Product Review Rating using Sentiment Analysis

Nuwanthi Yapa
Department of Computer Science
University of Moratuwa
nuwanthi.yapa.19@cse.mrt.ac.lk

***Abstract*—In this internet era people tend to share information and communicate with each other through internet especially through social media. Most of the business get more information about their customers and make their decisions according to peoples reviews. Therefore, the reviews in social media plays a vital role in this commercial world. Meanwhile through this channel people may express ideas to exaggerate their own products quality, bad ideas about good products and good ideas about bad products intentionally to increase or decrease customer perception towards some products. After all, the main problem is the data or information that we want in the web are not in structured manner, not 100% truthful and then how to extract the accurate sentiments out of them after they converted into a structured format and finally how to get advantage over those sentiments in real world. This paper is proposed a solution for this main problem. In this paper, extract large volume of data from Twitter and YouTube social media by using crawler and filters those truthful reviews over fake reviews after doing pre-processing on extracted reviews. Then the system analyzes word sense disambiguation and do feature extraction and get polarity about a product and get the knowledge by using the information and the results will show in the dashboard by using data mining techniques. Outputs of the system are product profile which gives product polarity by showing positive, negative or neutral rating and a forecasting for the product.**

***Index Terms*—Sentiment Analysis, fake reviews, positive reviews, negative reviews.**

I. Introduction

In modern commercialized world every small product has very competitive environment in the market. Marketing in visual media, printed media and social media make a huge impact on the revenue of the product and the brand name. E-Commerce is a very popular globally which let the customer to get to know about a product or a company easily. Research shows that over 80% of internet users spend most of their time on social media sites, such as Facebook, Twitter, Instagram and Google Plus. Using social media for marketing can enable small business looking to further their reach to more customers. It is also very important to monitor online reputation, making sure your brand is being viewed favorably and quickly jumping on any negative comments or reports. Social media reviews can be used to get a rating about the online reputation of the brand or product where positive reviews being favorable for developing brand image and negative comments degrade brand image. It is very useful to get an analyzed idea about social media reviews on a product or company. As a solution for the problem, the proposed was a rating system for companies and products. This system integrates reviews from social media sites (Facebook, Twitter, YouTube, Blog and Amazon), extract the meanings of integrated reviews and provide a positive or negative rating based on analyzed data. Filters can be added for the rating for better use. Based on given rating, top management can make decisions on product/ company reputation and create more productive marketing strategies. They can decide what the features to be developed, what is the most interested target market of their product, what the featured to be eliminated and how they should develop their marketing strategies. Organizations can use this system to get decisions by considering the social media data from the above aspects as mentioned earlier and further social scientist can use this system to analyze the way people think and what they talk about in different aspects. General people also can use this system to get to know about available products or brands and how positive or negative they are among people. One main feature which differentiate proposed approach from other systems is that, proposed approach does not just state only positive and negative, instead it provides continuous values from 0 - 1 that can be read from negative to positive. Other than that, our system provides sentiment information with time and the final application can be used to analyze historical data and thereby make decisions using data mining techniques.

II. Background and Motivation

This is a business world which has millions of substitutions for every tiny thing. Therefore, in this commercial world, everything become a competition. Nowadays marketing, advertising is the most trending way to persuade any customer to any product. Among all those marketing and advertising channels, social media are the best channel of advertising. By today, social media are open platform of getting feedbacks, reviews of customer to a company, product or a brand name. Since people can ask questions about products, can make comments on products, can make comparison on substitutes, consumers are totally aware of every substitute. Therefore feedbacks, reviews can make highly impact on those sales. So, keep alert on those customer feedbacks and reviews is very important when it comes to long run business. Because of the expansion of social media, it is very difficult to analyze every customer review and feedback. Therefore, this proposed solution came up with an idea to analyze those customer reviews and feedbacks in social media and make a rating for company wise, product wise or band name wise on making decisions. The motivation for this research was, the existing systems integrate the reviews from social media web sites without considering the fake reviews and the existing systems do not provide a feature wise analysis for a system. Even though companies can get positive and negative comments on their products and about the company from social media, it is a complex task to analyze reviews one by one and get overall idea about product or feature. It takes much time and accuracy may be low. There are product rating websites to be used, but they cannot be customized for a brand. Although we develop a generalized rating system, it can be customized easily. It takes more time to analyze reviews to decide on negativity or positivity, to get summary about particular feature, product or company and it becomes more complex if user want to get a rating according to age, gender and in a given period of time. In this proposed solution, filters can be applied, and rating will be filtered as user preferences.

