2014

Polarity trend analysis of public sentiment on YouTube

Amar Krishna
Iowa State University

Polarity trend analysis of public sentiment on YouTube

by

Amar Krishna

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:
Joseph Zambreno, Co-Major Professor
Leslie Miller, Co-Major Professor
PavanAduri
Kathryn Stolee

Iowa State University

Ames, Iowa

2014

DEDICATION

To

My Parents

and

My Brother and Sister in law

and

to my Nephew "Achintya"

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Dr. Joseph Zambreno for his continuous support and direction throughout my research work or my stay here in Iowa State University. He is a fantastic guide and things we discussed or things he explained made life easier for me. I appreciate his patience in explaining things to me which helped me a lot during my research. As an advisor he is the best a grad student can have.

I would also like to thank Dr. Sandeep Krishnan for his guidance and help during my research. He collaborated me on my research paper which in proceeding in COMAD 2013.

In addition, I would also like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience. I want to also offer my appreciation to those who were willing to participate in my surveys and observations, without whom, this thesis would not have been possible.

Finally, thanks to my family for their encouragement, patience, respect and love.

ABSTRACT

For the past several years YouTube has been by far the largest user-driven online video provider. While many ofthese videos contain a significant number of user comments, little work has been done to date in extracting trends from these comments because of their low information consistency and quality. In this paper we perform sentiment analysis of the YouTube comments related to popular topics using machine learning techniques. We demonstrate that an analysis of the sentiments to identify their trends, seasonality and forecasts can provide a clear picture of the influence of real-world events on user sentiments.Results show that the trends in users' sentiments are well correlated to the real-world events associated with therespective keywords.

CHAPTER 1

INTRODUCTION

With the rapid growth of social networking and the internet in general, YouTube has become by far the most widely used video-sharing service. The popularity of YouTube is because of ease of use and simplicity of these systems for the creation, collaboration and sharing of resources (images, videos) even from non-technical users [28]. Current YouTube usage statistics indicate the approximate scale of the site: at the time of this writing there are more than 1 billion unique users viewing video content, watching over 6 billion hours of video each month [1]. Also, YouTube accounts for 20% of web traffic and 10% of total internet traffic. YouTube provides many social mechanisms to judge user opinion and views about a video by means of voting, rating, favorites, sharing and negative comments, etc. It is important to note that YouTube provides more than just video sharing; beyond uploading and viewing videos, users can subscribe to video channels and can interact with other users through comments. YouTube is generally a comprise of implicit and explicit user-user interaction. This user-to-user social aspect of YouTube (the YouTube social network [2]) has been cited as one key differentiating factor compared to other traditional content providers. Mining the YouTube data makes more sense than any other social media websites as the contents here are closely related to the concerned topic (as it is a video content). While the sheer scope of YouTube has motivated researchers to perform data-driven studies [3, 4, 5], to our knowledge there has been no significant work related to identifying user sentiment of YouTube comments. Our work attempts to bridge this gap by mining the corpus of YouTube comments pertaining to a particular topic to identify trends in user sentiment. It is very obvious that subjectivity and sentiment analysis focuses on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations,

beliefs and speculations in natural language. We claim that the analysis of the sentiment contained in YouTube comments can act as the basis of an effective prediction model, with additional application to correlation between web buzz and stock market trends, box office results, and political elections. We performed our analysis over 6000 YouTube videos dealing with as many as 15 keywords (Obama, Federer, Dow Jones etc.) and collected more than 7 million comments.These set of comments acted as our representative dataset to perform the sentiment and prediction analysis, allowing us to shed light on the following aspects:

1. Temporal aspects of each comment (using the timestamp of each comment)
2. Relationship between the polarity (positive or negative) of each comment and real-world events
3. Sentiment trends over a particular window of time
4. Seasonality dependence of sentiment
5. Sentiment forecasting

We used theInternet Movie Database(IMDb) dataset designed and developed by Pang and Lee[8] as ourtrain set. We also created a small data set from the YouTube comments as our train set in order to keep the relevancy of the content intact. Analyzing the sentiment of each comment and their trending patterns can give us a clear picture of overall user base sentiment over a particular span of time. In my research, the method we use to find the polarity of each comment is the Naive Bayes classification technique. Naive Bayes model is the basic classification model used in this kind of emotions/sentiment analysis. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round and about 4" in diameter. Even if these features

depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers have worked quite well in many complex real-world situations [29]. We achieved an accuracy of about 81% using this classification method. Our future research will involve tweaking the algorithm to achieve an accuracy of almost 90%. The overall polarity of a particular user is used to find the average of sentiments over a period of time. Our data preprocessing step involved the cleaning of the dataset(as the YouTube Data is volatile, only about 60% of the comments were relevant to the topic) based on the time-stamp, language, demography etc. We did our statistical analysis using the standard tool R and prediction models were designed using the open source statistical tool Weka. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. By default, the system is set up to learn the forecasting model and generate a forecast beyond the end of the training data. Selecting the perform evaluation check box tells the system to perform an evaluation of the forecaster using the training data. That is, once the forecaster has been trained on the data, it is then applied to make a forecast at each time point (in order) by stepping through the data. These predictions are collected and summarized, using various metrics, for each future time step forecasted, i.e. all the one-step-ahead predictions are collected and summarized, all the two-step-

ahead predictions are collected and summarized, and so on. This allows the user to see, to a certain degree, how forecasts further out in time compare to those closer in time. Our preliminary analysis indicates that this approach is effective in identifying correlations between the trends in users' sentiments and real-world events. We also have compared the sentiment values of YouTube comments with that of several Tweets about the same keywords. Comparison of Tweets with the YouTube comments seems to be an unfair comparison, but it gives the insight how the sentiment about the same keyword varies from one social network to the other. Though the sentiment analysis graphs of Tweets about few keywords were same as of those YouTube comment graphs, there was a huge difference in case of other keywords.

CHAPTER 2

RELATED WORK

Several researchers have performed sentiment analysis of social networks such as Twitter and YouTube [4], [6], [7].These works deal with comments, tweets and other metadata collected from the social network profiles of users orof public events that have been collected and analyzed to obtain significant and interesting insights about the usage ofthese social network websites by the general mass of people. The work most closely related to ours is by Siersdorfer et al. [4]. They analyzed more than 6 million comments collected from 67,000 YouTube videos to identify the relationship between comments, views, comment ratings and topic categories. The authors show promising results in predicting the comment ratings of new unrated comments by building prediction models using the already rated comments. Pang, Lee and Vaithyanathan [8] perform sentiment analysis on 2053 movie reviews collected from the Internet Movie Database (IMDb). They examined the hypothesis that sentiment analysis can be treated as a special case of topic-based text classification. Their work depicted that standard machine learning techniques such as Naive Bayes or Support Vector Machines (SVMs) outperform manual classification techniques that involve human intervention. However, the accuracy of sentiment classification falls short of the accuracy of standard topic-based text categorization that uses such machine learning techniques. They reported that the simultaneous presence of positive and negative expressions (thwarted expectations) in the reviews make it difficult for the machine learning techniques to accurately predict the sentiments. Another work on the YouTube comments wasdone by SmitaShree and Josh Brolin[30] where the authors proposed an unsupervised lexicon-based approach to detect sentiment polarity of user comments in YouTube. They adopted a data driven approach and prepared a social media list of terms andphrases expressing

the user sentiment and opinion. But their results also showed that recall of negative sentiment is poorer compared to the positives, which may be due to the wide linguistic variation used in expressing frustration and dissatisfaction.

Research on mining YouTube performed by Lina McInerney et al.[31]discussed how social media can be used to radicalize a person. Their idea is:- Crawling, a global social networking platform, such as YouTube, has the potential to unearth content and interaction aimed at radicalization of those with little or no apparent prior interest in violent Jihadism. Their research explores whether such an approach is indeed fruitful. They collected a large dataset from a group within YouTube that wasidentified as potentially having a radicalizing agenda. They analyzed the data using social network analysis and sentiment analysis tools, examining the topics discussed and what the sentiment polarity (positive or negative) is towards these topics. In particular, they focused on gender differences in this group of users, suggesting most extreme and less tolerant views among female users.

Another work on YouTube was done by Orimaye Sylvester Oluboluet. al,[32],where the researchers focused on the YouTube videos uploaded on Yoruba language movies. In this paper, they present an automatic sentiment analysis algorithm for YouTube comments on Yoruba language movies. The algorithm uses SentiWordNet thesaurus and a lexicon of commonly used Yoruba language sentiment words and phrases. Another research on the videos available on web was done by Louis PhillipeMorencyet.al.[33]. Their paper addresses the task of multimodal sentiment analysis, and conducts proof-of-concept experiments that demonstrate that a joint model that integratesvisual, audio, and textual features can be effectively used toidentify sentiment in Web videos. The paper makes threeimportant contributions. First, it addresses for the first timethe task of tri-modal sentiment analysis, and shows that it isa feasible task that can

benefit from the joint exploitation ofvisual, audio and textual modalities. Second, it identifies asubset of audio-visual features relevant to sentiment analysis and presents guidelines on how to integrate these features.Finally, it introduces a new dataset consisting of real onlinedata, which will be useful for future research in that respective area.Another prominent work in sentiment analysis is by Bollen et al. [9]. The authors analyzed the Twitter feeds of users using two sentiment tracking tools to accurately predict the daily changes to the closing values of the Dow Jones Industrial Average (DJIA) Index. The authors report an accuracy of 86.7% and a reduction of more than 6% in the mean average percentage error. Cecilia OvesdotterAlm et al. [34] performed their research on the extracting emotions from text. This paper explores the text-based emotion prediction problemempirically, using supervised machinelearning with the SNoW learning architecture. The goal is to classify the emotional affinity of sentences in the narrative domain of children's fairy tales, forsubsequent usage in appropriate expressive rendering of text-to-speech synthesis. Other works have performed sentiment analysis of social networks such as Twitter to show that there exists a relationship between the moods of people to the outcome of events in the social, political, cultural and economic spheres[10], [12]. Another research on the social media sentiment analysis is done by A.Kowcika et al.[35] In their paper they propose a system which is able to collect useful information from the twitter website and efficiently perform sentiment analysis of tweets regarding the Smart phone war. The system uses efficient scoring system for predicting the user's age. The user'sgender is predicted using a well-trained Naïve BayesClassifier. Sentiment Classifier Model labels the tweet with a sentiment. KrisztianBalog et al.[36] proposed in his paper a method to collect useful information from the twitter website and efficiently perform sentiment analysis of tweets regarding the Smart phone war. The system uses efficient scoring system for predicting the user's age. Twitter Sentiment

