Classification of Consumer Reviews with Hybrid Approach Using Text Mining Technique

Milee Mode[1], Niti Khetra[2]
[1, 2] Dept of CE
[1, 2] Silver Oak College of Engineering & Technology, Gota, Ahmedabad, Gujarat, India

Abstract- *Many numbers of clients buys item, book travel tickets, purchase merchandise and number of devices using web. Here, customers share their perspectives about item, services, news, display of product and so on as web surveys, sites, remarks and so on. Numerous clients read survey data given on text mining to take choices, for example, purchasing items, watching its demo and going to restaurant and so on. It is difficult for web clients to check from large number of inputs.*

Critical and helpful data can be removed from audits through text mining techniques. We have used text mining and WordNet based technique from restaurant reviews and sentence weight score based on reviews. We have target to achieve better precision for result comparison as positive or negative survey with deep learning methods.

Keywords- Data Mining, Text Mining, Text Mining Process, Random Forest Algorithm, SVM.

I. INTRODUCTION

User reviews on E-commerce websites like Amazon.com and flipkart.com have a large influence on product reputation as they are heavily viewed by possible buyers before they decide to make instances of buying things for money [1]. Very often people use them to share reviews and comments for products. Gathering feedback from consumers is not only valuable for companies to analyze the market and improve their business, but also is very useful for customers to explore other products and business [2].

At constant time, advances in technology and changes in strategies for collecting explicit customer review are generating increasing volumes of unstructured textual data, making it difficult for managers to analyze and interpret this information.

Text mining tools and algorithms can help uncover customer attitudes and sentiments on products they have purchased and used [1].Text mining is a method enabling automatic extraction of information from textual data, is gaining in popularity. However, this method has performed below expectations in terms of depth of analysis of customer experience review and accuracy [3].

The proposed framework incorporates vital parts of client expertise, service methodologies and theories like co-creation processes, interactions and context. Understanding the overall perceptions of a product review it must be a positive, negative, balanced and constant [3].

II. RELAED BACKGROUND

A. Text Mining Process

Following step for Text mining process

Step 1: Text Preprocessing

Text preprocessing is the initial step of text mining which reads one text document at time and processes it. This step divides into following main three subtasks [4]

1.1 Tokenization

Tokenization is the process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens. The goal of the tokenization is the exploration of the words in a sentence. Textual data is only a textual interpretation or block of characters at the beginning. In information retrieval require the words of the data set. So we require a parser which processes the tokenization of the documents.

This may be trivial as the text is already stored in machine-readable formats. But still there are some problems that have been left, for e.g., the removal of punctuation marks as well as other characters like brackets, hyphens, etc. The main use of tokenization is identification of meaningful keywords. Another problem is abbreviations and acronym which need to be transformed into a standard form. [7]

1.2 Stop word removal

This step involves removing of HTML; XML tags from web pages and the process of removal of stop words like "a", "of" etc. are performed. [6]. Using Team Based Random Sampling method works by iterating over separate chunk of data which are randomly selected.

It then ranks terms in each chunk based on their in format values using the Kullback-Leibler divergence measure as shown in Equation.

$$dx(t) = Px(t).\log 2\ Px(t) / p(t)$$

Where Px (t) is the normalized term frequency of a term t within a mass x, and P(t) is the normalize term frequency of t in the entire collection. The final stop list is then constructed by taking the least informative terms in all chunks, removing all possible duplications. [9]

1.3 Stemming

These techniques are used to find out the root or stem of a word. Stemming is the process of converting the word to their stem. [6] Stemming finds the root or stem of the words that are phonologically related, i.e., removing the common suffixes, reducing the number of words, to accurately match stems.

Stemming Algorithms have been developed over the years to optimize the data. Porter's Algorithm is one of the efficient techniques for the English Language. [8]

Step 2: POS Tagging

Part of Speech tagging is the process of identifying the part of speech corresponding to each word in the text, based on both its definition, as well as its context. Using Stochastic Tagging, Stochastic taggers use probabilistic and statistical information to assign tags to words and choose most frequent tag in training text for each word.