III. Literature Review

This section review the existing sentiment analysis systems and similar researches carried out to develop comprehensive sentiment analysis systems on social media.

A. Fake Reviews

Evaluation text on web, analyze those results and use those results on various scenarios is most popular method to provide opinion on products, services and events in this technical world. Putting an opinion or comment on web is like a tip of iceberg. It leads to a huge effect for decision making of consumers as well as sellers. Nowadays most of the consumers take their product buying decisions depending on other consumers opinion because of dissemination of information in social media Meanwhile social media platform becomes a channel for spreading of misinformation, rumors, fake messages and propaganda. That information will interpret good product as bad or bad product as good. It may mislead the consumers who seek opinions on products or services.

There are so many ways to detect spam reviews about products on social media done by researchers. They have different approaches to detect fake/spam reviews. There is one approach which attempt to find irregular or discontinuous text flow, vulgar language or not related to a topic and check similarity between comments[?]. It depicts some features of fake reviews to differentiate it from legitimate comments. Those indicators are incoherent reviews with so many/much number of punctuation marks, new lines, stop words, non-ASCII characters and white-spaces, coherent reviews which do not have relevant content and inadequate reviews which have offensive words. The paper proposed a supervised learning approach and experiment different data sets of features to correctly classify reviews as spam or not with help of natural language processing techniques. This system mainly divided into three modules as Feature Extraction Module, Post-Comment Similarity Module and Topic Extraction Module (Fig.1).

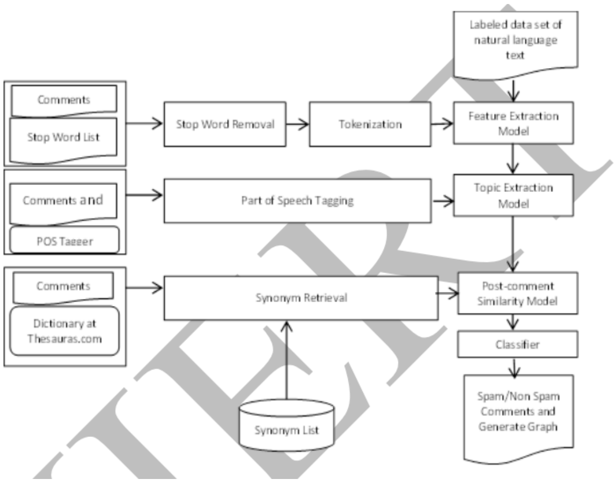

Fig. 1. System Architecture of Spam Review Detection System.

Another approach to find spam reviews is unsupervised iterative computer framework which considering both reviewers features and their interrelationships who work together[?]. And they worked on some assumptions.

- We assume that reviewers working together many times and always giving consistent opinions are suspicious
- We assume that reviewers who always represent supportive opinions with the labeled spam reviewers on some target product are suspicious

Based on number of features at tweet level and user level such as followers/follows, URLs, spam words, replies and hash tags there is another novel step to identify fake reviews proposed by another author[?]. He suggested a model with combination of three components, feature identification, preprocessing, learning algorithm and aggregate results. Below Fig. 2 depicts the model.

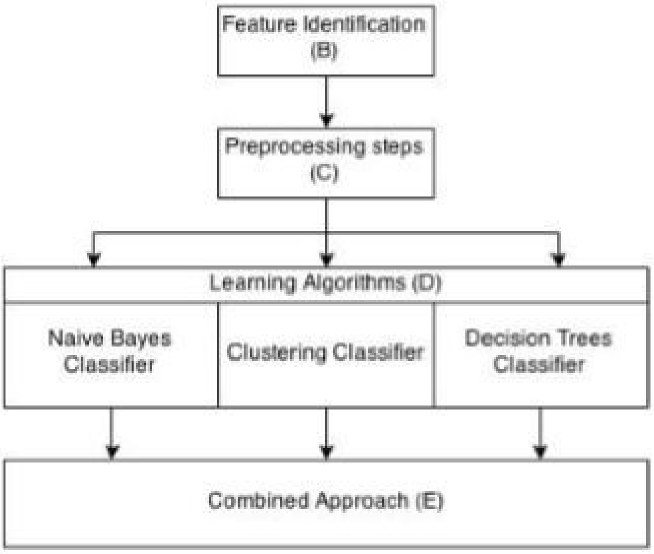

Fig. 2. Model of IJIERT Spam Review Detection System.

