Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 2533 – 2543

Online Product Recommendation System Using Sentiment Analysis and Spam Filtering

Miss. Mayuri Chaudhari, Prof. Nagaraju Bogiri

Computer Department
K J College of Engineering and Management Research
Pune, India

ABSTRACT

E-commerce has recently concentrated on the rise of internet retail platforms and can be recommended by multiple consumers who refer to product review opinions to pick their goods and products from their purchasing experience. But, if current product evaluation programs include customers' problems (eg. cost, power, structure, function, etc.) conventional recommendation systems do not recommend alternative products, it is also difficult to meet user requirements. In this paper, we suggest therefore a new product recommendation framework that analyses two forms of information: information on grievances and information on satisfaction from e-commerce review comments. Alternative things can be recommended to meet the needs and can resolve the product issue details as you browse. In this article, we would identify the extraction of complaint information by removing both negative information and positive information from product feedback and clarify alternative product recommendation approaches for complaints resolution and check the efficacy of complaint information extraction and alternative product recommendations.

Keywords— Machine Learning, Social Media, Text Mining, Text Classification, Sentiment Analysis, Online Reviews.

I. INTRODUCTION

Many businesses and software sectors store their data in Social networking creation provides the customer with an ability to share his or her views. That means the organization can't monitor the contents of the virtual universe now. Complaints in social media are submitted by customers who are not pleased by a company's services or goods. On the other hand, consumers are still optimistic for a commodity in the social media. This view could affect other potential clients, including positive or negative ones. Potential consumers can find out about a certain product before deciding to purchase goods.

An appraisal of the sentiment is expected to immediately decide whether the feeling is negative or positive. Feeling analyses are a subset of text mining that focuses in the text of a person's feeling, mood and attitude. The fundamental theory of sentiment analysis consists of categorizing the polarity of texts and determining whether they are positive or negative. Sentiment analyses are commonly used as rapid social network growth. For different places public opinion is becoming really critical. There have been some difficulties in collecting public examination.

Many product evaluation pages have recently been published on the Internet. It invites scientists to carry out a consumer review sentiment analysis. On product evaluations, customer opinion was evaluated in this paper.

II. RELATED WORK

In this paper [1], author proposes an approach which integrates content and usage information to detect fake product reviews. The proposed model exploits both product reviews and reviewers' behavioral traits interlinked by specific spam indicators. In this paper, fine-grained burst pattern detection is employed to better examine reviews generated over "suspicious" time intervals. Reviewer's past reviewing history is also exploited to determine the reviewer's overall "authorship" reputation as an indicator of their recent reviews' authenticity level.

This study [2] adopts a big data analytical approach to investigate the impact of online customer reviews on customer agility and subsequently product performance. authors develop a singular value decomposition-based semantic keyword similarity method to quantify customer agility using large-scale customer review texts and product release notes. Using a mobile app data set with over 3 million online reviews, our empirical study finds that review volume has a curvilinear relationship with customer agility. Furthermore, customer agility has a curvilinear relationship with product performance. this study contributes to innovation literature by demonstrating the influence of firms' capability of utilizing online customer reviews and its impact on product performance. It also helps reconcile inconsistencies found in literature regarding the relationships among the three constructs.

In this research [3], authors discovered that reviewers' posting rates (number of reviews written in a period of time) also follow an interesting distribution pattern, which has not been reported before. That is, their posting rates are bimodal. Multiple spammers also tend to collectively and actively post reviews to the same set of products within a shorttime frame, which we call cobursting. Furthermore, author found some other interesting patterns in individual reviewers' temporal dynamics and their co-bursting behaviors with other reviewers. Inspired by these findings, authors first propose a two-mode Labeled Hidden Markov Model to model spamming using only individual reviewers' review posting times. Authors then extend it to the Coupled Hidden Markov Model to capture both reviewer posting behaviors and co-bursting signals.

