hence the low pertinence of the mechanism of new data which is the reason of classifying data might be expensive or even prohibitive.

2. Machine-learning Approach

Machine learning methods rely on supervised classification approaches when sentiment detection is considered as a binary which are positive and negative. This approach takes labeled data to train classifiers. It becomes evident that features of the local context of a word are necessary to be taken into consideration such as negative and intensification.

Methods of Sentiment Analysis

The semantic Ideas of entities extracted from tweets can be used to measure the overall correlation of a group of entities with a given sentiment polarity. Polarity is the most vital form, which is if a text is positive or negative. However, sentiment analysis has methods in assigning polarity such as:

3. Natural Language Processing (NLP)

NLP mechanisms are relied on machine learning and more importantly statistical learning which takes a general learning algorithm along with a huge quantity of sample, of data to learn the rules. Sentiment analysis has been managed as a Natural Language Processing at various levels. Right from being a document level classification task, it is managed at the sentence level and recently at the phrase level as well. NLP is a field in computer science that includes building computers derive meaning from human language and input in an interactive way with the real world.

4. **Case-Based Reasoning (CBR)**

Case-Based Reasoning is the technique used to implement sentiment analysis. It is known by recalling the past successfully solved issues and use their solutions to solve the current deeply related issues. It is identified that some of the benefits of using it that it does not need an explicit niche model and so elicitation becomes a task of collecting care histories and these system scan learn by gaining new knowledge as cases.

5. **Artificial Neural Network (ANN)**

It is mentioned that Artificial Neural Network also known as neural network is a technique that interconnects group of artificial neurons. It will process information using the networks approach to computation

6. **Support Vector Machine (SVM)**

Support Vector Machine is used to find the sentiments of tweets. The training data should be collected from three different sentiment detection websites that mainly uses few pre-built sentiment lexicons to classify every tweet as positive or negative

APPENDICES (CODE)

```
# importing libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# cleaning the dataset (removing stopwords, stemming).
import re import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
corpus = []
for i in range(0, 1000):
 review = re.sub('[^a-zA-Z]', ' ', dataset['Review'][i])
 review = review.lower() review = review.split() ps = PorterStemmer()
 all_stopwords = stopwords.words('english') all_stopwords.remove('not')
 review = [ps.stem(word) for word in review if not word in set(all_stopwords)]
 review = ' '.join(review)
 corpus.append(review)

 # Creating the bag of words
from sklearn.feature_extraction.text import CountVectorizer
 cv = CountVectorizer(max_features = 1500)
 X = cv.fit_transform(corpus).toarray()
 y = dataset.iloc[:, -1].values

 # Splitting the dataset into the Training set and Test set
 from sklearn.model_selection import train_test_split
 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 0)
# Training the SVM
from sklearn.svm import SVC
classifier=SVC(kernel='linear', random_state = 0)
classifier.fit(X_train, y_train)
```

```
# Creating the Confusion Matrix of the classification model
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred)
print(cm)
accuracy_score(y_test, y_pred)

# for the visualization of positive sentiments on wordcloud we will store all the comments having polarity 1 in positive_comments
positive_comments=comments[comments['polarity']==1]
!pip install wordcloud
total_comments=''. join(positive_comments ['comment_text'])
wordcloud=WordCloud(width=1000,height=50 0,stopwords=stopwords).generate(total_comments)
plt.figure(figsize=(15,5)) plt.imshow(wordcloud) plt.axis('off')
```

RESULT

A. Twitter Retrieved

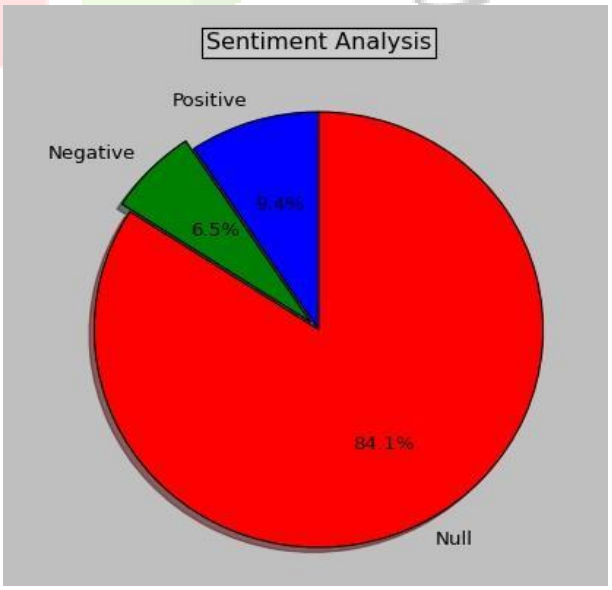
To link with Twitter API, developer required to agree in terms and conditions of development Twitter platform that has been given to get an authorization to use a data. The result from this process will be saved in form of a JSON file. The reason being, JSON (JavaScript Object Notation) is a lightweight data-interchange format that is very simple for humans to write and understand. Apart from that, JSON is easy for machines to create.

Nevertheless, the result will be classified into 2 categories, that are encoded and un-encoded. According to security issue for using a data, some of the results will be represented in an ID form like string ID.

Tweets from JSON file will be given the value of every word by matching with the lexicon dictionary. Due to limitation of words, lexicon dictionary cannot assign a value to every word, But as a particular language of python, which is able to analyze each tweet for getting a result.

B. Information Presented

The output will be shown in a pie chart which is showing a percentage of positive, negative and null sentiment hash tags



As shown in the Figure, the pie chart is representing of each percentage positive, negative and null sentiment hash tags in different color.

CONCLUSION & RECOMMENDATION

Twitter sentiment analysis is originated to analyze client's opinions toward the critical to success in the market. The program is utilizing a machine-learning approach which is more precise for analyzing a sentiment; along with natural language processing.

As a result, program will be classified sentiment into positive and negative, which is represented in the form of a pie chart and html page. However, the program has been lined up to be developed as a web application, due to constraints of Django which is only compatible with Linux server or LAMP. Hence, further progress of this element is recommended in future study.

REFERENCES

[1] M.Rambocas, and J. Gama, "MarketingResearch:TheRoleof SentimentAnalysis". The 5th SNA-KDD Workshop'11. Universityof Porto, 2013.

[2] A. K. Jose, N. Bhatia, and S. Krishna, "TwitterSentimentAnalysis". NationalInstituteof TechnologyCalicut,2010.

[3] P. Lai, "ExtractingStrongSentimentTrendfromTwitter". Stanford University, 2012.

[4] Y. Zhou, and Y. Fan, " A Sociolinguistic Study of American Slang," Theory and Practice in Language Studies, 3(12), 2209–2213, 2013. doi:10.4304/tpls.3.12.2209-2213

[5] 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, "Exploring the effects of emoticons," Information & Management, 45(7), 466–473, 2008.