BUSINESS INTELLIGENCE

FROM SOCIAL MEDIA

A THESIS
Submitted by

PRETISH CHACKO KURUVILA B110747CS
VINODH S B110099CS
VISHNURAJ V B110396CS

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Under the guidance of
Dr. GOPAKUMAR G

DEPARTMENT OF COMPUTER ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
NIT CAMPUS PO, CALICUT
KERALA, INDIA 673601

January 19, 2017

ACKNOWLEDGEMENTS

The success of our project was to a large extent due to the effort and guidance of our professors. We take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

We would like to thank our project guide Mr. Gopakumar for his guidance and encouragement with the project without which we wouldnt have been able to complete the project.

We would also like to thank Mr. Bharat Narayan, who took keen interest in our project and guided us all along, till its completion by providing us with all the necessary information and systems required for the project to work.

We are thankful as well to the members of the Computer Science Department for their help and support.

PRETISH CHACKO KURUVILA

VINODH S

VISHNURAJ V

DECLARATION

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text".

Place:
Date:

Signature :
Name :
Reg.No:

CERTIFICATE

*This is to certify that the thesis entitled: "***BUSINESS INTELLIGENCE FROM SOCIAL MEDIA***" submitted by Sri/Smt/Ms*

PRETISH CHACKO KURUVILA B110747CS
VINODH S B110099CS
VISHNURAJ V B110396CS

to National Institute of Technology Calicut towards partial fulfillment of the requirements for the award of Degree of Bachelor of Technology in Computer Science Engineering is a bonafide record of the work carried out by him/her under my/our supervision and guidance.

Signed by Thesis Supervisor(s) with name(s) and date

Place:
Date:

Signature of Head of the Department

Office Seal

Contents

Chapter

Abstract

The aim of this project is to develop a Visual Analytics Toolkit (VAT) that combines data mining with statistical techniques. VAT consists of a series of linked visualization views that is generated using social media consumer sentiment. Sentiment Analysis is used to derive the user sentiment from the raw data which is extracted from Twitter. This toolkit can be used in business intelligence systems where it is invaluable in fields such as social media advertising and predicting sales forecasts. if@

Chapter 1

Introduction

With the increasing amount of information coming from the internet as well as other sources such as logs and social networks, companies and institutions have need of tools that can take in all that information, filter the unnecessary parts and can show that data in a manner that is useful to the organization. To create such a tool requires a study in various areas such as Sentiment analysis, Parallel computing, Machine learning and Big Data.

Sentiment Analysis also known as Opinion Mining is the process of classifying sentences into positive, negative and neutral depending on the nature of the sentence and the manner in which it was meant.

Parallel Computing is based on the concept that one large problem can be split into multiple smaller problems to be solved separately which is then combined at the end. Here, we use multiple nodes to split the large amounts of data that needs to be processed so as to acquire the result quicker.

Machine Learning is the study of creating algorithms that can learn from, sort, and classify data on its own. We use it to classify the tweets upon retrieval into positive, negative, neutral and irrelevant.

Big Data is a generally used term to describe the processing extra-large data sets that cannot be processed through normal methods. Using Hadoop and Mahout we created the cluster to accomplish the training of the data that could not be done on a single node.

Chapter 2

Literature Survey

Early Business Intelligence (BI) systems operated using internal data sources as input. [1] suggested using the data from real time analysis and the development of web based intelligence systems was proposed by Chung and Chen [2]. Abbasi and Chen [3] suggested in their research the use of a framework advocating the usage of systems capable of displaying the information derived from Computer-Mediated Communication (CMC) text. One important method is Text mining which is the extraction of all relevant data from text. It is similar to the field of sentiment analysis which refers to the identification and extraction of select information as described by Pang and Lee [4] in their study. Liu [5] used two tools, OpinionFinder and SentiWordNet, to verify that the quantity of Word of Mouth (WOM) is the most important among valence , subjectivity, number of sentence, and number of valence words.

As Twitter became popular, much research was done in relation to it. Asur and Huberman [6] conducted a study and came to the conclusion that using social media one could predict the outcomes of events. Bollen et al. [7] after analysing tweets from Twitter daily then used a Self-Organizing Fuzzy Neural Network (SOFNN) to get the Dow Jones Industrial Average's closing values.