Authors [4] Aiming to improve the detection of opinion spams in mobile application marketplace, this study proposes using statistical based features that are modelled through the supervised boosting approach such as the Extreme Gradient Boost (XGBoost) and the Generalized Boosted Regression Model (GBM) to evaluate two multilingual datasets (i.e. English and Malay language). From the evaluation done, it was found that the XGBoost is most suitable for detecting opinion spams in the

English dataset while the GBM Gaussian is most suitable for the Malay dataset. The comparative analysis also indicates that the implementation of the proposed statistical based features had achieved a detection accuracy rate of 87.43 per cent on the English dataset and 86.13 per cent on the Malay dataset.

Authors propose [5] a novel hierarchical supervised-learning approach to increase the likelihood of detecting anomalies by analyzing several user features and then characterizing their collective behavior in a unified manner. Specifically, author model user characteristics and interactions among them as univariate and multivariate distributions. then stack these distributions using several supervised-learning techniques, such as logistic regression, support vector machine, and k-nearest neighbors yielding robust meta-classifiers. Authors perform a detailed evaluation of methods and then develop empirical insights. This approach is of interest to online business platforms because it can help reduce false reviews and increase consumer confidence in the credibility of their online information. this study contributes to the literature by incorporating distributional aspects of features in machine-learning techniques, which can improve the performance of fake reviewer detection on digital platforms.

This study [6] has identified different performance metrics that are commonly used to evaluate the accuracy of the review spam detection models. Lastly, this work presents an overall discussion about different feature extraction approaches from review datasets, the proposed taxonomy of spam review detection approaches, evaluation measures, and publicly available review datasets. Research gaps and future directions in the domain of spam review detection are also presented. This research identified that successfactorsofanyreviewspamdetectionmethodhaveinterdependencies. The feature’s extraction depends upon the review dataset, and the accuracy of review spam detection methods is dependent upon the selection of the feature engineering approach. Therefore, for the successful implementation of the spam review detection model and to achieve better accuracy, these factors are required to be considered in accordance with each other.

Nowadays [7], online reviews play an important role in customer’s decision. Starting from buying a shirt from an e-commerce site to dining in a restaurant, online reviews has become a basis of selection. However, peoples are always in a hustle and bustle since they don’t have time to pay attention to the intrinsic details of products and services, thus the dependency on online reviews have been hiked. Due to reliance on online reviews, some people and organizations pompously generate spam reviews in order to promote or demote the reputation of a person/product/organization. Thus, it is impossible to identify whether a review is a spam or a ham by the naked eye and it is also impractical to classify all the reviews manually. Therefore, a spiral cuckoo search based clustering method has been introduced to discover spam reviews. The proposed method uses the strength of cuckoo search and Fermat spiral to resolve the convergence issue of cuckoo search method. The efficiency of the proposed method has been tested on four spam datasets and one Twitter spammer dataset.

Nowadays [8] with the increasing popularity of Internet, online marketing is going to become more and more popular. This is because; a lot of products and services are easily available online. Hence, reviews about all these products and services are very important for customers as well as organizations. Unfortunately, driven by the will for profit or promotion, fraudsters used to produce fake reviews. These fake reviews written by fraudsters prevent customers and organizations reaching actual conclusions about the products. These fake reviews or review spam must be detected and eliminated so as to prevent deceptive potential customers. In this paper, we have applied supervised learning technique to detect review spam. The proposed work uses different set of features along with sentiment score to build models and their performance were evaluated using different classifiers.

This article [9] proposes a feature framework for detecting fake reviews that has been evaluated in the consumer electronics domain. The contributions are four fold:(i) Construction of a dataset for classifying fake reviews in the consumer electronics domain in four different cities based on scraping techniques;(ii)definition n of a feature framework for fake review detection; (iii) development of a fake review classification method based on the proposed framework and(iv) evaluation and analysis of the results for each of the cities understudy.

In this paper [10, a review processing method is proposed. Some parameters have been suggested to find the usefulness of reviews. These parameters show the variation of a particular review from other, thus increasing the probability of it being spam. This method introduced classifies the review as helpful or non-helpful depending on the score assigned to the review.

In [11] paper, Spam campaigns spotted in popular product review websites (e.g., amazon. com) have attracted mounting attention from both industry and academia, where a group of online posters are hired to collaboratively craft deceptive reviews for some target products. The goal is to manipulate perceived reputations of the targets for their best interests. Detailed The pair wise features are first explicitly utilized to detect group colluders in online product review spam campaigns, which can reveal collusions in spam campaigns from a more fine-grained perspective.

