Social Analysis

A Data Science Theoretical Approach to Big Data Analytics

[1]Ramin Agha Amin, [2] Kunwar Babar Ali

[1]M. Tech(SE) Student, Noida International University, Greater Noida, UP 203201
[2]Assistant Professor, Dept of CSE, Noida International University, Greater Noida, UP 203201

Abstract: Current analysis of social sciences approaches in data analysis can be divided in to four part which are 1.text analysis, 2.social network analysis,3.social complexity, and also 4. social simulations However, when it comes to organizational and societal units of analysis, there exists no way to explain the concept, articulate the model, analyze the data, explanations, and prediction social media interactions as people' associations with thoughts, values, identities, opinions and so on. To solve this limitation, based on the analysis of data science and ML (Machine Learning, this paper presents a new way to data analytics called Social Analysis: A Data Science Theoretical Approach to Big Data Analytics

Index Terms **– Social Analysis, Customer Behavior, Sentimental Analysis, Data Science, Python, Machine Learning, Analysis, Trend Analysis, Forecasting, Opinion Prediction.**

1. INTRODUCTION

With increased development in marketing and the customer-oriented advertising concepts, the workshops, companies, firms and organizations find it vital to know the customer's behavior towards its product. The companies have to get response and reply from the customers in order to learn and grow themselves. With the evolution of it, the social media plays a huge role in expressing opinions and the views of users and social networks are acting a key role in this context. The firms may find it very problematic to get the customer response effectively and efficiently through questionnaires procedures or face to face interviews and because of that the companies may start to look for other approaches of getting the customer feedback and sentiments. That's when the necessity of a sentiment, emotions, feeling, and behavior analysis system comes into existence as it is really operative to know the customer's idea about a specific brand by analyzing what he or she posts on social media, for an example on Facebook. This has a vast corporate value because companies discovery info on how the market think about them extremely seriously.

This data is one of the most genuine and accurate customer responses an organization can get. But gathering this data is quite a challenging mission because the bulk of data that needs to be processed is gigantic. After a proper filtering [1], the data can be analyzed and the ideas can be given a positive, negative or a neutral division. If this information is used to offer a service to corporations, it can be depicted for the client giving them a decent idea of the success or fame of their brand.

It would be really valuable for the corporations to know what type of folks like their brands. For an example, if it is possible to analyze and identify that the people who like Apple products are attentive in Music and Painting then Apple can target those market parts when they are evolving their subsequent product and they can widen up their market to provide to new parts of customers. As a consequence of that, the profiling of the customers has been done with their interests and deliver these data to the related mobile firms with the intention of making their marketing operations successful. Also, the trend of the fame of the relevant brands and with that information the companies can identify the increasing or reducing interest in customers on their brand. If graphical information on these data can be provided, then the firm can decide when to release their next product what are their fruitful and unproductive products.

2. LITERATURE REVIEW

In the current context, Social Media has become a significant part in one's life and if we can analyze a person's social media behavior, we will be able to profile them and predict their behavior under certain circumstances. The application we developed has this feature of user profiling which will be helpful for the firms to make their decisions based on target markets and target customer segments. For an example if we are seeing the 'Apple Inc' company we can recognize apple users and their preferences and this will help Apple to design their next product or services in a more user-friendly and easier to use way. In other words, now it is not necessary to ask questions from customers to get their opinion towards certain products, we can get their view by analyzing their online conduct such as comments and status updates on social media for example Facebook and Twitter. 'Real-World Behavior Analysis through a Social Media Lens' research proposes that there is a robust connection among a person's online behavior and the real-world behavior. The researchers say that it is likely to predict the real-world actions of a person by analyzing his online behavior. In their context, they have designated the community based on characteristics such as race, ethos, nation, etc. In this request, we can hand-pick the community as a set of users who uses a certain product or a brand because we are designing the application to help IT based companies. Data collection can be happening through Facebook. Then the text processing is to be performed on the collected data. In the research, they have used a diverse technique to analyze data and we are executing a sentiment and semantic analysis on the collected data. They have used correlational analysis to recognize word groups whose occurrence of mention was most meaningfully related statistically to the magnitude of social action during the similar period, then used multivariate regression analysis to allocate quantities to them for predictive analysis. In this application, we will be able to predict the user's view and attitude towards certain brands and products in the future. The prediction of event, Extraction of Attitude, Detection of key people and Mood analysis are some of the results of the analysis and these are very useful for this application as well.

Social media are actually scalable communications technologies that revolution Internet based communications into a communicating dialogue environment [13]. On the demand side, the users and the creators, and manufacturer are increasingly rotating to various types of social media to search for data and information and to make personal and business choices concerning views/ideologies, services/products, personalities/politicians, and public services. On the ``supply-side", terms such as ``Enterprise 2.0" [3] and ``social business' 'are being used to describe the emergence of private enterprises and public institutions that strategically adopt and use

social media channels to increase organizational effectiveness, enhance operational efficiency, empower employees, and co-create with stakeholders. The organizational and societal adoption and use of social media is generating large volumes of unstructured data that is termed Big Social Data. New organizational positions such as Social Media Manager, Chief Listening manager, Chief Digital Officer, and Chief Data Scientist have emerged to meet the technological developments, organizational evolution, market expectation, and societal transformations.

However, the current state of art and practice regarding social media engagement is encountered with numerous technological problems, scientific questions, operational management issues, managerial challenges, and training deficiencies. As such, not many organizations are generating competitive advantages by extracting meaningful facts, actionable insights and effective outcomes from Big Social Data analytics. also, there are unsolved problems regarding how Big Social Data integrates with the existing data sets of an organization (that is, data from internal enterprise systems) and its relevance to the organization's key performance factors. To address these diverse but interrelated issues, this paper presents a novel data science approach to Big Data Analytics in general and Big Social Data Analytics in particular for Facebook, Twitter and other social media channels. Specially, this paper introduces a research program situated in the domains of Data Science [5] [7] and Computational Social Science [8] with practical applications to Social Media Analytics in organizations [4], [9], [10]. It addresses some of the important theoretical and methodological limitations in the emerging paradigm of Big Data Analytics of social media data [11].

