A Survey On Multi-Layer Text Classification With Voting For Consumer Reviews

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, 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 Classification, Text Mining Process.

I. INTRODUCTION

Data mining is defined as to put out information from large data set. It is the process of finding the useful information from very huge size of dataset. The term data mining is appropriately named as '**Knowledge mining from data**' or "**Knowledge mining**". [1]

It is an interdisciplinary field and it involves an integration of techniques like database technology, statistics, learning of machine, recognition of pattern, information retrieval and spatial or space related data analysis. Data mining promise to efficiently and very important from large databases. The popular name of data mining is Knowledge Discovery from data or KDD.

Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. The following definition is given: Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. [1]

Data mining two types:

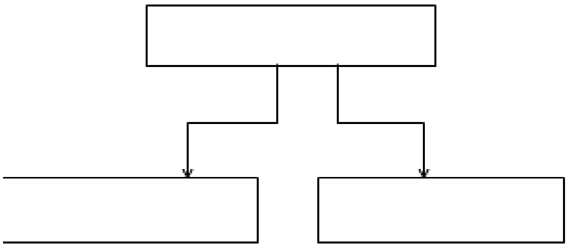

Fig. 1 Types of Data Mining.

Web Data Mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process. [3]

Text mining is the knowledge discovery from textual data or textual data exploration to uncover useful but hidden information. Text mining also known as text data mining or text analytics is the process of discovering high quality information from the textual data sources.

High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output.

II. 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 [2]

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. [11]

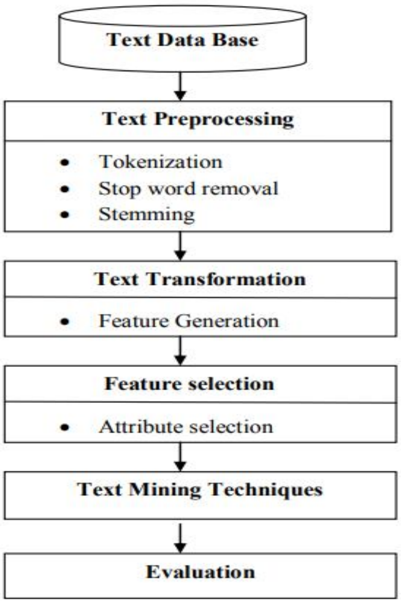

Fig. 2 Text Mining Process. [7]

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. [7]. 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).log2 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. [13]

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. [7] 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. [12]

Step 2: TEXT TRANSFORMATION

Text transformation means to convert text document into the bag of words or vector space document model notation, which can be used for further effective analysis. [7]

Step 3: 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. [7] The main idea of Feature Selection (FS) is to select subset of features from the original documents. [5]

Step 4: TEXT MINING METHODS

There are different text mining methods as in data mining had been proposed such as clustering, classification, information retrieval, topic discovery, summarization, topic extraction. [7]

Step 5: EVALUATION

An important issue of Text categorization is how to measures the performance of the classifiers. [5] This phase includes evaluation and interpretation of results in terms of calculating precision and recall, accuracy etc. [7]

B. Text Mining Classifier

2.1 Naïve Bayes Classifier

The Naive Bayes classifier is perhaps the simplest and the most widely used classifier. It models the distribution of documents in each class using a probabilistic model assuming that the distribution of different terms are independent from each other. Even though this so called "naive Bayes" assumption is clearly false in many real world applications, naive Bayes performs surprisingly well. [8] The basic idea in NB is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. [6]

2.2 SVM (Support vector Machine)

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, 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]

2.3 K-Nearest Neighbor

KNN is a classification algorithm is used for text classification. As given in KNN classifies dataset or objects by voting several labeled training data with their smallest distance from each dataset or object. It uses the local neighborhood to predict the class of an object. The majority vote of its neighbors decides the class of an object. The object is assigned to the class most common among its k nearest neighbors. [9]

2.4 Voting schema

This algorithm is based on method of classifier committees and is based on idea that given task that requires expert opinion knowledge to be performed. k experts opinion may be better than one if their individual judgments are appropriately combined. Different combination rules are present as the simplest possible rule is majority voting (MV)If two or three classifiers are agree on a class for a test document, the result of voting classifier is that class. This method is easy to implement and understand but it takes long time for giving result. [5]

2.5 Neural Network Classifiers

Neural networks are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the generative classifiers. [10]

III. 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] Nikita Jain , Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER", Nov 2013 , eISSN: 2319-1163 | pISSN: 2321-7308.

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

[3] BrijendraSingh ,Hemant Kumar Singh , "WEB DATA MINING RESEARCH: A SURVEY" , 2010 IEEE[copyright] , 978-1-4244-5967-4.

[4] LokeshKumar ,ParulKalra Bhatia, "TEXT MINING: CONCEPTS, PROCESS AND APPLICATIONS", 2010,ISSN-2229-371X.

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

[6] Vidhya. K. A , G. Aghila , "Text Mining Process, Techniques and Tools : an Overview", July-December 2010, Volume 2, No. 2, pp. 613-622 .

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

[8] Mehdi Allahyari , SeyedaminPouriyeh , Mehdi Assef, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez,

KrysKochut , “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques”, 2017.

[9] NiharRanjan, Abhishek Gupta, IshwariDhumale, PayalGogawale ,RugvedGramopadhye , “A SURVEY ON TEXT ANALYTICS AND CLASSIFICATION TECHNIQUES FOR TEXT DOCUMENTS” , Vol. 5, Issue, 11, pp. 5952-5955, November, 2015.

[10] Charu C. Aggarwal, ChengXiangZhai,” A SURVEY OF TEXT CLASSIFICATION ALGORITHMS”, 2012, /978-1-4614-3223-4_6.

[11] TanuVerma, Renu, Deepti Gaur, “Tokenization and Filtering Process in RapidMiner”, Volume 7– No. 2, April 2014 , ISSN : 2249-0868.

[12] ArjunSrinivasNayak , 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.

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