In [12] paper, Online product reviews have become an important source of user opinions. Due to profit or fame, imposters have been writing deceptive or fake reviews to promote and/or to demote some target products or services. Such imposters are called review spammers. In the past few years, several approaches have been proposed to deal with the problem. In this work, take a different approach, which exploits the burrstones nature of reviews to identify review spammers.

In [13] paper, Online reviews on products and services can be very useful for customers, but they need to be protected from manipulation. So far, most studies have focused on analyzing online reviews from a single hosting site. How could one leverage information from multiple review hosting sites? This is the key question in our work. In response, develop a systematic methodology to merge, compare, and evaluate reviews from multiple hosting sites. Focus on hotel reviews and use more than 15million reviews from more than3.5million users spanning three prominent travel sites.

In [14] paper, Users increasingly rely on crowd sourced information, such as reviews on Yelp and Amazon, and liked post sand ads on Facebook. This has lento market for black hat promotion techniques via fake (e.g., Sybil) and compromised accounts, and collusion networks. Existing approaches to detect such behavior relies mostly on supervised (or semi-supervised) learning over known (or hypothesized) attacks. They are unable to detect attacks missed by the operator while labeling, or when the attacker changes strategy.

In [15] paper, Online reviews have become an increasingly important resource for decision making and product designing. But reviews systems are often targeted by opinion spamming. Although fake review detection has been studied by researchers for years using supervised learning, ground truth of large scale datasets is still unavailable and most of existing approaches of supervised learning are based on pseudo fake reviews rather than real fake reviews. Working with Dianping1, the largest Chinese review hosting site, present the first reported work on fake review detection in Chinese with filtered reviews from Damping's fake review detection system.

In [16] paper, Online reviews are quickly becoming one of the most important sources of information for consumers on various products and services. With their increased importance, there exists an increased opportunity for spammers or unethical business owners to create false reviews in order to artificially promote their goods and services or smear those of their competitors. In response to this growing problem, there have been many studies on the most effective ways of detecting review spam using various machine learning algorithms. One common thread in most of these studies is the conversion of reviews to word vectors, which can potentially result in hundreds of thousands of features.

In [17] paper, it providing an efficient and effective method to identify review spammers by incorporating social relations based on two assumptions that people are more likely to consider reviews from those connected with them as trustworthy, and review spammers are less likely to maintain a large relationship network with normal users. The contributions of this paper are two-fold: (1) elaborate how social relationships can be incorporated into review rating prediction and propose a trust based rating prediction model using proximity as trust weight; and (2) design a trust-aware detection model based on rating variance which iteratively calculates user-specific overall trustworthiness scores as the indicator for spam city.

In [18] paper, to detect fake reviews for a product by using the text and rating property from a review. In short, the proposed system (ICF++) will measure the honesty value of a review, the trustiness value of the reviewers and the reliability value of a product. The honesty value of a review will be measured by utilizing the text mining and opinion mining techniques. The result from the experiment shows that the proposed system has a better accuracy compared with the result from iterative computation framework (ICF) method.

In [19] paper, Online Social Networks (OSNs), which captures the structure and dynamics of person-to-person and person-to-technology interaction, is being used for various purposes such as business, education, telemarketing, medical, entertainment. This technology also opens the door for unlawful activities. Detecting anomalies, in this new perspective of social life that articulates and reflects the off-line relationships, is an important factor as they could be a sign of a significant problem or carrying useful information for the analyzer.

In [20] this paper, mangoes are graded in four types like Green Mango, Yellow Mango and Red Mango which are based on machine learning method. This system considers RGB values size and shape of mangoes. Following analysis is used to obtain good probability. This helps to train system to identify appropriate maturity of mangoes. This research is conducted on two machine learning method i.e. Naive Byes and SVM (Support Vector Machine).

