International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13282-13287
© Research India Publications. http://www.ripublication.com

Optimal Product Recommendation from Real-time Data Feed

Nithin N V
M.Tech Student, Department of Information Science and Engineering,
Ramaiah Institute of Technology, Autonomous Institute Affiliated to VTU Bangalore, Karnataka, India.

Prof. George Philip C
Associate Professor, Department of Information Science and Engineering,
Ramaiah Institute of Technology, Autonomous Institute Affiliated to VTU Bangalore, Karnataka, India.

ABSTRACT

Online reviews have become one of the most important aspects of any business. Posting reviews online for products bought or services received has become a trendy approach for people to express opinions and sentiments about a product. Social media like Twitter contains rich information about people’s preferences and holds enormous amount of data about many products. In the existing system, when a person wants to buy a product online, he or she will typically start by searching for reviews and opinions written by other people. A buyer can be misled by false recommendations of products, which may confuse people to draw a conclusion about the product. Also, the high volume of reviews for a single product makes it harder for users as well as manufacturers to choose the best reviews and understand the truth underlying the quality of a product. Sentiment analysis provides a way of evaluating user experience. The work done so far in this field, helps in understanding emotions and identifying age groups from twitter. Classifying the tweets into positive, negative and neutral is a challenging task. The aim of the project is to provide recommendations that can be used by customers to buy a product online and also help manufacturers improve the overall product features.. The project mainly focuses on collecting the real-time tweets from Twitter app, using tweepy. The TextBlob library is used in categorizing tweets into positive, negative and neutral by sentimental polarity scores. The categorized data is evaluated by classifier algorithms like Naïve Bayes, Maximum Entropy and Support Vector Machine. The outcome of each of the algorithms is the various metrics like accuracy, precision, recall, and F-measure. By comparing the average accuracy obtained from the algorithms, a user can decide which brand of the product is good to buy. Also, manufacturers can gauge how a product is faring in the market. The proposed system is implemented in Python language using the Jupyter Lab 5.5.0 development platform. The novelty in the proposed system is that real-time data will be analyzed from the most recent twitter feeds. The proposed model can be used to review any product for which sufficient tweets are available, as it gives optimal results by applying all three algorithms.

INTRODUCTION

With the quick growth of the Internet, product related reviews which were word-of-mouth conversations earlier, have now migrated to online social media. One’s surroundings and relationships are getting closer through social media and has now-a-days become part of one’s life. The decisions made in day-to-day activities are now often based on social media opinions. Researchers are more influenced towards social media when they have to search for new information resources. So they utilize the social media to gather information and observe interesting inferences, that are happening around them. Twitter is a social media platform, where one can find customer’s reviews and suggestions regarding a particular product. People find this medium, pretty good to share about their day-to-day activities. Because of this, researchers are quite happy to gather information about small or big problems from this huge platform. The concern of the customer is to know what other people are experiencing about the items or services of competitors. Detailed and timely advice from the statements of the customer helps in ahead a competitive advantage and help in new product invention for companies. Such information helps in target key advertising campaigns. Sentiment Analysis helps corporates to get consumers opinion in real-time. This real-time data helps them to design new advertising strategies, improve item features and predict the chances of item failure.

The motivation of the project is giving recommendations to a customers about a products they are interested in. It means analysing the reviews and feedback of existing customers of the product and to help an user to decide which brand of a product is good to buy. Also, manufacturers can gauge how a product is faring in the market.