Another approach is to find spam reviews by using supervised classification method with machine learning techniques and text and natural language. Iteration computation framework is another approach to detect fake/spam reviews based on coherent examination based on flow smoothness information between sentences. It defines reviewers coherent metrics to analyze coherent of the review in the granularity of sentence. It uses word transition probability, conditional probability, word concurrence probability. This author suggests this approach based on some assumptions. The spam reviews have a connection with products or services which are either positive or negative depending on spammers intention. Most surrounding reviews are giving the biggest effect rather than outliers for consumers decisions. Spammers use untruthful sentiment words to express their deceptive idea to create positive or negative feeling in consumers minds.

Another approach is based on n-gram techniques. The problem is modelled as the classification problem and Nave Bayes (NB) classifier and Least Vector Machine (LS-SVM) are used on three representation. Another article proposed the degree of relevance Review Pertinence. It measures the pertinence of review by considering not only the similarity between a review and its corresponding product but also the correlation among reviews.

According to all the papers there are so many approaches detecting fake reviews but still in research level. In all those research papers have been addressed detecting spam reviews by either identifying fake reviewers and their reviews as fake or identifying fake reviews. But there is a gap because neither one of the papers does not talk about both, identifying fake reviews which are posted by original reviewers and identifying fake reviews which are posted by fake reviews.

B. Word Sense Disambiguation

Here we examine different approaches and techniques to the problem of Word sense disambiguation (WSD). In this section we present a review from most classical approaches to most recent and novel approach to the problem of WSD. There have been many researches have happened in the domain of WSD.

In supervised WSD it uses machine-learning techniques for inducing a classifier from manually sense-annotated data sets. Usually, the classifier (often called word expert) is concerned with a single word and performs a classification task in order to assign the appropriate sense to each instance of that word. The training set used to learn the classifier typically contains a set of examples in which a given target word is manually tagged with a sense from the sense inventory of a reference dictionary. Generally, supervised approaches to WSD have obtained better results than unsupervised methods.

A Naive Bayes classifier is a simple probabilistic classifier based on the application of Bayes theorem. It relies on the calculation of the conditional probability of each sense Si of a word w given the features fj in the context. The sense S which maximizes the following formula is chosen as the most appropriate sense in context:

$$\hat{S} = \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} P(S_i \mid f_1, \dots, f_m) = \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} \frac{P(f_1, \dots, f_m \mid S_i) P(S_i)}{P(f_1, \dots, f_m)}$$
$$= \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} P(S_i) \prod_{j=1}^{m} P(f_j \mid S_i),$$

where m is the number of features, and the last formula is obtained based on the naive assumption that the features are conditionally independent given the sense (the denominator is also discarded as it does not influence the calculations). The probabilities P(Si) and P (fj — Si) are estimated, respectively, as the relative occurrence frequencies in the training set of sense Si and feature f j in the presence of sense Si. Zero counts need to be smoothed because it leads to zero probabilities.

The decision tree algorithm is used to solve nonlinear classification problems. This algorithm constructs a top-down tree type structure recursively. Furthermore, it models a group for all the known values of the testing property using features that have gained maximum information to classify samples by testing all features like ID3, C45.

Maximum entropy is a general technique for estimating probability distributions from data. The overriding principle in maximum entropy is that when nothing is known, the distribution should be as uniform as possible, that is, have maximal entropy. In maximum entropy we use the training data to set constraints on the conditional distribution. Each constraint expresses a characteristic of the training data that should also be present in the learned distribution. We let any real-valued function of the document and the class be a feature, fi (d, c). Maximum entropy allows us to restrict the model distribution to have the same expected value for this feature as seen in the training data, D. Thus, we stipulate that the learned conditional distribution P (c—d) must have the property: In practice, the document distribution P (d) is unknown, and we are not interested in modeling it. Thus, we use our training data, without class labels, as an approximation to the document distribution, and enforce the constraint:

$$\frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} f_i(d, c(d)) = \sum_{d} \mathrm{P}(d) \sum_{c} \mathrm{P}(c|d) f_i(d, c) .$$

$$\frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} f_i(d, c(d)) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sum_{c} \mathrm{P}(c|d) f_i(d, c).$$

Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck that is, the lack of large-scale resources manually annotated with word senses. These approaches to WSD are based on the idea that the same sense of a word will have similar neighboring words. They can induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters. They do not rely on labeled training text and, in their purest version, do not make use of any machinereadable resources like dictionaries, ontologies etc. However, the main disadvantage of fully unsupervised systems is that, as they do not exploit any dictionary, they cannot rely on a shared reference inventory of senses.

A first set of unsupervised approaches is based on the notion of context clustering. Each occurrence of a target word in a corpus is represented as a context vector. The vectors are then clustered into groups, each identifying a sense of the target word. The similarity between two words v and w can then be measured geometrically, for example, by the cosine between the corresponding vectors v and w:

$$sim(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\,\mathbf{v}\,||\,\mathbf{w}\,|} = \frac{\sum_{i=1}^{m} v_i w_i}{\sqrt{\sum_{i=1}^{m} v_i^2 \sum_{i=1}^{m} w_i^2}},$$

C. Feature Extraction

When considering about reviews of products it is important to focus on features of products as users review products according to features of the product. As an example, review of a mobile phone can be given in many aspects such as camera, operating system, battery life etc. Feature selection in sentiment analysis is a significant role for identifying relevant attributes and it increases the accuracy of classification.

There are different kinds of features can be identified from literature review on sentiment analysis and those can be categorized as below.

Morphological Types There are 3 types of morphological features can be identified as

- Semantic
- Syntactic
- Lexicon structural

Semantic type of features in morphological types based on semantic orientation and contextual information. In contextual information method it is used to add text at sentence level.SO (Semantic Orientation) technique consists of point wise mutual information (PMI) and latent semantic analysis (LSA), those techniques assign polarity rank for each word or phrase.

The point-wise mutual information technique provides a formal way to model the mutual information between the features and the classes. This method was derived from the information theory.

In sentiment analysis proper feature selection technique is important and it does a significant role to identifying relevant features in a review and increasing accuracy. There are four main feature selection categories as

- NLP Based
- Statistical based
- Clustering based
- Hybrid

NLP based feature selection techniques are mainly operate on 3 basic principles. Those approaches are based on with POS tagging of words in sentiments

1) Noun, Adjectives, Adverbs usually describes a feature of product.
2) Terms occurring near subjective of sentences can be select as a feature.
3) If take P- product and F-feature , then in a sentence if it includes F of P ,P has F ,then we can select features from those phrases.

Those NLP based techniques has high accuracy than other techniques but the accuracy of it depend on the accuracy of POS tagging as it is the main technique used here to extract features.

Statistical based techniques further divided into 3 types as univariate, multivariate and hybrid.Univariate statistical based method takes attributes separately and examples of univariate type includes occurrence frequency, log likelihood information gain (IG),chi-square. Those univariate techniques have computational efficiency, but it ignores interaction of attributes. When consider about multivariate techniques it uses genetic algorithms, recursive feature elimination and decision trees. Comparing to univariate methods multivariate methods need high computational power.

Hu et al.,[2016] applied hybrid technique for data extraction such as POS tagging combined with WordNet dictionary of NLTK.Frequent feature set identification was done using Association Miner CBA.

D. Sentiment Analysis

When comes to the sentiment analysis, there are 3 main classification levels in sentiment analysis can be identified as , document-level SA, sentence-level SA and aspect-level SA.In document level SA it aims to classify document and giving polarity for whole document as considering it as a one topic. In sentence level Sentiment Analysis, it aims to express polarity of each sentence in documents as each sentence is whether negative or positive. In document-level SA and sentence-level SA there is not a fundamental difference as sentences are just short documents.document-level SA and sentence-level SA does not gives necessary details and to obtain the necessary details we have to go to aspect level SA.By reading more research papers and after went through online surveys it is identified as there are many sentiment Analysis algorithms were proposed in last few years.