III. PROPOSED METHODOLOGY

A new proposed framework consists in representing a set of reviews data provided as FIN (Fake Information Networks) and solving the issue of spam detection in a problem of FIN classification. In particular, to show the reviews data set as a FIN where the reviews are linked through different types of nodes (such as functionality and users). Then a weighting algorithm is used to calculate the importance (or weight) of each function. These weights are used to calculate the latest review labels using supervised and unsupervised procedures. Based on our observations, defining two views for features (review-user and behavioral-linguistic), the classified features as review behavioral have more weights and yield better performance on spotting spam reviews in both semi-supervised and unsupervised approaches. The feature weights can be added or removed for labeling and hence time complexity can be scaled for a specific level of accuracy. Categorizing features in four major categories (review-behavioral, user-behavioral, review-linguistic, user-linguistic), helps us to understand how much each category of features is contributed to spam detection.

System Architecture:

The Fig.1 shows the proposed system architecture.

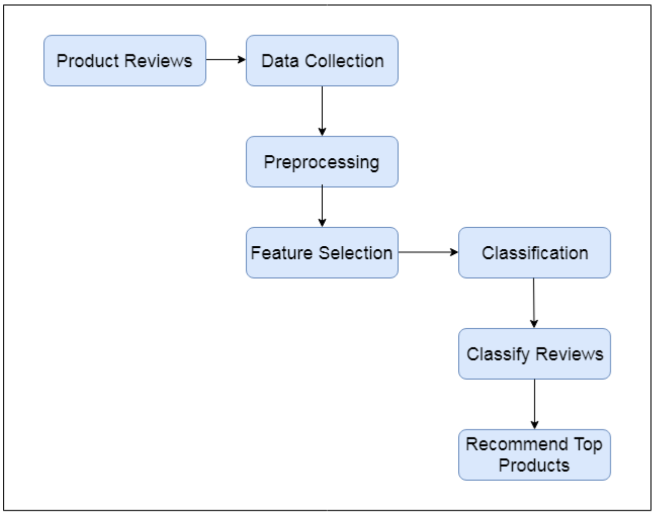


Fig 1. System Architecture

- SpamDup framework that is a novel network based approach which models review networks as heterogeneous information networks.
- A new weighting method for spam features is proposed to determine the relative importance of each feature and shows how effective each of features are in identifying spams from normal reviews.
- SpamDup framework improves the accuracy against the state-of-the art in points of time complexity, which extremely depends to the number of features utilized to detect a spam review.

The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network and to map the problem of spam detection into a FIN classification problem. In particular, model review dataset as in which reviews are connected through different node types. The fig. 2 shows the flowchart of SpamDup framework.

Mathematical Model:

User-Behavioral (UB) based features:

Burstiness: Spammers, usually write their spam reviews in short period of time for two reasons: first, because they want to impact readers and other users, and second because they are temporal users, they have to write as much as reviews they can in short time.

$$x_{BST}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0, \tau) \\ 1 - \frac{L_t - F_t}{\tau} & (L_i - F_i) \in (0, \tau) \end{cases} \quad (1)$$

Where,

$L_i - F_i$ describes days between last and first review for τ = 28.

Users with calculated value greater than 0.5 take value 1 and others take 0.

User-Linguistic (UL) based features:

Average Content Similarity, Maximum Content Similarity: Spammers, often write their reviews with same template and they prefer not to waste their time to write an original review. In result, they have similar reviews. Users have close calculated values take same values (in [0; 1]).

Review-Behavioral (RB) based features:

Early Time Frame: Spammers try to write their reviews a.s.a.p., in order to keep their review in the top reviews which other users visit them sooner.

$$x_{ETF}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0, \delta) \\ 1 - \frac{L_t - F_t}{\delta} & (L_i - F_i) \in (0, \delta) \end{cases} \quad (2)$$

Where,

$L_i - F_i$ denotes days specified written review and first written review for a specific business. We have also δ = 7. Users with calculated value greater than 0.5 takes value 1 and others take 0.

- Rate Deviation using threshold: Spammers, also tend to promote businesses they have contract with, so they rate these businesses with high scores. In result, there is high diversity in their given scores to different businesses which is the reason they have high variance and deviation.
-

$$x_{DEV}(i) = \begin{cases} 0 & Otherwise \\ 1 - \frac{r_{ij} - avg_{e \in E*j} r(e)}{4} > \beta_1 \end{cases} \quad (3)$$

Where,

β_1 is some threshold determined by recursive minimal entropy partitioning. Reviews are close to each other based on their calculated value, take same values (in [0; 1)).

