Project Report

on

“Twitter Sentiment Analysis”

Submitted to the

AISSMS IOIT, Pune

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

SAMRUDDHI DEORE

2021-2022

Abstract

The entire world is transforming quickly under the present innovations. The Internet has become a basic requirement for everybody with the Web being utilized in every field. With the rapid increase in social network applications, people are using these platforms to voice them their opinions with regard to daily issues. Gathering and analyzing peoples' reactions toward buying a product, public services, and so on are vital. Sentiment analysis (or opinion mining) is a common dialogue preparing task that aims to discover the sentiments behind opinions in texts on varying subjects. In recent years, researchers in the field of sentiment analysis have been concerned with analyzing opinions on different topics such as movies, commercial products, and daily societal issues. Twitter is an enormously popular microblog on which clients may voice their opinions. Opinion investigation of Twitter data is a field that has been given much attention over the last decade and involves dissecting "tweets" (comments) and the content of these expressions. As such, this paper explores the various sentiment analysis applied to Twitter data and their outcomes.

Keywords—Twitter; sentiment; Web data; text mining; SVM; Bayesian algorithm; hybrid; ensembles

C) CONTENTS

CONTENTS

INTRODUCTION

Sentiment analysis is also known as “opinion mining” or “emotion Artificial Intelligence” and alludes to the utilization of natural language processing (NLP), text mining, computational linguistics, and bio measurements to methodically recognize, extricate, evaluate, and examine emotional states and subjective information. Sentiment analysis is generally concerned with the voice in client materials; for example, surveys and reviews on the Web and web-based social networks.

As a rule, sentiment analysis attempts to determine the disposition of a speaker, essayist, or other subjects in terms of theme via extreme emotional or passionate responses to an archive, communication, or occasion. The disposition might be a judgment or assessment, full of emotion (in other words, the passionate condition of the creator or speaker) or an expectation of enthusiastic responses (in other words, the impact intended by the creator or buyer). Vast numbers of client surveys or recommendations on all topics are available on the Web these days and audits may contain surveys on items such as on clients or fault-findings of films, and so on. Surveys are expanding rapidly, on the basis that individuals like to provide their views on the Web.

Large quantities of surveys are accessible for solitary items which make it problematic for clients as they must peruse each one in order to make a choice. Subsequently, mining this information, distinguishing client assessments and organizing them is a vital undertaking. Sentiment mining is a task that takes advantage of NLP and information extraction (IE) approaches to analyze an extensive number of archives in order to gather the sentiments of comments posed by different authors [1, 2]. This process incorporates various strategies, including computational etymology and information retrieval (IR) [2]. The basic idea of sentiment investigation is to detect the polarity of text documents or short sentences and classify them on this premise. Sentiment polarity is categorized as “positive”, “negative” or “impartial” (neutral). It is important to highlight the fact that sentiment mining can be performed on three levels as follows [3]:

· Document-level sentiment classification: At this level, a document can be classified entirely as “positive”, “negative”, or “neutral”.

· Sentence-level sentiment classification: At this level, each sentence is classified as “positive”, “negative” or unbiased.

· Aspect and feature level sentiment classification: At this level, sentences/documents can be categorized as “positive”, “negative” or “non-partisan” in light of certain aspects of sentences/archives and commonly known as “perspective-level assessment grouping”.

The main objective of this paper is to study the existing sentiment analysis methods of Twitter data and provide theoretical comparisons of the state-of-art approaches. The paper is organized as follows: the first two subsequent sections comment on the definitions, motivations, and classification techniques used in sentiment analysis. A number of documentlevel sentiment analysis approaches and sentence-level sentiment analysis approaches are also expressed. Various sentiment-analysis approaches used for Twitter are described including supervised, unsupervised, lexicon, and hybrid approached. Finally, discussions and comparisons of the latter are highlighted.

