Opinion Mining using Machine Learning Techniques

Helen Josephine V.L., Research Scholar, Bharathiar University, Coimbatore, India, helenjose.cbe@gmail.com

Dr. Duraisamy S, Assistant Professor, Department of Computer Science, Chikkanna Govertnment Arts College, Tirupur, India, sdsamy.s@gmail.com

Abstract **Enormous volume of unstructured data are available in the form of reviews or comments in the social media and the product website. The buyers or the users are enforced to do absolute inspection on these information before choosing any product or service. A machine learning approach is necessary to mine these opinions which help the customer and the organization to make proper decision. This research paper analyses the opinions about the mobile learning system. This paper also examines the classification accuracy of K Nearest Neighbour algorithm. The classification accuracy of K Nearest Neighbour algorithm is compared with Multinomial Naive bayes probabilistic classification algorithm and random forest data mining algorithm. The other classification metrics precision, recall and F-measure also compared for these various machine learning algorithm.**

***Keywords* — Machine Learning, Opinion Mining, K Nearest Neighbour, Random Forest, Multinomial Naive Bayes**

I. Introduction

Sentiment analysis also called opinion mining, is the field of computational study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes[1]. Opinions expressed in a set of source documents about an object need to be mined, by extracting attributes of the object from the review comments and determining whether the opinions are positive, negative or neutral.

The rapid growth in the communication technology has led to easy access of information. There is a swift increase in the usage of mobile devices in all the fields. Private and Government sectors also aim to enable qualify education for all the students through mobile device. It because popularized the pedagogical methods such as learning through mobile devices [2]. Various mobile learning systems are available and also the users' opinions about these systems are aired in the social blogs or review Websites.

In sentiment analysis, there are two categories of information namely opinion (subjective information) and facts (objective information). Facts are the statements which explain the nature or qualities about the product or events. But the opinions describe the appraisals, attitudes and emotions regarding to the entity [3]. The major research has been done on objective nature of the product but recently more focus on the opinions and emotions on the product.

Opinion mining seeks to identify the opinion conveyed in the text document either by applying information retrieval or computational linguistics. The opinion expressed on the topic is given significance rather than the topic itself [4], [5]. Sentiment analysis or opinion mining analyses to extract the subjective information in source materials by applying natural language processing, computational linguistics and text analytics. Opinion classification is broadly studied in Natural language processing. For the given a set of review documents X, it determines whether each document x_i (i=1 to n) expresses a positive, negative or neutral opinion / sentiment on an object. This is identical to supervised classification method in data mining.

In this paper, we implement the machine learning algorithm to analysis the sentiment in free mobile learning system reviews. This analysis assists the service provider / manufacturer to enhance the mobile learning system and also helps the user to choose appropriate mobile learning system without spending much time. Mobile learning is not e-learning that is facilitated with mobile technology, and also the practices other components and explorations across multiple contexts [6]. Mobile learning system helps the learners to learn any subject, at any time, any place and just when the knowledge is required.

In this paper opinion or sentiment classification has been performed on mobile learning system review data set. This methodology not only focuses the opinion words but also giving importance to corpus words which are frequently used in the documents under review. This paper also explains the methodology to remove words which are commonly used in the dataset of mobile learning system reviews. Singular value decomposition methods have been used to rank the corpus and prepare the data for opinion mining. This paper is organized into the following sections. Section II explains about the various techniques

DOI : 10.18231/2454-9150.2018.0641
© 2018, IJREAM All Rights Reserved.

used in the field of opinion mining. Section III briefly describes the materials and methods and classification algorithms, section IV describes the experiment results obtained and discussion. Finally Section V concludes the paper.

II. Literature Survey

Opinion mining helps to analyze the customer reviews and extract knowledge. In general, sentiment analysis or opining mining has been investigated mainly in three levels. (i) Document Level (ii) Sentence Level (iii) Entity and aspect or feature level. The first two determines the overall sentiment or opinion of the product or services, where as the third one find out the products feature level opinion. Several techniques help to mine the opinion in document level. Each reviews has been considered as one document. NLP (Natural Language Processing), Data Mining Classification Algorithms, Machine Learning Algorithms and Artificial Neural Network Algorithms place major role in this research area. Semantics based sequential characteristic such as unigram and bigram was created. Optimum feature sets for classification was identified. SVM classifiers are constructed by Swaminathan et al 2010. C4.5 algorithm used to identify the classification strategy of movie popularity and PART classifier protocol was proposed by Asad et.al 2012. Supervised classification algorithm has been used to mine the feed back of the students' comments [7].