As shown in the above Fig.4 sentiment classification techniques can be divided into 2 main categories as Machine learning approach and Lexicon based approach. The sentiment classification of machine learning can be further divided into two categories as supervised ML and unsupervised ML.Here supervised ML needs a large no of labeled training data set and unsupervised machine learning used when it is difficult to find labeled training data set. Lexicon based approach can be categorized into two categories as dictionary-based approach and lexicon based approach, those methods depends on finding opinion lexicon that is used to analyze the text.

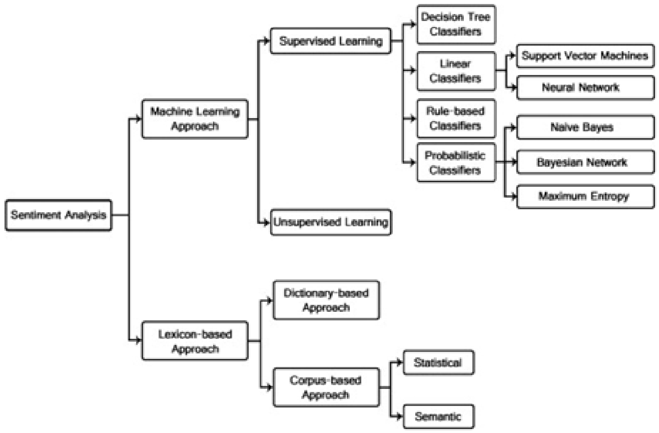

Fig. 3. Sentiment Analysis approaches

E. Product Profiling and Forecasting

Data mining use well researched statistical principles to discover the patterns of data. It can identify patterns, forecast trends and support decision making. This is done using data mining algorithms and using data mining tools such as Weka. Here we mainly focus about data mining algorithms as we prefer to use algorithms in our system. The data set which is to be used in data mining algorithm should go through the sentiment analysis tool. Mainly there are three algorithms which are used to discover the patterns in data sets.

Basic idea about clustering is gathering data with similar characteristics into one cluster. Clustering is an unsupervised learning method which can be used for both sentiment analysis and prediction models. k means clustering is widely used in many researches. Clustering is being discussed in [19]. In this method, twitter data is used as input for the algorithms and classified into two categories positive and negative. Then performed how much percentage of data falls for each category. After that sentiment is performed. These clusters are being used for prediction purposes of new test data sets.

Linear regression model is basically used for numerical classification. Below linear function shows the weight for each attribute. sur and Huberman predict movie box office

$$x = \beta_0 w_0 + \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_n w_n$$

sales using tweets related to a particular movie.it examine the distribution of the tweets of different movies in different period of time.also they consider the linked urls in tweet and retweets in their model. They gather data continuously for three weeks and plot gathered data in a graph for visualization. Then get the correlation between urls with tweets and retweets with box office performance. Get the tweet rate and construct a linear regression model among the variables considered.

In ref linear regression model is used for predicting election results. Tweets mentioning political parties are collected and ranking them based on tweet volume. Then get the ranking of the election results and check whether both are identical.

IV. Proposed Approach

After analyzing and getting details on the existing sentiment analysis systems and systems that are yet in research level, new approach is suggested to develop reliable and effective solution to get the maximum use of consumers sentiments on products or brands which is called Sentiment Analysis on Social Media.

To analyze data, data needs to be in correct format. There are thousands and millions of sentiment data in the web, especially in social media sites that can be used to get valuable conclusions. But they are not in a correct format or not in a structured way to get hundred percent usage from them. So, it needs to convert them to a correct format and use them as we want. This is the first part of proposed solution, which is developing a crawler to get data what we want and store them back as we can use them anytime. Twitter social media is used to analyze sentiment. The crawler that developed for Twitter social media are capable of crawling current data and getting past data. On the other hand, this crawler is capable of getting information about the people who update the statuses for product profiling purposes.

After having a large source of data which is in a structured manner, the next thing that must take place in the project Sentiment Analysis for Social Media is analyze sentiments. Since sentiments are in different languages in different ways, English language is focused through this solution. So, analyzing sentiments is kind of analyzing a natural language and therefore this part is about natural language processing. For this we use Natural Language Toolkit, also known as NLTK which is a leading platform for building Python programs to work with human language data. There are different ways that we can use to analyze sentiment data using this toolkit, but none of them gives hundred percent accuracy.