Sentiment analysis is a means of assessing written or spoken languages to decide whether articulation is positive, negative or neutral and to what degree. The current analysis tools in the market are able to deal with tremendous volumes of customer criticism reliably and precisely. In conjunction with contents investigation, sentiment analysis discovers customers' opinions on various topics, including the purchase of items, provision of services, or presentation of promotions. Immense quantities of client-created web-based social networking communications are being persistently delivered in the forms of surveys, online journals, comments, discourses, pictures, and recordings. These correspondences offer significant opportunities to obtain and comprehend the points of view of clients on themes such as intrigue and provide data equipped for clarifying and anticipating business and social news, such as product offers [4], stock returns [5], and the results of political decisions [6]. Integral to these examinations is the assessment of the notions communicated between clients in their content interchanges.

"Notion examination" is a dynamic area of research designed to enhance computerized understanding of feelings communicated in content, with increases in implementation prompting more powerful utilization of the inferred data. Among the different web-based social networking platforms, Twitter has incited particularly far-reaching client appropriation and rapid development in terms of correspondence volume.

Twitter is a small-scale blogging stage where clients generate 'tweets' that are communicated to their devotees or to another client. At 2016, Twitter has more than 313 million dynamic clients inside a given month, including 100 million clients daily [7]. Client origins are widespread, with 77% situated outside of the US, producing more than 500 million tweets every day [8]. The Twitter site positioned twelfth universally for activity in 2017 [9] and reacted to more than 15 billion API calls every day [10]. Twitter content likewise shows up in more than one million outsider sites [8]. In accordance with this enormous development, Twitter has of late been the subject of much scrutiny, as Tweets frequently express client's sentiment on controversial issues. In the social media context, sentiment analysis and mining opinions are highly challenging tasks, and this is due to the enormous information generated by humans and machines [11].

LITERATURE SURVEY

Opinions are fundamental to every single human action since they are key influencers of our practices. At whatever point we have to settle on a choice, we need to know others' thoughts. In reality, organizations and associations dependably need to discover users' popular sentiments about their items and services. Clients use different types of online platforms for social engagement including web-based social networking sites; for example, Facebook and Twitter. Through these webs based social networks, buyer engagement happens progressively. This kind of connection offers a remarkable open door for advertising knowledge. Individuals of every nationality, sexual orientation, race and class utilize the web to share encounters and impressions about virtually every feature of their lives. Other than composing messages, blogging or leaving remarks on corporate sites, a great many individuals utilize informal organization destinations to log opinions, express feelings and uncover insights about their everyday lives. Individuals compose correspondence on nearly anything, including films, brands, or social exercises.

These logs circulate throughout online groups and are virtual gatherings where shoppers illuminate and impact others. To the advertiser, these logs provide profound snippets of insight into purchasers' behavioral inclinations and present a continuous opportunity to find out about client emotions and recognitions, as they happen without interruption or incitement.

Be that as it may, recent explosions in client-produced content on social sites are introducing unique difficulties in capturing, examining and translating printed content since information is scattered, confused, and divided [12]. Opinion investigation is a method of information mining that can overcome these difficulties by methodically separating and dissecting web-based information without causing delays. With conclusion examination, advertisers are able to discover shoppers' emotions and states of mind continuously, in spite of the difficulties of information structure and volume. The enthusiasm in this study for utilizing sentiment analysis as an instrument for promoting research instrument is twofold.

Sentiment analysis critically encourages organizations to determine customers' likes and dislikes about products and company image. In addition, it plays a vital role in analyzing data of industries and organizations to aid them in making business decisions

REQUIREMENT SPECIFICATION AND ANALYSIS

- Import Necessary Dependencies
- Read and Load the Dataset
- Exploratory Data Analysis
- Data Preprocessing
- Splitting our data into Train and Test Subset
- Transforming Dataset using Count Vectorizer
- Model Building and Evaluation