From an academic research stand-point, Social Set Analysis addresses two major limitations with the current state of the knowledge in data Science: (i) a vast majority of the literature is on twitter datasets with only 5% of the papers analyzing and studying Facebook data raising representativeness, validity and methodology concerns [11], and (ii) mathematical structuring of social data hasn't evolved beyond the four dominant approaches [12] of text analysis (information extraction and classification), social complexity analysis (complex systems science), social network analysis (graph theory), social simulations (cellular automata and agent-based modelling). To put it honestly, currently we don't have deep academic knowledge of the most dominant action on social media platforms performed by hundreds of millions of unique users every day: ``like" on Facebook. In fact, as Claudio CiofRevilla (2013), one of the founding parents of the field of Computational Social Science, observed: Reliance on the same mathematical structure every time (e.g., game theory, as an example), for every research problem, is unfortunately a somewhat common methodological pathology that leads to theoretical decline, rejection and a sort of inbreeding visible in some areas of social science research. Dimensional empirical features of social phenomena-such as discreteness-continuity, deterministic-stochastic, fnite and infnite, contiguous-isolated, local-global, long-term versus short-term, synchronic-diachronic, independence-inter-dependence, among others-should determine the choice of mathematical formulation.

2.1 Classification of User Behavior and Info Propagation on a Social Media Network

Users of these online social media application often play vital roles that can be presumed from observations of users' online activities. Researchers have presented five types of users with common observed online behaviors, where all these clients also depict correlated profile properties. Their main focus is on the multimedia shared by users of social media networks and they all bring out the psychological side of the social media online and offline. The research focuses on two main facets. (i) the correlations that exist between users' (demographic and psychological) profiles and the latent roles that emerge from their online behavior, and (ii) how the characteristics of broadcasts influence their popularity. In this approach the users are categorized based on their behavior similarities and clustering algorithms are used in order to group users. They group users according to features of their behavior on Facebook, but then also review the demographic and psychological data associated with each cluster to interpret the inferred formal roles. To gather data from Facebook they have used a set of volunteers and they were asked to answer to an online survey which will gather data about their personalities. This approach differs from our approach because we are only gathering data directly from a Facebook application that we created to get the status updates of the users. After gathering data, they are analyzed and the users are clustered based on behavior features and the variation in broadcast popularity is measured with properties of the broadcast in order to identify common characteristic of successful broadcasts. They have preprocessed data. A common data mining approach to extracting entity groupings is the application of the standard K-means algorithm. This research was done mainly to group and cluster the users. They have not focused on doing sentiment and semantic analysis on the gathered data. They were using a different technique to identify the frequently repeated words and cluster the users accordingly. Our approach goes beyond this as we are clustering the users after doing a sentiment and semantic analysis which will understand what the user has really meant through his/her status updates. This approach measures the popularity of the posts by analyzing the number of likes and comments these status updates have received but, in our approach, we are not taking the popularity of the posts to account. [3]

2.2 Recognizing Personality Traits Using Facebook Status Updates

Through this research the researchers have made an attempt to understand user traits by referring to Facebook status updates. They have referred the five basic personal traits namely extraversion, neuroticism (the opposite of emotional stability), agreeableness, conscientiousness, and openness to experience for this study. Further, given a particular status the researchers have focused on four distinct features such as LIWC (Linguistic Inquiry and Word Count) features, Social Network features, Time-related features and other features of that single status. As many traits can be possessed by an individual, they have trained a binary classifier to separated users who display the personal trait from people who do not. They have compared the performance of three learning algorithms trained on these features, namely Support Vector Machine with a linear kernel (SVM), Nearest Neighbor with k=1 (kNN) and Naive Bayes (NB). Results being yielded, they were able to determine that depending on the trait focused on, the successful classifier varied. A shortage that can be identified in this attempt is, it doesn't provide any insights to business intelligence relating to any business domain. [4]

2.3 Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews

This research has been done aiming one of the most challenging tasks which is to achieve a higher level of sentiment classification using Natural Language Processing techniques. With the modifications done to the Maximum Entropy algorithm the researchers have classified sentiments taken from movie reviews published by various type of people. The "Customer Behavior Analysis for Social Media" also uses the Maximum Entropy algorithm with the help of Apache OpenNLP to classify the sentiments gathered from social media.

Since people's preferences may change due to many reasons and the level of knowledge of the reviewers about the industry could vary, the researchers have worked on doing a personalized classification. Expert reviews would be easier to classify as positive or negative and on the other hand it will be very hard to separately identify reviews of normal viewers. While customizing the classification according to the latter mentioned problem, the personalization has been done to remove the unfairness of using a single method for every user with different tastes and likes.