As the next part of the solution, it needs to identify the crawled data are valid or not, because there can be fake reviews which are posted by fake reviewers or fake reviewers posted by original reviewers. For that we are crawling information both users and reviews. This module divided into two sub components,

- Identifying fake reviewers
- Identifying fake reviews

To achieve this, it needs to identify some features of both fake reviewers and fake reviews. To get these features and features behavior on fake reviewers and fake reviews,enough fake data set will be trained. By using those features we are implementing an algorithm to give score for each feature and aggregating those scores, finally it will give probability of being fake for each one.

In first part by using fake probability that higher than predefined fake probability margin, those reviewers are mentioned in the database, and it can be identified those as fake reviewers. For the second part, input reviews will be fed which are not posted by fake reviewers as we identified. Then carry out the process and assign score for being fake to each feature of review and aggregate the results and give probability of being fake for each review. For further procedure, it will give lesser probability of being fake according to the identified fake probability margin.

After removing the fake reviews, it needs to identify the correct sense of the reviews as a word may have several senses. It needs to certify the reviews are going to analyze are relevant to the product. To achieve this, word sense disambiguation is used. Hybrid approach will be proposed which combines several supervised learning techniques along with the bigram model to achieve this task. A model is trained the combined classifier by feeding data sets of each sense of the search words which are going to use to search products and according to the trained data classifier will then disambiguate a coming review.

From the module of word sense disambiguation, it returns the data of appropriate product removing other ambiguate data crawled by crawlers. Then using those relevant data and doing a feature descriptions extraction from those data. In sentiment analysis proper feature selection technique is important and it does a significant role to identifying relevant features in a review and increasing accuracy. So, decided to implement a part of feature extraction for the project. From many feature extraction techniques chose NLP based feature descriptions extraction method to implement the module.

In the next part of the project it is going to implement a module of analyzing sentiments using machine learning techniques. Under machine learning techniques chose supervised learning technique to implement sentiment analysis module which predict the polarity of given input text. Under Supervised learning technique decided to use two classifiers as probabilistic classifiers and linear classifiers. Under probabilistic classifiers chose 3 classifiers to implement the module as naive bayes classifier, bernoulli naive bayes classifier and multinomial naive bayes classifier to implement classifiers and under linear classifiers chose to use linear SVM classifier and logistic regression classifier. Those 5 classifiers will be trained using same training data set and then it predicts 5 outputs from those 5 classifiers.

Then as the next part of the project, its going to create a dashboard to show the results from the above two parts which are the crawler and sentiment analysis using python. Here it is going to display how the sentiment polarity differs for a selected item with the time using a graph. Using this it can be seen how it is changing the user sentiment polarity of a specified brand or product with time changes by the users of this system.

Output of the sentiment analysis is used as the input of the data mining. Preprocessed data will be used in this module. This gives mainly two outputs to the dashboard which are product profiling and forecasting. Extracted data should be preprocessed by removing hashtags, url and replacing missing values.

Sentiment data and user data will be sent to the data mining as the input. User data consist of demographic variables such as date of birth, location and profession of the user and the sentiment data is the tweet, which is labelled as positive, negative and neutral. Database consists of set of tables with entity attributes and with relationships. Mainly two outputs are to the user- Product Profile and the Forecasting.

Product profile gives the how ratings change over the time. Previous and current data are used here. Output is a given as graph, rating of the product against time. This output can be taken through filters such as age group, location and profession. This gives a better view for the product owners and product users to get an idea about a product.

Second output is forecasted rating. Previous and current data is analyzed through time series analysis and forecasting methods are used on the time series analysis to give the forecasted output. In proposed system ARIMA model, nave forecasting, simple moving average, weighted moving average, exponential moving average, adaptive rate smoothing methods are used for trend products and holt winter method for seasonal products. Considering the accuracy of each method, highest accurate method is selected and used for the implementation.

All these outputs are integrated into the dashboard for visualization purpose. This system gets data from crawler and input those to sentiment analysis tool and use the output of sentiment analysis tool and data mining processes to show how the sentiments are changing over the time and how user sentiments can be used in decision making processes in businesses through product profiling, trend analysis and forecasting.

Anyone who interested in searching what people say about a product or brand can use this system. This is especially very useful for organizations producing products and services to know what people are talking about their products and services and were they positive or negative, who are the competitors they have and what they can do to improve the reputation their products and services among customers and what features should they include with them in order to make the customers positive with their products.