Review-Linguistic (RL) based features:

Number of first Person Pronouns, Ratio of Exclamation Sentences containing '!': First, studies show that spammers use second personal pronouns much more than first personal pronouns. In addition, spammers put '!' in their sentences as much as they can to increase impression on users and highlight their reviews among other ones. Reviews are close to each other based on their calculated value, take same values (in [0; 1]).

Algorithm:

Algorithm 1: Spam review detection using behavioral features method

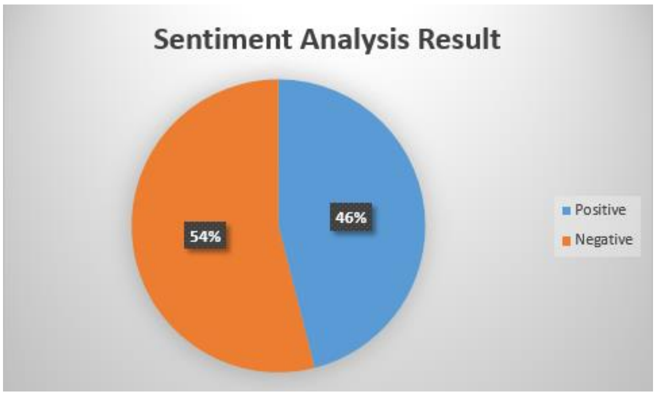
```
Input: review Ri , τ = 0.5,0.55,0.6 //threshold value for labelling the review
Output: Spam or Not-Spam
  for each review Ri in review dataset do
    // behavior features (F1, F2, F3, …, F13)
    for each behavior feature Fi calculate normalize value do
      // variable V is calculating normalize value of F
      Vi = calculate normalize value Fi
      Sum += Vi
    end for
    // calculating average score
      Average Score = Sum / 13
    for each value Vi do
      // calculating drop score
      DropScore = (Sum - Vi) / 12
      if | Average Score – DropScore | >= 0.05 then
        assign weight Wi ← 2
        Total Weight += 2
      else
        assign weight Wi ← 1
        Total Weight += 1
      end if
    end for
    for each value Vi do
      // calculating total spam score
      Score += Wi * Vi
    end for
    Spam Score = Score / Total Weight
    if Spam Score > τ then
      label Ri ← Spam
    else
      label Ri ← Not-Spam
    end if
  end for
```

IV. RESULT AND DISCUSSION

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and Jdk 1.8. The application is web application used tool for design code in Eclipse and execute on Tomcat server.

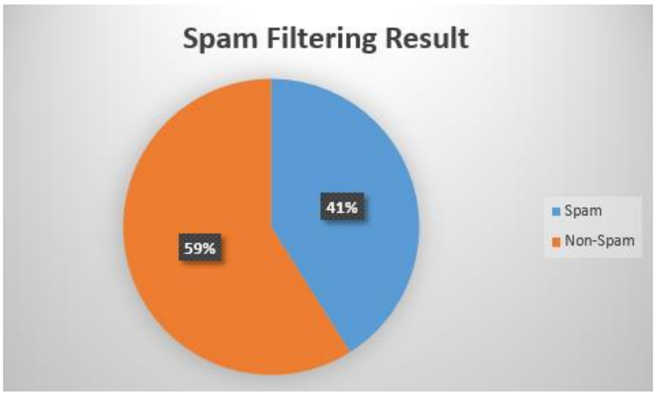
Sentiment Result:

We use sentiment analysis for positive and negative classification.



Spam Filtering Result:

Experimental evaluation outcomes shows the Amazon API uses for product review dataset with higher percentage of spam reviews have better performance because when fraction of spam reviews increases, probability for a review to be a spam review increases and as a result more spam reviews will be labeled as spam reviews.



