Customer Behavior Analysis for Social Media

S. D. Kularathne, R.B. Dissanayake, N.D. Samarasinghe, L.P.G. Premalal, S. C. Premaratne

Faculty of Information Technology, University of Moratuwa, Sri Lanka

Abstract— *It is essential for a business organization to get the customer feedback in order to grow as a company. Business organizations are collecting customer feedback using various methods. But the question is 'are they efficient and effective?' In the current context, there is more of a customer oriented market and all the business organizations are competing to achieve customer delight through their products and services. Social Media plays a huge role in one's life. Customers tend to reveal their true opinion about certain brands on social media rather than giving routine feedback to the producers or sellers. Because of this reason, it is identified that social media can be used as a tool to analyze customer behavior. If relevant data can be gathered from the customers' social media feeds and if these data are analyzed properly, a clear idea to the companies what customers really think about their brand can be provided.*

Keywords— Machine Learning, Natural Language Processing, Word Sense Disambiguation, Sentiment Analysis, Trend Analysis, Forecasting, Opinion Prediction, Association Rule Mining .

I. INTRODUCTION

With the development in marketing and the customer oriented marketing concepts, the companies and organizations find it essential to know the customer's attitude towards its brand. The companies need to get feedback from the customers in order to learn and grow themselves. With the development of technology, the internet plays a huge role in expressing ideas and opinions of people and social networks are acting more prominent in this context. The companies may find it very difficult to get the customer feedback effectively through questionnaires or face to face interviews and because of that the companies may start to look for other ways of getting the customer feedback and sentiments. That's when the need of a sentiment analysis system comes into play as it is really effective to know the customer's opinion about a certain brand by analyzing what he/she posts on social media, for an example on Facebook. This has a huge business value because companies find information on how the market perceives them extremely valuable. These information is one of the most genuine and accurate customer feedback a company can get. But collecting this data is quite a challenging task because the amount of data that needs to be processed is massive. After a proper filtering [1], the data can be analyzed and the opinions can be given a positive, negative or a neutral polarity. If this information is used to provide a service to companies, it can be graphically represented for the client giving them a good idea of the success or popularity of their brand.

It would be really useful for the companies to know what type of people like their brands. For an example, if it is possible to analyze and identify that the people who like Apple products are interested in Music and Art then Apple can target those market segments when they are developing their next product and they can widen up their market to cater to new segments of customers. As a result of that, the profiling of the customers has been done with their interests and provide these data to the relevant mobile companies with the intention of making their marketing campaigns successful. Furthermore, the trend of the popularity of the relevant brands and with that information the companies can identify the increasing or decreasing interest in customers on their brand. If graphical information on these data can be provided, then the company can decide when to release their next product what are their successful and unsuccessful products.

The focus of this research is on doing customer behavior analysis for consumer electronic brands Apple, Sony and Samsung.

II. RELATED WORK

2.1 Real-World Behavior Analysis through a Social Media Lens

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 are developing has this aspect of customer profiling which will be beneficial for the companies to make their decisions based on target markets and target customer segments. For an example if we are catering the 'Apple Inc' company we can identify apple users and their preferences and this will help Apple to design their next product in a more user friendly 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 opinion by analyzing their online behavior such as comments and status updates on social media such as Facebook and Twitter.

'Real-World Behavior Analysis through a Social Media Lens' research suggests that there is a strong correlation between a person's online behavior and the real-world behavior. The researchers say that it is possible to predict the real-world behavior of a person by analyzing his/her online behavior. In their context, they have selected the community based on characteristics such as race, culture, country, etc. In this application, we can select the community as a set of users who uses a certain product or a brand because we are designing the application to serve IT based companies. Data collection can be done through Facebook. Then the text processing is to be performed on the gathered data. In the research, they have used a different method to analyze data and we are performing a sentiment and semantic analysis on the gathered data. They have used correlational analysis to identify word categories whose frequency of mention was most significantly related statistically to the magnitude of social action during the same period, then used multivariate regression analysis to assign coefficients to them for predictive analysis. In this application, we will be able to predict the user's opinion and attitude towards certain brands and products in the future. Event prediction, Attitude extraction, Key people detection and Mood analysis are some of the results of the analysis and these are very useful for this application as well. [2]

2.2 Characterizing User Behavior and Information Propagation on a Social Multimedia Network

Members of these online social media communities often play distinct roles that can be deduced from observations of users' online activities. Researchers have presented five clusters of users with common observed online behaviors, where these users also show correlated profile characteristics. They are focusing mainly on the multimedia shared by users of social media and they bring out the psychological facet 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]

2.3 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]

2.4 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]

2.5 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]

2.6 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]

2.7 Comparison

The similar work carried out by others is described in this chapter with their limitations and accuracy. In the 'Real World Behaviour 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.