LITERATURE SURVEY

Will discusses various papers which covers methodologies for an optimal product recommendation system. Haruna Isah [1], discusses about how user made content from social media platforms can provide early clues about product allergies, adverse events and product duplicating. The global scourge of fake drug and cosmetic items poses an excessive danger to public safety. One strategy for fighting fake items is through the effective communication and tracking of early cautioning sign of product antipathies, side effects, drug resistance and diseases occurred. The approach to these glitches is by applying text mining and sentiment analysis techniques on drug and cosmetic product related text data. M. Mazhar Rathore [2], discuss the processing of an rich amount of numerous social networks data to look after the earth events, diseases, incidents, user trends, and views to make future real-time decisions and facilitate imminent planning. The framework contains five layers namely data collection, data processing, application, communication, and data storage. The system organizes Spark at the top of the Hadoop ecosystem in order to run real-time analysis. Therefore, collecting such real-time geo-social data is a very challenging task. It requires a special computational environment and advanced computing techniques with intelligent management in order to provide in-time/real-time analysis. Sonia Mashal [3], discuses about how the emotion analysis on Twitter data pose problem of insufficient characters, since the authors tries to express all their emotions within that short length, and also the huge volume of slang increases the pre-processing required. The author only focuses on the task of emotion analysis and classification, using machine learning techniques like Naïve Bayes and SVM. The data used is from Twitter. The data was automatically collected by using emotion hashtags for five emotion categories namely happy, sad, angry, fear and surprise. The datasets needed for classification tasks are huge, which provide a better coverage of the vocabulary, that may not occur with less amount of Twitter data. Sudhir Kumar Sharma [8], discusses about how the opinions of public about political competitors can be investigated to gauge results of a election. Earlier, organizations, governments and business entities used to gather information on focused groups for acquiring native opinions and their perspectives. Twitter as a social media communicating web application with microblogging features, has a huge and continuously growing user database. Along with, the application gives a rich informational collection, in the form of messages that are generally short status keep posted from users that are expressed in not more than 140 characters in length. Shubham Goyal [9], discusses how everyone in the modern age is included with some web based life platform, and how the public mood enormously reflected in the social media today. People life loaded up with emotions and opinions. One can't envision the world without them. They lead the human life by impacting the manner in which way we think, what we do and how we act. It has changed the manner in which we share data.. The receivers of the information do not only consume the available content on web, but in turn, actively interpret this content and produce new fragments of information. The author proposes to use this source of data and foresee the sentiment or feelings of public towards a subject. Food price crisis is being considered here and public opinion is predicted for the food crisis topic only.

From the above discussion, influence of social media in the day-to-day life of people are understood. Based on the research and survey done in the field of social media like twitter, it can be concluded that sentiment analysis helps to know about the consumers views about any product by analysing social media reviews or experiences of the consumers. The proposed system aim to correctly identify the sentiment of the tweets and produce comparatively accurate results of the searched item in the system with a supervised base model using classifier algorithms like Naïve Bayes, Max. Entropy and SVM. Hence, the system can give optimal recommendations to the consumers and for manufactures.

Assumptions made in the paper is carried out on collecting past 38 to 48hrs of real-time data. For the next 24hrs, may be the data will differ/vary. The results & accuracy may also differ. After discussing functional and non-functional requirements, let's move to the detailed description of the analysis and design part of the system.

ANALYSIS

Textblob[4] is a Python library for processing textual data. It delivers a simple API for diving into common natural language processing (NLP) tasks. Textblob is used to analyse the collected tweets in the proposed model and also to categorize the tweets into different categories. The mathematical model involves the classifier algorithms like Naïve Bayes, Maximum Entropy and Support Vector Machine to analyse the categorized data and to train and test the model to obtain the results of each classifier with various metrics like accuracy, precision, recall and f-measure. This helps to know how best the collected data is trained and tested for the categorized data, with cross validation process. The details of the metrics which are used in this approach are later explained in chapter 7. The mathematical models used are briefly explained below:

Naïve Bayes

The Naïve Bayes (NB) [17]"classifier is based on Bayes rule, a practical Bayesian learning model that is easy to understand and implement. The Bayes rule allows us to determine this probability of any event. Nave Bayes Classifier[5] makes usage of all the features in the feature vector and inspects them distinctly as they are equally independent of each other. There is an inbuilt library for Naïve Bayes in NLTK. The detailed description of this algorithm is as shown below: "