CONCLUSION

Sentiment Analysis is a case study that looks at the feeling, mood, entropy or feelings of people. This paper addresses a basic issue of the study of feelings and the classification of feelings of polarity. Data was compiled from online product reviews of Amazon.com. A method known as the categorization of emotion polarity and POS along with thorough explanations of each phase was proposed. These measures include pre-processing, pre-filtering, partitioning, data consistency. Functionality that include machine learning expertise. Much work has been done in opinion mining and consumer evaluation in the form of a study of documents, sentences, and features. Opinion Mining can become a most interesting field of study for potential preferences by using a number of found function expressions derived from the reviews. More novel and successful approaches need to be invented to address the existing difficulties of mining opinion and sentiment analysis.

References

[1] Dematis, E. Karapistoli, and A. Vakali, "Fake review detection via exploitation of spam indicators and reviewer behavior characteristics," in Proc. Int. Conf. Current Trends Theory Pract. Inform. Cham, Switzerland: Edizioni Della Normale, 2018, pp. 581–595.

[2] S. Zhou, Z. Qiao, Q. Du, G. A. Wang, W. Fan, and X. Yan, "Measuring customer agility from online reviews using big data text analytics," J. Manage. Inf. Syst., vol. 35, no. 2, pp. 510–539, Apr. 2018.

[3] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Bimodal distribution and co-bursting in review spam detection," in Proc. 26th Int. Conf. World Wide Web (WWW), 2017, pp. 1063–1072.

[4] M. Hazim, N. B. Anuar, M. F. A. Razak, and N. A. Abdullah, "Detecting opinion spams through supervised boosting approach," PLoS ONE, vol. 13, no. 6, 2018, Art. no. e0198884.

[5] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting review manipulation on online platforms with hierarchical supervised learning," J. Manage. Inf. Syst., vol. 35, no. 1, pp. 350–380, Jan. 2018.

[6] N. Hussain, H. Turab Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review," Appl. Sci., vol. 9, no. 5, p. 987, 2019.

[7] C. Pandey and D. S. Rajpoot, ''Spam review detection using spiral cuckoo search clustering method,'' Evol. Intell., vol. 12, no. 2, pp. 147–164, Jun. 2019.

[8] R. Narayan, J. K. Rout, and S. K. Jena, "Review spam detection using opinion mining," in Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Singapore: Springer, 2018, pp. 273–279.

[9] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," Inf. Process. Manage., vol. 56, no. 4, pp. 1234–1244, Jul. 2019.

[10] R. Ghai, S. Kumar, and A. C. Pandey, ‘‘Spam detection using rating and review processing method,’’ in Smart Innovations in Communication and Computational Sciences. Singapore: Springer, 2019, pp. 189–198.

[11] Ch. Xu and J. Zhang,” Combating product review spam campaigns via multiple heterogeneous pairwise features”, In SIAM International Conference on Data Mining, 2014.

[12] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Exploiting bustiness in reviews for review spammer detection”, In ICWSM, 2013.

[13] j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos, “True view: Harnessing the power of multiple review sites”,In ACM WWW, 2015.

[14] Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, “Towards detecting anomalous user behavior in online social networks”, In USENIX, 2014.

[15] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao,” Spotting fake reviews via collective PU learning”, In ICDM, 2014.

[16] M. Crawford, T. M. Khoshgoftaar, and J. D. Prusa,” Reducing Feature Set Explosion to Faciliate Real-World Review Sapm Detection”, In Proceeding of 29th International Florida Artificial Intelligence Research Society Conference, 2016.

[17] H. Xue, F. Li, H. Seo, and R. Pluretti,” Trust-Aware Review Spam Detection”,IEEE Trustcom/ISPA.,2015.

[18] E. D. Wahyuni , A. Djunaidy,” Fake Review Detection From a Product Review Using Modified Method of Iterative Computation Framework”, In Proceeding MATEC Web of Conferences, 2016.

[19] R. Hassanzadeh,” Anomaly Detection in Online Social Networks: Using Datamining Techniques and Fuzzy Logic”, Queensland University of Technology, Nov, 2014.

[20] G.D. Upadhye, D.Pise,“Grading of Harvested Mangoes Quality and Maturity Based on Machine Learning Techniques”,IEEE International conference on smart city and Emerging Technology,2018.