While Customer Behavior Analysis for Social Media gather sentiment data from Social media using applications this project has gathered data from www.imdb.com. Classified reviews have been taken for the analysis directly from the website and been saved into files respectively to the movie. This project uses a small corpus and is not recognizing semantic and linguistic features which decrease the accuracy of the results. [5]

2.4 TWITTER DATA COLLECTING TOOL WITH RULE-BASED FILTERING AND ANALYSIS MODULE

The approach of this research is to gather data from twitter using the twitter API and then use a custom-built tool to do the data analysis. The tool continuously gathers data from tweeter and stores it in a database. The data gathering approach that we have implemented is very similar in many ways. The contrasting difference being that we gather data from Facebook instead of Twitter. Since Facebooks privacy policy is rather complex in comparison to that of Twitter the data need to be collected using a Facebook app. Using an app, it is possible to get the users to allow us to take their data using Facebook permissions. The app having taken permission sends http requests to the Facebook graph API and the data is returned in JSON, this information is stored in a database at the server and is then further filtered and used for data mining applications. [6]

2.5 SEMANTIC ANALYSIS BASED ON ONTOLOGIES WITH SEMANTIC WEB STANDARDS

This research introduces a method to do semantic analysis of natural language text. They have used ontologies in order to perform this task and they discuss the merits of using ontologies in semantic analysis. They have used a semantic representation based on disclosure representation theory. The research suggests using RDF and OWL because Semantic Web Technology has been emerging as the tool for representing and processing semantic and ontologies. Their semantic representation is special because it has a graph representation. They have this feature because their system works with the Semantic Web (the data representation model is in the graph form). They have developed ontologies for representing natural language semantics. They have used a lexicon in order to perform semantic analysis and a lexicon requires an ontology. Mainly they have focused on Japanese. In the research each sense of ambiguity is explicitly represented with an RDF graph. Still the accuracy of this approach can be increased and we are hoping to do that in our approach. [7]

2.6 COMPARISON

The similar work carried out by others is described in this chapter with their limitations and accuracy. In the 'Real World Behavior Analysis through a Social Media Lens' research they have used the frequency of the words mentioned in order to assign coefficients for predictive analysis. This does not show whether the relevant status update is positive or negative towards the relevant product. They are measuring the frequency of the words mentioned only. In our approach we are giving the polarity of the status update which will be more effective than the frequency of mention in order to predict the real behavior of the customers.

'Recognizing Personality Traits Using Facebook Status Updates' research is mainly based on identifying the personal traits by analyzing the Facebook status updates using several algorithms. They have no business value in the application. In our approach, we are not analyzing the personal traits but we are providing a business value to the application by giving the companies an opportunity to measure how much the customers are interested in their company and products.

'Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews' research only analyzes the sentiments on the movie reviews which does not have the semantic aspect in analyzing. This will not give an accurate result as this does not take the linguistic value into consideration. In our approach, we are performing a sentiment analysis as well as a semantic analysis which will increase the accuracy of the results. Furthermore, we are using a bigger corpus which has thousands of status updates in order to train the sentiment engine which will increase the accuracy.

2.7 Objectives

Summary of main objectives are:

- To analyze the behavior of individuals on social media and based on those behaviors implement the organizational goals like prediction of supply and demand.
- Putting data to work for customer behavior, and Using Population informatics for Samsung, Apple and Sony.
- Recommendation and prediction of sales.

3. THE DOMAIN

Facebook And twitter Data About The Brans, Apple, Sony And Samsung, The aim of the system is to provide critical market information using social media. Therefore, it is vital that the information is gathered from social media sources. Data gathering from Facebook is done through a Facebook application via the Graph API and data gathering from Twitter is done using the Twitter4J library with the Twitter Streaming API. [8][9][10]

4. WORD SENSE DISAMIGUATION

When one word has more than one meaning there is an ambiguity. To realize the precisely sense of the word, it is vital to do a word sense disambiguation. When given a data set on the relevant products, Apple, Sony and Samsung, there is an issue in understanding whether it really discussing about the apple brand or apple fruit. In order to make the analysis more accurate the sense of the word 'apple' is disambiguated as it has two senses (i.Company sense, ii. Fruit Sense). The Word Sense Disambiguation has been done by using the "Naïve Bayes" Classifier and by means of a unique Keyword search algorithm. [11]

company_keywords_list={}
Fruit_keywords_list={}
If item in company keyword list is in the given sentence:

Sentence score +=1
If item in fruit keyword list is in the given sentence:
Sentence score-+1
If sentence score > 0:
The sentence has a company sense
If sentence score <0:
The sentence has a fruit sense
If sentence score =0:
The sense of sentence is undecided

5. SENTIMENT ANALYSIS

Sentiment Analysis is a Natural Language Processing module which uses Machine Learning Methods. This process is used to classify views with their divergence for a given set of features. Machine Learning plays a enormous role in this process. These algorithms are categorized into 3 main types, namely Supervised Learning, Unsupervised Learning and Semi-supervised learning. The goal of this research was to choose the most precise machine learning algorithm and do combinations to that to give more precise sentiment analysis consequences. For that, two Supervised learning techniques and one semi-supervised learning technique has been applied and compared with each other by using the evaluation results.

5.1 SUPERVISED LEARNING TECHNIQUES

Under these methods, the Maximum Entropy Algorithm and the Naïve Bayes Algorithm are the most commonly and widely used two Algorithms for Natural Language Processing projects. These algorithms are used by their authors to give the best outputs for their customers [12]. One of the prerequisites is a proper amount to train the algorithms. A fully marked or labeled data are essential to train these classifiers. After testing with numerous accessible text corpuses, The Twitter Sentiment Corpus by Niek Sanders has been used. This Corpus is made using a large set of tweets extracted from Twitter, and each tweet is tagged with their polarity as 'positive', 'negative' or 'neutral' and tweets which include many unknown characters and symbols or irrelevant languages are tagged as 'irrelevant'. These tweets are connected to products such as Apple, Microsoft and Google which makes corpus very much appropriate for the "Customer Behavior Analysis for Social Media", since the development itself addresses such products.
After selecting the training corpus, the next job was to extract the structures and send it to the algorithm to train. This is where the Machine Learning Algorithms are required.