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Where $p(C_k|x)$ is posterior probability, which is the statistical probability that a hypothesis is true calculated in the light of relevant observations. P(x) is the predictor prior probability, that is the probability as assessed before making reference to certain relevant observations. $P(x|C_k)$ is the likelihood of underlying labels observed in the training data, that is the probability of predictor given class. $P(C_k)$ is the prior probability of class, defined as probability distribution that would express one's beliefs about this quantity before some evidence is taken into account. "

Maximum Entropy

In Maximum Entropy Classifier[6], no assumptions are taken regarding the relationship between features. The principle in Maximum Entropy is to model all that is known and assume nothing about that which is unknown. This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label. There is also an inbuilt library for Maximum Entropy in NLTK. The detailed description of this algorithm is as shown below:

$$P_\lambda(y|X) = 1/Z(X)exp\left\{\sum_i \lambda_i f_i(X,y)\right\}$$

‘X’ is the feature vector and ’y’ is the class label i.e., positive, negative and neutral. Z(X) is the normalization factor and λi is the weight coefficient.

Support Vector Machine (SVM)

SVM Classifier[7], uses large margin for classification. It separates the tweets using a hyperplane. There is also an inbuilt library for SVM in scipy. The detailed description of this algorithm uses a discriminative function defined as:

$$g(X) = w^T \phi(X) + b$$

‘X’ is the feature vector, ’w’ is the weights vector and ’b’ is the bias vector. ‘Ø’ is the non-linear mapping from input space to high dimensional feature space. ’W’ and ‘b’ are learned automatically on the training set. ”

UML DESIGN

System design is the process of describing the architecture and models for a system to satisfy the specified requirements.

This design part includes:

1. System Architecture Diagram
2. Use Case Diagram

System Architecture

The system architecture fig 4.1 shows the overall design of the proposed system. In the proposed architecutre, process starts from the collection of tweets from the Twitter application using tweepy. It considers only the tweets, which discribes the feelings/suggestions/expriences of the user on any particular product in social meida.

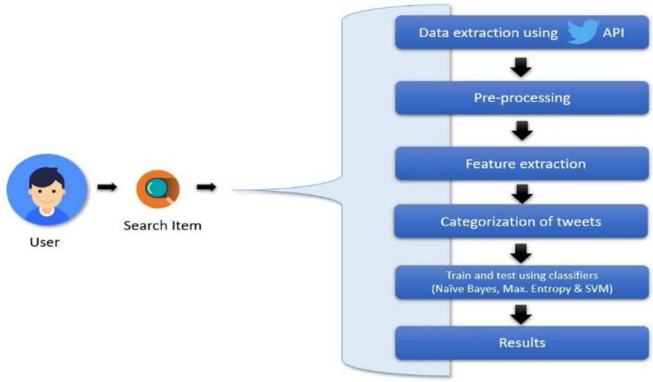

User
Search Item
Data extraction using API
Pre-processing
Feature extraction
Categorization of tweets
Train and test using classifiers (Naïve Bayes, Max. Entropy & SVM)
Results

Figure 4.1 System Architecture

System Architecture

After collecting the raw data, the pre-processing step will start to clean the data by removing the punctuation marks, unwanted text, and converts usernames and extra white spaces to a specified format. After pre-processing, feature extraction will be carried out to extract the special and informative words. Next, the collected data is categorized into positive, negative and neutral on polarity basis using an open source library called TextBlob. After that, by applying the classifier algorithms like Naïve Bayes, Maximum Entropy & SVM to train and test the model. The outcome of each of the algorithms is the various metrics like accuracy, precision, recall, and F-measure. By comparing the average accuracy obtained from the algorithms, a user can decide which brand of the product is good to buy. Also, manufacturers can gauge how a product is faring in the market.

Use case diagram

Use case diagram fig 4.2 represents the pictorial representation of the several tasks involved in the process, like possible use cases, the relationship between the use cases and the actors. Use cases find the functionalities of the system. Here, actor/user is the customer or company and tweepy is used to extract the tweets from the Twitter app. Now the pre-processing of the collected data is done to clean the noisy/raw data. Feature extraction is carried out to extract the useful and special words from the collected data. The library textblob is used to, categorize the tweets into positive, negative and neutral statement. Now classifier algorithms are applied to the train and test the model, The results of each of the algorithms is the various metrics like accuracy, precision, recall, and F-measure. By comparing the average accuracy obtained from the algorithms, a user can decide which brand of the product is good to buy. Also, manufacturers can gauge how a product is progressing in the market.

