A UNIQUE METHODOLOGY IN CLASSIFYING USER REVIEWS USING TEXT MINING: SURVEY

Aagam Shah[1], Komal Kothari[2], Umang Thakkar[3]
Department of Computer Engineering,
Silver Oak College of Engineering and Technology,
Gujarat Technological University, Ahmedabad, India

Abstract: In this outgrowing usage of internet, digital medium has picked up the highest reach in the sleeves of any individual in forms of social media websites. Amidst the growth of digital data, it has become essential to deal with it in prospects of mining it efficiently to accurately handle the data and storing it in efficient ways. Data mining is a field which is picking pace tremendously with the increasing demand. Data mining includes Text mining which is used to evaluate the content of the data and further classify it. With the increasing user content on the web, it becomes essential for the business heads to analyze the hidden potential reviews for a product in order to match the needs of the customer. This feat can be achieved by concepts of text mining. Statistical analysis can help in evaluating the classified content and plot the market trends graphically which can help in analyzing the business outcomes. Whereas another field of text mining is sentimental analysis which can judge the sentiments of a person from the text content provided. It becomes essential at the time of evaluating success ratio of any product. In this survey, we are analyzing the current existing systems and algorithms developed for classification of text.

IndexTerms - **Text *mining, Customer Reviews, Statistical Analysis, Classified Reviews, Sentimental Analysis, Machine Learning.***

I. INTRODUCTION

In today's digitized modern world, users get benefitted by different products over the world quite hassle-free just on one click with the boon of e-commerce websites. And with the growing flock of websites, enormous data on the web is generated too. For big corporate heads like Amazon, Facebook, mining the user's data has become a tedious task, since it is in large numbers. Data mining is a field which deals with big data to extract knowledge from it and to store it efficiently. Amongst it, the field of text mining has drawn attention as it is related to analyze the content of the data. The objective of Text Mining is to exploit information contained in textual documents in various ways, including discovery of patterns and trends in data, associations among entities, predictive rules, and etcetera [1].

The data on the web is in raw form and it is most essential for the business heads to analyze the content in order to grow their market sales. This data contains the true evaluation of product as well as unbiased reviews. These reviews towards products help the business heads in understanding the needs of user and also promoting the product as per user's requirement. Such sort of substance provided by web is named as client produced content. Client created content contains a great deal of significant and essential data about the product as well as business administrations. Text mining algorithms are used to understand these data and the sentiments of the user. There are two main branches of text mining which includes sentimental analysis and statistical analysis. Sentimental analysis focuses on understanding the motive of the user after the text written while statistical analysis is done to evaluate the data content and make market strategies to grow business. The content generated on the web can sometimes be not trustworthy since there is no control on the nature of this substance on the web and thus, these elevate fraudsters to compose counterfeit reviews to defame the business administrations, to provide misguiding reviews, to generate irrelevant content regardless of the product or service, to advertise unrelated content, and so on. These fake surveys anticipate clients and associations achieving genuine decisions about the product and services that a company has to offer in real.

In this case, Review Analysis has become vital to generate authenticated and unbiased reviews which help in avoiding fraudulent activities used to promote business by publishing fake reviews. This can be achieved solely by **text mining**.

Hereby in this survey we focus on learning the existing systems about text mining and its algorithms. And then after properanalysis, we may further extend our ideas towards implementing an efficient classification algorithm using machine learning.

II. THEORETICAL REVIEW

Text mining is the evaluation of information contained in regular descriptive content in a text. The utilization of such analysis to tackle business issues and to understand the requirements of users is called text analytics [2]. Text mining can empower a relationship to decide potentially imperative business bits of information from content based substance, for instance, word reports, email and postings by means of online systems administration media streams like Facebook, Twitter and LinkedIn [2]. It becomes essential for the business heads to analyze and understand the needs of their clients by mining people's needs and demands from their reviews. This client reviews help them to expand the market deals by targeting the correct prerequisite of the customers. For achievement of such undertaking, text mining is utilized and further that raw content is cleaned, mined and investigated. The flow chart for the above said methodology is Knowledge Discovery in Database (KDD). The term KDD, majorly focuses on obtaining knowledge from raw and unprocessed data and also provides data mining techniques and its implementations. KDD alludes to the general procedure of finding valuable information from unprocessed data. It includes the survey of knowledge obtained after evaluation, as well as on generating patterns to efficiently display the outcome to settle on the choice of what qualifies as learning. It additionally incorporates the decision of encoding plans, preprocessing, examining, and projections of the information before the data mining step is performed in the model. The KDD flow process is not only important to data mining but it also attracts the researchers towards implementing this methodology in machine learning, data visualization, statistics, databases, artificial intelligence, and information acquisition for master systems. The sole objective of the KDD procedure is to extract useful knowledge from unrecognized data with regards to humongous databases. It does this by utilizing Data mining techniques to distinguish what is regarded useful information, as

indicated by the determinations of measures and thresholds, utilizing a database even if there is any requirement of preprocessing of raw data, sub sampling (classifying) of datasets, and changes of that database.