III. THE DOMAIN – FACEBOOK AND TWITTER DATA ABOUT THE BRANDS, 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]

IV. WORD SENSE DISAMBIGUATION

When a word has more than one senses there is an ambiguity. To realize the exactly sense of the word, it is essential to do a word sense disambiguation.

When given a data set on the relevant brands, Apple, Sony and Samsung, there is an issue in understanding whether it really talks 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 is done by using the Naïve Bayes Classifier and using 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

V. SENTIMENT ANALYSIS

Sentiment Analysis is a Natural Language Processing component which uses Machine Learning Techniques. This process is used to categorize opinions with their polarity for a given set of features. Machine Learning plays a huge role in this process. These algorithms are categorized into three main types, namely Supervised Learning, Unsupervised Learning and Semi-supervised learning. The aim of this research was to select the most accurate machine learning algorithm and do integrations to that to give more accurate sentiment analysis results. For that, two Supervised learning techniques and one semi-supervised learning technique has been implemented and compared with each other by using the evaluation results.

5.1 Supervised Learning Techniques

Under these techniques, 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 utilized by their authors to give the best outputs for their clients [12]

One of the prerequisite is a proper corpus to train the algorithms. A fully annotated or labeled data are required to train these classifiers. After testing with several available 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 related to products such as Apple, Microsoft and Google which makes corpus very much suitable for the "Customer Behaviour Analysis for Social Media", since the project itself addresses such products.

After selecting the training corpus, the next task was to extract the features and send it to the algorithm to train. This is where the Machine Learning Algorithms are required.

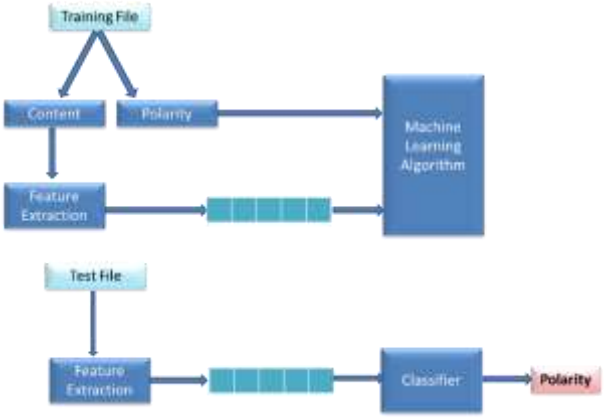


Chart: Flow chart of a classifier which uses machine learning algorithms

Maximum Entropy Classification

This classifier has been used with the help of Apache OpenNLP library. The library has a tool called the "Document Categorizer" which runs the Maximum Entropy Algorithm for classification of text for pre-given categories. 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 probability measures for each feature [13]. The algorithm has a weighted categorizing method which gives a particular weight to each and every feature in a given 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 classify them. After categorizing them as positive, negative, neutral and irrelevant, they are sent back to the database for further

analysis and in order to access easily. This tools output is having a satisfactory level of accuracy

Naïve Bayes Classification

The Naïve Bayes Algorithm is much similar to Maximum Entropy Algorithm. It is based on the Bayes rule and it uses a probabilistic learning method [14]. For this classifier, an unlabeled set of sentiments which the polarity is already identified accurately has to be collected at first. Then this collection of sentiments has to be separated with respect to their polarity. Then the sentiments are added to separate arrays in order to train them by sending through the Naïve Bayes Algorithm. The training for each set is done independently.