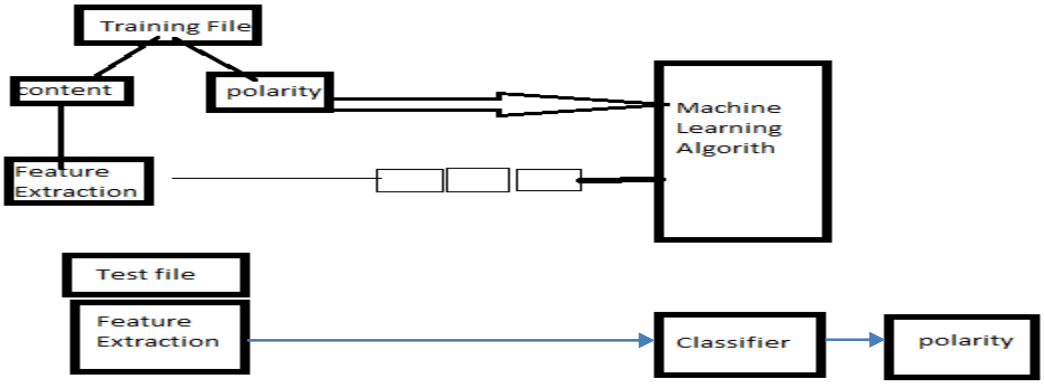


Chart: Flow chart of a classifier which uses machine learning algorithms

Maximum Entropy Categories

This categorization has been used with the assistance of Apache OpenNLP library. The library has a tool called the "Document Categorizer" which runs the Maximum Entropy Algorithm for organization of text for pre-given groups. After giving a tagged corpus, using this tool the corpus has been trained. The algorithm checks each and every feature in the corpus and gives a weight to the feature along with its polarity tagged and gives likelihood measures for each feature [13]. The algorithm has a weighted classifying method which gives a specific weight to each and every feature in a specified document. After that the algorithm checks the probability of occurrence of each feature in the throughout the document to classify them. The formula used by the algorithm to get the probability of each factor is as below,

$$P(c|d,\lambda) \stackrel{def}{=} \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c' \in C} \exp \sum_i \lambda_i f_i(c',d)}$$

The training of the algorithm takes some while because the optimization process takes time. After the training of the tool is completed, it has the ability to check and categorize a given sentiment or a set of sentiments. The pre-processed Tweets and statuses are sent through this tool to categorize them. After categorizing them as the positive, the negative, the neutral and unrelated, they are sent back to the database for more analysis and in order to access simply. This tools output is having an acceptable level of correctness.

Naïve Bayes Classification

The Naïve Bayes Algo is much comparable to Maximum Entropy Algorithm. It is built on the Bayes rule and it uses a probabilistic learning way [14]. For these categories, an unlabeled set of sentiments which the polarity is already recognized accurately has to be composed at first. Then this group of sentiments has to be detached with respect to their polarity. Then the sentiments are added to isolated arrays in order to train them by sending through the Naïve Bayes Algorithm. The training for each set is done independently.

In the same way, negative texts, neutral texts and unrelated properties are also trained. The collection of text selected for this training is the similar from the similar Corpus which was used for Maximum Entropy training, since it is essential for the difference between the two algorithms to be trained using the similar corpus. The Sander's corpus is having tweets with their polarity tagged in front of them as revealed in chart "The flow chart of a classifier which uses machine learning algorithms". So, to use it with the Naïve Bayes Algorithm, the polarities had to be removed. After the training is completed, a set of sentiments can be passed through the tool and their polarities.

The correctness of the results from the Naïve Bayes classifier was lesser than the accuracy of the Maximum Entropy classifier. Treating each and every property autonomously when taking probability measures is which has led it to give a more accuracy level than others. Then, for further improvements of the Sentiment Analysis tool, Maximum Entropy classification has been used.[16]

5.2 SEMI - SUPERVISED LEARNING TECHNIQUES

Unlike in Supervised learning techniques which need a fully annotated or labeled data for training, this technique also takes the use of unlabeled data. This method falls between supervised learning and unsupervised learning techniques. Getting fully labeled data is costly, but unlabeled data can be extracted easily. Thus, using a semi supervised learning technique has an advantage over supervised techniques because of that. "SentiWordNet 3.0" lexical database for English, which has been created by the University of Princeton, and which is publicly available has been used in this approach. This corpus contains a large set of sentiments labeled with their positive and negative scores [15]. When the tool has been implemented after training it using SentiWordnet, not only the polarity, but the polarity score can also be given for a given sentiment. Shown below are the outputs given by the tool for the same set of test data used in the previous two techniques.

This technique does not provide very correct results, but since it can give a polarity score as well, the outputs of the system has been used for many vital analyses in this research such as trend analysis.

5.3 ALGORITHM FOR EMOTION DETECTION

Emoticons are extensively used today by people communicating via social media over non-verbal textual communication methods. So, emoticons are very significant features when inspection the polarity of a sentiment. Most of the other classifying ways remove these emoticons before they analyze the feeling, but an emoticon (feeling symbols) can alter the polarity of a sentiment [16]. Therefore, to increase the correctness of the sentiment analysis, an emoticon recognizing algorithm has been combined with the system. Technologies like as Java and SQL programming languages have been used to create this tool.

Fig: Emoticons that are identified

There are three category types of emoticons feeling mostly used by social media users today. They are the positive or better feeling emoticons, like "□, :-D , :* , <3 , B) , O:) ", the negative emoticons or feeling icons like "□ , :'(, :/ " , and also the neutral once like " □" [17]. In this algo three arrays are used to store these three types of emoticons. If needed, no of emoticons can increase to these arrays because new emoticons get introduced in time to time and step by step After getting the relevant emoticons into the arrays, now the algorithm can check for the availability of the feelings previously analyzed by the above sentiment/feeling analysis techniques. If an emoticon is recongnized, then the algorithm sees through the arrays to find the group belong to. When filtering is done, the polarity column of the database and the polarity of the smiley faces recognized are checked. If both had the match, then the database can be updated as very negative or extremely positive. If the smiley and the polarity showed are both neutral, then there is no update performed or done. If the smiley and the polarity do not match, the data has to be updated.