Analysis: The Good the Bad and the OMG!, paper by Efthymios Kouloumpis et al.[37] deals with the utility of linguistic features for detecting the sentiment of Twitter messages. They evaluatethe usefulness of existing lexical resources as well as featuresthat capture information about the informal and creative language used in micro-blogging. Another sentiment analysis of web text was done using the blog posts by Gilad Mishne et al.[38]. In their paper they addressed the task of estimating this state-of-mind from the text written by bloggers. To this end, we build models that predict the levels of various moods according to the language used by bloggers at a given time; our models show high correlation with the moods actually measured, and substantially outperform a baseline.Similarly, research done by Albert Bifet and Eibo Frank [39] in their papertitled"Sentiment Knowledge Discovery in Twitter Streaming Data"focuses on the challenges that Twitter data streams pose, focusing on classification problems, and then considers these streams for opinion mining and sentiment analysis. To deal with streaming unbalanced classes, we propose a sliding window Kappa statistic for evaluation in time-changing data streams.Similar work [40],[41],[42],[43] on twitter was done by several researchers mentioned in the following papers(reference numbers).Determining the Sentiments of Opinions by Soo-Min Kim and Eduard Hovy[45] is the paper where they discuss a system that contains a module for determining word sentiment and another for combining sentiments within a sentence. We experiment with various models of classifying and combining sentiment at word and sentence levels, with promising results.

Melville et al. [13] identify two approaches for solving sentiment analysis problems: knowledge-based approacheswhich use domain specific background knowledge and machine learning based approaches, as used by Pang et al. [8]. They take advantage of both approaches and show better classification performance as compared to using either of the approaches in

isolation. Nasukawa et al. [14] achieved high precision (75%-95%) by breaking down the sentiment analysis of the entire text into analysis of sentiments associated with the specific subjects in the text. Spertus [15] identifies features and rules based on the syntax and semantics of the sentences to identify 64% of the abusive messages and 98% of the non-abusive messages from the test set. Sentiment classification of web pages is one type of web-content classification. There have been several other works on other types of web-content classification. There are some works on genre classification where web content(e.g., news articles) is used to identify the genre to whichthe content belongs[16], [17], [18]. Web page classification is another topic that has beenexplored by several researchers [19-24].The authors collect several measures from the web pages including titles of the pages, URLs, hyperlink structure, etc. to classify web pages according to the topics they discuss. While web classification is considered to be more difficult than general text classification, the authors identify severalmeasures that can improve the classification accuracy.

One of the most prominent works in web page classification was done by Daniele Riboni in the paper “Feature Selection for Web Page Classification”[44]. They conducted various experiments on a corpus of 8000 documents belonging to 10 Yahoo! categories using Kernel Perception and Naive Bayes classifiers. Their experiments show the usefulness of dimensionality reduction and of a new structured oriented weighing technique. They also introduce a new method for representing linked pages using local information that makes hypertext categorization feasible for real-time applications.Other classification works are like the one done by Eibe Frank et al.[46] In their paper they propose an appropriate correction by adjusting attribute priors. This correction can be implemented as another data normalization step, and they show that it can significantly improve the area under the ROC curve. They also show that the modified version of

MNB is very closely related to the simple centroid-based classifier and compare the two methods empirically. Another work on the sentiment analysis of social media is done using multimodal approach, discussed in the paper by Diana Maynard et al.[47]. They examine a particular use case, which is to help archivists select material for inclusion in an archive of social media for preserving community memories, moving towards structured preservation around semantic categories. The textual approach they take is rule-based and builds on a number of sub-components, taking into account issues inherent in social media such as noisy ungrammatical text, use of swear words, sarcasm etc.

Our work differs significantly from the above works. Previouswork on the YouTube has focused on analyzing the number of likes and dislikes for a particular comment, whereas we focus on the polarity of each comment. To the best of our knowledge, there has not been any prior work that focuses on sentiment analysis of YouTube comments. While previous works have performed sentiment analysis at a given point of time, our focus is on the changes in the trends of users' (commenters') sentiments based on the contents of the videos (queries based on different keywords). Further, we focus on exploring the sentiment forecasts as expressed through sentiment scores of comments.We also compare the sentiment values of YouTube comments with that of the tweets (tweets harvested for the keywords used in the YouTube mining).

CHAPTER 3

OUR APPROACH

3.1. Research Questions

In our research we discuss following questions and try to give answer to each one of them.

RQ1. How does the sentiment for a particular keyword (video) trend over a particular window of time?

RQ2. How does the sentiment embedded in user comments relate to the performance of the videos (keywords) inthe given time period?

RQ3. How well can we forecast the users' sentiments for the next 26 weeks following the last timestamp of eachdataset?

RQ4. Are the sentiments associated with the comments a good indicator of the correlation between the web buzzand real world events in politics, sports, etc.?

3.2. Data Collection Process and Algorithm

We modeled the data by automating queries and keyword based searches to gather videos and their correspondingcomments. Python scripts using the YouTube APIs were used to extract information about each video (comments and their timestamps). We collected 1000 comments per video (YouTube allows a maximum of 1000 comments per video to be accessed through the APIs), and used keywords like "Federer", “Nadal", “Obama" etc., to collect the data for specific keywords. The timestamp and author name of each video were also collected. The final dataset used for the sentiment analysis had more than 3000 videos and more than 7 million comments. We performed data pre-processing on the collected comments. YouTube comments comprise of several languages depending on the demography of the commenter. However, to simplify the sentiment analysis, we modified the data collectionscripts to collect only English comments.

From the collected English comments, only comments in the standardUTF-8 encoding were selected in order to remove comments with unwanted characters. The steps below explain the procedureto collect the comments with their respective timestamps and author names for the keywords specified by the user. In steps 2-4, the Google APIs for YouTube are used to configure the query with the number of videos to be fetched, the language of interest for comments, the search keyword, and how the comments are to be sorted. Step 5 collects the IDs of the videos related to the specified keyword. Steps 6 and 7 collect the comments associated with these videos and extract the timestamps, author names and comment text fromthe comment entries. All the comments for a single keyword are aggregated into one dataset which is used as the test set as explained in the following:

Step 1: Prompt the user to specify the search keyword (keywords) and number of videos (numVideos)

Step 2: Set maxNumVideos = max(50; numVideos) (As Google limits the maximum number of videosfetched in one iteration to 50)

Step 3: Set up the YouTube client to use the YouTube-Service() API to communicate with the YouTube servers

Step 4: Use the YouTubeVideoQuery() API to set the query parameters like language, search keyword,etc

Step 5: Perform successive queries to get the videoID of each video related to the keyword

Step 6: Collect the comments associated with each videoID using the GetYouTubeVideoCommentFeed() API (maximum limit of comments per video is 1000)

Step 7: Extract the comments with their respective timestamps and author names

Source Code

```
importgdata.youtube
importgdata.youtube.service
import codecs
```

```
import sys
# Pass the keyword(s) on the command-line
iflen(sys.argv) != 3:
error_msg = "Usage error: " + sys.argv[0] + " [keywords] [num_videos]"
sys.exit(error_msg)
keywords = sys.argv[1]
num_videos = int(sys.argv[2])
max_results = 50
ifnum_videos<max_results:
max_results = num_videos
# Create the youtube client
client = gdata.youtube.service.YouTubeService()
client.developer_key='AI39si7znTOoAY_1ofaYG-
A41Qaarn4bgRHgEgCfNJltdFxlEXOU94fjIkhn4oOoyy8UoA-
1f2Nf0aTg0pxomofo1RRw0HYDFA'
query = gdata.youtube.service.YouTubeVideoQuery()
# Configure the API query
query.vq = keywords
query.lr = 'en'
query.orderby = 'viewCount'
#query.orderby='relevance'
query.racy='include'
query.max_results = max_results

# Perform successive queries to get the video IDs we need
start_index = 50
video_IDs = []
whilestart_index<= num_videos:
query.start_index = start_index
print "grabbing videos from index ", start_index, " to ", start_index+max_results-1
feed = client.YouTubeQuery(query)
```

```
for entry in feed.entry:
entry_text = entry.id.text
video_IDs.append(entry_text.split("/")[-1])
start_index += max_results
# This function grabs a given video's comment feed
defcomments_generator(client, video_id):
comment_feed = client.GetYouTubeVideoCommentFeed(video_id=video_id)
whilecomment_feed is not None:
for comment in comment_feed.entry:
yield comment;
next_link = comment_feed.GetNextLink()
ifnext_link is None:
comment_feed = None
else:
comment_feed = client.GetYouTubeVideoCommentFeed(next_link.href)

# For each ID, grab the comment data (this will be the slowest part)
# I'm not sure if I'm handling the unicode correctly
outfile = codecs.open('FINAL_DATA_NASDAQ_STOCK.txt', 'a', encoding='utf-8')
for id in video_IDs:
comment_feed = comments_generator(client, id)
for comment in comment_feed:
comment_text = comment.content.text
comment_text = comment_text.replace('\n', ' ')
post_date = comment.published.text.split("T")[0]
outstring = post_date + '$' +' '+ comment_text.decode('utf-8') + '\n'
outfile.write(outstring)
```

The above code takes the keyword "Nasdaq" as input and outputs the comments with their respective timestamp and further exports it to the 'FINAL_DATA_NASDAQ_STOCK.txt'.

3.3. Sentiment Analysis

We follow the standard sentiment classification approach given in [25]. The approach is shown in Figure 3-1. We usethe Naive Bayes classification technique for sentiment analysis.A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood.In other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods. An advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.The classifier is trained on the IMDb database (based on the analysis done in the research paper [8]).In the paper, the authors examine the effectiveness of applying machine learning techniques to the sentiment classification problem. A challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner. For example, the sentence “How could anyone sit through this movie?” contains no single word that is obviously negative. The authors selected only reviews where the author rating was expressed either with stars or some numerical value (other conventions varied too widely to allow for automatic processing). Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral.