III. Methodology

This research analyzed the mobile learning reviews which is available in the android website. Intensive study has been done only on the learners' opinion of free Mobile learning system. Out of these reviews 100 positive, 100 negative and 100 neural reviews are used. Three different types of machine learning algorithms are used to perform the opinion classification.

All the reviews are collected from the online website and stores as .csv file which contains two columns namely review comments and the classification values ie positive, negative, or neutral. Initial pre-processing has been performed on the data set. For this dataset document term matrix is formed. Each row contains the individual review. Every column is the term with in that particular review. Each cell of the matrix denotes the frequencies of words which specify the number of times that term occurs in the review document. Stopwords and stemming along with spell check places important role in minimizing the number of term or the column of the document matrix.

To reduce the dimensionality of the document term matrix, Singular value decomposition (SVD) is used [8]. It is also used to find the important word presented in the review document [9]. To strengthen the importance of the data, the outlier word has been identified based on the word frequency. If the frequency is below 10 and above 75 need to be removed. These terms are identified by using Term frequency and Inverse Document Frequency methods [10].

The final matrix values are loaded into the Jupyter notebook in Anaconda. Python is the data science language is used to mine the opinion of the data set. Scikit learn and pandas are the open source library which can be imported in python. It contains many machine learning algorithm and data visualization tools. With the help of this package confusion matrix has been evolved. Classification accuracy of K Nearest Neighbour, Multinomial Naive Bayes and Random forest algorithm were obtained. Other metrics precision, recall and f-measure also calculated. Eventually, the accuracy of these three algorithms is compared.

A. K Nearest Neighbour

KNN is the simple and sophisticated approach to classification. It has been used in many applications in the field of data mining, image processing and many others including text classification [11]. This algorithm first calculates the similarity between test set and all the samples in the training data to get K nearest samples [12]. Similarity between the point s identified the distance between the points by using either Euclidean distance or some other distance formula. KNN selection is based on distance weighted voting. KNN is the supervised text classification algorithm and the result is efficient if the training set is large. Consider the vector X and set of M labelled instances {xi,yi} for i= 1 to M. The classifier classifies the class label of X in any one of the predefined N classes. This algorithm finds the k nearest neighbours of X and classifies the object of X by a majority vote of its neighbours. KNN classifier applies Euclidean distance as the distance metric [13]. Some other enhanced KNN classifier using Hamming distance metric.

Euclidean distance formula [14]

$$dist(p,q) = \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2} \qquad \text{[Eq. 1]}$$

Hamming Distance

$$D_H = \sum_{i=1}^{k}\left|p_i - q_i\right| \qquad \text{[Eq. 2]}$$

Manhattan Distance – Calculate the distance between real vectors using the sum of their absolute difference.

Minkowski Distance – Generalization of Euclidean and Manhattan distance.

This algorithm has the following disadvantages (i) It requires distance computation of k nearest neighbours. Intensive computation is necessary, especially for the large training dataset. Because of this issue classifying unknown records are relatively expensive. (ii) Noisy and irrelevant features degrade the classification accuracy.

B. Multinomial Naive Bayes

Naive Bayes classifier evaluates the class conditional probability by understanding that the attributes are conditionally independent. Conditional independence is evaluated as follows

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad \text{[Eq. 3]}$$

P(c|X)=P(x1|c) * P(x2|c)*.....*P(xn|c)*P(c) [Eq. 4]

Where, P(c|x) – posterior probability of the target class for given attribute
P(c) - Prior probability class
P(x|c) - likelihood which is the probability of predictor given class
P(x) - prior probability of predictor

Naive bayes is a machine learning approach method to predict the likelihood that an event will occur given evidence that's present in dataset [15]. It is also called conditional probability in the world of statistics. Naive bayes is a group of algorithms based on principles of Bayes theorem with naïve (strong) assumption, that every feature is independent of the others, in order to predict the category of a test data set. They are probabilistic classifiers, therefore will calculate the probability of each category using Bayes theorem, and the category with the highest probability will be output[16][17]. Naive Bayes algorithm has three different varieties (1) Multinomial Naive Bayes (2) Barnali Naive Bayes (3) Gaussian Naive Bayes Out of three multinomial navie bayes outperforms for the categorical and described discrete frequency counts in other words word counts or something like that. Barnali navie bayes performs good for making predictions from binary features. Thirdly Gaussian naive bayes approach is good for making predictions from normally distributed features [18].