5.4 THE ALGORITHM OF CONTRADICTION DETECTION

Words like "but", "though", "still", "even though", "yet", "anyway", "however", "nevertheless", "nonetheless", "anyhow", "notwithstanding" and "despite that" can affect the final polarity of an assumed sentence. For example, there can be a Tweet like, "Apple is a good company but iPhone 5 is a very bad product." The first part of this status is positive, and the 2nd part is negative. So, the total polarity of the sentence has to be perceived as neutral. In most cases, the classifiers fail to distinguish these types of statuses and tweets put by social media users and this could reduce the accuracy of the sentiment analysis module.

In this research an algorithm has been applied to detect and analyze such cases. If there is a meaning changing word in a sentence, it is taken out to analyze disjointedly. The word is sifted out as, "<space>word" with a space in front of the word to prevent the error of having such word in the start of a sentence.

This algorithm separates the sentence to parts at the meaning changers. For the above given instance, the status will be separated as the before part, "Apple is a good company" and the part after, "iPhone 5 is a very bad product" by the word "but".

After the separation is done, each part is sent however the sentiment analysis classifier separately and the polarity is taken. The final polarity of the sentence is finalized as,[13]

- If Positive & Positive as Positive
- If Negative & Negative as Negative
- If Positive & Negative as Neutral
- If Negative & Positive as Neutral
- If Neutral & Positive as Positive
- If Neutral & Negative as Negative
- If Positive & Neutral as Positive
- If Negative & Neutral as Negative
- If Neutral & Neutral as Neutral

6. TREND ANALYSIS AND FORECASTING

Trend analysis was done for three representations of data: 1.total interest, 2. total sentiment score and 3. total positive score. The total interest is used to show the interest the public shows to a brand regardless of the sentiment. So, the graph shows how many times a brand has been spoken about with respect to time. The second graph displays how good or badly a brand is received by the market, therefore the number of times a brand has been spoken about positively and negatively is considered and an total score is plotted on the graph. The third graph shows how positively a brand has been spoken about, this graph only displays how many times a brand has been spoken about positively, negative and neutral comments are disregarded. When creating the graphs, the data was combined into months in order to create a sliding window for data. It was decided to practice "time series forecasting" for prediction of future behavior of the products on which the trend analysis was done. Being a famous method in domains like stock market forecasting, the models should make reasonably well in a scenario like this. Because of the dynamic nature of data on social media and because of the fact that consumer electronic brands have a nature of fast rising and falling in terms of fame and because the number of social media users is also always increasing, these two effects may cancel each other out or increase each other, hereafter it was decided that a trend cannot be easily recognized for a data set like this. So, two replicas of time series forecasting were implemented, an Exponential Smoothing Method and the Regression Model for Forecasting. The former because of the fact that the data set has a fast-changing flora and the most recent observations have a higher impact on the future values and the dataset may not display a trend, the latter because a trend cannot be recognized and if one exists it will be more precise. When doing forecasting the system assesses each of these models for all the historical data and takes the model which displays the lowest Mean Squared Error. So, as more and more data are collected the structure will be able to use the forecasting model that is best fit. [18][13][19]

7. THE ASSOCIATION RULE MINING

This technique is used to derive the user attributes that accounts for a particular opinion (positive, negative and neutral). In laymen terms this method will determine users with certain set of qualities will tend to give a certain opinion to a given brand on social media platforms. Due to the fact that this method is about getting the common item sets in a given data set apriory algorithm has been used for this process. The data set wants to be pre-processed appropriate manner so that the algorithm can take and compute them accordingly [13]. Three disconnected brand wise data sets were used for the three brands which were attentive, namely 1. Apple, 2. Sony and 3. Samsung. Every data set has four columns as shown below;

- age category (young, middle & old)
- gender category(male & female)
- relationship status category (single, married & in a relationship)
- opinion (negative, neutral & positive) towards that brand

The number of rows of each brand wise data set is given below;

- Samsung – 1057 rows
- Apple – 833 rows
- Sony – 961 rows

The results of the calculation are a set of connotation orders that indicate general trends exist in the data set. This method was applied for 3 different products and three sets of instructions (for opinion positive, for negative and also neutral) were produced for each brand; so all together there are 9 rules:

These rules consists of two components which are "IF" and "THEN". Having no items common these two are also known as predecessor and consequential. These can be plotted to get a better understanding.

E.g. Rules generated for brand Samsung on getting positive opinions from users Rules support, confidence, lift

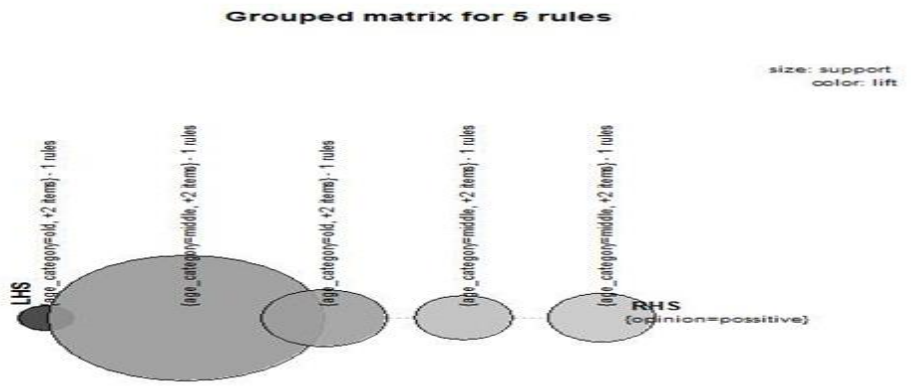
{age_category=old, gender=male,rel_status=married} => {opinion=possitive} 0.009, 0.833, 2.062

{age_category=middle, gender=female,rel_status=married} => {opinion=possitive} 0.103, 0.615, 1.520

{age_category=old, gender=female,rel_status=single} => {opinion=possitive} 0.05, 0.6, 1.85

{age_category=middle, gender=female,rel_status=in a relationship} => {opinion=possitive} 0.026, 0.536, 1.325

{age_category=middle, gender=female,rel_status=single} => {opinion=possitive} 0.032, 0.514, 1.274 When plotted graphically they look like as shown below;



8. PREDICTING THE OPINIONS USING CLASSIFIERS

A standard binary classifier is used to build three classification models for each of three brands focused. Each model is trained using relevant training data gathered from social media platforms.