Apart from this, after knowledge extraction statistical analysis is used for business perspective while sentimental analysis of the data is done to understand the real context of the text. Statistical analysis is utilized to discover designs in unstructured and semi-organized client information that can be used to make a more positive client experience and boost up sales [3].

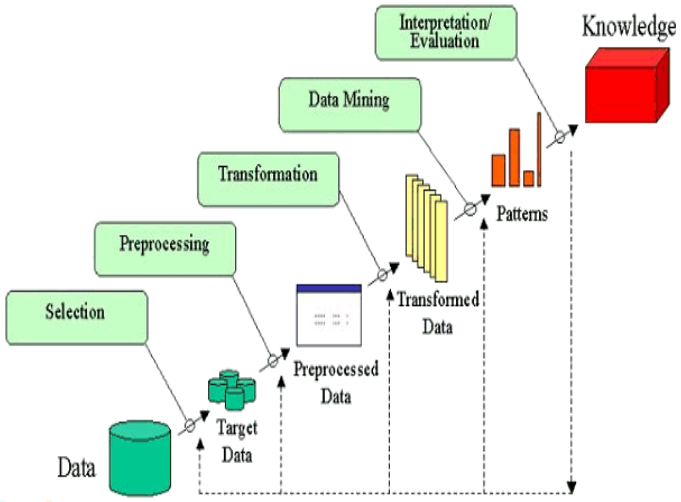

Figure 1: KDD Process

III. DEFINATIONS AND TERMINOLOGIES RELATED TO TEXT MINING

3.1 Classification: Classification is the method toward finding a model (or feature) that represents and recognizes information classes or ideas. The model is determined to evaluate data objects for which the class labels are known (i.e., the sample dataset). And further, the model is trained to obtain the missing values by prediction techniques in unknown class labels [4].

3.2 Regression: Regression is used to predict missing or Unavailable numerical data values rather than (discrete) class labels. **Regression analysis** is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data. [4].

3.3 Decision Tree: Decision tree is a graphical representation of classification algorithms applied. It represents a tree like structure, where each root node shows a test on an attribute value, each branch stands for the output result, and tree leaves denotes class labels [4].

3.4 Clustering: Clustering is an approach similar to that of set theory. It splits data into different subsets where data of similar kind is combined in a single subset. No two subsets can have similar object of a dataset [4].

IV. ALGORITHMS FOR CURRENT USER REVIEW MINING:

As Text Mining is an essential task in classifying user reviews to achieve statistical and sentimental analysis, several types of algorithms are developed to analyze and understand the raw data collected from user reviews and then further to classify them according the need of the companies. Among all different text mining algorithms working in dynamic environment, some algorithms are better and improved and results in better performance. Our approach is to combine best algorithms that are capable to provide authenticated and classified outcome of analyzed user reviews.

4.1. Sentiment Analysis and Classification Based On Textual Reviews [5]

According to this study paper, Mining is utilized to help individuals to obtain knowledge from raw data. Sentimental analysis focuses on recognizing person's opinions from the given text. This analysis is valuable in social media monitoring to naturally describe the general feeling or state of mind of users as reflected in web-based social networking toward a particular brand or company and decide if they are reviewed positively or negatively on the web. And for regulating the task of classifying a single topic textual review, and for expressing a positive or negative evaluation, a new algorithm called Sentiment Fuzzy Classification where parts of speech tags are used to improve the classification accuracy on the benchmark dataset of Movies reviews.

Advantages:

- Compares past and previous algorithms
- Provides accuracy for multi-theme document
- Classifies by identifying words according to their emotions in positive, negative or neutral categories
- Uses functional words as stop words to identify different meaningful words

Disadvantages:

- Limits itself to bag full of words and does not improvise by self learning
- It may change meaning of the entire sentence in order to classify a sentence by eliminating stop words.

4.2. Statistical And Sentimental Analysis Of Consumer Product Reviews [6]

This research paper is strongly written after observing the outgrowing field of big data commerce and its need to classify further to boost up the e commerce giants like Amazon, Flipkart and so on. Tech giants which possess user data and their reviews for different products are needed to be classified further for analysis purpose and growth of business sector. They are not only beneficial for the consumers but also for the product manufacturers. Online reviews have the potential to provide an insight to the buyers about the product like its quality, performance and recommendations; thereby providing a clear picture of the product to the future buyers. To match this need, this paper provides a method to filter user reviews on basis of sentimental analysis in two categories which are positive and negative. In this research, data analysis of a large set of online reviews for mobile phones is conducted.

Advantages:

- It focuses not only on classification of the text into positive and negative sentiment but have also included sentiments of anger, anticipation, disgust, fear, joy, sadness, surprise and trust.
- Classifies the user reviews according to brand names and then distinguishes them according to the ratings and review length.
- Sentimental analysis done with the help of inbuilt *Syuzhet* package is accurate enough for classifying the reviews.

Disadvantages:

- This paper solely focuses on classifying the reviews and not on providing authenticated reviews.
- Sentimental analysis done may only be beneficial to the business heads but not the customers.