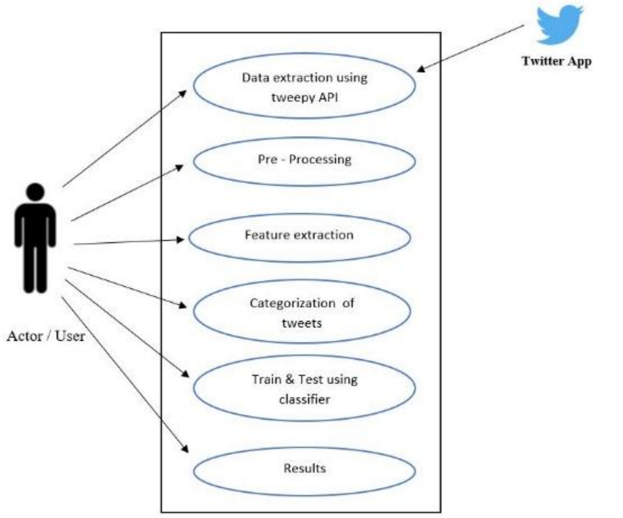


Figure 4.2 Use case diagram.

IMPLEMENTATION

System implementation involves the process of converting the system design into a working system. The main purpose of system implementation is to design and create system models that meet the required specifications. Python language is used, which can run on any operating system or platform, it includes standard libraries. Provides an interactive mode, made easy to test short snippets of code. It is mainly used for many applications like natural language processing, machine learning, and data analysis.

Environment Setup

To setup the environment the following software's and packages are required:

- Anaconda Navigator 5.4 version software.
- Numpy package
- NLTK library package
- Sklearn package
- Pandas package
- Collections package
- OS package
- Textblob package
- Tweepy package
- Scipy package

Data Collection

Twitter app is used to extract the data by using the tweepy library. Twitter provides authentication mechanism to allow access to user data. Twitter currently uses oauth authorisation mechanism which allows users to grant third parties access to their data without using their username and password. By registering an application, Twitter provides a consumer key and consumer secret, which will uniquely identify the Twitter application. Using the consumer key, consumer secret, access token and access token secret helps in connecting to the Twitter app. Once the authentication has been done, the streaming API instance filter real time public tweets containing keyword i.e., search item. Streaming data is stored in a csv file for further processing. Next is the pre-processing stage to clean the data.

Pre-Processing

Pre-processing is fundamental to all Natural Language Processing (NLP) Task. Pre-processing is one of the most important step, for accurate information retrieval, and to have a proper data set. Pre-processing is implemented to clean the noisy data. This process covers the removal of punctuations, additional white spaces, and converts the usernames to at_user, upper case letters to lower case letters, #word to word.

Feature Extraction

Feature extraction is one of the text pre-processing step. Normally tweets are in the form of unstructured text data and it usually requires a transformation of text into a representation of processable format. The feature extraction is based on tokenizing the text, which means the process of splitting sentences into list of tokens.

Categorization of tweets

There are many techniques available to extract natural text and find out the sentiments like happy, sad, bored, angry etc. Tweets are not always positive or negative. It could be neutral also. The sentiment scores can be calculated using sentiment compound polarity using TextBlob library. It is based on the sentiment type and categorizes as positive, negative and neutral. The following routine explains clearly how the TextBlob classifies the tweets into different categories. If the sentiment score of the tweet is greater than zero, it is classified into positive statement, equal to zero means, it is categorised as neutral tweet and remaining tweets are categorized as negative tweets. Calculates the polarity like:

If : sentimental polarity of tweet is > 0 => positive.

If else : sentimental polarity of the tweet is = 0 => neutral.

Else : remaining tweets are => negative.