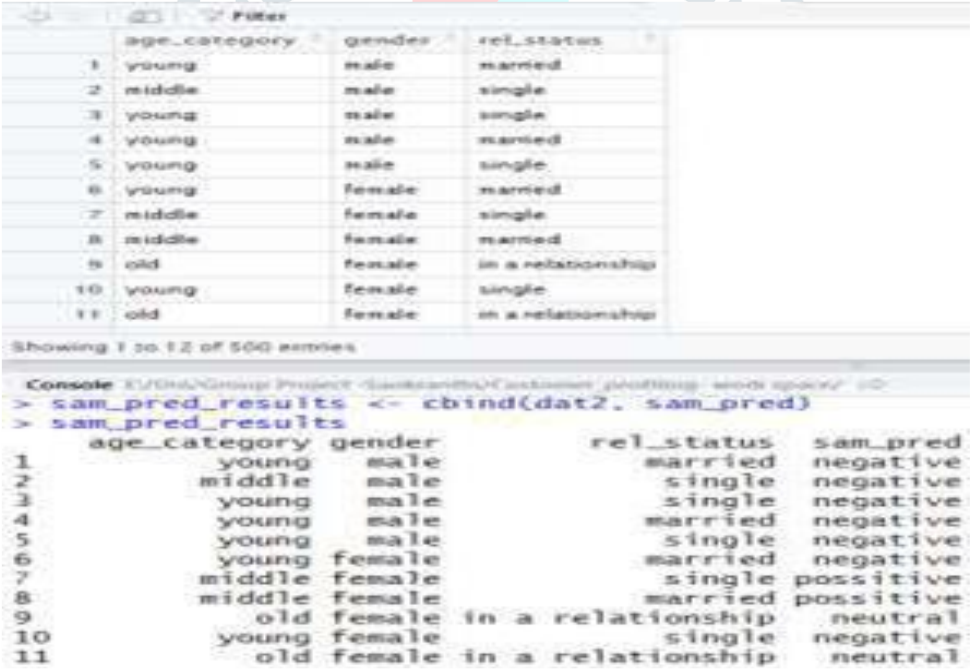
The tree classifier for brand Samsung is given below; n= 845

node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 845 526 negative (0.3775147 0.2650888 0.3573696)
2) gender=male 432 229 negative (0.4699075 0.2638888 0.2662046) *
3) gender=female 413 226 possitive (0.2808717 0.2663438 0.4527945)
6) age_category=young 147 87 negative (0.4081633 0.2789116 0.3139252) *
7) age_category=middle,old 266 125 possitive (0.2104263 0.2593985 0.5300752)
14) age_category=middle 222 109 possitive (0.2522523 0.2386387 0.5090091) *
15) age_category=old 44 16 possitive (0.0000000 0.3636364 0.6363638)
30) rel_status=in a relationship 4 1 neutral (0.00000000.7500000 0.2500000) *
31) rel_status=single 40 13 possitive (0.0000000 0.3260000 0.6740000) *

The classifier generates predictions when a set of user data passed in to it. The results are shown below;



	age_category	gender	rel_status
1	young	male	married
2	middle	male	single
3	young	male	single
4	young	male	married
5	young	male	single
6	young	female	married
7	middle	female	single
8	middle	female	married
9	old	female	in a relationship
10	young	female	single
11	old	female	in a relationship

Showing 1 to 12 of 500 entries

```
> sam_pred_results <- cbind(dat2, sam_pred)
> sam_pred_results
   age_category gender          rel_status  sam_pred
1         young   male             married  negative
2        middle   male              single  negative
3         young   male              single  negative
4         young   male             married  negative
5         young   male              single  negative
6         young female             married  negative
7        middle female              single possitive
8        middle female             married possitive
9           old female in a relationship   neutral
10        young female              single  negative
11          old female in a relationship   neutral
```

9. RESULTS AND EVALUATION

9.1 WORD SENSE DISAMBIGUATION

As the word sense dis-ambiguation is perfomed by using a key word search method we need to make sure it is precise to analyze the whole the data as the output of this component is used in the next steps of the system.

Table: Word Sense Disambiguation evaluation

Test File		Tagged number	Actual number	Difference	Accuracy
Test 1: Dell.txt	company related sentences	57	55	2	
	fruit related sentences	35	33	2	97%
	undecided sentences	4	8	4	
Test 2: Apple.txt	company related sentences	95	78	17	
	fruit related sentences	0	0	0	82%
	undecided sentences	11	28	17	
Test 3: Apple.txt	company related sentences	22	16	8	
	fruit related sentences	69	52	19	72%
	undecided sentences	18	40	29	
Test 4: Apple.txt	company related sentences	117	94	23	
	fruit related sentences	69	52	17	80%
	undecided sentences	27	68	41	