Stochastic taggers are popular because of their higher degree of accuracy [14].

Step 3: SVM Classifier

The application of Support vector machine (SVM) method to Text Classification has been propose by. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. [5]

SVM classifier method is outstanding from other with its effectiveness to improve performance of text classification combining the HMM and SVM where HMMs are used to as a feature extractor and then a new feature vector is normalized as the input of SVMs,[5]

So the trained SVMs can classify unknown texts successfully, also by combing with Bayes use to reduce number of feature which as reducing number of dimension. SVM is more capable to solve the multi-label class classification. [5]

Step 4: Feature Extraction

In this Phase mainly perform extract a good subset of words to represent text or sequences. Also used to recognize & classify “significant” vocabulary item from the text

- N-gram

Sometime one word not represent people’s opinion like positive or negative that time multiple n words are considered as one feature. Then multiple n words are considered as one feature. Multiple forms of N-grams have been used including unigram, bigram, trigrams etc [2].

An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n−1) – order Markov model. [10]

N-gram model predicts xi based on xi-(n-1),…,xi-1 . In probability terms, this is

$$P(xi \mid xi\text{-}(n\text{-}1),\ldots,xi\text{-}1$$

Example the word “TEXT” would be composed of the following N-grams:

bi-grams: _T, TE, EX, XT, T_
tri-grams: _TE, TEX, EXT, XT_, T_ _
quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _

In general, a string of length k, padded with blanks, will have k+1 bi-grams, k+1tri-grams, k+1 quad-grams, and so on

Step 5: filter features by Correlation Co-efficient

Correlation is a statistical method used to assess a possible linear association between two continuous variables. It is simple both to calculate and to interpret [12].

Correlation is measured by a statistic called the correlation coefficient, which represents the strength of the putative linear association between the variables in question [11].

Step 6: Feature Selection

This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations and minimum search space required. [6] The main idea of Feature Selection (FS) is to select subset of features from the original documents. [5]

- Random Forest algorithm

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.[13]

Step 7: Shannon Entropy Model

The measure of information entropy associated with each possible data value is the negative logarithm of the probability mass function for the value. [16]

It has been used in a variety of application. in particular, Shannon entropy is often stated to be the origin of the mutual informationmeasure. Shannon's original work has resulted in many alternative measures of information or entropy. [17]

B. Tool

2.1 RapidMiner

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, validation and optimization[15].

RapidMiner is developed on an open core model. The RapidMiner (free) Basic Edition, which is limited to 1 logical processor and 10,000 data rows, is available under the AGPL license. [15]

III. PROPOSED SYSTEM

3.1 Proposed Flow

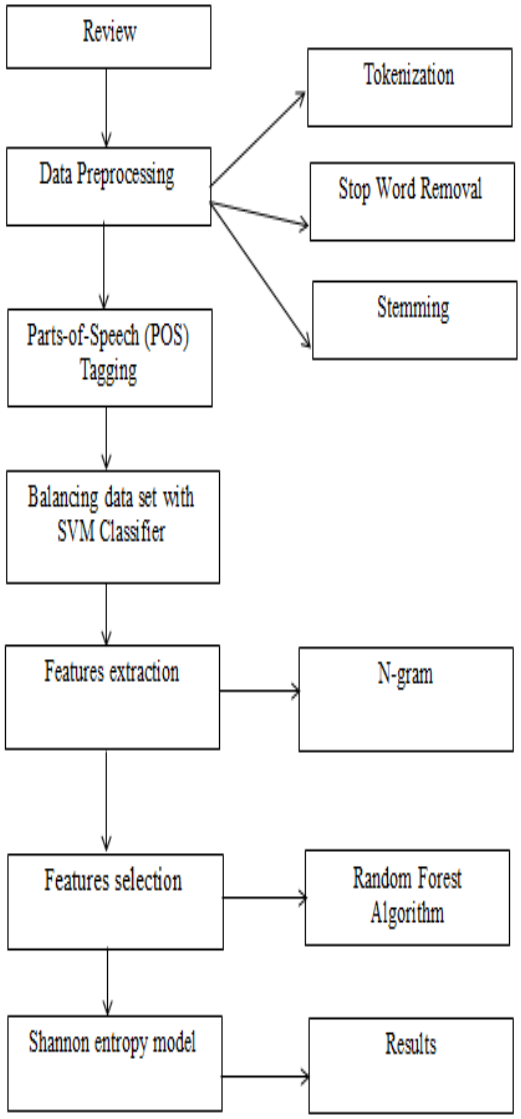

Fig 1 Proposed flow

3.2 Algorithm

Step (1).- For each comment

Step (2).- tokenization each sentences into word.

Step (3).- after the tokenization remove the most frequently word using Random selection stop word elimination.

Step (4).- Then apply the process of Kullback–Leibler divergence measures.

$$D_{\mathrm{KL}}(P||Q) = -\sum_i P(i)\,\log\frac{Q(i)}{P(i)}$$

Where expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P. KL define only if for all i, Q(i) = 0 implies P(i) = 0 (absolute continuity). Whenever P(i) =0 the contribution of the i-th term is interpreted as zero because limx->0 x log(x) = 0

Step (5).- Apply Porter Stemming Algorithm is to reduce words from different grammatical forms to get the root form of word.

Step (6).- Then POS tagging using Stochastic Tagging is the process of identifying the part of speech corresponding to each word in the text.

Step (7).- Now imbalance data set will be balanced with SVM Classifier and it classify with Positive, Negative, Balanced and Constant.

Step (8).- If it is any word is not classify in SVM Classifier then apply the Features Extraction using N-gram it used to extract a good subset of words to represent text.

Step (9).- Then filter features by Correlation is used to measure and describe the relationship between two variables.

Correlation Coefficient ρX,Y between two random variables X and Y with expected values μX and μY and standard deviations σX and σY is defined as

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

Where E is the expected value operator, cov means covariance, and corr is a widely used alternative notation for the correlation coefficient.

Step (10).- Features Selection using Random Forest Algorithm.

Step (11).- Check the results using Shannon Entropy Model.

$$H(X) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$

Where H(X) having a symbol of set {A, B, C, D} where the symbol occurrence frequencies are {A=0.5, B=0.2, C=0.1, D=0.3}, pi is the probability of given symbol and to calculate log2 from another log base use (e.g. log10 or loge).

IV. EXPERIMENT & RESULT

In this paper, experiment is carried out using tool Rapidminer 8.1 we are taking into consideration the data set Iris, given as sample data inside the repository panel of the tool. We apply validation tool on the data set which in turn contains testing operations. In testing column, we took apply model and performance tools respectively. After connecting all operators we execute the tool which in turn shows the accuracy as result.

Table 1 proposed flow result

Accuracy	Precision	Recall
0.90	0.89	0.9082
0.93	0.88	0.9778

Now, we see to obtain the result form proposed flow that yields the best results. In Table 1 show the result of precision and recall for proposed flow output. In training dataset, Proposed Flow achieves accuracy 0.93 when the precision is 0.88 and recall 0.9773. In testing dataset, Proposed Flow achieves accuracy 0.90 when the precision is 0.89 and recall 0.9082.

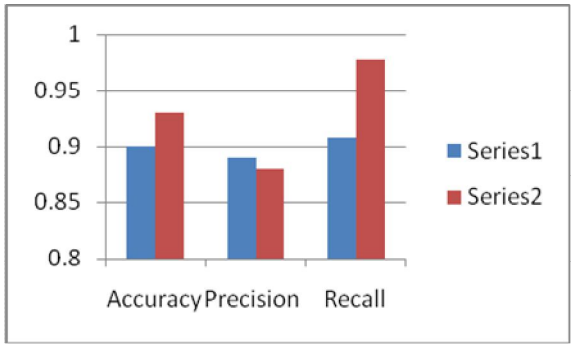

Fig 2 proposed flow result

Furthermore, another experiment is carried out to obtain the number of fold with the best result. The result is shown as Fig.2. We can see from the graph that, when the result reaches the highest point.

Table 2 number of features for accuracy

No of Features	Accuracy	Run time (sec)
7699	0.7681	34.76 sec
9436	0.7493	44.92 sec

Now, we aim to obtain the number of features that yield the best results. This is achieved by running Random Forest classifier. A dataset of 2000 reviews is used, with the maximum number of features set to 7699 and 9436. Table 2 shows the accuracy against the number of features. We can see from the above Table 2 that Random Forest classifier achieves the best accuracy 0.7681 when the number of features reaches to 7699 in testing dataset. In training dataset Random Forest classifier achieves the best accuracy 0.7493 when the number of features reaches to 9436.

V. CONCLUSION

With the reviewed set of papers, i have found that multiple features need to be focused while learning from dataset of reviews. Yelp reviews have been the standard reviews dataset, using which we have tried to get better results. SVM classifier has performed classification by finding the hyper-plane that differentiates the two classes very well because the data is in tag format. We have built the algorithm by mixing text mining algorithms with natural language processing. We have found that it will give better results.

REFERENCES

[1] L. Jack , Y.D. Tsai , "Using Text Mining of Amazon Reviews to Explore User-Defined Product Highlights and Issues", Int'l Conf. Data Mining | DMIN'15.

[2] Yan Zhu, Melody Moh, Teng-Sheng Moh ," Multi-Layer Text Classification with Voting for Consumer Reviews", 2016 IEEE, 978-1-4673-9005-7/16.

[3] Francisco Villarroel Ordenes , Jamie Burton , Babis Theodoulidis , Thorsten Gruber, Mohamed Zaki , "Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach", march 2017, 278-295.

[4] Amrut M. Jadhav, Devendra P. Gadekar "A Survey on Text Mining and Its Techniques", Volume 3 Issue 11, November 2014.

[5] Vandana Korde, "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY", Vol.3, No.2, March 2012.

[6] Sathees Kumar B , Karthika R , "A SURVEY ON TEXT MINING PROCESS AND TECHNIQUES", Volume 3 Issue 7, July 2014, ISSN: 2278 – 1323.

[7] Tanu Verma, Renu, Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", Volume 7– No. 2, April 2014 , ISSN : 2249-0868.

[8] Arjun Srinivas Nayak , Ananthu P Kanive , Naveen Chandavekar , Dr. Balasubramani R , "Survey on Pre-Processing Techniques for Text Mining" , Volume 5 Issues 6 June 2016, Page No. 16875-16879, ISSN: 2319-7242.

[9] Dr. S. Vijayarani , Ms. J. Ilamathi , Ms. Nithya , "Preprocessing Techniques for Text Mining - An Overview" ,Vol 5(1),July-16, ISSN:2249-5789.

[10] William B. Cavnar, John M. Trenkle, "N-Gram-Based Text Categorization", MI 48113-4001.

[11] https://en.wikipedia.org/wiki/Correlation_and_dependence

[12] M.M Mukaka, "Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research", September 2012, 69-71.

[13] https://en.wikipedia.org/wiki/Random_forest

[14] Braja Gopal Patra, Niloy Mukherjee, Arijit Das, Soumik Mandal, Dipankar Das and Sivaji Bandyopadhyay , "Identifying Aspects and Analyzing their Sentiments from Reviews" , 2014[copyright] IEEE, 978-1-4799-9900-2.

[15] Abhishek Kori, "Comparative Study of Data Classifiers UsingRapidminer", 2017, Volume 5, Issue 2, ISSN: 2321-9939.

[16] https://en.wikipedia.org/wiki/Entropy_(information_theory)

[17] P.A. Bromiley, N.A. Thacker, E. Bouhova-Thacker ,"Shannon Entropy, Renyi Entropy, and Information", No. 2004-004.