A NOVEL APPROACH IN USER REVIEWS ANALYSIS USING TEXT SUMMARIZATION AND SENTIMENT ANALYSIS: SURVEY

Komal Kothari[1], Aagam Shah[2], Satvik Khara[3], Himanshu Prajapati[4]

[1,3,4]Department of Computer Engineering, [2]Technical Consultant
[1]Silver Oak College of Engineering and Technology
[1]Gujarat Technological University, Ahmedabad, India

Abstract: With the increasing usage of digital media, the textual content is rapidly increasing over the web in recent years. These textual data possess a rich source of information which is beneficial for many. Tech giant portals like Yelp or Amazon, contains millions of user reviews in textual format which contains valuable reviews against a product or a business. It is nearly inhumane task to manually surf over all these reviews and this is where Text Summarization comes into effect. Text Summarization is a technique to scope larger texts into a short meaningful summary retaining the original intent of the text paragraph. Implementing the Text Summarization methods on user generated reviews can provide a better understanding in much lesser time. In this survey, we are examining different models of text summarization and sentiment analysis on our customer review dataset in order to provide a better methodology for the customers of the review-sites to quickly understand peer opinions and make more informed decisions.

Index Terms - Text Summarization, Customer Reviews, Sentimental Analysis, Summarized Reviews, Machine Learning, User Reviews.

I. INTRODUCTION

Amidst the hike in internet usage, there has been a tremendous growth in the digital data over the years. Easy access of internet benefits the user to post textual as well as multimedia content on social media platforms. The data gathered over the web is highly varied. People share opinions, reviews over the web for the products and businesses they have used. This data is usually in textual format. User reviews are much useful for each business to grow. User reviews help in understanding the correct need of the customer and the problems they face. E-Commerce websites, Travel and Food Industry, Entertainment Industry, and many more rely on these user reviews as prime source of market value analysis and publicity. User reviews are also beneficial for the potential customer in order to get familiar with the product from previous experiences of other users. Customers rely on these user reviews provided on such online portals and analyze these reviews in order to make any decision towards buying the product or subscribing any service. Customer opinions are much important as it:

1) Stimulates the decision to buy
2) Increases confidence in new customers
3) Helps build loyalty for existing customers
4) Improves SEO positioning of e-commerce
5) Brings credibility to products and the company

The internet explosion has taken word of mouth publicity to the next level. Consumers are benefitted with every details about the company and the products on their fingertips in form of user reviews. The customers now have the perk to express their views and experience to their fellow consumers, in a way which is highly influential and powerful. Online evaluation and star ratings give a voice to the customers, increase buyer's confidence and can significantly improve sales. All things considered, online ratings and reviews give mind boggling worth and advantages that organizations can't bear to overlook. But what if there are millions of these reviews? Review portals like Yelp, Amazon, etc., possess millions of user reviews accustomed to different products and services. These reviews contain much important content but it is almost impossible for any human being to read and analyze these reviews. It is indeed very important to understand all these review in most convenient manner, and in order to analyze these reviews, it is where Text Summarization comes into the play!

1.1.   Why Text Summarization?

Text Summarization, as the name suggests, is the method of summarizing the textual format of data. [1] Text summarization refers to the procedure of minimizing the long paragraphs of text into a meaningful summary. The idea is to create a coherent and fluent summary having only the important information from the original text document. When the text summarization methods are applied on the user reviews, it marginally reduces the reading time for the users as they can get the quick overview of the original review. Text Summaries make the selection process much easier when the user is struggling to understand the features of the product/services. In order to further understand the real meaning of the client generated review, sentiment analysis is also used to examine and understand the intentions of the user mentioned in the review.

1.2.   Why Sentiment Analysis?

Sentiment analysis is a technique that extracts emotions from the raw texts. Sentiment analysis focuses on understanding the motive of the user after the text written in the review. It provides positive negative characteristics of the text. This can be very useful in understanding the user reviews.

In order to analyze the review more rapidly and to understand the emotions behind the review, **text summarization** and **sentiment analysis** are the crucial methods.

Hereby in this survey, we focus on learning the existing systems about text summarization and sentiment analysis along with its algorithms in the market. And then after proper analysis, we may further extend our ideas towards implementing an efficient text summarization and sentiment analysis algorithm using machine learning techniques.

II. THEORETICAL OVERVIEW

With such a huge volume of textual data revolving around in the digital world, it is the need of the hour to research and build new machine learning algorithms that can automatically shorten the length of the texts and provide accurate summaries which will fluently pass the intended messages.

Text Summarization is the study of extracting important information from the texts and accommodating it into shorter and meaningful abbreviation. Moreover, using text summarization methods effectively reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area. And thus machine learning methods like Automatic text summarization is much needed. [2] **Automatic Text Summarization** is the automatic creation of literature abstracts, using different techniques like word frequency and phrase frequency to extract important sentences from the document for summarization purposes. Different methods are used like the presence of key words, words used in the title appearing in the text, and the location of sentences, to extract significant sentences for automatic text summarization.