There was a drastic reduction in the number of prediction errors. After further study Romero et al. [8] analyzed the manner in which hashtags on a network created using interactions among Twitter users were spread, and their

results confirmed the complex contagion principle from sociology. OConnor et al. [9] in their study using several surveys on consumer confidence and political opinion found that these surveys were linked with the number of times the sentiment word occured in the tweets. The results showed how useful text streams could be as a replacement for or addition to traditional polling. Shi et al. [10] studied the behaviour of Twitter users in re-tweeting, and the results showed that weak ties are more probably lead to retweeting.

In their study in classification using algorithms and mapreduce, Ayma et al. [11] showed that when using a cluster the speed with which the data is processed increases with the amount of data used and that by increasing the number of nodes the processing time did not have to necessarily drop. Among the various classification algorithms, the SVM classifier is considered as one of the best and most efficient. In cases of large problems using parallel SVM based on mapreduce is even better in terms of the computation time as Zhanquan Sun and Geoffrey Fox [12] found out in their research on the topic.

Chapter 3

Problem Definition

To develop a real-time application that calculates user sentiment about a company using input from social networks and visualizes it in a manner that will help organizations in making better informed decisions in regards to their products, advertising campaigns and sales forecasts.

Chapter 4

Design

The user can select a company from a provided list of companies if he/she wishes to see the visualized results from data analyzed over a period of time or he/she can search Twitter using any keyword which can be entered in the trends search option. If the latter option is used data is acquired at that instant for processing.

4.1 Data retrieval

In the case of streaming analysis,we use the Twitter API to retrieve the input data for the application. The API returns the data in JSON format. A schema is then created with 24 attributes(tweet id, tweet, created at, screen name, name, followers count, friends count, statuses count, time zone, retweet count, location, user mentions1, user mentions2, hashtags1, hashtags2, hashtags3, hashtags4, URL1, URL2, latitude, longitude, place name, place type, timestamp), after which the data is stored in CSV(comma-separated-value) format.

In the case of processed data, tweets were extracted on a regular basis for a period of two weeks and stored.

4.2 Preprocessing

Before processing data for sentiment analysis we need to remove the noise contained in it. This preprocessing stage involves, data filtering, tokenization,

removing stop words. We implement a machine learning approach so as to classify and assign sentiments to the extracted tweets. This is the same for both the streaming and the stored data.

4.3 Training and Testing Phase

A feature list is first generated containing frequently reoccurring words from the training data. A feature vector is then created for each and every tweet using the contents of the feature list. Along with the polarity value the feature vector is then fed into the Naive Bayes Classifier [11] and Support Vector Machine [12] to train.

The feature vector for each test tweet is created, and these feature vectors run through the trained classifier to get the polarity and to find the accuracy of the classifier. The obtained polarities of all the tweets are tabulated. Using the resulting table we plot graphs showing the sentiments expressed in regards to a particular company. Figure 4.1 shows the flow diagram of the system:

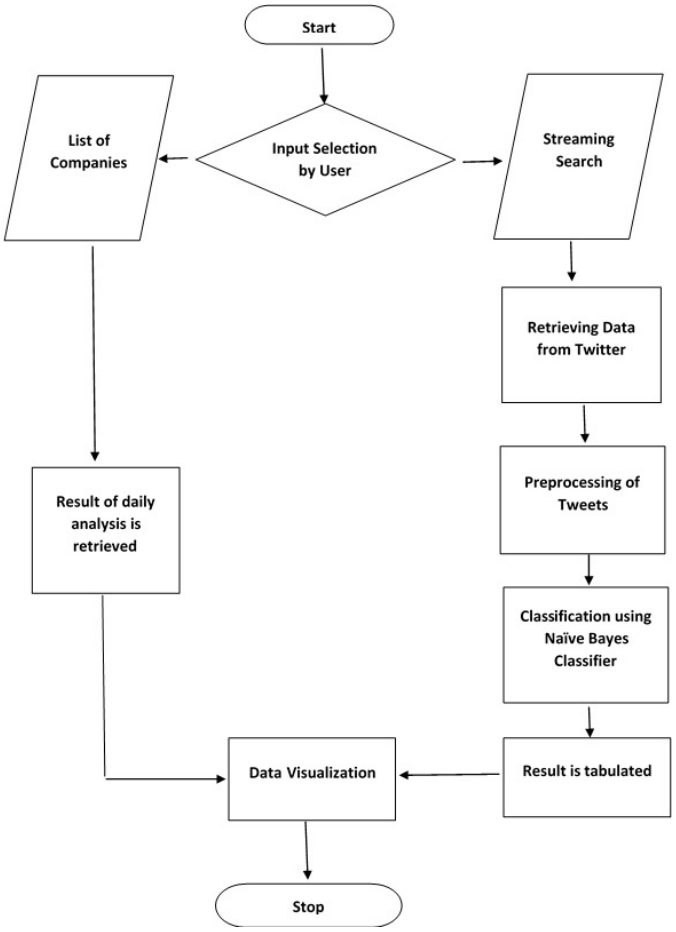

Figure 4.1: Flow Diagram of System

Chapter 5

Implementation

5.1 Data Acquisition

Data was retrieved from twitter and using their respective APIs. Schema for the acquired data was created in a local system.