C. Random Forest

Random Forest is supervised classification algorithm and it also called an ensemble algorithm. Ensembled algorithms are those which combine more than one algorithms of same or different kind for classifying objects. The basic idea behind a random forest is to combine many decision trees into a single model. Individually, predictions made by decision trees may not be precise, but combined together, the predictions will be closer to the mark on average [19]. Predictions have inconsistency because they will be widely spread around the right answer. Random forest algorithm can divide into two stages. (i) Random forest creation (2) Perform prediction by using random forest classifier. Random forest runtimes are relatively fast, and they are capable to deal with missing and unbalanced data. Random forest classifier won't overfit the model, even more trees in the forest [20]. The disadvantage of Random Forest are that when used for regression they cannot predict beyond the range in the training data., and that they may over-fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon your own data set.

Random vectors θk created for each 'k'th tree in the forest. It is independent of the past random vector θ_1,……$\theta_{k-1}$and with same distribution. The tree is grown using the random vector θ_k and the training set, which results in the classifier h(x, θ_k), where x is the input vector[19]. As a result a random forest consists of set of trees with classifiers got from independent identically distributed random vectors; and each tree casts a vote for the class at input x. as the number of trees increase, the generalization error converges to

PX,Y(P_θ(h(X,$_\theta$)=Y)-maxj≠Y P_θ(h(X,$_\theta$)=j)<0) [Eq. 5]

Where X, Y are random vectors and is generalization error probability over the P(X,Y) space. In random split selection θ be made up of a number of independent random integers between 1 to K. The nature and dimensionality of θ determined on its use in tree construction.

Reduction of the correlation leads to increase the accuracy of the random forest. The correlation is reduced by the randomness used while maintaining the strength. Each node is built into tree using randomly selected inputs. Random forest is an effective tool in prediction.

IV. Result and Discussions

In this proposed work, the basic preprocessing and other exclusive hybrid techniques SVM, term frequency and outlier removal has been applied to the dataset. The cleaned dataset has been fed into the existing machine learning techniques Multinomial Naïve Bayes and Random forest. The entire dataset has been split into training dataset and testing dataset. For each algorithm, model has been produced based on the training set. Using the model the testing dataset opinion has been predicted and compared with the actual target value.

Confusion matrix has been prepared based on the actual and predicted value for the positive, negative and neutral values. This problem is considered as multi class (3 class) problem. The target values are positive(A), negative(B) and neutral(C). Based on the confusion matrix value classification accuracy and other metrics namely precision, recall and f-measure has been evaluated and the result has been tabulated in Table I and Table II K Nearest Neighbour algorithm yields classification accuracy 55%. Multinomial Naïve Bayes and Random forest offers 59%

and 63% respectively. Out of all machine learning algorithms Random forest provides best result. . Fig. 1 illustrates the bar graph of classification accuracy of various machine learning algorithm.

TABLE I : CLASSIFICATION ACCURACY PRODUCED BY DIFFERENT SUPERVISED MACHINE LEARNING ALGORITHM

Different Data Mining Classifier	Classification Accuracy (in %)
KNN	55
Multinomial Naive Bayes	59
Random Forest	63

TABLE II : PRECISION, RECALL AND F-MEASURE OBTAINED FORM DIFFERENT SUPERVISED MACHINE LEARNING ALGORITHM

Algorithm	Precision	Recall	F-Measure
KNN	0.54	0.55	0.54
Multinomial Naïve Bayes	0.6	0.58	0.59
Random Forest	0.64	0.62	0.63

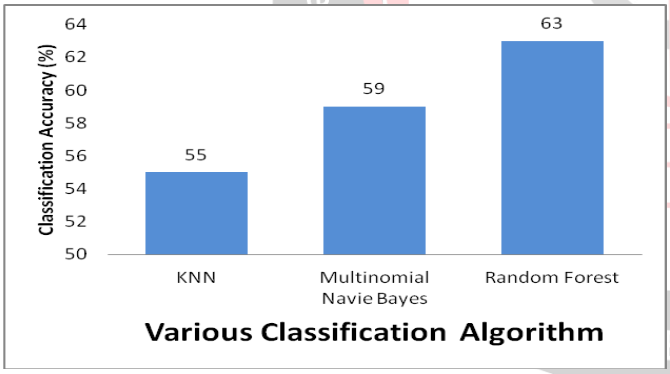

Fig.: 1 Classification accuracies of various supervised machine learning algorithm

First precision and recall were computed for each class label to analyze the individual performance of the class labels. Then theses values are averaged to find overall precision and recall.

The following formulae has been used to calculated precision, recall and f-measure respectively

$$\text{Precision}_{Positive} = \frac{TP\_Positive}{TP\_Positive + FP\_Positive} \quad \text{[Eq. 6]}$$

$$\text{Re}\,call_{Positive} = \frac{TP\_Positive}{TP\_Positive + FN\_Positive} \quad \text{[Eq. 7]}$$

$$F - Measure = \frac{2 * \text{Precision} * \text{Re}\,call}{\text{Precision} + \text{Re}\,call} \quad \text{[Eq. 8]}$$

Fig. 2 shows the graph for Precision, Recall and F-Measure of various data mining classification algorithm.

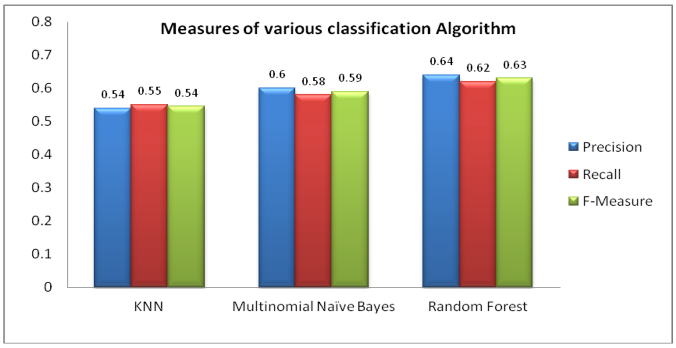

Fig.: 2 Metrics of various classification algorithm

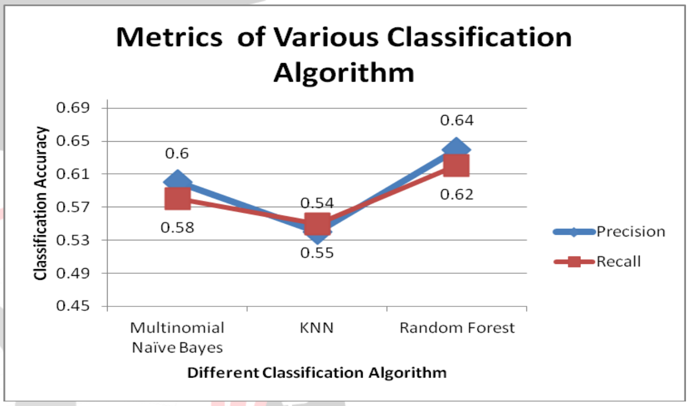

Fig.: 3 Precision and Recall of Multinomial Naïve Byes, KNN, Random Forest

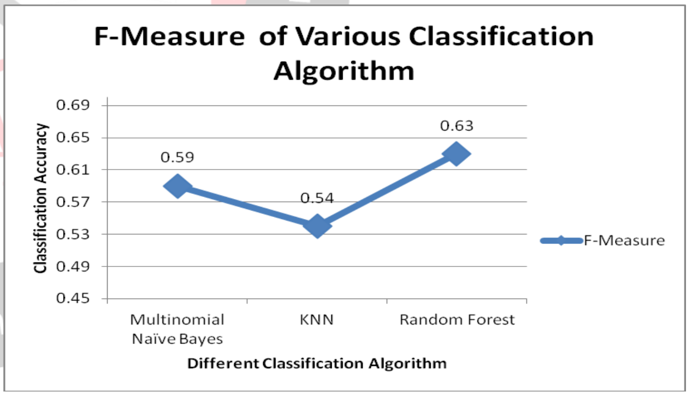

Fig.: 3 F-Measure of Multinomial Naïve Byes, KNN, Random Forest

Random Forest algorithm obtained 0.64 precision, 0.62 recall and 0.63 F-Measure. The experiment was conducted with three different supervised machine learning algorithm to mine the polarity of the mobile app reviews. Compare to K Nearest Neighbour and Multinomial Naïve Bayes, the Random Forest algorithm provides best result for the dataset.

V. Conclusion

In this research paper, the performance of the different classification algorithm has been investigated. Mobile

DOI : 10.18231/2454-9150.2018.0641
© 2018, IJREAM All Rights Reserved.

learning app reviews were pre-processed by basic and advanced techniques. SVD, TF_IDF were used to reduce the dimensionality of the words and to identify the frequent and non-frequent words. By using this pre-processed dataset, the efficiency of K Nearest Neighbour, Multinomial Naïve Bayes and Random Forest algorithms' accuracy and ot her classification metrics precision, recall and f-measure has been evaluated. Out of the three algorithms Random forest algorithm produces improved results.

REFERENCES

[1] Liu, B. 2010. Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F. J. Damerau, Eds. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921,

[2] M. Sharples, P. Lonsdale, J. Meek, P.D. Rudman and G. Vavoula, "An Evaluation of MyArtSpace: a Mobile Learning Service for School Museum Trips", Proceedings of the 6th Annual Conference on Mobile Learning, pp. 238-244, 2007.

[3] Lei Zhang and Bing Liu, "Aspect and Entity Extraction for Opinion Mining" https://www.cs.uic.edu/~lzhang3/paper/ZhangLiu-AEEE.pdf

[4] Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4):315 346.

[5] Bing Liu . Exploring User Opinions in Recommender Systems. Proceeding of the second KDD workshop on Large Scale Recommender Systems and the Netflix Prize Competition, Aug 24, 2008, Las Vegas, Nevada, USA.

[6] Sharples. M, Lonsdale. P, Meek. J, Rudman. P. D, & Vavoula. G. 2007. "An Evaluation of MyArtSpace: a Mobile Learning Service for school Museum trips". In Proceedings of mLearn 2007, Melbourne, Australia.

[7] V. Dhanalakshmi,Dhivya Bino,A. M. Saravanan2016 "Opinion mining from student feedback data using supervised learning algorithms", 3rd MEC International Conference on Big Data and Smart City (ICBDSC)DOI: 10.1109/ICBDSC.2016.7460390

[8] Vikas K Vijayan, K. R. Bindu, Latha Parameswaran, "A comprehensive study of text classification algorithms", Advances in Computing Communications and Informatics (ICACCI) 2017 International Conference on, pp. 1109-1113, 2017.

[9] Roberta Akemi Sinoara, João Antunes and Solange Oliveira Rezende (June 2017), Text mining and semantics: a systematic mapping study, Journal of the Brazilian Computer Society 2017 23:9 https://doi.org/10.1186/s13173-017-0058-7

[10] Awad M., Khanna R. (2015) Support Vector Machines for Classification. In: Efficient Learning Machines. Apress, Berkeley, C

[11] Khu P. Nguyen, Huy Q. Phan, "Feasible settings for the adaptive latent semantic analysis: Hk-LSA model", Computational Intelligence and Applications (ICCIA) 2017 2nd IEEE International Conference on, pp. 219-224, 2017.

[12] Han Jia-wei, M. Kamber. Data Mining-Concepts and Techniques, Second Edition. Machine Industry Publisher, 2007.

[13] Yun-lei Cai, Duo Ji ,Dong-feng Cai, A KNN Research Paper Classification Method Based on Shared Nearest Neighbor, Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan

[14] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, Martin Vetter. Euclidean Distance Matrices: Essential theory, algorithms, and applications. IEEE Singnal Processing Magazine Volume 32 Issue : 6.

[15] Abhay B Rathod, Sanjay M Gulhane, Shailesh R Padalwar, A "Comparative study on distance measuring approches for permutation representations", International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), 2016

[16] Haiyi Zhang, Di Li, Naïve Bayes Text Classifier, IEEE International Conference on Granular Computing (GRC 2007)

[17] Neha Sharma, Manoj Singh, Modifying Naive Bayes classifier for multinomial text classification, International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2016

[18] Carlos Bustamante, Leonardo Garrido, Rogelio Soto, Comparing Fuzzy Naive Bayes and Gaussian Naive Bayes for Decision Making in RoboCup 3D, Advances in Artificial Intelligence, MICAI 2006

[19] Leo Breiman, Random Forests, Machine Learning, Springer Link, https://doi.org/10.1023/A:10109334, October 2001, Volume 45, Issue 1, pp 5–32

[20] Data Mining Concepts and Techniques,Jiawei Han, Micheline Kamber Morgan Kaufman Publishers, 2003.

Helen Josephine V L is a Research Scholar in Bharathiar University, Coimbatore, and she is working as a Assistant Professor in the Department of Computer Applications, CMRIT, Bangalore. Her research interest includes Machine Learning, Web mining, Opinion mining and Sentiment analysis.

Dr. S. DURAISAMY is Assistant Professor of Department of Computer Science in Chikkanna Government Arts College. He obtained Ph.D in Computer Science in 2008. He has produced 12 Ph.D candidates and guiding many research scholars. He has published more than 80 articles in national and international journals. His area of interest includes Software Engineering, Software Testing and Data Mining.

IJREAM International Journal for Research in Engineering Application & Management