In the same manner, negative texts, neutral texts and irrelevant features are also trained. The collection of text selected for this training is the same from the same Corpus which was used for Maximum Entropy training, since it is necessary for the comparison between the two algorithms to be trained using the same corpus. The Sander's corpus is having tweets with their polarity tagged in front of them as shown in chart "*Flow chart of a classifier which uses machine learning algorithms"*. Therefore, 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 accuracy of the results from the Naïve Bayes classifier was lower than the accuracy of the Maximum Entropy classifier. Treating each and every feature independently when taking probability measures is which has lead it to give a higher accuracy level than others. Therefore, for further improvements of the Sentiment Analysis tool, Maximum Entropy classification has been used.

5.2 Semi - Supervised Learning Techniques

Unlike in Supervised learning techniques which require a fully annotated or labeled data for training, this technique also takes the use of unlabeled data. This technique 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 accurate results, but since it can give a polarity score as well, the outputs of the system has been used for many important analyses in this research such as trend analysis.

5.3 Emoticon Detection Algorithm

Emoticons are widely used today by people communicating via social media through non-verbal textual communication methods. Therefore, emoticons are very important features when checking the polarity of a sentiment. Most of the other classifying methods remove these emoticons before they analyze the sentiment, but an emoticon can change the polarity of a sentiment completely [16]. Therefore, to increase the accuracy of the sentiment analysis, an emoticon detection algorithm has been integrated with the system. Technologies such as Java and SQL have been used to create this tool.

Fig: Emoticons that are identified

There are 3 main types of emoticons commonly used by social media users today. They are positive emoticons, such as "☺, :-D , :* , <3 , B) , O:) ", negative emoticons such as "☹ , :'(, :/ " , and also neutral emoticons such as "😐" [17]. In this algorithm, three separate arrays are used to store these three types of emoticons. If needed, more emoticons can be added to these arrays because new emoticons get introduced time to time

After getting the relevant emoticons into the arrays, now the algorithm can check for the availability of the sentiments previously analyzed by the above sentiment analysis techniques. If an emoticon is identified, then the algorithm checks through the arrays to find the category belong to.

When filtering is completed, the polarity column of the database and the polarity of the smiley identified are checked. If both the polarities match, then the database can be updated as very negative or very positive. If the smiley and the polarity identified are both neutral, then no update is done. If the smiley and the polarity do not match, the database has to be updated.

5.4 Contradiction Detecting Algorithm

Words like "but", "though", "still", "even though", "yet", "anyway", "however", "nevertheless", "nonetheless",

"anyhow", "notwithstanding" and "despite that" can affect the final polarity of a given sentence. For example, there could 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 second part is negative. Therefore, the total polarity of the sentence has to be detected as neutral. In most cases, the classifiers fail to detect these type 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 implemented to detect and analyze such cases. If there is a meaning changing word in a sentence, it is taken out to analyze separately. The word is filtered out as, "<space>word" with a space in front of the word to prevent the error of having such word in the beginning of a sentence.

This algorithm splits the sentence to parts at the meaning changers. For the above given example, 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 though the sentiment analysis classifier separately and the polarity is taken. The final polarity of the sentence is finalized as,

- 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

VI. TREND ANALYSIS AND FORECASTING

Trend analysis was done for three representations of data, total interest, total sentiment score and total positive score.

The total interest is used to depict the interest the public shows towards a brand regardless of the sentiment. Therefore, the graph shows how many times a brand has been talked about with respect to time. The second graph shows how positively or negatively a brand is received by the market, therefore the number of times a brand has been talked about positively and negatively is considered and an overall score is plotted on the graph. The third graph shows how positively a brand has been talked about, this graph only shows how many times a brand has been talked about positively, negative and neutral comments are ignored. When generating the graphs, the data was aggregated into months in order to create a sliding window for data.

It was decided to use time series forecasting for prediction of future behavior of the brands on which the trend analysis was done. Being a popular method in domains like stock market forecasting, the models should perform 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 rapidly rising and falling in terms of popularity and because the number of social media users is also constantly increasing, these two effects may cancel each other out or boost each other, hence it was decided that a trend cannot be easily identified for a data set like this.