Training and Testing of tweets

Machines are faster at processing and storing data compared to humans. This is referred to as leverage of speed. Machines can also learn when there is enough relevant data, known as training data. The content of the training data often referred as labeled or human labeled data format, is designed to train specific machine learning models. Normally machine learning model needs to be tested in the real world to measure how robust its predictions are. Here mainly Naïve Bayes, Maximum Entropy and Support vector classifier algorithms are used to train and test the collected data.

RESULTS

This chapter discusses the results of the implemented system. The system is trained, tested and evaluated by different classifiers and from each algorithm the obtained metrics factors like accuracy, precision, recall, and F-measure are observed. From the comparison of average accuracy obtained from the algorithms, an user to decide which brand of a product is good to buy. Also, manufacturers can scale how a product is faring in the market. The metrics[21] are helpful to obtain the results based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values.

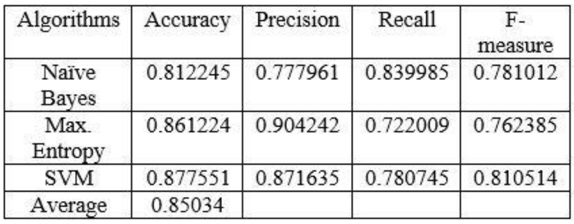
Accuracy : (True Positive +True Negative) / (True Positive + False Positive +False Negative + True Negative).

Precision : (True Positive) / (True Positive + False Positive)

Recall : (True Positive) / (True Positive + False Negative)

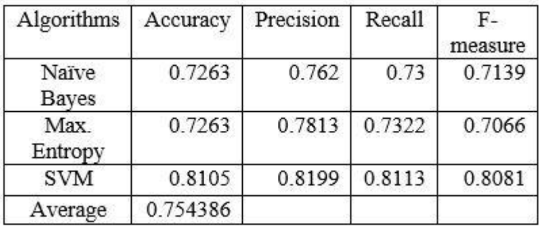
F-measure : 2* (Recall * Precision) / (Recall + Precision)

After categorizing the tweets, the tweets are fed to the training model. The evaluations are done and rechecked by cross-validation process. In this process, first the positive and negative tweets are mixed and then it is randomly shuffled. This is essential because in cross-validation, if the shuffling is not done, then the test chunk might either have only negative or only positive tweets data. Following are the results which are obtained from the different classifiers that are tabulated for products A, B, C and D. Also graphs are plotted to have better understanding of the results.



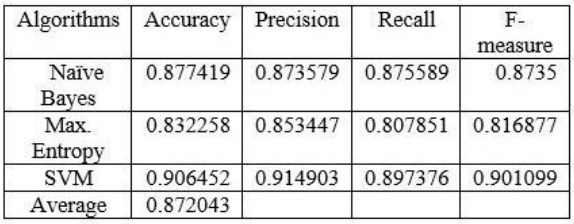
Algorithms	Accuracy	Precision	Recall	F-measure
Naïve Bayes	0.812245	0.777961	0.839985	0.781012
Max. Entropy	0.861224	0.904242	0.722009	0.762385
SVM	0.877551	0.871635	0.780745	0.810514
Average	0.85034			

Table No. 1: Result values of product A



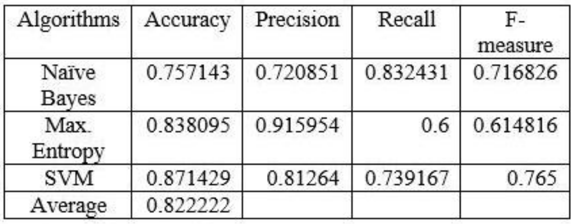
Algorithms	Accuracy	Precision	Recall	F-measure
Naïve Bayes	0.7263	0.762	0.73	0.7139
Max. Entropy	0.7263	0.7813	0.7322	0.7066
SVM	0.8105	0.8199	0.8113	0.8081
Average	0.754386			