5.2 Data Preprocessing

The data that we acquired contains noise which need to be preprocessed. As the first step we start out with filtering. To visualize the results all the unrequired data such as HTML code and image links are removed leaving just the tweet, timestamp, latitude, longitude, retweets and follower count. The file is then converted into CSV format.

5.3 Sentiment Analyzer using Machine Learning

Using the Hadoop cluster consisting of 4 nodes with Mahout (Machine learning library), two separate classification models were created one for live streaming data using the Naive Bayes classifier and one for the preprocessed data using a training set containing 1.5 million tuples. The cluster was used, as the model generation could not be done on a single node and required a group of computers working in tandem. It split the workload among each node which improved the processing time.

5.4 Data Visualization

Using the classified data we generate an opinion graph that plots the number of tweets extracted in regards to the company against the day of the week it was extracted on. With it a count of positive, negative and neutral opinions expressed daily regarding the company is plotted as well. The precise location from where the tweets were posted is drawn on a map of the world along with user's statistics which is an indicator of how popular the company is among tweeters. Along with these a tag cloud that shows the frequently occurring words in the positive and negatively classified sentences is displayed as a means of identifying areas of user discontent in the company.

5.5 Challenges Faced

In the course of our project, we came up against quite a few problems. Due to the large amount of data that had to be used to generate a classifier model it was deemed not possible to do so on a single system. As a result of which we created a cluster to help us in doing so.

5.5.1 Cluster Setup

The cluster creation had its own issues where a large amount of time was spent in trying to install a cluster manager, but due to the numerous errors that popped up, its installation could not be completed either. This forced us to install each and every component of the Hadoop ecosystem manually one by one on all the nodes in the cluster, though it did help us in gaining a better understanding of the environment. Due to unfixed issues with the framework in which the data node would abruptly stop or the UI would not appear we had to repeatedly format and install each component again.

5.5.2 Data Extraction

Twitter had imposed a limit on the amount of data(1 to 40 percent of the actual tweets) one could retrieve using the Streaming API. Therefore the amount of relevant data that could be used after days of extraction was comparatively small in size. The only way in which complete access to all the tweets that were generated could be achieved was if we purchased it at a rate of 3000 dollars per month using the Firehose API.

5.5.3 Training Data

A part of every machine learning classifier model development is that a preferably large amount of training data is given to improve the classification of the tweets. Due to the fact that twitter had restricted the amount of time a person could store tweets in any form online, we had to manually train the tweets ourselves.

Chapter 6

Result

6.1 Classifier

An analytics toolkit that can classify tweets with an accuracy of 80.43 percent for the SVM classifier and 82.86 percent for Naive Bayes classifier which is the current standard in terms of accuracy in sentiment analysis is obtained.This was obtained using a training set of 1.5 million tuples for Naive Bayes and 10,000 tuples for SVM.

6.2 Visualization

The classified data is then visualized in a variety of ways using the information that comes with the tweet such as location, time created and hashtags to find out what the customer thinks about the product and whether it be positive or negative identifies the areas they are happy or discontent with.

6.2.1 Opinion Count Graph

The graph plots out the number of opinions received on every day of the week in regards to the stored data and in the case of the data that is taken in at that point of time the graph shows the rate at which the tweet comes in.