Therefore, two models of time series forecasting were implemented, an Exponential Smoothing Method and a Regression Model for Forecasting. The former because of the fact that the data set has a fast changing nature and the most recent observations have a higher impact on the future values and the dataset may not show a trend, the latter because a trend cannot be identified and if one exists it will be more accurate. When doing forecasting the system evaluates each of these models for all the historical data and takes the model which shows the lowest Mean Squared Error. Therefore, as more and more data is accumulated the system will be able to use the forecasting model that is best fit. [18][19][20][21]

VII. ASSOCIATION RULE MINING

This approach is used to derive the user attributes that accounts for a particular opinion (positive, negative and neutral). In laymen terms this approach will determine users with certain set of attributes will tend to give a certain opinion towards a given brand on social media platforms. Due to the fact that this approach is about getting the frequent item sets in a given data set apriory algorithm has been utilized for this process. The data set needs to be pre-processed appropriate manner so that the algorithm can accept and compute them accordingly [22].

Three separated brand wise data sets were used for the three brands which were focused, namely Apple, Sony and Samsung. Every data set has four columns as shown below;

- age category (young, middle & old)
- gender (male & female)
- relationship status (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 computation are a set of association rules that indicate general trends exist in the data set. This method was applied for three different brands and three sets of rules (for opinion positive, negative and neutral) were generated for each brand; so all together there are nine sets of rules.

This rules consists of two components "if" and "then". Having no items common these two are also known as antecedent and consequent. 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, 2.063

{age_category=middle,gender=female,rel_status=married} => {opinion=possitive} 0.104, 0.615, 1.521

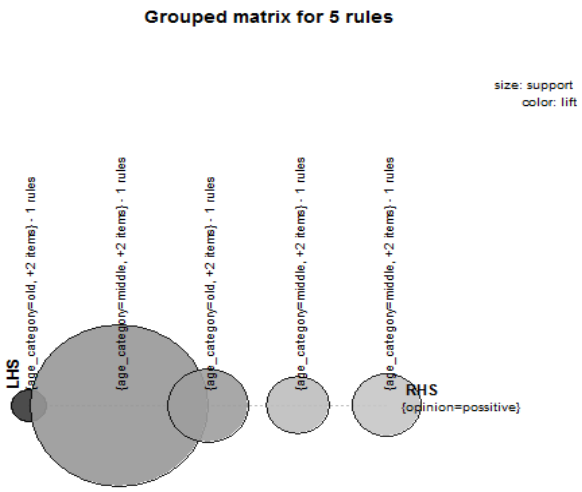
{age_category=old,gender=female,rel_status=single} => {opinion=possitive} 0.04, 0.6, 1.485

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

0.028, 0.536, 1.326

{age_category=middle,gender=female,rel_status=single} => {opinion=possitive} 0.032, 0.515, 1.275

When plotted graphically they look like as shown below;



VIII. OPINION PREDICTION USING CLASSIFIER

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 (0.3775148 0.2650888 0.3573964)

2) gender=male 432 229 negative (0.4699074 0.2638889 0.2662037) *

3) gender=female 413 226 possitive (0.2808717 0.2663438 0.4527845)

6) age_category=young 147 87 negative (0.4081633 0.2789116 0.3129252) *

7) age_category=middle,old 266 125 possitive (0.2105263 0.2593985 0.5300752)

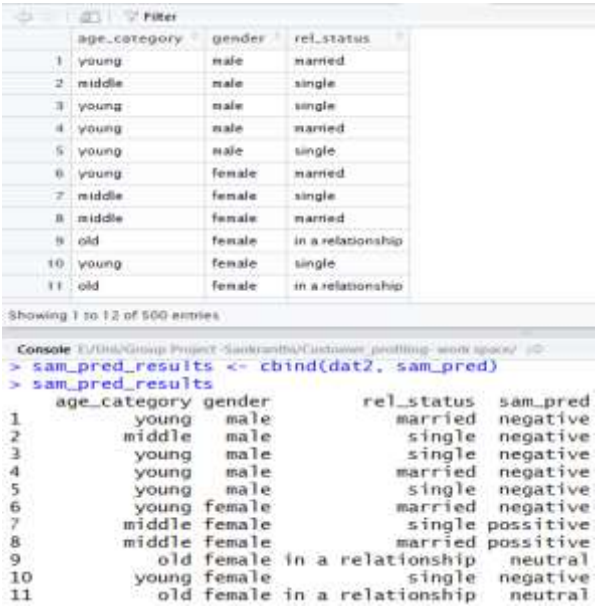
14) age_category=middle 222 109 possitive (0.2522523 0.2387387 0.5090090) *