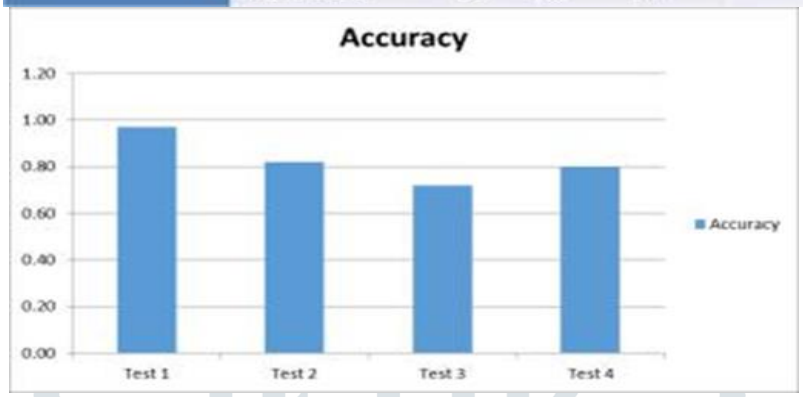


Chart: Accuracy of Word Sense Disambiguation

9.2 SENTIMENT ANALYSIS

In a sentiment Analysis which is analysis of feelings, the correctness of the classifier used a very vital part. so, for the Client Behavior Analysis for Social Media, a proper classifier with a better level of accuracy was required. To select this, after implementation and application of classifiers, an evaluation had to be performed.

In this, three different machine learning techniques has been implemented, namely; Maximum Entropy Classifier, Naïve Bayes Classifier, Classifier which uses SentiWordnet. So, to evaluate them, they had to be tested in similar circumstances. In this point, same test data sets had to be tested when checking their accuracy. For the assessment of Naïve Bayes and Maximum Entropy Classifiers, the same teaching of data sets has been used as well to compare them more clearly, but for the other classifier the SentiWordnet Corpus was used to train since the technique had a different learning way.

Niek Sander's Lexical Corpus has been used to teach the 2 Supervised Learning Classifiers. Tests have been done using numerous data sets which the polarities of the feelings are known. Most data sets compressed, extracted tweets from the Twitter Crawler and checked by user manually and the polarities were recognized before classifying using the tools. Some training sets were extracted using hashtags. For positive feelings, tweets with the #happy were crawled and for negative, #unhappy or #negative were crawled and used for testing. Another test set was created using a feedback form delivered among colleagues, which contained a set of sentences for them to indicate the polarity.

To check the accuracy of the classifiers, an algorithm has been created. This algorithm can compare the results of the tools with the actual results and provide a percentage accuracy level for each tool.

Table : Evaluation of Sentiment Analysis Classifiers

Test No.	Total Sentiments	Maximum Entropy Classifier		Naïve Bayes Classifier		SentiWordnet Classifier	
		Matchings	Accuracy %	Matchings	Accuracy %	Matchings	Accuracy %
1	10	6	60.00	5	50.00	5	50.00
2	15	11	73.33	5	33.33	9	60.00%
3	340	254	74.72	86	43.24	125	36.76
4	551	484	87.84	221	59.48	181	32.88
5	1000	773	77.30	615	61.50	389	38.90

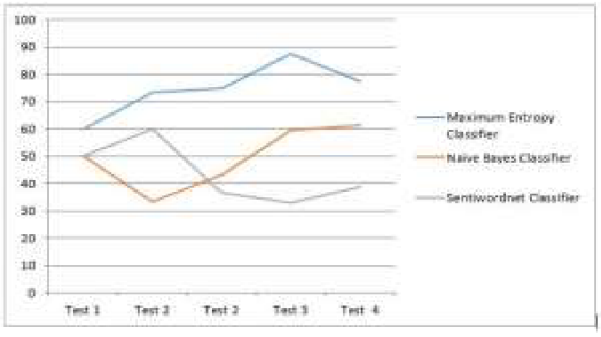


Chart : Evaluation of Sentiment Analysis classifiers

From the outcomes above, it can be clearly explained that the Maximum Entropy Classifier provides a better accurateness when comparing with the other two approaches. So, the Maximum Entropy Classifier has been united with the Emoticon Detection Algorithm and evaluated again. The same test data sets were used for this assessment as well to analyze the accuracy changes clearly. The outcomes were as shown below

Table: Evaluation of Emoticon Detection Algorithm

Test Number	Total Sentiments	Maximum Entropy Classifier		Maximum Entropy Classifier with Emoticon Detection	
		Matchings	Accuracy %	Matchings	Accuracy %
1	10	6	60.00	7	70.00
2	15	11	73.33	11	73.33
3	340	254	74.72	272	79.11
4	551	484	87.84	490	88.92
5	1000	773	77.30	796	79.60

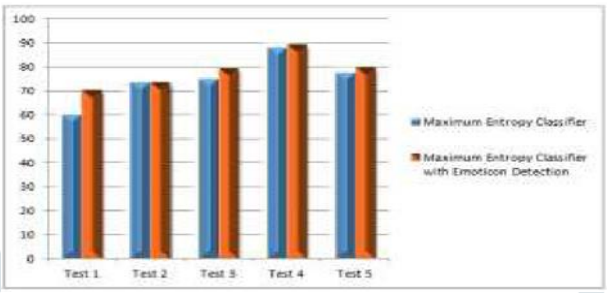
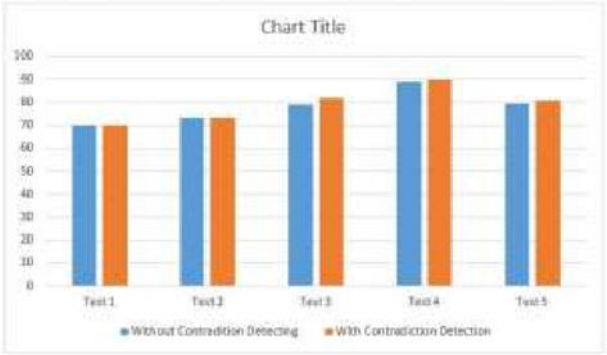