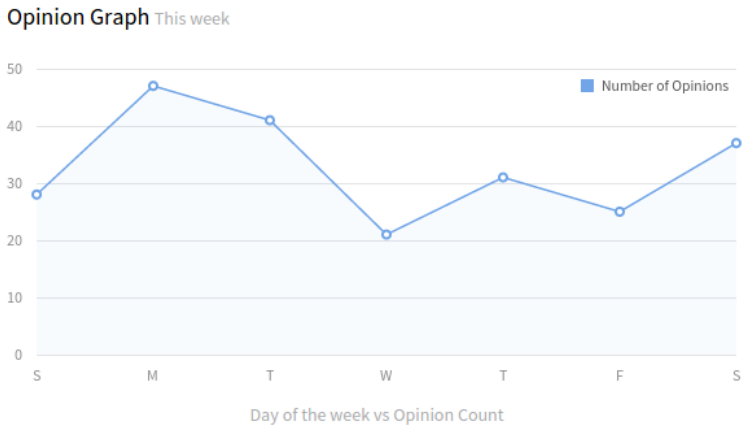

Figure 6.1: Day of the week vs Opinion Count for Apple

6.2.2 Polarity Graph

This graph lists the number of positive, negative and neutral opinions received on each day the data was extracted.

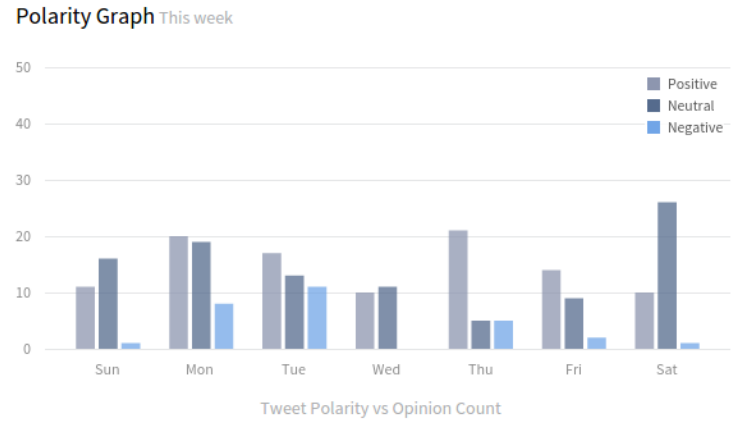

Figure 6.2: Tweet Polarity vs Opinion Count for Apple

6.2.3 Tweet Locator

The location map that plots the area from where the tweet originated.

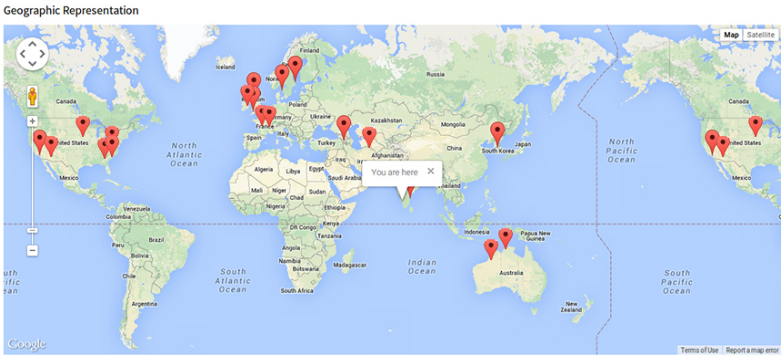

Figure 6.3: Tweet Locator for Apple

6.2.4 Company Statistics

A comparison is done between the data that is retrieved for the company selected and the tweets extracted for the rest in terms of the number of positive, negative and neutral opinions.

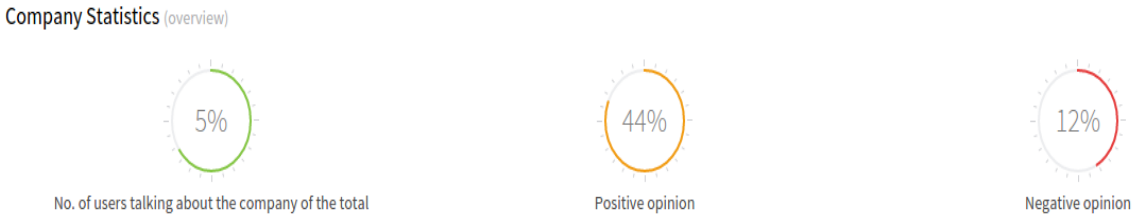

Figure 6.4: Company Statistics for Apple

6.2.5 Tag Cloud

A collection of words occurred repeatedly in the tweets are grouped together according to their tweet polarity. Which can be considered as a good indicator for positive and negative attribute of the company.