Table No. 2: Result values of product B



Algorithms	Accuracy	Precision	Recall	F-measure
Naïve Bayes	0.877419	0.873579	0.875589	0.8735
Max. Entropy	0.832258	0.853447	0.807851	0.816877
SVM	0.906452	0.914903	0.897376	0.901099
Average	0.872043			

Table No. 3: Result values of product C



Algorithms	Accuracy	Precision	Recall	F-measure
Naïve Bayes	0.757143	0.720851	0.832431	0.716826
Max. Entropy	0.838095	0.915954	0.6	0.614816
SVM	0.871429	0.81264	0.739167	0.765
Average	0.822222			

Table No. 4: Result values of product D

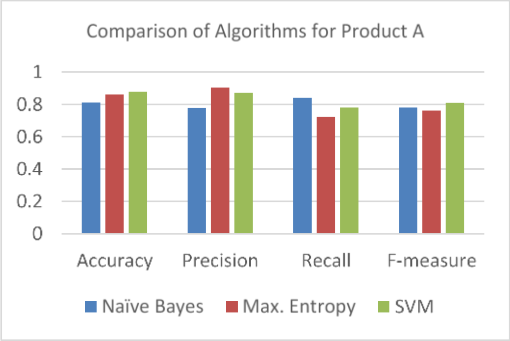


Fig 1: Comparison graph of algorithms for product A.

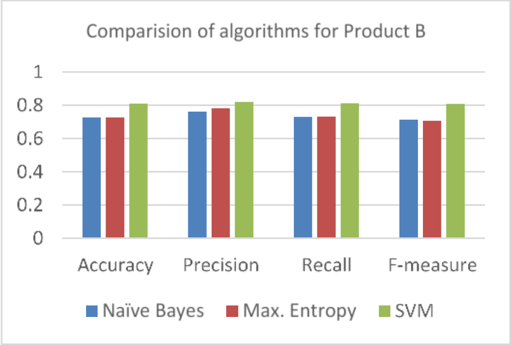


Fig 2: Comparison graph of algorithms for product B.

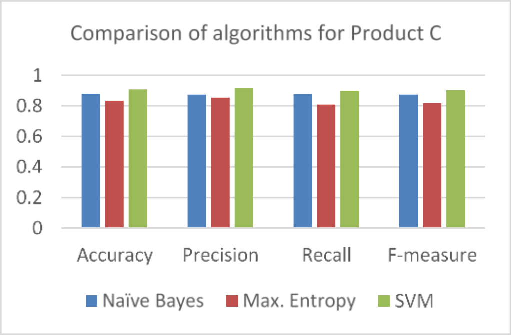


Fig 3: Comparison graph of algorithms for product C.

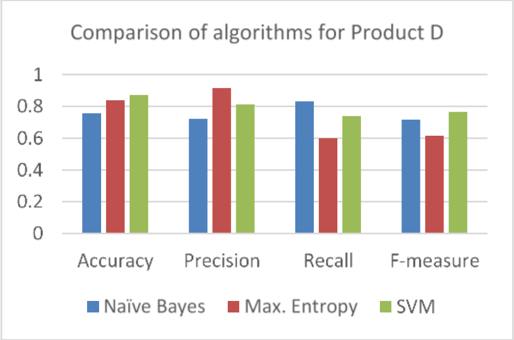
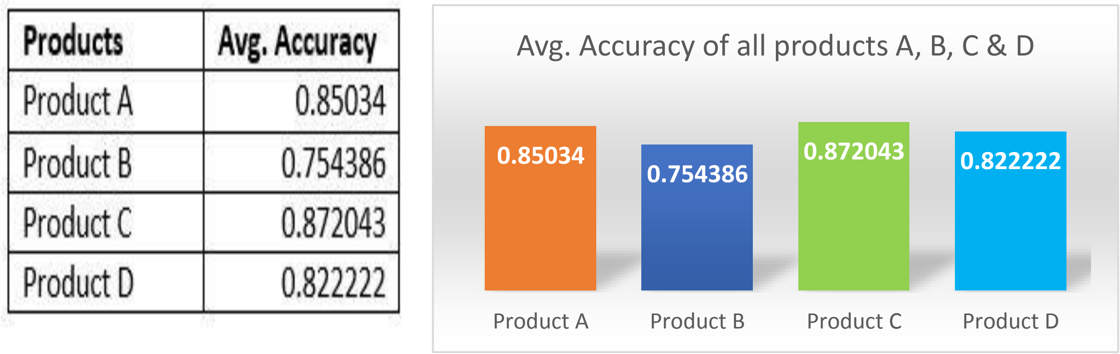


Fig 4: Comparison graph of algorithms for product D.