The training set consists of 5000 positive and 5000 negative movie reviews respectively. The comments we collected for each keyword is used as the test data for classification. The size of the test data varies for each keyword and ranges from as low as 10000 (Dow Jones data) to as high as 300,000 (Obama). The Naive Bayes classifier is trained on the comments from the training set and is then used to calculate the overall sentiment for each comment in the test set. A comment is considered as a bag of independent words (i.e., the ordering of the words is not considered). The positive and negative comments in the train dataset are stored in two separate dictionaries, which we refer to as positive dictionary (positive comments) and negative dictionary (negative comments). For each comment, the polarity/sentiment of each word is calculated by calculating the number of times the word appears in the positive and negative dictionaries. For each word, the positive polarity is number of times the word appears in the positive dictionary divided by the total number times it appears in both the positive and the negative dictionaries. Similarly, the negative polarity is the number of times the word appears in the negative dictionary divided by the total number of times it appears in both the dictionaries. Figure 2-2 shows an example of how the overall sentiment of a comment in the test set is calculated. The word “like" in the figure appears 292 times in the positive dictionary and 460 times in the negative dictionary. Thus, the positive polarity of the word “like" is 0.39 and negative polarity is 0.61. Similarly the positive and negative polarity of each word is calculated. Adding the positive polarities of all the words in the comment gives the positive polarity of the entire comment. Similarly, adding the negative polarities of all the words in the comment gives us the negative polarity of the comment. The comment is then classified as positive if the positive polarity of the comment is greater than the negative polarity, and negative otherwise.

For this example, since the negative polarity is greater than the positive polarity, the sentence is classified as negative.

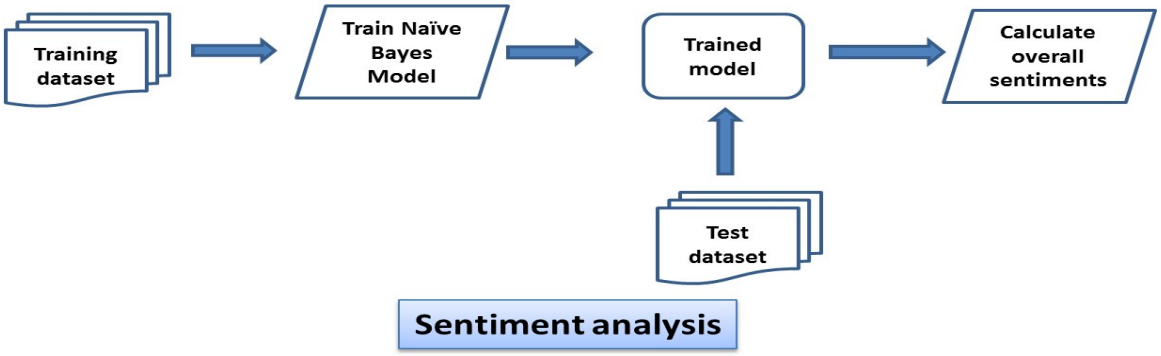

Figure 3-1: Standard Sentiment Classification Approach

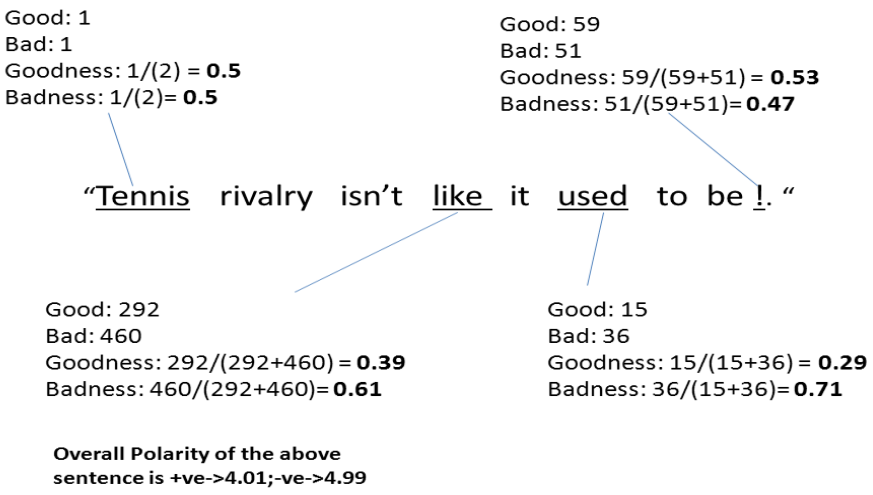

Figure 3-2: Calculation of overall sentiment in a test

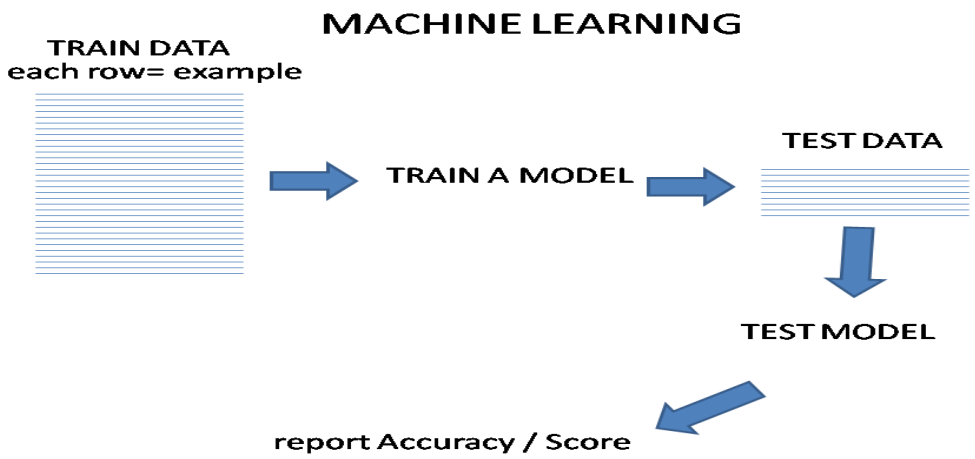

Figure 3-3: overview of the entire sentiment analysis process

The Figure 3-3 gives an overview of the entire sentiment analysis process(including the how to decide the train set and test set).

Source code used for the polarity detection of each YouTube comment

```
import math
from decimal import Decimal
#read in positive and negative lines from files
poslines= open(r'rt-polarity1.pos', 'r').read().splitlines()
neglines= open(r'rt-polarity.neg', 'r').read().splitlines()
#there is a total of 5331 positives and negatives.
#lets take first N as training set, and leave rest for validation
N= 4800
poslinesTrain= poslines[:N]
neglinesTrain= neglines[:N]
poslinesTest= poslines[N:]
neglinesTest= neglines[N:]
```

```
#create the train set and the test set by attaching labels to text to form a
#list of tuples (sentence, label). Labels are 1 for positive, -1 for negative
#if you don't get this look up text comprehensions in Python
trainset= [(x,1) for x in poslinesTrain] + [(x,-1) for x in neglinesTrain]
testset= [(x,1) for x in poslinesTest] + [(x,-1) for x in neglinesTest]
#count the number of occurances of each word in positives and negatives
poswords={} #this dictionary will store counts for every word in positives
negwords={} #and negatives
forline,label in trainset: #for every sentence and its label
for word in line.split(): #for every word in the sentence
        #increment the counts for this word based on the label
                #the .get(x, 0) method returns the current count for word
                #x, of 0 if the word is not yet in the dictionary
if label==1: poswords[word]= poswords.get(word, 0) + 1
else: negwords[word]= negwords.get(word, 0) + 1
#evaluate the test set
wrong=0 #will store number of misslassifications
forline,label in
final,totpos, totneg= 0.0, 0.0,0.0
for word in line.split():
     #get the (+1 smooth'd) number of counts this word occurs in each class
    #smoothing is done in case this word isnt in train set, so that there
    #is no danger in dividing by 0 later when we do a/(a+b)
                #it's a trick: we are basically artifically inflating the counts by 1.
    a= poswords.get(word,0.0) + 1.0
    b= negwords.get(word,0.0) + 1.0
  #increment our score counter for each class, based on this word
totpos+= a/(a+b)
totneg+= b/(a+b)
 #create prediction based on the counter values
prediction=1
```

```
iftotneg>totpos: prediction=-1
        if prediction!=label:
wrong+=1
print '%.2f | %.2f | prediction=%d ' % (totpos, totneg,prediction)
else:
print '%.2f | %.2f | prediction=%d ' % (totpos, totneg,prediction)

#print the error rate
print 'error rate is %f' % (1.0*wrong/len(testset),).
```

The above mentioned source code calculates the polarity on each sentence and outputs the negative and positive polarity associated with it.

CHAPTER 4

RESULTS

After identifying the sentiments for each comment using the Naive Bayes classifier, we performed the analysis of the trends in sentiments. For analyzing the sentiment trends, we aggregated the sentiments for comments on a weekly basis and calculated the mean sentiment for each week. We expressed the data as a time series model and used the statistics tool R for finding the trends in the comments for each keyword. For forecasting the future sentiment values, we used the 'forecasting' module of Weka (a data mining tool) [26]. The forecasting module uses the existing dataset to forecast the sentiments for 26 weeks into the future. This section addresses the research questions RQ1 and RQ2. We use the decompose() function in R to model the comment sentiments as time series data, and to give the overall trend, the seasonality (repeated pattern), and the random trend in the data. The function first determines the trend component using a moving average (if filter: a parameter passed to the decompose function) is NULL, a symmetric window with equal weights is used), and removes it from the time series. Then, the seasonal figure is computed by averaging, for each time unit, over all periods. The seasonal figure is then centered. Finally, the error component is determined by removing trend and seasonal figure (recycled as needed) from the original time series. Figure 4-1 shows the sentiment trends for the keyword “Roger Federer". The figure shows the sentiment trend when the mean of the sentiment values per week is used as input. The figure is divided into four layers. The uppermost layer (observed) gives the observed mean values per week. The second layer (trend) gives the overall trend. The third layer gives the seasonal component of the trend, which is the repeated pattern in the data. The lowermost layer gives the random component in the trend. We see that during the initial stages of the graph the sentiment was mostly positive (because those were the peak years

of Roger Federer's career, winning as many as 6 Grand Slams in a span of 3 years), but in the year 2009-2011 (middle stages of the graph) he won only one Grand Slam, and hence the sharp decrease in the sentiment trends. For the Federer dataset, during 2011, the trends have hit their lowest values because Federer did not win a single Grand Slam in 2011. In the seasonality graph we see that during summer every year (time when Wimbledon Championships are held), Federer always gets a sharp increase in the sentiment values (as he performs well at the Wimbledon Championships, and has been to semi-finals for almost 10 years in a row). Figures 4-2 to 4-8 show the trend decomposition graphs for the keywords: Obama, Romney, Nadal, Gangnam, Adele and Oprah Winfrey. Their respective trends are discussed in Section 4.2. Their respective graphs depict the same parameters (observed, trend, seasonality, random) as explained in the case of Federer.