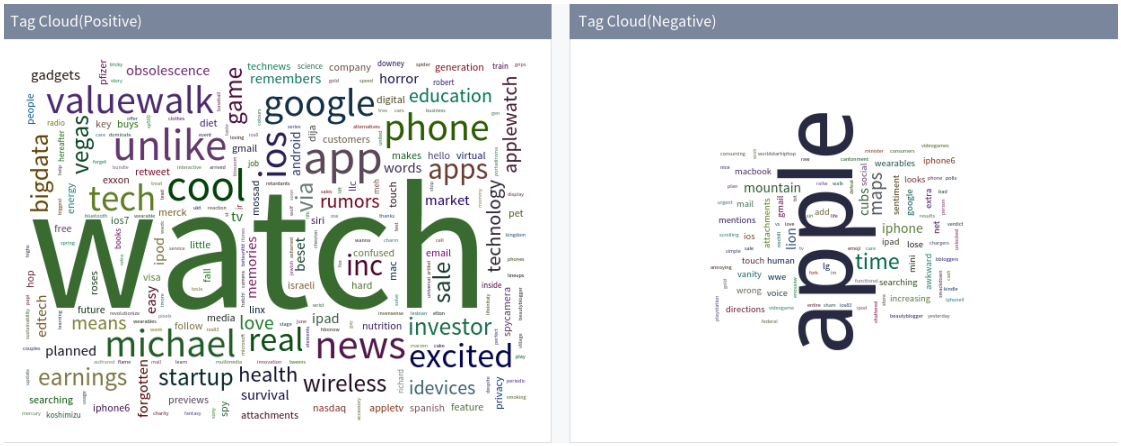

Figure 6.5: Tag Cloud (left:positive, right:negative)

Chapter 7

Conclusion

A Visual Analytics Toolkit is developed that can perform an analysis of live data i.e. Streaming Analysis as well as that of stored data and visualizes it in a manner that can be utilized by anyone in an organizations that seeks to derive insights from data. It can be used to identify trends within a business and a market that can affect an organization. Businesses are able to identify factors that affect their product quality and what their customers think about them. The toolkit is designed in such a manner that anyone can operate it and get their results accurately as well as instantaneously

Also in the development of this project we used lexical analysis as well as machine learning. We learned that the usage of lexical analysis would require a large amount of time to classify the data as compared to machine learning, where after the initial training of data using the classifier the classification could be done in a matter of seconds. Creating a cluster, running tasks in parallel, use of open-source frameworks such as Apache Hadoop were things we studied in detail and implemented in the development of this Toolkit.

Bibliography

[1] B. H. WATSON H. J. REYNOLDS A. M. WIXOM and HOFFER J. A. Continental airlines continues to soar with business intelligence. Inf. Syst. Manage, 19:102–112, 2008.

[2] W. CHUNG and H. CHEN. Web-based business intelligence systems: a review and case studies. Handbooks in Information Systems: Business Computing, G. Adomavicius and A. Gupta Ed.:373–396, 2009.

[3] A. ABBASI and CHEN. Cybergate: A design framework and system for text analysis of computer mediated communication. MIS Quart, 32(4):811–837, 2008.

[4] B. PANG and L LEE. Opinion mining and sentiment analysis. 2008.

[5] Y. Liu. Word of mouth for movies: Its dynamics and impact on box office revenue. J. Market, 70(3):74–89, 2006.

[6] S. ASUR and B. A. HUBERMAN. Predicting the future with social media. Proceedings of the ACM International Conference on Web Intelligence, pages 128–137, 2010.

[7] J. MAO H. BOLLEN and X. ZENG. Twitter mood predicts the stock market. J. Comput. Sci., 2(1):1–8, 2010.

[8] D. M. MEEDER B. ROMERO and J. KLEINBERG. Differences in the mechanics of information diusion across topics: Idioms, political hashtags, and complex contagion on twitter. Proceedings of the 20th ACM International World Wide Web Conference., 2011.

[9] B. BALASUBRAMANYAN R. ROUTLEDGE B. R. OCONNOR and N. A. SMITH. From tweets to polls: Linking text sentiment to public opinion time series. Proceedings of the International AAAI Conference on Weblogs and Social Media, 2010.

[10] Z. RUI H. SHI and A. B. WHINSTON. Information sharing in social broadcast: Evidences from twitter. Proceedings of the 21st Workshop on Information Systems and Economics., 2010.

[11] P. Happ D. Oliveira R. Feitosa G. Costa A. Plaza P. Gamba V. A. Ayma, R. S. Ferreira. Classification algorithms for big data analysis, a map reduce approach. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-3/W2, 2015.

[12] Zhanquan Sun and Geoffrey Fox. Study on parallel svm based on mapreduce.