Products	Avg. Accuracy
Product A	0.85034
Product B	0.754386
Product C	0.872043
Product D	0.822222

Fig. 6: Avg. Accuracy table and comparison graph for products A, B, C & D

From Fig 6, it can be seen that product C has gained greater average accuracy using three classifiers. And also product C has gained more accuracy compared to other products. As noticed, the four products have taken some variations in accuracy rate in the real world i.e., Twitter. By comparing the average accuracy obtained from the classifiers, a user can decide which brand of the product is good to buy. Also, manufacturers can gauge how a product is faring in the market.

CONCLUSION

Twitter has useful information about many products. Twitter platform is helpful to find opinions about a product very quickly and easily. This project mainly focuses on helping a costumer to buy a product. Collection of raw tweets form the Twitter app using tweepy is done first. The collected raw data, is pre-processed to clean noisy data. And then categorized into positive, negative and neutral on polarity basis using an open source library called textblob. After that, classifier algorithms like Naïve Bayes, Maximum Entropy & SVM are used to train and test the model. The results of each of the algorithms are the various metrics like accuracy, precision, recall, and F-measure. By comparing the average accuracy obtained from the algorithms, a user can decide which brand of the product is good to buy. Also, companies can measure how the product is faring in the market. By this, customers have clear vision about the product. The implemented system will help the company for the brand management and also to the users to buy the best product online. Finally, the purpose of the entire project is to advice customer to buy quality and genuine products and help companies to locate the best reviews and understand the true underlying quality of a product. By this it is helpful in increasing the overall product quality.Realising the system was a very challenging task. Some of the important features that can be included in future are, by increase the volume of the collected twitter data to improve results. Designing an automated tool for the developed system. With these enhancements, the system should be able to provide good insights to the system in future.

REFERENCES

[1] Haruna Isah, Paul Trundle, Daniel Neagu, "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis", IEEE 14th UK Workshop on Computational Intelligence (UKCI), pp. 2162-7657, October 2014.

[2] M. Mazhar Rathore, Anand Paul, Awais Ahmad, "Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions". IEEE International Conference on Communications (ICC), pp. 1938-1883, May 2017.

[3] Sonia Xylina Mashal, Kavita Asnani, " Emotion Analysis of Social Media Data using Machine Learning Techniques", IOSR Journal of Computer Engineering, PP 17-20, 2017.

[4] Textblob is a python library for processing textual data. Sentiment analysis, classification and more; https://pypi.org/project/textblob/

[5] Hai yi Zhang; Di Li ,"Naïve Bayes Text Classifier approach", IEEE International Conference on Granular Computing, pp. 7695-3032, November 2007.

[6] Behrouz Behmardi, Raviv Raich, Alfred O. Hero, "Entropy estimation using the principle of maximum entropy" pp. 1520-6149, May 2011.

[7] Kwang In Kim, Keechul Jung, Se Hyun Park, Hang Joon Kim, "Support vector machines for texture classification", pp. 1542 – 1550, November 2002.

[8] Ajay Deshwal, Sudhir Kumar Sharma, "Twitter Sentiment Analysis using Various Classification Algorithms", 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), pp.978-1-5090-1489-7, September 2016.

[9] Shubham Goyal "Sentimental Analysis of Twitter Data using Text Mining and Hybrid Classification Approach", International Journal of Advance Research, Ideas and Innovations in Technology, Vol. 2, Issue z5, pp. 2454-132X, 2016.