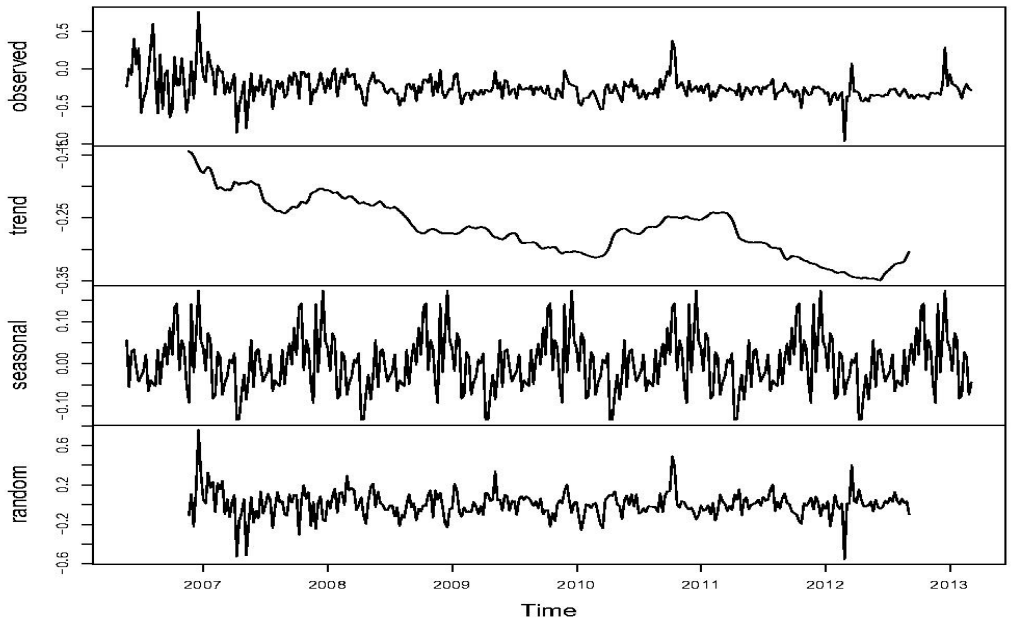

Figure 4-1: Trend decomposition graphs for Roger Federer

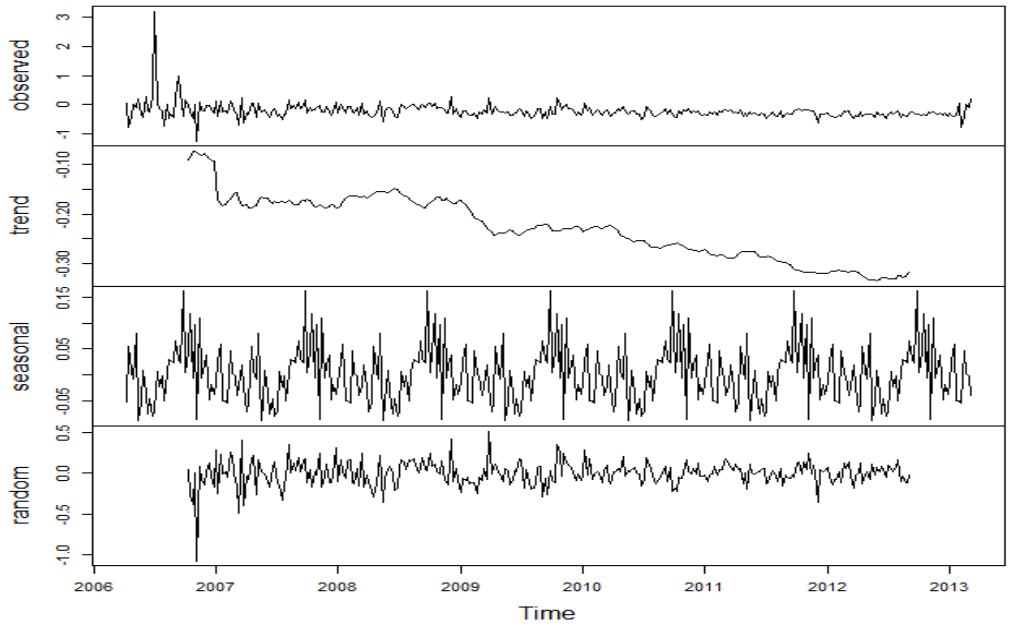

Figure 4-2: Trend decomposition graphs for Nadal

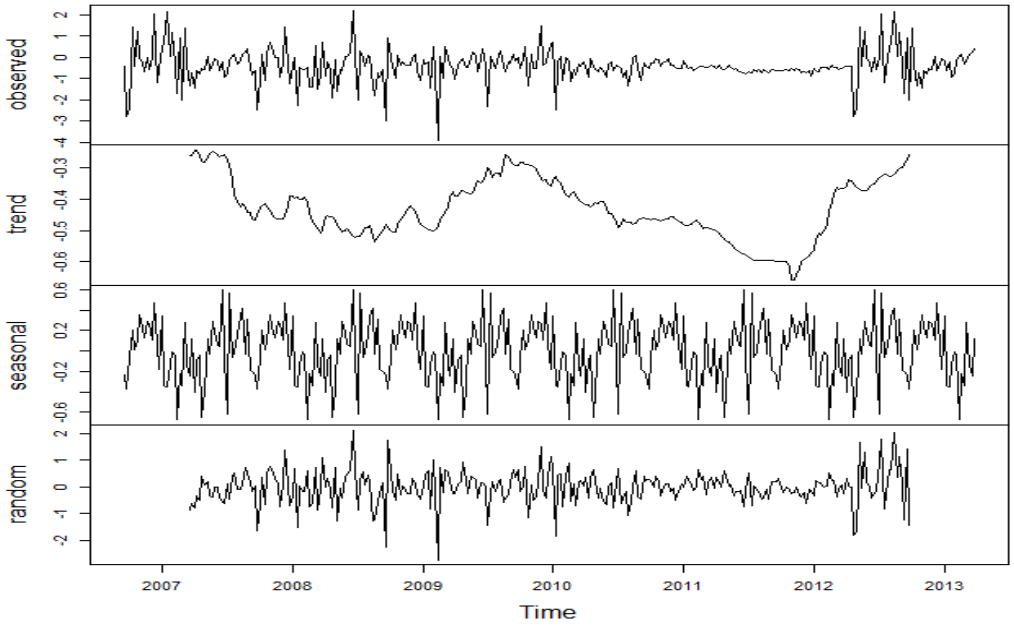

Figure 4-3: Trend decomposition graphs for Romney

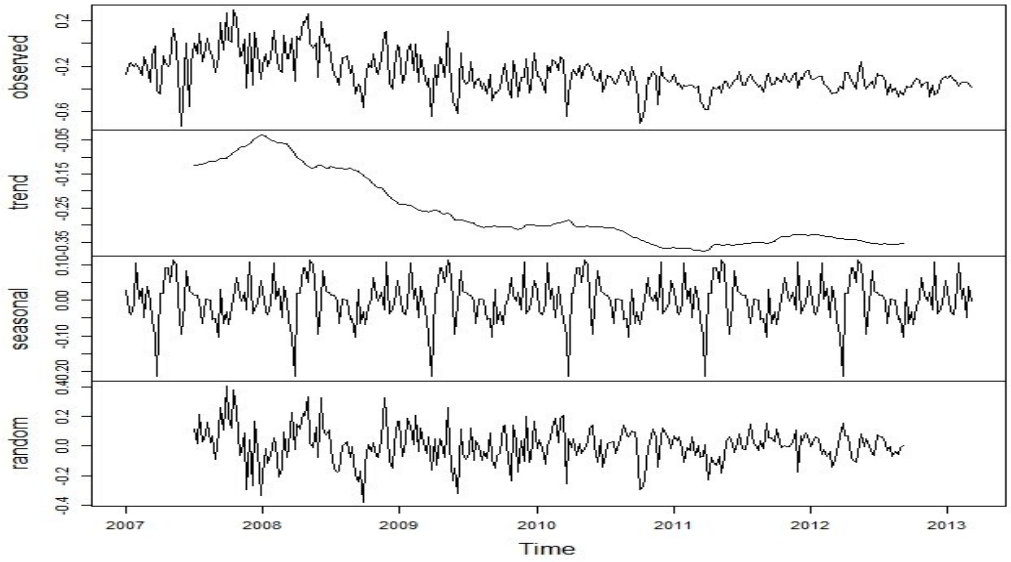

Figure 4-4: Trend decomposition graphs for Obama

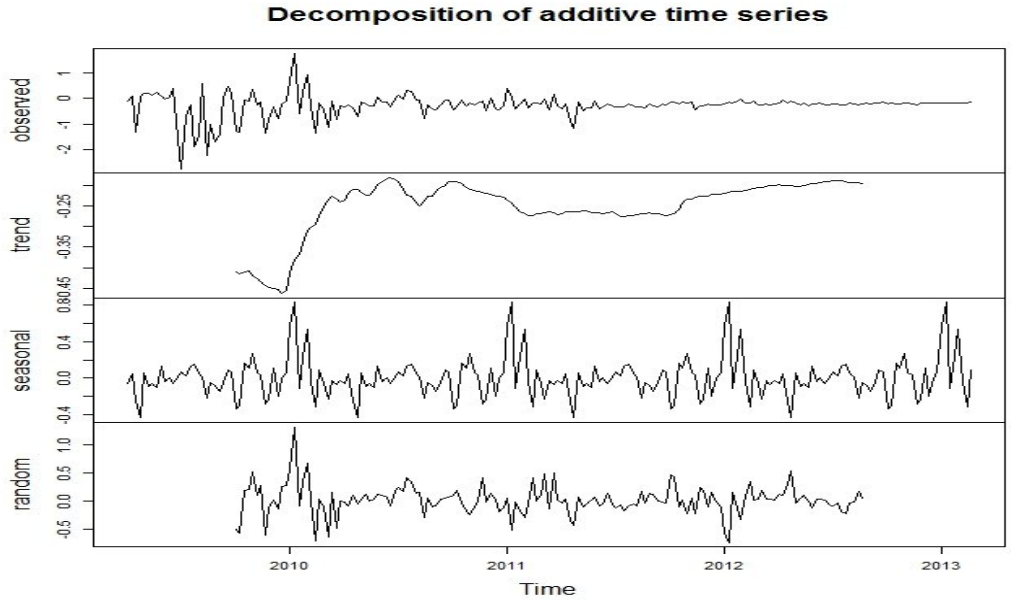

Figure 4-5: Trend decomposition graphs for Gangnam

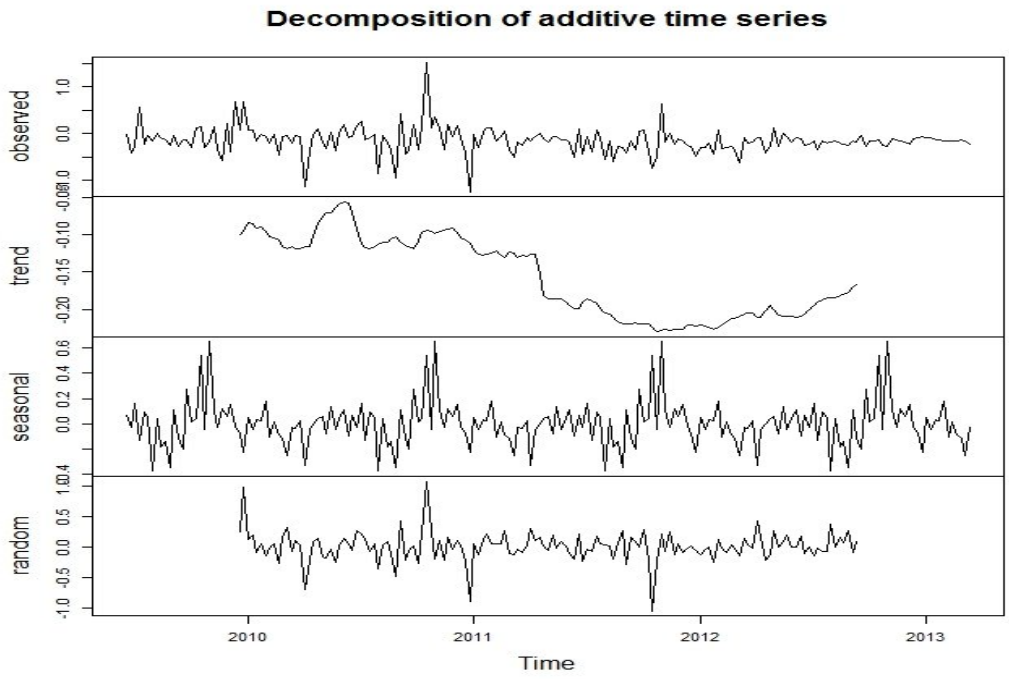

Figure 4-6: Trend decomposition graphs for Adele

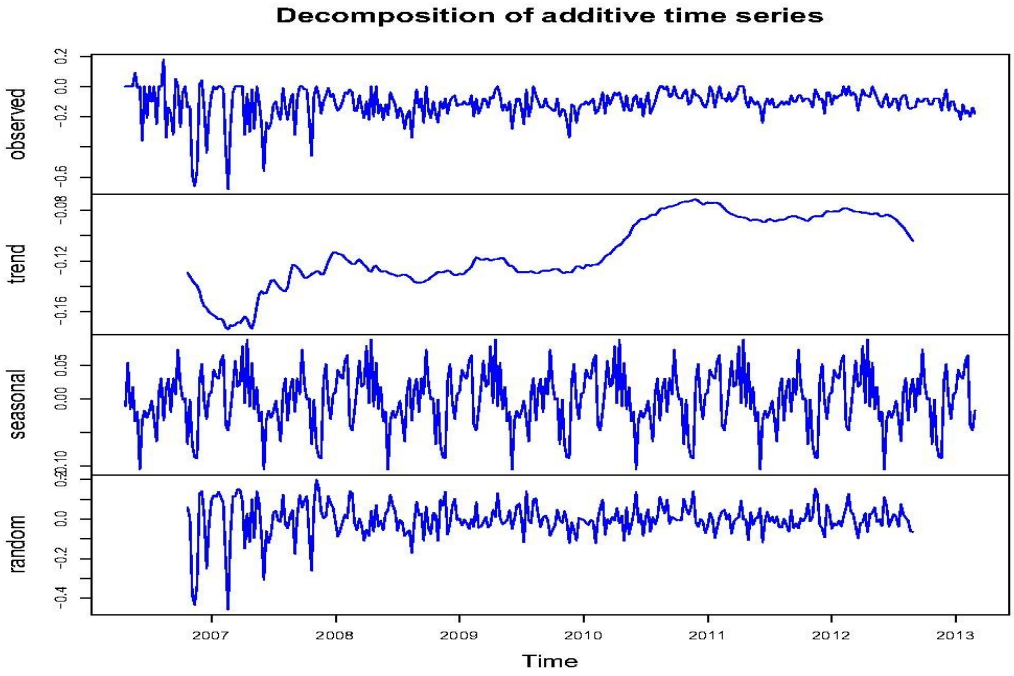

Figure 4-7: Trend decomposition graphs for Oprah Winfrey

We also collected tweets to do the same analysis as above. As per the twitter restrictions of grabbing only 3500 tweets for a particular user, we are presenting the mean graph obtained from the Oprah Winfrey tweets below:-

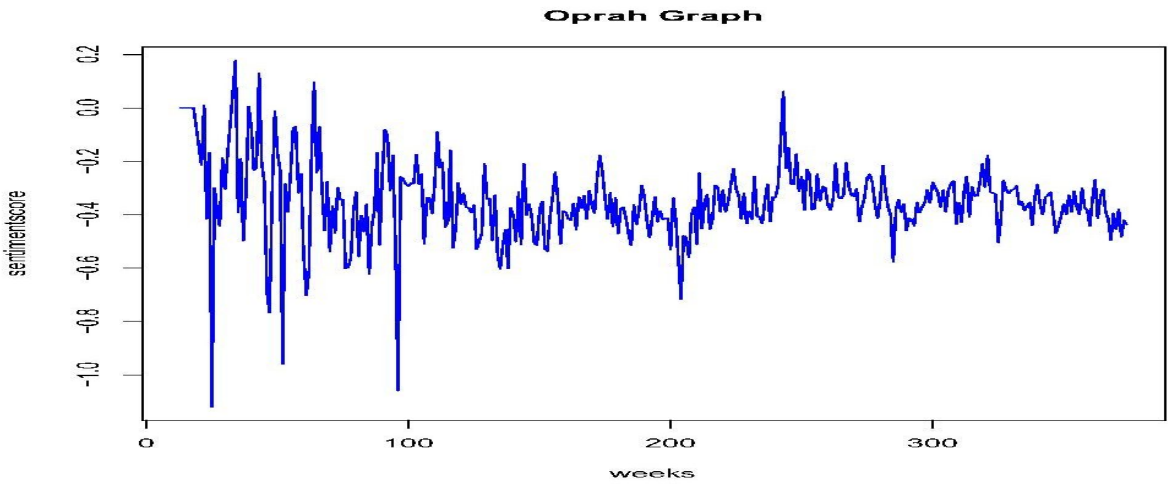

Figure 4-8: Mean graph of the sentiment scores of tweets

4.1. 26 weeks forecasting using Weka

This section discusses the results for research question RQ3. Time series forecasting is the process of using a model to generate predictions (forecasts) for future events based on known past events. Time series data has a natural temporal ordering - this differs from typical data mining/machine learning applications where each data point is an independent example of the concept to be learned, and the ordering of data points within a data set does not matter. For forecasting the sentiment values into the future, we use the forecasting module provided by Weka [27]. Weka provides a time series forecasting environment for Weka. It includes a wrapper for Weka regression schemes that automates the process of creating lagged variables and date-derived periodic variables and provides the ability to do closed-loop forecasting. We perform the forecasts for all the data sets (presenting the results only for Federer and Obama) using SMO

regression, which is a support vector machine approach. This step can load or import a time series forecasting model created in Weka's time series analysis and forecasting environment and use it to generate a forecast for future time steps beyond the end of incoming historical data. This differs from the standard classification or regression scenario covered by the Weka Scoring plugin, where each incoming row receives a prediction (score) from the model, in that incoming rows provide a "window" over the recent history of the time series that the forecasting model then uses to initiate a closed-loop forecasting process to generate predictions for future time steps. Each dataset consists of the comments, sentiment values and their corresponding timestamps collected as part of the data collection approach described in Section 3.2. Weka allows the forecast module to use the entire training data to build a model, and using it to forecast the values for a specified time period into the future. Since the data is aggregated over a week, we forecast the sentiments for 26 weeks into the future. We use the standard settings of the forecast module except for the base machine learner, which we set to SMO regression. This means that the forecaster trains an SMO regression model using the data provided. The trained model is used to make forecasts for each week in the current dataset. Weka allows evaluation of the forecast using several metrics. We select the mean Absolute Error (MSE) and the Root Mean Square Error (RMSE) metrics. The default confidence level of 95% is selected. The system uses the known target values in the training data to set the confidence bounds. A 95% confidence level means that 95% of the true target values fall within the interval. The 26 week forecast results for the Federer and Obama data are shown in Figures 4-9 and 4-10 respectively. We average the MAE and RMSE values for the 26 weeks (26 values). For the Federer data, the average MAE value for the 26 weeks is approximately 0.08 and the average RMSE value is 0.13. For the Obama dataset, the average MAE value is 0.089 and the average RMSE value is 0.125.

The low RMSE values indicate that using SMO regression enables us to forecast the future sentiment values accurately. The relative measures give an indication of how the well forecaster's predictions are doing compared to just using the last known target value as the prediction. They are expressed as a percentage, and lower values indicate that the forecasted values are better predictions than just using the last known target value. A score of >=100 indicates that the forecaster is doing no better (or even worse) than predicting the last known target value. Note that the last known target value is relative to the step at which the forecast is being made - e.g. a 12-step-ahead prediction is compared relative to using the target value 12 time steps prior as the prediction (since this is the last "known" actual target value).

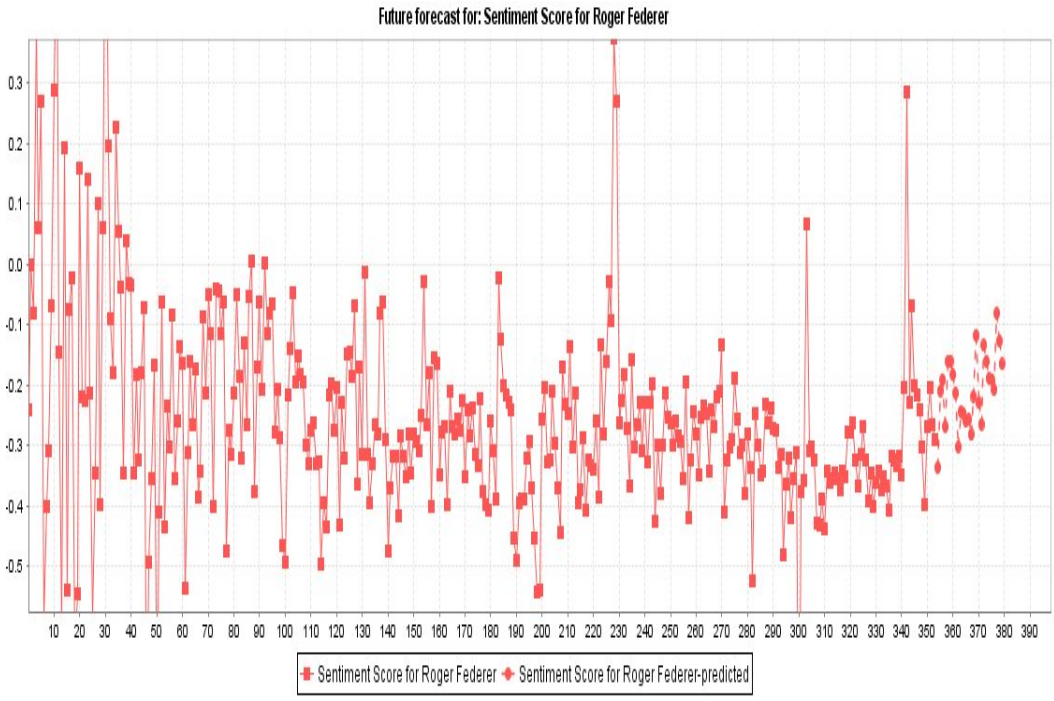

Figure 4-9: 26 week forecast results for the Federer

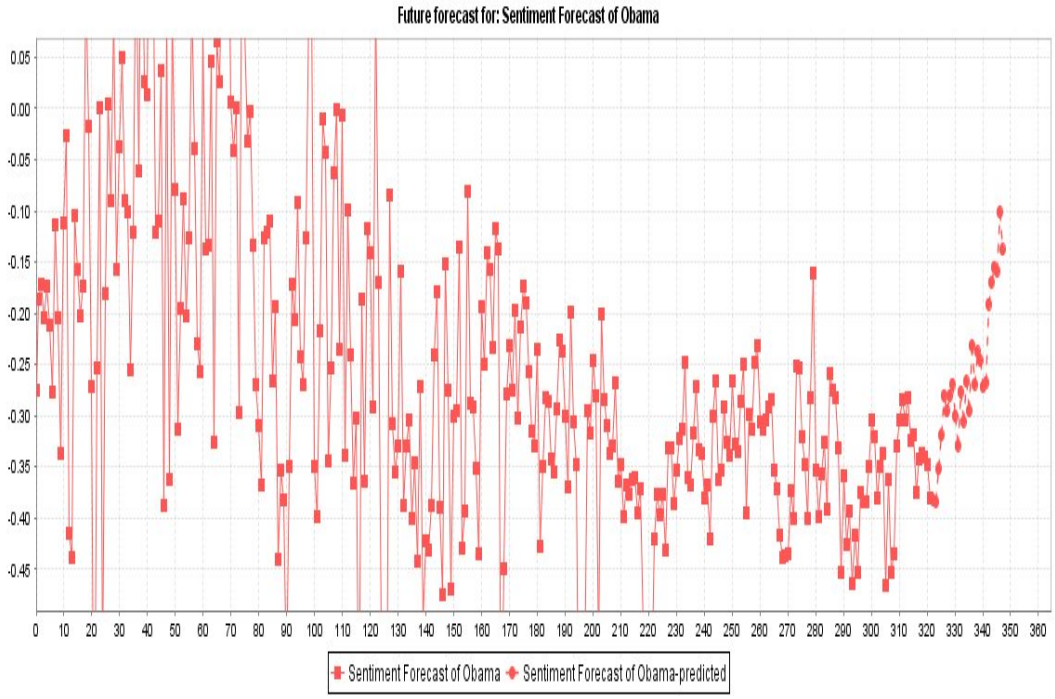

Figure 4-10: 26 week forecast results for the Obama

4.2. Comparing the Trends (Real World Dependencies)

This section addresses research question RQ4. Figures4-11 to 4-14 show how the sentiment trends vary when it comes to comparison between two opponents in the same field of interest. Figure 4-11 illustrates one of the greatest rivalries in Tennis i.e., Federer vs. Nadal. The graphs are used as an example to illustrate that user sentiments were complementing in the case of both these players. We can see that during 2006-2007, which is point 'a' in the plot, users had positive sentiments for Federer as he was on top of his career. Nadal became his greatest opponent by defeating him in French Open Finals for two years in a row (which depicts more positive sentiment for Nadal with respect to Federer). Between 2007-2009 (point 'b'), users' sentiments for Federer tend to decrease but stay consistent in case of Nadal (he defeated Federer in Australian Open and Wimbledon in the year 2009). Nadal's trend shows better consistency

than Federer's trend. During the year 2011-2012 (point 'c') (Dec 2011), Nadal's trend shows a steep decrease as he lost to Djokovic (another rival) in several tournaments and also was out of action for almost 8 months to injury. This observation suggests that user/commenter sentiment is highly dependent on and correlated to the performance of the respective keywords (in this case, players) in their field. Similar observations can be seen for the competition between Obama and Romney for the presidential elections. Figure 4-12 shows the trends for both parties. Point 'a' on the plot shows that during the year 2008, users' sentiments for Obama were highly positive as compared to that of Romney's as Romney was not in contention. Point 'b' depicts the a low sentiment for Obama because of the economic recession. Point 'c' shows Obama is way ahead of Romney after he started his campaign for a second term as a President. Point 'd' shows a narrow increase in Romney's sentiment over Obama's as a result of his performance in First presidential debate. Figure 4-13 presents the sentiment trends for keywords like Dow Jones and how they are closely related with the real-time fluctuations in the Dow Jones index. We collected the data for Dow Jones index for 10 years and mapped it along with the sentiment trend graph obtained from the YouTube comments sentiment values. The figure depicts the trends in users' sentiments with respect to how the stock market behaved from 2008 to 2012. Figure 4-14 depicts a comparison of figures and shows that with the change in Federer's rankings, the sentiment trends also have been correspondingly changing. During 2007, Federer was ranked number 1, and in 2008 he went down to the second rank, as observed in the figure.

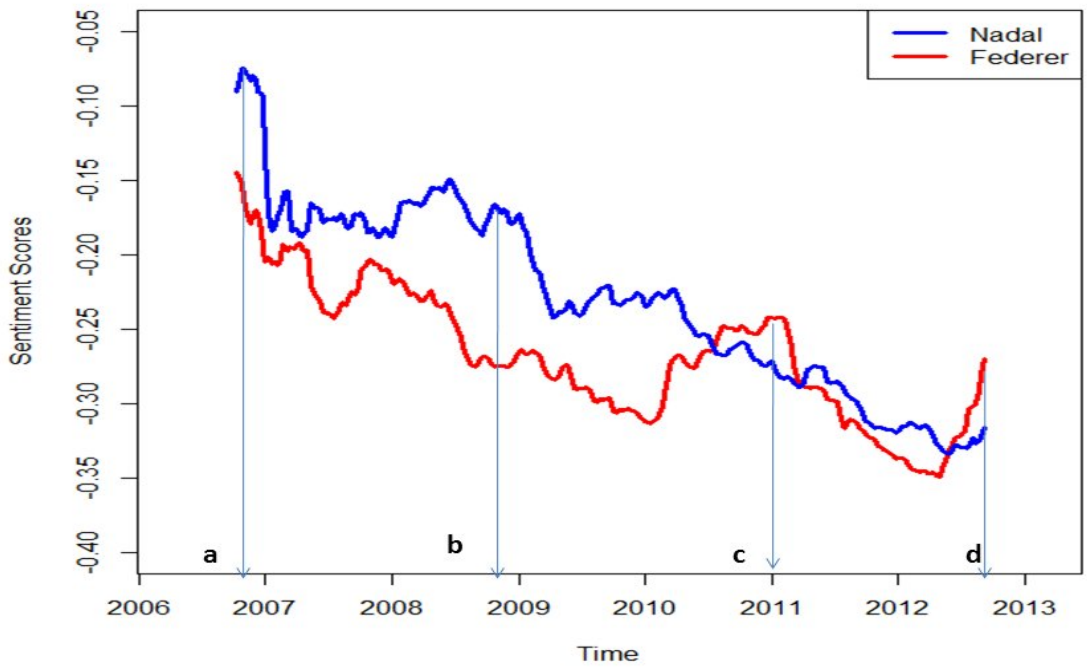

Figure 4-11: Sentiment trends for Federer vs. Nadal

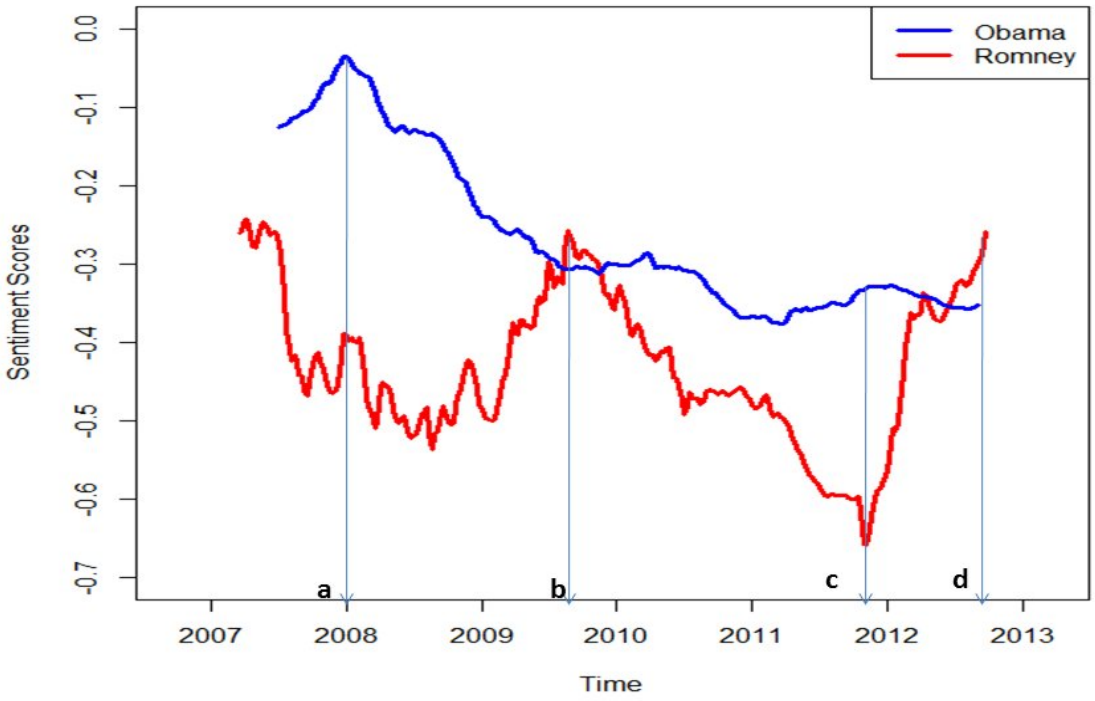

Figure 4-12: Sentiment trends for Obama vs. Romney

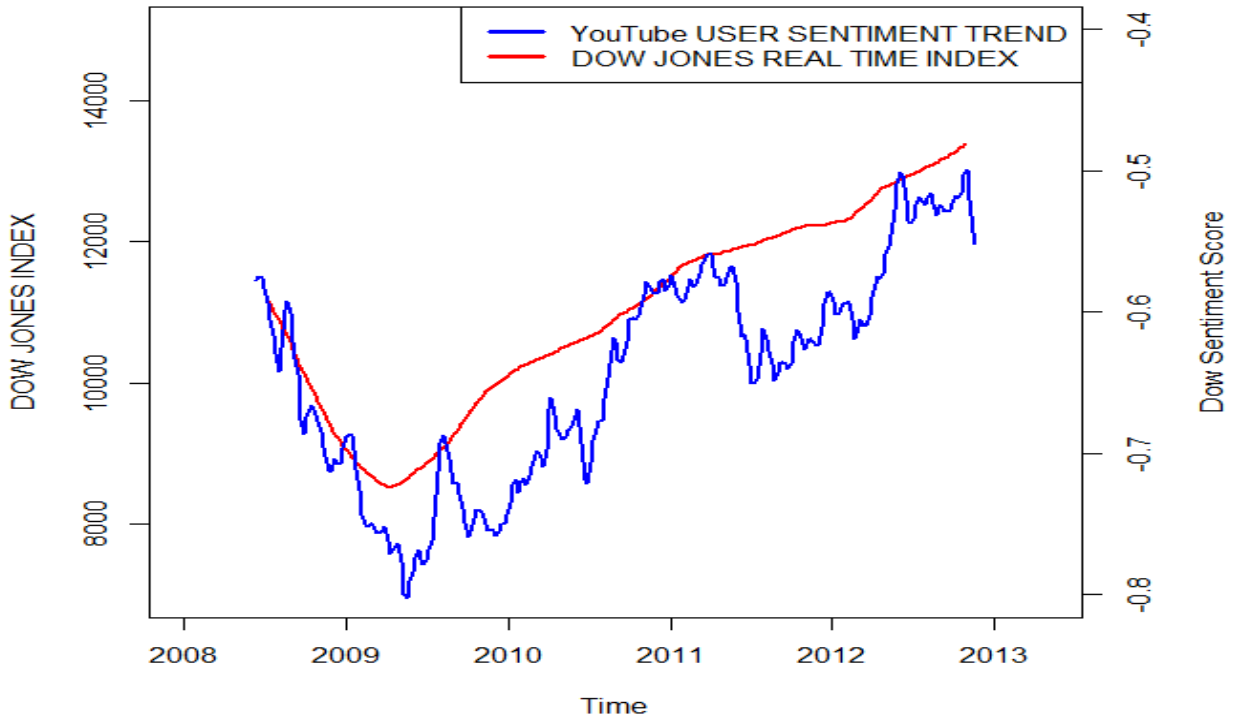

Figure 4-13: Sentiment trends for Dow Jones

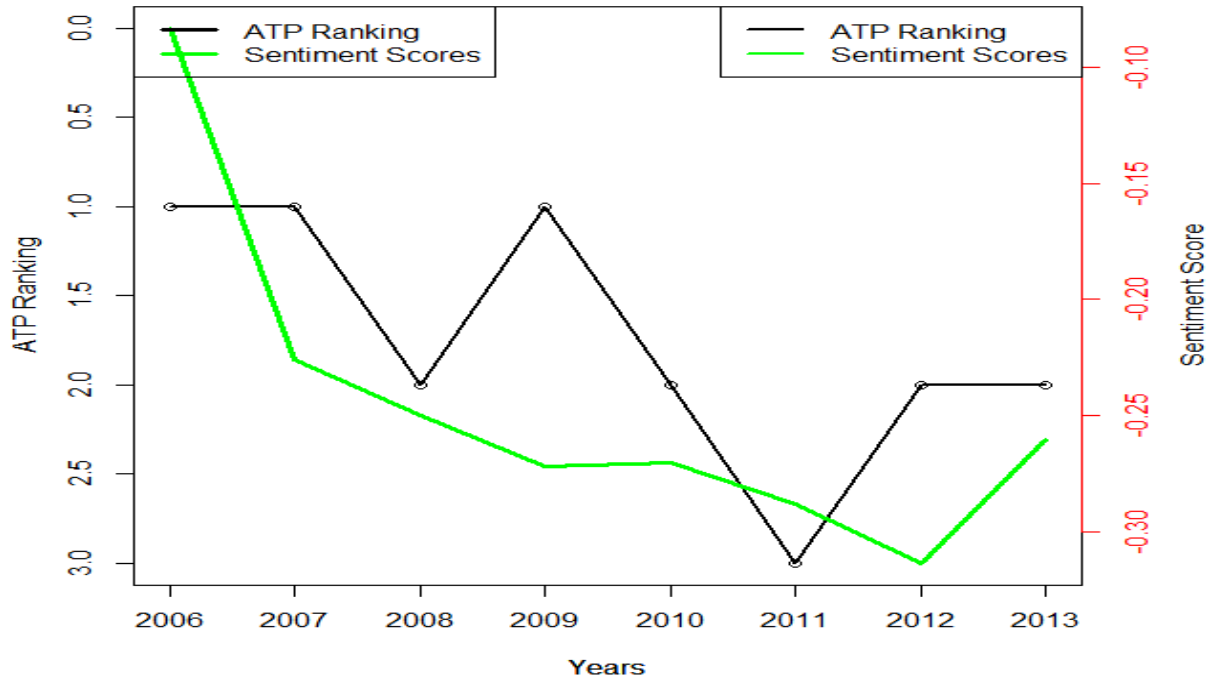

Figure 4-14: Change in sentiment trends with change in Federer's ranking

CHAPTER 5

THREATS TO VALIDITY

5.1. Construct Validity

Before discussing what construct validity refers to , we will give a brief explanation about it. To understand the traditional definition of construct validity, it is first necessary to understand what a construct is. A construct, or psychological construct as it is also called, is an attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories. For example, "overall English language proficiency" is a construct. It exists in theory and has been observed to exist in practice.

Construct validity has traditionally been defined as the experimental demonstration that a test is measuring the construct it claims to be measuring. Such an experiment could take the form of a differential-groups study, wherein the performances on the test are compared for two groups: one that has the construct and one that does not have the construct. If the group with the construct performs better than the group without the construct, that result is said to provide evidence of the construct validity of the test. An alternative strategy is called an intervention study, wherein a group that is weak in the construct is measured using the test, then taught the construct, and measured again. If a non-trivial difference is found between the pretest and post-test, that difference can be said to support the construct validity of the test. Numerous other strategies can be used to study the construct validity of a test, but more about that later. A possible threat to the construct validity of this study is the limited number of keywords we have analyzed. Our goalin the paper is to identify whether the sentiment trends associated with the YouTube users' comments are strongly correlated with the real world events. We have collected

data from different domains such as sports and politics since these are the domains where people have strong opinions. While analyzing more keywords might give additional insight into the trends, our data provides a representative picture of the current trends for the most popular keywords on YouTube. Dealing with more numbers of keywords and videos will give us the chance to plot more points on the time series plot which will give us the insight to reduce the current noises we have in the graphs.

In the future, we will continue to collect data for keywords from different domains to investigate whether we can continue to accurately perform sentiment analysis and forecast future trends. We will be collecting more than 40 million comments to get a good picture of the trends.

5.2. Internal Validity

Internal validity refers to how well an experiment is done, especially whether it avoids confounding (more than one possible independent variable [cause] acting at the same time). The less chance for confounding in a study, the higher its internal validity is.Therefore, internal validity refers to how well a piece of research allows you to choose among alternate explanations of something. A research study with high internal validity lets you choose one explanation over another with a lot of confidence, because it avoids (many possible) confounds.Inaccuracies in our data collection process could be one of the possible threats to this study. We have performedmanual and automated inspections on our dataset and the data collection scripts to verify its accuracy.

5.3. Conclusion Validity

In many ways, conclusion validity is the most important of the four validity types because it is relevant whenever we are trying to decide if there is a relationship in our

observations (and that's one of the most basic aspects of any analysis). Perhaps we should start with an attempt at a definition:

Conclusion validity is the degree to which conclusions we reach about relationships in our data are reasonable. Whenever you investigate a relationship, you essentially have two possible conclusions -- either there is a relationship in your data or there isn't. In either case, however, you could be wrong in your conclusion. You might conclude that there is a relationship when in fact there is not, or you might infer that there isn't a relationship when in fact there is (but you didn't detect it!). So, we have to consider all of these possibilities when we talk about conclusion validity.Although conclusion validity was originally thought to be a statistical inference issue, it has become more apparent that it is also relevant in qualitative research. For example, in an observational field study of homeless adolescents the researcher might, on the basis of field notes, see a pattern that suggests that teenagers on the street who use drugs are more likely to be involved in more complex social networks and to interact with a more varied group of people. Although this conclusion or inference may be based entirely on impressionistic data, we can ask whether it has conclusion validity, that is, whether it is a reasonable conclusion about a relationship in our observations.

In this study, we used the Naive Bayes algorithm to perform sentiment analysis. The use of a single machine learningalgorithm to perform the analysis could be a possible threat to the conclusion validity of this study. However, Naive Bayes has been successfully used by researchers in the past to perform sentiment analysis due to its simplemodel yet high performance for most datasets. In the future, we will continue to evaluate the performance of severalother machine learners such as k-nearest neighbor, Support Vector Machines (SVMs), etc.

5.4. External Validity

External validity refers to "whether the casual relationships can be generalized to different measures, persons, settings and times. However, in an applied discipline, the purpose of which includes working to improve the health of the public, it is also important that external validity be emphasized and strengthened. So, *external*validity refers to the approximate truth of conclusions the involve generalizations. Put in more pedestrian terms, external validity is the degree to which the conclusions in your study would hold for other persons in other places and at other times.

The extent to which the observations from this study can be generalized to other similar studies on other social networking sites is a possible external validity threat to this study. YouTube is the largest and most widely used videostreaming website. The users of YouTube span across different countries. This could mean that the sentiments associated with the collected comments give a close estimation of the actual sentiments of the public. We are also in the process of collecting data from other social networking sites such as Twitter, Facebook and LinkedIn and aim to investigate the extent to which these results can be generalized across multiple social networking sites. We have made our dataset public so that other researchers can validate the results of this study and conduct other types of analyses. All our source codes used in this research are available on Google Code.

CHAPTER 6

CONCLUSION AND OUTLOOK

In this paper we investigate the comments associated with YouTube videos and perform sentiment analysis of each comment for keywords from several domains. We identify whether the trends, seasonality and forecasts of the collected videos provide a clear picture of the influence of real-world events on users' sentiments. We perform sentiment analysis using the Naive Bayes approach to identify the sentiments associated with more than 7 million comments. Analyzing the sentiments over a window of time gives us the trends associated with the sentiments.Results show that the trends in users' sentiments is well correlated to the real-world events associated with therespective keywords. Using theWeka forecasting tool, we are able to predict the possible sentiment scores for 26 weeks into the future with a confidence interval of 95%. While previous studies have focused on the comment ratings and their dependencies to topics, to the best of our knowledge, our work is the first to study the sentiment trends in YouTube comments, with focus on the popular/trending keywords. Our trend analysis and prediction results are promising, and data from other prominent social networking sites such as Twitter, Facebook, Pinterest, etc. will help to identify shared trend patterns across these sites.

In future we will be working on comparing the sentiment trends across various social networks like Twitter , Facebook etc. Currently our on-going research deals with the sentiment trends related to the tweets and its comparison with the YouTube trends. We are also using k-nearest neighbor classifiers instead of Naive Bayes Classifier. Our research lays out a good platform for researchers to focus on social media sentiment analysis with more precision . Topics like social media text classification using different classification techniques (higher accuracy rate) is one of the most prominent area for future research.

CHAPTER 7

REFERENCES

1. YouTube. Youtube statistics. http: //www.youtube.com/yt/press/statistics.html, 2013.

2. MirjamWattenhofer, Roger Wattenhofer, and Zack Zhu. The Youtube social network. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM, 2012.

3. Alan Mislove, MassimilianoMarcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pages 29-42, 2007.

4. Stefan Siersdorfer, SergiuChelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: analyzing and predicting Youtube comments and comment ratings. In Proceedings of the 19th International Conference on World Wide Web, pages 891-900, 2010.

5. FabricioBenevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. Identifying video spammers in online social networks. In Proceedings of the International Workshop on Adversarial Information Retrieval on the Web, 2008.

6. Ashish Sureka, PonnurangamKumaraguru, AtulGoyal, and SidharthChhabra. Mining Youtube to discover extremist videos, users and hidden communities. In 6th Asia Information Retrieval Societies Conferences, AIRS, pages 13-24, 2010.

7. France Cheong and Christopher Cheong. Social Media Data Mining: A social network analysis of tweets during the 2010-2011 Australian Floods. In Pacific Asia Conference on Information Systems, PACIS, 2011.

8. Bo Pang, Lillian Lee, and ShivakumarVaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002.

9. Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Journal of Computer Science, 2(1):1-8, 2011.

10. Brendan O'Connor, RamnathBalasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM, 2010.

11. EfthymiosKouloumpis, Theresa Wilson, and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM, 2011.

12. Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM, 2011.

13. Prem Melville, WojciechGryc, and Richard D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1275-1284, 2009.

14. Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture, pages 70-77, 2003.

15. Ellen Spertus. Smokey: automatic recognition of hostile messages. In Proceedings of the 14th National Conference on Artificial Intelligence, pages 1058-1065, 1997.

16. Sven Meyer zuEissen and Benno Stein. Genre classification of web pages: User study and feasibility analysis. In Advances in Artificial Intelligence, pages 256-269, 2004.

17. Maya Dimitrova, Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Web genre visualization. In Proceedings of the Conference on Human Factors, 2002.

18. E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In Proceedings of the 18th Conference on Computational Linguistics - Volume 2, pages 808-814, 2000.

19. Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. Web classification using support vector machine. In Proceedings of the 4th International Workshop on Web Information and Data Management, pages 96-99, 2002.

20. Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, and Wei-Ying Ma. Web-page classification through summarization. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 242-249, 2004.

21. Min-Yen Kan. Web page classification without the web page. In Proceedings of the 13th International World Wide Web Conference, pages 262-263, 2004.

22. Min-Yen Kan and Hoang Oanh Nguyen Thi. Fast webpage classification using url features. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pages 325-326, 2005.

23. Pavel Calado, Marco Cristo, Marcos Andre Goncalves, Edleno S. de Moura, Berthier Ribeiro-Neto, and NivioZiviani. Link-based similarity measures for the classification of web documents. Journal of the American Society for Information Science and Technology, 57(2):208-221, January 2006.

24. Marco Cristo, Pavel Calado, Edleno Silva de Moura, NivioZiviani, and Berthier A. Ribeiro-Neto. Link information as a similarity measure in web classification. In String Processing and Information Retrieval, 10th International Symposium, SPIRE, pages 43-55, 2003.

25. Andrej Karpathy. Sentiment analysis example. http://karpathy.ca/mlsite/lecture2.php, 2011.

26. Mark Hall, Eibe Frank, Geo_rey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. SIGKDD Explorer Newsletter, 11:10-18, November 2009.

27. WEKA Wikipages. Time series analysis and forecasting with weka. http://wiki.pentaho.com/ display/DATAMINING/Time+Series+Analysis\ linebreak+and+Forecasting+with+Weka, 2013.

28. X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of internet short video sharing: YouTube as a case study. In Technical Report arXiv:0707.3670v1 cs.NI, New York, NY, USA, 2007. Cornell University, arXiv e-prints. (YouTube traffic statistics)

29. Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, 2006. (available online PDF (http:/ / citeseerx. ist. psu. edu/ viewdoc/ download?doi=10. 1. 1. 122. 5901& rep=rep1& type=pdf)) (Naive Bayes classification)

30. User Sentiment Detection: A YouTube Use Case by SmitaShree Choudhary and Josh Breslin

31. Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation By Lina McInerney et al. Published in Social Network Analysis and Mining, 2009. ASONAM '09.

32. Sentiment Analysis amidst Ambiguities in YouTube Comments on Yoruba Language (Nollywood) Movies ByOrimaye Sylvester Olubolu et. al.

33. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web By Louis-Philippe Morency et al.

34. C. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 347-354, Vancouver, Canada, 2005.

35. Sentiment analysis of Social Media by A Kowcika et al.

36. K. Balog, G. Mishne, and M. de Rijke. Why are they excited? identifying and explaining spikes in blog mood levels. In Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006), 2006.

37. Twitter Sentiment Analysis: The Good the Bad and the OMG! byEfthymiosKouloumpis et al.

38. Capturing Global Mood Levels Using Blog Posts by GiladMishne et al.

39. Sentiment Knowledge Discovery in Twitter Streaming Data by Albert Bifet and Eibo Frank

40. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, pages 10-17, 2010.

41. A. Go, L. Huang, and R. Bhayani. Twitter sentiment classification using distant supervision. In CS224N Project Report, Stanford, 2009

42. S. Petrovic, M. Osborne, and V. Lavrenko. The Edinburgh Twitter corpus. In

43. #SocialMedia Workshop: Computational Linguistics in a World of Social Media ,pages 25-26, 2010.

44. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding micro-blogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pages 56-65, 2007

45. Feature Selection for Web Page Classification by Daniele Riboni

46. Determining the Sentiments of Opinions by Soo-Min Kim and Eduard Hovy

47. Naive Bayes for Text Classification with Unbalanced Classes by Eibe Frank and Remco R. Bouckaert

48. Multimodal Sentiment Analysis of Social Media by Diana Maynard et al.