4.3. An Efficient Machine Learning Bayes Sentiment Classification method based on Review Comments [7]

In this research paper, author made creativity by mixing blend of accuracy as well as classification. A noteworthy concern while joining semantic information bases for sentiment mining is that the words chose does not settle trait significance and couldn't ground positive and negative utilization of vague terms. These worries frequently make it hard to order the conclusion words from client audit remarks. This paper shows a novel technique called Machine Learning Bayes Sentiment Classification (MLBSC) to enhance the grouping precision by framing classes (i.e., positive, nonpartisan and negative) in light of the separated words from client audit remarks.

Advantages:

- The MLBSC method adapts itself from sample data and thus helps in self modification of user reviews to extract information.
- Sentiment classification becomes accurate with the help of naïve Bayesian method.

Disadvantages:

- Sentimental class labels are filled using probability algorithms which in turn creates more variance in statistical analysis.

4.4 Opinion Mining for Thai Restaurant Reviews using K-Means Clustering and MRF Feature Selection [8]

In this survey paper, author discussed opinion mining on many Thai restaurant reviews in an unsupervised way is a challenging task to survey feedbacks of the customers on their products and services. This is greatly useful for proprietors to enhance their business. In this paper, author propose a opinion mining on Thai eatery audits utilizing K-Means grouping and MRF feature selection. The proposed technique starts with content preprocessing for breaking surveys into words and expelling stop words, trailed by content change for making watchwords and producing input vectors.

Advantages:

- MRF feature helps in recognizing relevant features from a bunch of different features and thus reducing the features set.
- This reduction in feature sets save a lot of computational time.
- K-mean method increases performance of clustering restaurant reviews.

Disadvantages:

- This approach does not focus on providing detailed classification. Clustering here is done on unsupervised data which is not liable.
- Authenticity of user review data processed is not guaranteed.

4.5. Online Product Review Summarization [9]

Author shows a new way for organizing user reviews for the betterment of marketers and customers. Review synopsis is a procedure of extracting and gathering audits which are posted on websites. Review summarization helps to hike critical information about any item on a lesser time. This framework is most appropriate for client and business heads. It gives a stage to see all reviews as well as to keep up records of a survey. From shopper surveys, critical angles are distinguished and on that viewpoint sentiment classification id applied, lastly apply the ranking algorithm to decide the specific item ranking.

Advantages:

- Provides all detailed review analysis at the same platform.
- Classification based on sentiments of the user reviews is beneficial for adjourning the ranking algorithm significantly.

Disadvantages:

- Authentication of reviews is not guaranteed since it has no measures to check the liability of the content provided.

V. CONCLUSION

By studying many research and survey papers, we came to know different types of text mining algorithms and furthermore decided to merge them to get better efficiency rather than to perform individually. For enhancing such performance, we came up with novel approach to select algorithms dynamically based on some condition and situation in which they are better. Moreover we also need to gather some sample datasets of user reviews and then further improvise its efficiency by machine learning. We decided to take user reviews of food app like Zomato, Yelp, Swiggy, etc as a sample dataset.

REFERNCES

[1] Text Mining and its Business Applications - CodeProject. (2014). Codeproject.com. Retrieved 7 February 2018, from https://www.codeproject.com/Articles/822379/Text-Mining-and-its-Business-Applications

[2] What is text mining (text analytics)? - Definition from WhatIs.com. (2018). Search Business Analytics. Retrieved 9 February 2018, from http://searchbusinessanalytics.techtarget.com/definition/text-mining

[3]What is statistical analysis? - Definition from WhatIs.com. (2018).WhatIs.com. Retrieved 17 February 2018, from http://whatis.techtarget.com/definition/statistical-analysis

[4] Jaiwan Han, Micheline Kamber, Jian Pie, "Classification: Advanced Methods," in Data Mining Concepts and Techniques, 3rd edition. The United States

[5] Sentiment analysis and classification based on textual reviews - IEEE Conference Publication. (2018). Ieeexplore.ieee.org. Retrieved 17 February 2018, from http://ieeexplore.ieee.org/document/6508366/

[6] Statistical and sentiment analysis of consumer product reviews - IEEE Conference Publication. (2018). Ieeexplore.ieee.org. Retrieved 19 February 2018, from http://ieeexplore.ieee.org/document/8203960/

[7] An efficient machine Learning Bayes Sentiment Classification method based on review comments - IEEE Conference Publication. (2018). Ieeexplore.ieee.org. Retrieved 19 February 2018, from http://ieeexplore.ieee.org/document/8249985/

[8] Opinion mining for thai restaurant reviews using K-Means clustering and MRF feature selection - IEEE Conference Publication. (2018). Ieeexplore.ieee.org. Retrieved 19 February 2018, from http://ieeexplore.ieee.org/document/7051469/

[9] Summarizing customer review based on product feature and opinion - IEEE Conference Publication. (2018). Ieeexplore.ieee.org. Retrieved 19 February 2018, from http://ieeexplore.ieee.org/document/7860894/