15) age_category=old 44 16 possitive (0.0000000 0.3636364 0.6363636)

30) rel_status=in a relationship 4 1 neutral (0.0000000 0.7500000 0.2500000) *

31) rel_status=single 40 13 possitive (0.0000000 0.3250000 0.6750000) *

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



Filter

	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

Console

```
> 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
```

IX. RESULTS AND EVALUATION

9.1 Word Sense Disambiguation

As the word sense disambiguation is done by using a key word search method we need to make sure it is accurate to analyze all the data as the output of this component is used in the next steps of the system.

Table: Word Sense Disambiguation evaluation

Text File		Tagged number	Actual number	Difference	Accuracy
Text 1: Dell.txt	company related sentences	57	55	2	
	fruit related sentences	35	33	2	97%
	undecided sentences	4	8	4	
Text 2: Apple.txt	company related sentences	95	78	17	
	fruit related sentences	0	0	0	82%
	undecided sentences	11	28	17	
Text 3: Apple.txt	company related sentences	22	16	6	
	fruit related sentences	69	52	19	72%
	undecided sentences	18	40	29	
Text 4: Apple.txt	company related sentences	117	94	23	
	fruit related sentences	69	52	17	80%
	undecided sentences	27	68	41	

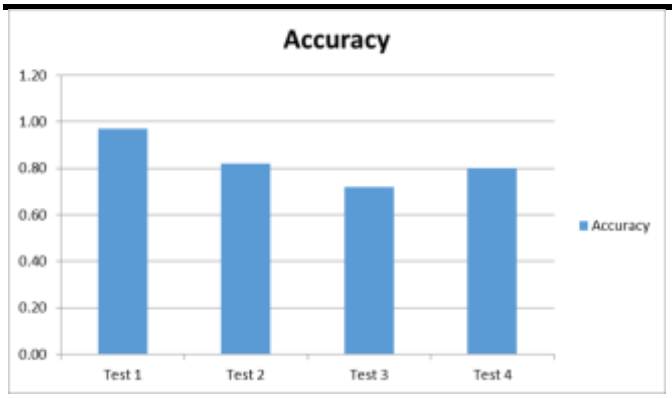


Chart: Accuracy of Word Sense Disambiguation

9.2 Sentiment Analysis

In a sentiment Analysis, the accuracy of the classifier used a very important component. Therefore, for the Customer Behavior Analysis for Social Media, a proper classifier with a higher level of accuracy was needed. To choose this, after implementation of classifiers, an evaluation had to be done.

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 conditions. In this case, same test data sets had to be given when checking their accuracy. For the evaluation of Naïve Bayes and Maximum Entropy Classifiers, the same training data sets have 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 method.

Niek Sander's Lexical Corpus has been used to train the two Supervised Learning Classifiers. Tests have been done using several data sets which the polarities of the sentiments are known. Most data sets included extracted tweets from the Twitter Crawler and manually checked and the polarities were identified before classifying using the tools. Some training sets were extracted using hashtags. For positive sentiments, tweets with #happy were crawled and for negative, #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		Naive 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

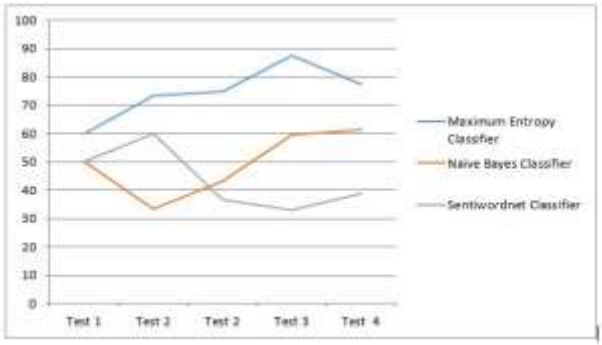


Chart : Evaluation of Sentiment Analysis classifiers

From the results above, it can be clearly clarified that the Maximum Entropy Classifier provides a better accuracy when comparing with the other two approaches. Therefore, the Maximum Entropy Classifier has been integrated with the Emoticon Detection Algorithm and evaluated again. The same test data sets were used for this evaluation as well to analyze the accuracy changes clearly. The results 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