DESIGN AND IMPLEMENTATION

```
import pandas as pd
import nltk
tweets=pd.read_csv('Tweets.csv')
tweets.head()
```

```
tweet_idairline_sentimentairline_sentiment_confidencenegativereasonnegativereason_confide
nceairlineairline_sentiment_goldnamenegativereason_goldretweet_counttexttweet_coordtwee
t_createdtweet_locationuser_timezone0570306133677760513neutral1.0000NaNNaNVirgin
AmericaNaNcairdinNaN0@VirginAmerica What @dhepburn said.NaN2015-02-24 11:35:52 -
0800NaNEastern Time (US &
Canada)1570301130888122368positive0.3486NaN0.0000Virgin
AmericaNaNjnardinoNaN0@VirginAmerica plus you've added commercials t...NaN2015-02-24
11:15:59 -0800NaNPacific Time (US &
Canada)2570301083672813571neutral0.6837NaNNaNVirgin
AmericaNaNyvonnalynnNaN0@VirginAmerica I didn't today... Must mean I n...NaN2015-02-24
11:15:48 -0800Lets PlayCentral Time (US &
Canada)3570301031407624196negative1.0000Bad Flight0.7033Virgin
AmericaNaNjnardinoNaN0@VirginAmerica it's really aggressive to blast...NaN2015-02-24
11:15:36 -0800NaNPacific Time (US & Canada)4570300817074462722negative1.0000Can't
```

Tell1.0000Virgin AmericaNaNjnardinoNaN0@VirginAmerica and it's a really big bad thing...NaN2015-02-24 11:14:45 -0800NaNPacific Time (US & Canada)

```
tweets.shape
```

```
(14640, 15)
```

# Data Pre Processing

```
tweets_df=tweets.drop(tweets[tweets['airline_sentiment_confidence']<0.5].index,axis=0)
tweets_df.shape
```

```
(14404, 15)
```

```
X=tweets_df['text']
y=tweets_df['airline_sentiment']
```

# Clean Your Text Data

```
from nltk.corpus import stopwords
nltk.download('stopwords')
import string
from nltk.stem import PorterStemmer


stop_words=stopwords.words('english')
punct=string.punctuation
stemmer=PorterStemmer()

import re
cleaned_data=[]
for i in range(len(X)):
  tweet=re.sub('[^a-zA-Z]',' ',X.iloc[i])
  tweet=tweet.lower().split()
  tweet=[stemmer.stem(word) for word in tweet if (word not in stop_words) and (word not in punct)]
  tweet=' '.join(tweet)
  cleaned_data.append(tweet)
```

```
cleaned_data[:10]
```

```
['virginamerica dhepburn said',
 'virginamerica today must mean need take anoth trip',
 'virginamerica realli aggress blast obnoxi entertain guest face amp littl recours',
 'virginamerica realli big bad thing',
 'virginamerica serious would pay flight seat play realli bad thing fli va',
 'virginamerica ye nearli everi time fli vx ear worm go away',
 'virginamerica realli miss prime opportun men without hat parodi http co mwpg grezp',
 'virginamerica well',
 'virginamerica amaz arriv hour earli good',
 'virginamerica know suicid second lead caus death among teen']
```

```
y
0          neutral
2          neutral
3         negative
4         negative
5         negative
            ...   
14634     negative
14636     negative
14637      neutral
14638     negative
14639      neutral
Name: airline_sentiment, Length: 14404, dtype: object
```

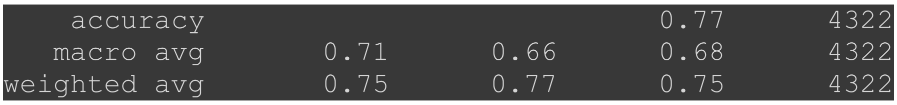
```
sentiment_ordering = ['negative', 'neutral', 'positive']

y = y.apply(lambda x: sentiment_ordering.index(x))

y.head()
0    1
2    1
3    0
4    0
5    0
Name: airline_sentiment, dtype: int64
```

```
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=3000,stop_words=['virginamerica','unit'])
X_fin=cv.fit_transform(cleaned_data).toarray()
X_fin.shape
(14404, 3000)
```

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
model_1 = MLPClassifier()
model_2 = KNeighborsClassifier()
model_3 = SVC()
model_4 = DecisionTreeClassifier()
model_5 = RandomForestClassifier()