Chart: Evaluation of Emoticon Detection Algorithm

The "contradictory detection" is another addition done to the Maximum Entropy Classifier to improve the level of accuracy. The evaluation results before and after the combination of this algorithm is shown below,

Table: Evaluation of Contradiction Detecting Algorithm

Test Number	Total Sentiments	Maximum Entropy Classifier with Emoticon Detection		Maximum Entropy Classifier with Emoticon Detection & Contradiction Detecting Algorithm	
		Matchings	Accuracy %	Matchings	Accuracy %
1	10	7	70.00	7	70.00
2	15	11	73.33	11	73.33
3	340	272	79.11	279	82.05
4	551	490	88.92	494	89.66
5	1000	796	79.60	809	80.90



Furthermore, the Maximum Entropy Classifier has been tested by changing the size of the data set used for the training of the tool to check the effect of having a proper corpus to get accurate results. The same Sander's Corpus has been trained by taking several samples from it and making different files with different number of sentiments. The test data set used for Test 2 in the previous evaluations was used for this. The results of this evaluation are shown below;

Table: Evaluation of Maximum Entropy Classifier by changing the training data set size

Test Number	Test Data Set Size	Training Data Set Size	Accuracy %
1	340	100	33.12
2	340	500	42.56
3	340	1500	61.26
4	340	3682	79.60

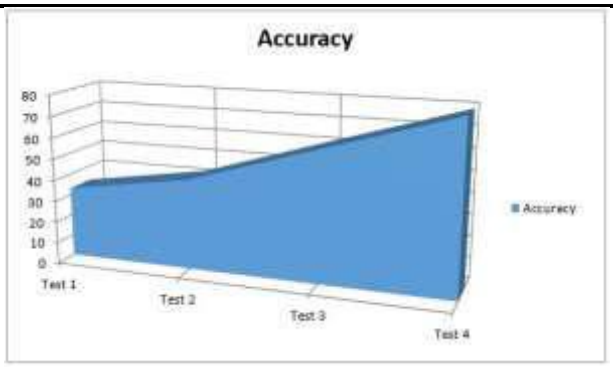


Chart: Accuracy change with respect to size of training data set size

10. FUTURE WORK

We need to increase the accuracy of all the components as the current average accuracy of each component is around 70%-90%. The sarcasm identifier should be improved. We have catered only 4 brands from this research and if the number of brands can be increased then it will be very useful to many companies in their decision-making process. Our system does not have a language detection option and it will be very user friendly if the language detection option is implemented. As we are using only English status updates and tweets it is not completely accurate to make predictions. To increase the accuracy of the result we need to use as many languages as possible.

REFERENCES

[1] Benjamin Flesch, "Social Set Visualizer: A Set Theoretical Approach to Big Social Data Analytics of Real-World Events" Copenhagen Business School, Denmark and 2Westerdals Oslo School of Arts, Comm & Tech, Norway, 2015. https://ieeexplore.ieee.org/document/7364036/figures#figures

[2] R. E. Montalvo, ''Social media management,'' Int. J. Manage. Inf. Syst., vol. 15, no. 3, pp. 91–96, 2011

[3] C. Vollmer and G. Precourt, Always On: Advertising, Marketing, and Media in an Era of Consumer Control (Strategy + Business). New York, NY, USA: McGraw-Hill, 2008.

[4] A. McAfee, Enterprise 2.0: New Collaborative Tools for Your Organization's Toughest Challenges. Boston, MA, USA: Harvard Business Press, 2009.

[5] R. Vatrapu, ''Understanding social business,'' in Emerging Dimensions of Technology Management. India: Springer, 2013, pp. 147–158.

[6] W. S. Cleveland, ''Data science: An action plan for expanding the technical areas of the field of statistics,'' Int. Statist. Rev., vol. 69, no. 1, pp. 21–26, 2001. [Online]. Available: http://dx.doi.org/10.1111/j.1751-5823.2001.tb00477.x

[7] M. Loukides, What Is Data Science? Sebastopol, CA, USA: O'Reilly Media, 2012.

[8] N. Ohsumi, ''From data analysis to data science,'' in Data Analysis, Classification, and Related Methods. Berlin, Germany:Springer, 2000, pp. 329–334. [Online]. Available: http://dx.doi.org/ 10.1007/978-3-642-59789-3_52

[9] D. Lazer et al., ''Computational social science,'' Science, vol. 323, no. 5915, pp. 721– 723, 2009.

[10] J. Sterne, Social Media Metrics: How to Measure and Optimize Your Marketing Investment. New York, NY, USA: Wiley, 2010.

[11] M. Sponder, Social Media Analytics: Effective Tools for Building, Interpreting, an Using Metrics. New York, NY, USA: McGraw-Hill, 2011.

[12] Z. Tufekci. (2014). ''Big questions for social media big data: Representativeness, validity and other methodological pitfalls.'' [Online]. Available: http://arxiv.org/abs/1403.7400

[13] S. D. Kularathne (2017) "Customer Behavior Analysis for Social Media".achademida journal

[14] A. A. Mohammad, K. C. Sun, H. Liu and K. Sagoo, "Real-World Behavior Analysis through a Socila Media Lens".

[15] F. T. ODonovan, C. Fournelle and S. Gaffigan, "Characterizing User Behavior and Information Propagation On A Social Multimedia Network".

[16] G. Farnadi, S. Zoghbi, M. F. Moens and M. De Cock, "Recognizing Personality Traits Using Facebook Status Updates".

[17] N. Mehra, S. Khandelwal and P. Patel, "Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews".

[18] Changhyun Byun, Hyencheol Lee, Yanggon Kim and Kwangmi Ko Kim, "Twitter data collecting tool with rule-based filtering.

[19] Alexander Hogenboom, Danieella Bal and Flavius Frasincar, "Exploiting Emoticons in Sentiment Analysis".analysis module".