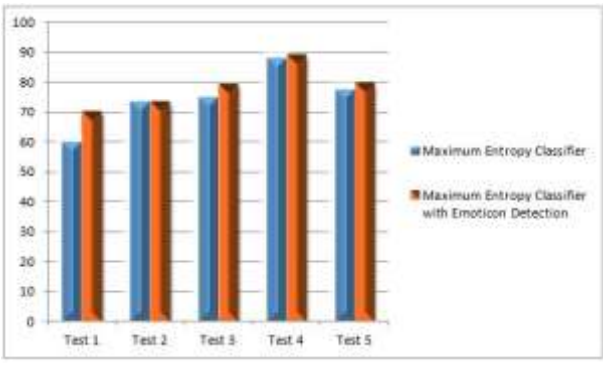


Chart: Evaluation of Emoticon Detection Algorithm

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 integration 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

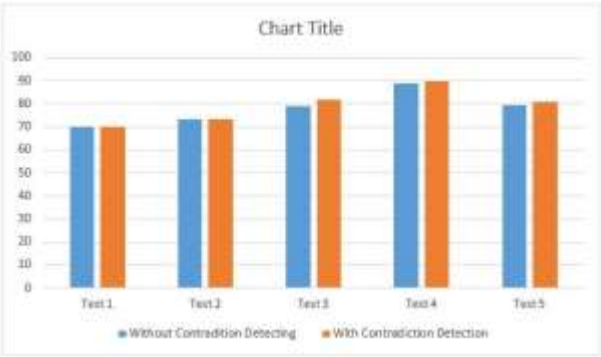


Chart: Evaluation of Contradiction Detecting Algorithm

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

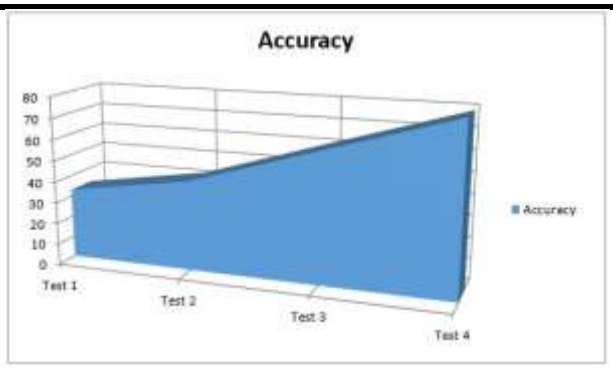


Chart: Accuracy change with respect to size of training data set size

9.3 Association Mining

Samsung

Rows of test data set - 50

	predicted		
	negative	neutral	positive
negative	0	0	0
neutral	0	0	12
positive	0	0	38

recall-positive= 38/(38) = 100%
precision-positive= 38/(12+38)=76%

Rows of test data set -100

	predicted		
	negative	neutral	positive
negative	13	0	0
neutral	5	0	18
positive	2	0	62

recall-positive= 62/(2+62) = 96.8%
precision-positive= 62/(18+62) = 77.5%

Rows of test data set -200

	predicted		
	negative	neutral	positive
negative	13	2	12
neutral	8	1	21
positive	16	0	77

recall-positive = 77/ (77+16) = 82.5%
precision-positive = 77/ (12+21+77) = 70%

X. DISCUSSION

It is very important fact to analyze how people think in different context about different things. This becomes more important when it comes to the business world because businesses depend on their customers and they always try to make products or services in order to fulfill customer

requirements. So knowing what they want and what they think and talk about existing products, services and brands is more useful for businesses to make decisions, identify competitors and analyze trends.

In the current context people are not bothered about giving their feedback to companies in traditional ways such as interviews, questionnaire, feedback forms, etc. But they are using social media mostly to express their opinion on various things they meet and use in their day to day life. That is why we created a system to analyze this social media information and give valuable feedback to the relevant companies. We gathered social media data such as tweets from twitter and status updates, comments, likes and personal information from Facebook. We filtered the gathered data and selected data that are relevant to the specific brands that we are catering to and we did a word sense disambiguation and sentiment analysis on these data to identify whether these status updates and tweets have a positive or a negative opinion towards the brands. After that we did a trend analysis which will give valuable information for the companies to make their decision because through the trend analysis they can identify where they became successful and where they failed in the past. In order to give the companies more information about their customers we have done a customer profiling as well. We expressed the associations of the customers of the relevant brands. Here we have tried to stereotype the customer segments who are interested in specific brands so that the companies can target these customers when creating new products.

10.1 Achievement of Objectives

The objectives of the research are as follows:

- Create a Facebook application and a Twitter data crawler which would enable to extract user data such as comments, statuses, interests, age, etc. and extract relevant information
- Data extraction preprocessing
- Semantic analysis (Word sense disambiguation)
- Sentiment Analysis with Emoticon detection and Contradiction detection
- Product profiling and Customer profiling
- Trend analysis and forecasting

We have successfully achieved all the objectives in order to reach our target. Our system can extract social media data and give a highly accurate result of the sentiment of these social media data and it analyses the trends and perform customer profiling successfully.

10.2 Problems encountered

The understanding of the research requirements was very low initially and this costed us a lot of time as with unclear requirements were not able to identify clear cut components of the research, but with the support from the supervisor we were able to understand the requirements better and perform and finish our tasks on time. As all the components were new concepts for the team, we had to do more research on each component and finally we managed to get a thorough knowledge on all the components and even design our own algorithms to perform the tasks.

10.3 Further 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] S. R. Yerva, Zolt' an Mikl 'os and K. Aberer, "It was easy, when apples and blackberries were only fruits".
[2] A. A. Mohammad, K. C. Sun, H. Liu and K. Sagoo, "Real-World Behavior Analysis through a Socila Media Lens".
[3] F. T. ODonovan, C. Fournelle and S. Gaffigan, "Characterizing User Behavior and Information Propagation On A Social Multimedia Network".
[4] G. Farnadi, S. Zoghbi, M. F. Moens and M. De Cock, "Recognizing Personality Traits Using Facebook Status Updates".
[5] N. Mehra, S. Khandelwal and P. Patel, "Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews".
[6] Changhyun Byun, Hyencheol Lee, Yanggon Kim and Kwangmi Ko Kim, "Twitter data collecting tool with rule-based filtering and analysis module".
[7] N. Arkawa, "Semantic Analysis based on Ontologies with Semantic Web Standards".
[8] https://dev.twitter.com/faq, "Twitter API and Twitter4j".
[9] Twitter4J, "http://twitter4j.org/en/index.html".
[10] F. Developers, "https://developers.facebook.com/docs/graph-api".
[11] http://aclweb.org/aclwiki/index.php?title= Word_sense_disambiguation, "Word Sense Disambiguation".
[12] Bo Pang, Lillian Lee and Shivakumar Vaithayanthan, "Thumbs Up? Sentiment Classification using Machine Learning Techniques".

[13] Adam L Berger, Stephen A and Bella Pietra, "A Maximum Entropy Approach to Natural Language Processing".
[14] J. Alfons, V. David and N. Hermann, "Bridging the gap between Naive Bayes and Maximum Entropy Text Classification".
[15] S. Baccianella, Andrea Esuli and Fabrinzo Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining".
[16] Alexander Hogenboom, Danieella Bal and Flavius Frasincar, "Exploiting Emoticons in Sentiment Analysis".
[17] Eshrag Refae and Verena Rieser, "CanWe Read Emotions from a Smiley Face? Emoticon-based Distant Supervision for Subjectivity and Sentiment Analysis of Arabic Twitter Feeds".
[18] Terry Therneau, Beth Atkinson and Vrian Ripley, "Recursive partitioning for classification, regression and survival trees".
[19] R. Lindek, "Forecasting models".
[20] Everette S Gardner, "Exponential Smoothing: The State of the Art - Part 2".
[21] Sanjay Kumar Paul, "Determination of Exponential Smoothing Constant to Minimize Mean Square Error and Mean Absolute Deviation".
[22] Michael Hahsler, Christian Buchta, Bettina Gruen, Kurt Hornik and Christian Borger, "Data Mining Association Rules and Frequent Item sets".