X_train,X_test,y_train,y_test=train_test_split(X_fin,y,test_size=0.3)

model_1.fit(X_train,y_train)
model_2.fit(X_train,y_train)
model_3.fit(X_train,y_train)
model_4.fit(X_train,y_train)
model_5.fit(X_train,y_train)
```

```
y_pred_1=model_1.predict(X_test)
y_pred_2=model_2.predict(X_test)
y_pred_3=model_3.predict(X_test)
y_pred_4=model_4.predict(X_test)
y_pred_5=model_5.predict(X_test)

from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred_1))
print(classification_report(y_test,y_pred_2))
print(classification_report(y_test,y_pred_3))
print(classification_report(y_test,y_pred_4))
print(classification_report(y_test,y_pred_5))
```

```
precision    recall  f1-score   support

           0       0.84      0.84      0.84      2720
           1       0.53      0.53      0.53       915
           2       0.63      0.64      0.64       687

    accuracy                           0.74      4322
   macro avg       0.67      0.67      0.67      4322
weighted avg       0.74      0.74      0.74      4322

              precision    recall  f1-score   support

           0       0.82      0.54      0.65      2720
           1       0.32      0.67      0.43       915
           2       0.58      0.51      0.54       687

    accuracy                           0.56      4322
   macro avg       0.57      0.57      0.54      4322
weighted avg       0.68      0.56      0.59      4322

              precision    recall  f1-score   support

           0       0.80      0.93      0.86      2720
           1       0.67      0.47      0.55       915
           2       0.78      0.59      0.67       687

    accuracy                           0.78      4322
   macro avg       0.75      0.66      0.70      4322
weighted avg       0.77      0.78      0.77      4322

              precision    recall  f1-score   support

           0       0.80      0.80      0.80      2720
           1       0.48      0.46      0.47       915
           2       0.56      0.58      0.57       687

    accuracy                           0.69      4322
   macro avg       0.61      0.61      0.61      4322
weighted avg       0.69      0.69      0.69      4322

              precision    recall  f1-score   support

           0       0.81      0.91      0.86      2720
           1       0.63      0.46      0.53       915
           2       0.70      0.61      0.65       687
```



```
    accuracy                           0.77      4322
   macro avg       0.71      0.66      0.68      4322
weighted avg       0.75      0.77      0.75      4322
```

RESULT

Upon evaluating all the models we can conclude the following details i.e.
Accuracy: As far as the accuracy of the model is concerned SVM performs better than Random Forest which in turn performs better than MLP.

We, therefore, conclude that the SVM is the best model for the above-given dataset.
KNN is the worst fit for the above dataset.

CONCLUSION

In this project, diverse techniques for Twitter sentiment analysis methods were discussed, including machine learning, ensemble approaches and dictionary (lexicon) based approaches. In addition, hybrid and ensemble Twitter sentiment analysis techniques were explored. Research outcomes demonstrated that machine learning techniques; for example, the SVM and MLP produced the greatest precision. SVM classifiers may be viewed as standard learning strategies, while dictionary (lexicon) based techniques are extremely viable at times, requiring little efforts in the human-marked archive. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques.

In general, it was expected that ensemble Twitter sentiment-analysis methods would perform better than supervised machine learning algorithms, as they combined multiple classifiers and occasionally various features models. However, hybrid methods also performed well and obtained reasonable classification accuracy scores, since they were able to take advantage of both machine learning classifiers and lexicon-based Twitter sentiment-analysis approaches.

REFERENCES

List all the material used from various sources for making this project proposals

[1] A. DuVander, "Which APIs are handling billions of requests per day?," Programmable Web, 2012.

[2] I. Twitter, "Twitter IPO Prospectus," ed, 2013.

[3] Alexa.com, "Website Traffic Ranking," ed, 2017.

[4] A. DuVander, "Which APIs are handling billions of requests per day?," Programmable Web, 2012.

[5] A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Comput. Surv., vol. 49, no. 2, pp. 1-41, 2016.

[6] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportuniti