Data Mining Algorithms for a Feature-Based Customer Review Process Model with Engineering Informatics Approach

A K Varudharajulu and Y Ma*
University of Alberta, Canada
*yongshen@ualberta.ca

Abstract. The data generated from online communication acts as potential gold mines for discovering knowledge for end users. Large amount of data is also generated in the form of web documents, emails, blogs, and feedback, etc. Text analytics and opinion mining are used to extract human thoughts and perceptions from unstructured texts. This paper proposes a method that focuses on analysing different classification and clustering algorithms aimed at extracting and consolidating opinions of customers from social media sites like Facebook, Twitter and through surveys, at multiple levels of granularity to monitor and measure customer satisfaction. This is an automated approach, in which algorithms aid in the process of knowledge assimilation and the analytics. Domain experts ratify the knowledge base and provide training data sets for the system to intuitively gather more instances for ratification. The system identifies opinion expressions as phrases containing opinion words, opinionated features and also opinion modifiers. These expressions are categorized as positive, negative or neutral. Opinion expressions are identified and categorized using localized linguistic techniques. Opinions can be congregated at any desired level of specificity i.e. feature level or product level, user level or service level, etc. It has been found that J48 classification algorithm and simple k-means clustering algorithm are most suitable for restaurant customer reviews.

1. Introduction
The World Wide Web has become the most popular way to obtain information, such as travel reviews. According to the study conducted in 2014 [1], 81% of the Internet users have searched related comments before buying a commodity at least once. 32% have provided a rating on a product, service, or person via an online rating system, and 30% (including 18% of online senior citizens) have posted an online comment or review regarding a product or service. It should be noted that these online investigations had a significant impact on the customer’s decisions. In addition, Customer reviews lead to feedback for the manufacturers or service providers to benefit from the user’s ideas to improve their products and services appropriate for the target group [2], which increase customer satisfaction [3]. Key objectives of the proposed feature-based process model are: (1) to enable real-time extraction of customer comments from social media and reflect uniformly the opinions of customers, and then provide a systematic basis for management review and decision making for the operation; (2) to create a quality improvement cycle with the help of a software toolkit such that the management can measure the effectiveness of process changes, and new innovations on packages and menu items; (3) to generate positive assurance on service industry quality management generically and eventually to create economic benefits in the forms of higher satisfaction rate, more customer visits, sustained business growth and reputable brand name in
the sector. In the field of machine learning, there are many existing classification and clustering algorithms. This paper focuses on selecting the right algorithms for data mining of restaurant customer reviews through careful and comprehensive analysis, the analyzing process of different algorithms is discussed in methodology (Section 3).

2. Literature review

Social media reviews are very noisy and full of all kinds of spelling, grammatical, and punctuation errors. Most natural language processing (NLP) tools such as parts-of-speech (POS) taggers and parsers need clean data to perform accurately. Thus, a significant amount of pre-processing is needed before any analysis [4]. To overcome the need of pre-processing during the data mining process, in this reported work, opinion mining model has been adopted [5][6]. There have been some datasets consisting of restaurant customer reviews [7][8]. Sentiment analysis is commonly performed since the management needs to qualify customers’ reviews towards the restaurant’s issues, activities, and features; Weka [9] is a useful tool for sentimental analysis [10][11]. Typically, customer reviews can be analysed through the steps of word parsing and tokenization (extract terms), stop-words removal (elimination of non-value adding words for sentiment analysis), lemmatization and stemming (convert all infections to their root word), term selection/feature extraction (remove the terms that have poor prediction ability). Then the cleaned customer feedback texts are classified (assign a class -positive/negative), e.g. J48 algorithm [12]. Such algorithms could result in different classification quality depending on the match to the dataset used. In comparative analysis of classification algorithms for diabetes dataset [13] and Swahili language Tweets [14], Naïve Bayes algorithm was more accurate and suitable while performance of J48 was better in an analysis done in bank dataset [15] and credit card fraud detection [16]. The next step in data mining is to design the clustering model. The dataset was categorized into three by using k-means clustering [17][18] based on the sentiment of the comments as Improve, Grow and Continue (i.e. keep going). This paper studies classification [19][20][21] and clustering [22] algorithms for a feature-based customer review process model.

3. Methodology

3.1. Data collection

Restaurant customer reviews data collected from the internet is used as sample data for classification and clustering. The collected data may have unstructured data for classification and clustering like emoticons, or hashtags, which were pre-processed and removed using one of the Weka tool’s filters string to vector. Weka is a collection of open source Machine Learning algorithms used for pre-processing, classifiers, clustering, association rules, and visualization [23]. It is a Java-based tool used in the field of data mining, created by researchers at the University of Waikato in New Zealand. It uses flat text files to describe the data. It can work with a wide variety of data files including its own “.arff”. The pre-processed training data is used to build a classifier model and compared. So, the time taken to build the model excludes the time taken for pre-processing.

3.2. Classification Algorithms.

J48 decision tree algorithm builds a decision tree after analyzing the training instances. In the decision tree, the topmost node is the root node and the leaf nodes are the class values. In general, the nodes are attributes and the branches are decisions. While classifying the test instances using the constructed decision tree, it starts from the root node and traverses to leaf node based on the condition at each node and assigns a class value. Naïve Bayes Algorithm is a simple probabilistic classifier that uses Bayes theorem to calculate the probability for each instance by counting the frequency and combinations of values in a given dataset. It assumes all attributes to be independent given the value of the class variable [24].
3.2.1. Bayes Theorem [25].

\[
P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}
\]

(1)

\(P(C/X)\) is the posterior probability of class \((C, \text{target})\) given predictor \((X, \text{attributes})\). \(P(C)\) is the prior probability of class. \(P(X|C)\) is the likelihood which is the probability of predictor given class. \(P(X)\) is the prior probability of predictor. \(P(X)\) is constant for all classes.

3.2.2. Performance factors. This section includes the performance factors that have been used in the experimental analysis of classification algorithms [26]. Confusion matrix in Table 1 can be used to identify checking statuses of correlations between the predicted results and the actual judgements as shown below: True Positive (TP): When the prediction is YES and the actual value is also YES; True Negative (TN): When the prediction is NO and the actual value is also NO; False Negative (FN): When the prediction is NO but the actual value is YES; False Positive (FP): When the prediction is YES but the actual value is NO.

| Table 1. Confusion matrix |
|---------------------------|
|                         | Predicted Yes | Predicted No |
| Actual Yes               | TP            | FN           |
| Actual No                | FP            | TN           |

Correctly Classified Instances, i.e. Sum of TN and TP, measure the numbers of instances whose predicted and actual class values are equal out of the total instances. Incorrectly Classified Instances are measured with the number of instances whose predicted and actual class values are different i.e. Sum of FN and FP.

Precision, \(\frac{TP}{(TP + FP)}\), gives what fraction of the predicted positive values are actually positive. Recall, \(\frac{TP}{(TP + FN)}\), gives what fraction of actually positive values is predicted positive. F-Measure is the weighted average of Precision and Recall.

\[
F = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}
\]

Experimental analysis. Experiments are performed on Restaurant data by using Classification algorithms in WEKA tool. Total no of instances is 2000.

| Table 2. Result based on accuracy and time |
|-------------------------------------------|
| Algorithm | Correctly classified instances | Percentage of Correctly classified instances | Incorrectly classified instances | Percentage of incorrectly classified instances | Time taken to build a model (in seconds) |
|-----------|------------------------------|--------------------------------------------|-------------------------------|----------------------------------|-----------------------------------|
| J48       | 1921                         | 96.05                                      | 79                            | 3.95                             | 0.13                              |
| Naïve Bayes | 1475                       | 71.75                                      | 565                           | 28.25                            | 0.85                              |

3.2.3. Results. The sample data was tested over two classification algorithms which alternatively work better in different fields. As the result shows in Table 2, though Naïve Bayes equally takes lesser time to build a model, J48 classifies most of the instances correctly (see Figure 1) and increases the accuracy. Hence the J48 algorithm was used in our model. The different performance factors for Positive and Negative customer reviews using J48 and Naïve Bayes algorithms are in Table 3. Figure 2 gives a graphical representation of the various performance factors for classification algorithms in our case.
Simple k-means is an iterative clustering algorithm in which items are moved among clusters till all instances are covered and the desired cluster is reached. Hierarchical clustering algorithm finds the similarity or dissimilarity between every pair of objects in the data set by calculating the distance between objects. Then group the objects into a binary, hierarchical cluster tree. At the next step, it determines where to cut the hierarchical tree into clusters by pruning branches off the bottom of the hierarchical tree so as to partition data. Farthest First clustering algorithm is a variant of k-means, it places the cluster center at the point further from the present cluster. The points that are farther are clustered together first. This feature of the farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed. Make_Density_Based clustering algorithm can be used when the clusters are irregular. A cluster is a dense region of points that are separated by low-density regions from the tightly dense regions. The make density-based clustering algorithm can also be used in noise and when outliers are encountered. The points with the same density and present within the same area will be connected to form clusters. The above clustering algorithms were tested with our dataset to find which is suitable for our product model by comparing various performance factors. Table 4 shows the time taken by each algorithm to cluster our dataset for various numbers of clusters.

Table 3. Performance factors for Positive and Negative using J48 and Naïve Bayes algorithms

|          | J48          | Naïve Bayes |
|----------|--------------|-------------|
|          | Positive     | Negative    | Positive | Negative |
| TP Rate  | 0.962        | 0.959       | 0.730    | 0.705    |
| FP Rate  | 0.041        | 0.038       | 0.295    | 0.270    |
| Precision| 0.959        | 0.962       | 0.712    | 0.723    |
| Recall   | 0.962        | 0.959       | 0.730    | 0.705    |
| F-Measure| 0.961        | 0.960       | 0.721    | 0.714    |
| MCC      | 0.921        | 0.921       | 0.435    | 0.435    |
| ROC Area | 0.990        | 0.990       | 0.808    | 0.807    |
Figure 2. Performance factors for positive and negative comments (J48 vs. Naïve Bayes) algorithms

Table 4. Time taken (in seconds) to form Clusters

| Cluster number | Simple k-means | Hierarchical | Farthest First | ke Density Based |
|----------------|----------------|--------------|----------------|------------------|
| 10             | .221           | .363         | .224           | .230             |
| 20             | .225           | .356         | .227           | .237             |
| 50             | .227           | .344         | .234           | .244             |

4. Results and Conclusion
For clustering, as the cluster size increases, the time taken for clustering in Make_Density_Based and Farthest First clustering algorithms increases. Hierarchical clustering algorithm takes more time when number of clusters is less. In all cases, simple k-means clustering algorithm remains consistent. So, Simple k-means is preferred for less no of clusters and Hierarchical clustering for large size of clusters. For classification algorithms, the two classification algorithms which alternatively work better in different fields; Naïve Bayes takes lesser time to build a model, J48 classifies most of the instances correctly and increases the accuracy. Hence the J48 algorithm is recommended to be used in our model.

5. Future Work
The major future research scope is in sentiment analysis using the above classification and clustering methods. The areas identified are: (1) Sentiment Analysis on abbreviations and emoji; (2) Improving sentiment word identification algorithm; (3) Successful handling of bipolar sentiments; (4) Apply the results of sentiment analysis in marketing research; (5) Expand the technique to collective intelligent research; and (6) Generation of highly content lexicon database.

Acknowledgments
The authors would like to acknowledge the financial support: MITACS Accelerate grant (IT09867, PI: Yongsheng Ma) and NSERC Discovery grant (NSERC RGPIN 5641 Ma, PI. Yongsheng Ma). All the research works were carried out at the University of Alberta.

References
[1] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” Found. Trends Inf. Retr., vol. 1, no. 2, pp. 91–231, 2006.
[2] F. Hemmatian and M. K. Sohrabi, “A survey on classification techniques for opinion mining and sentiment analysis,” Artif. Intell. Rev., 2017.
[3] ISO, “ISO 9000:2015, Quality management systems — Fundamentals and vocabulary,” 2015.
[4] L. Dey and S. M. Haque, “Opinion mining from noisy text data,” Int. J. Doc. Anal. Recognit., vol. 12, no. 3, pp. 205–226, 2009.
[5] S. Pippal, L. Batra, A. Krishna, H. Gupta, and K. Arora, “Data mining in social networking sites: A social media mining approach to generate effective business strategies,” Int. J. Innov. Adv. Comput. Sci., vol. 3, no. 2, pp. 22–27, 2014.
[6] D. B. Fernandez and S. Lujan-Mora, “Comparison of applications for educational data mining in Engineering Education,” 2017 IEEE World Eng. Educ. Conf., pp. 81–85, 2017.
[7] “Restaurant Review _ Customer speaks _ Authenticate Indian Restaurant.”
[8] “BBB Business Profile, DriveTime | Reviews and Complaints.”
[9] R. Duriqi, V. Raca, and B. Cico, “Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA,” Embed. Comput. (MECO), 2016 5th Mediterr. Conf. on. IEEE, pp. 96–101, 2016.
[10] V. S. Kumbhar, “Hotel Websites Facebook Data Analysis Using Weka,” in Second International Conference of Commerce and Management (ICCM), 2017.
[11] K. Dave, K. Dave, S. Lawrence, S. Lawrence, D. M. Pennock, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” Proc. 12th Int. Conf. World Wide Web, pp. 519–528, 2003.
[12] S. Wakade, C. Shekar, K. J. Liszka, and C.-C. Chan, “Text Mining for Sentiment Analysis of Twitter Data,” IJE, p. 5, 2012.
[13] S. Joshi, B. Pandey, and N. Joshi, “Comparative analysis of Naive Bayes and J48 Classification Algorithms,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 12, pp. 813–817, 2015.
[14] H. Seif, “Naive Bayes and J48 Classification Algorithms on Swahili Tweets: Performance Evaluation,” Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 1, p. 1, 2016.
[15] T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” Int. J. Comput. Sci. Appl. ISS0 0974-1011, vol. 6, no. 2, pp. 256–261, 2013.
[16] F. Alam and S. Pachauri, “Comparative Study of J48 , Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA,” Adv. Comput. Sci. Technol., vol. 10, no. 6, pp. 1731–1743, 2017.
[17] S. Revathi and T. Nalini, “Performance Comparison of Various Clustering Algorithm,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 2, pp. 2277–128, 2013.
[18] O. Abbas, “Comparisons Between Data Clustering Algorithms,” Int. Arab J. Inf. Technol., vol. 5, no. 3, pp. 320–325, 2008.
[19] M. N. Amin and A. Habib, “Comparison of Different Classification Techniques Using WEKA for Hematological Data,” Am. J. Eng. Res., no. 43, pp. 2320–847, 2015.
[20] T. C. Sharma and M. Jain, “WEKA Approach for Comparative Study of Classification Algorithm,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 2, no. 4, pp. 1925–1931, 2013.
[21] A. Naik and L. Samant, “Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime,” Procedia Comput. Sci., vol. 85, pp. 662–668, Jan. 2016.
[22] G. Li and F. Liu, “Sentiment analysis based on clustering: A framework in improving accuracy and recognizing neutral opinions,” Appl. Intell., vol. 40, no. 3, pp. 441–452, 2014.
[23] G. Kaur and A. Chhabra, “Improved J48 Classification Algorithm for the Prediction of Diabetes,” Int. J. Comput. Appl., vol. 98, no. 22, pp. 13–17, 2014.
[24] A. Goyal and R. Mehta, “Performance Comparison of Naïve Bayes and J48 Classification Algorithms,” Int. J. Appl. Eng. Res., vol. 7, no. 11 SUPPL, pp. 1389–1393, 2012.
[25] R. T. Bayes, P. Laplace, and S. H. Jeffreys, “Bayes ’ theorem.” pp. 1–12, 2018.
[26] M. Meenakshi and G. Geetika, “Survey on Classification Methods using WEKA,” Int. J. Comput. Appl., vol. 86, no. 18, pp. 16–19, 2014.