2.1.   Text Summarization can be divided into two major approaches: -

1) Extraction-based Summarization

2) Abstraction-based Summarization

1. Extractive Summarization: This approach is widely used and it rely on extracting various parts from the textual paragraph, such as pulling key phrases and sentences, as per their importance in the provided text and compile them together in order to create a summary of the most important context of the original paragraph. The extraction is made based on the pre-defined equation without modifying anything in the original text. Thus, the major task which is of utmost importance in following these approach is to choose and extract the right sentences for summarization.

Input document → sentences similarity → score sentences → select sentences with higher rank.

2. Abstractive Summarization: This approach uses advanced machine learning algorithms and NLP techniques to obtain a new summary from the pile of texts. In this approach, it is also possible that some parts of the summary may not even present in the original text but they are added using NLP techniques in order to provide shorter and meaningful summary. This method includes paraphrasing and when applied with deep learning principles, this method can overcome the grammatical errors from the extractive method. The text summarization algorithms required to perform abstraction summarization are very difficult to develop; and that is because the use of extraction methods are more popular.

Input document → understand context → semantics → create your own summary.

Sentiment Analysis also referred as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Besides, these systems also extract attributes of the expression e.g.:

- o   *Polarity*: if the speaker expresses a *positive* or *negative* opinion,

- o   *Subject*: the thing that is being talked about,

- o   *Opinion holder*: the person, or entity that expresses the opinion.

At present, sentiment analysis is a topic of extraordinary interest and development since it has numerous practical applications. Since publicly and privately accessible data over Internet is continually growing, a large number of texts expressing opinions are accessible in review sites, blogs, social media and forums.

This raw data over the web could be automatically converted into well-structured information of public reviews about products or services, or any topic that individuals can express their views about with the help of sentiment analysis framework. This data can be

exceptionally helpful for business applications like marketing analysis, public relations, product reviews, product surveys, net advertiser scoring, product feedback, and client support.

Sentiment Analysis is the automated technique of learning the sentiment or emotion of the provided user opinion in the textual format. This method can help in understanding and analyzing the product reviews as it bifurcates them into special tags such as *positive, neutral, negative.* These tags are used in learning the real motive of the writer for the provided review. Sentiment classifiers are further used to train the model not only to understand how and what the clientele are talking about a product or a service in the reviews but also to further classify the aspects of the review for which category it is intended for.

III. TASKS AND APPLICATIONS OF TEXT SUMMARIZATION

Tasks followed in Text Summarization are as follows:

1) Construction of an intermediate representation of the input text
2) Scoring the sentences based on the representation
3) Selection of a summary comprising of a number of sentences

Text Summarization has hundreds of applications but we shall limit here with few main applications [3]:

1. User Review Understanding – Text Summarization helps in understanding the user reviews as the entire content of the review can be converted into bullet points and a short abbreviation for faster acknowledgement.

2. News – Text Summarization can help in generating headlines, creating an introduction and even embed caption on pictures from the brief news article provided.

3. Social Media Posting – Social Media content is preferred to be apt and concise. Thus, Text Summarization can be used to transform long blog posts into shorter ones for the social media audience.

4. Conversation Summary – Meeting Recording can be converted into textual format and then only the important points can be extracted from them.

5. Content Writing – Text Summarization can help in generating an outlined summary from the provided topic and points which is beneficial in content writing.

IV. ALGORITHMS FOR CURRENT USER REVIEW SUMMARIZATION:

As Text Summarization is an essential task in preparing a short overview of the textual review, and sentiment analysis being essential in understanding the real emotion behind that review, several types of algorithms are developed to analyze and understand the raw data collected from user reviews and then further to summarize them. Among all different text summarization methods working in dynamic environment, some algorithms are better and improved and results in better performance. Our approach is to examine these techniques that are capable to provide efficient summary and overall better understanding of the analyzed user reviews.

4.1. A Sentiment-Based Hotel Review Summarization [4]

This research work by Debraj Ghosh explains about the summarization of hotel reviews obtained from the travelers. Author says that users also share their perspective about product, hotel, news, and other topics on Web in the form of reviews, blogs, etc. and he feels that it is not easy for any peer web users to read and understand the contents from such a large number of reviews. Hence, he suggests that the important and useful information can be extracted from the reviews through opinion mining and summarization process. He has obtained about 78.2% of accuracy of hotel review classification as positive or negative review by machine learning method. The classified and summarized hotel review information helps the online users to understand the review contents easily during a short time.

Advantages:
- The abstract overviews obtained are pretty good in terms of readability.
- The web user or aspiring traveler gets an overall idea of hotel based on only ten lines of reviews generated by the system.
- This will also be used to create more customized reviews relating to a topic like room, service, food, and also list of hotels by their price and rating.

Disadvantages:
- Efficiency of current algorithm is still low (~78.2% on average).
- Few commonly used sentences occur twice or more in the output summary, which needs to be eliminated.
- Quite a few of the overviews contain similar lines but constructed in different manner, which must have been easily avoided.

4.2. Towards Opinion Summarization of Customer Reviews [5]

This study is strongly presented after observing the current work in the field of summarization along with the earlier methods that were used. In this paper, after the effective comparison, the author has introduced a research plan to use neural networks on user-generated travel reviews to generate summaries that take into account shifting opinions over time. The author outlines future directions in summarization to address all the issues that are being pointed earlier in the survey. The author suggests that by resolving the existing problems, it will be easier for users of review-sites to make more informed decisions.

Advantages:
- In this paper, the background and theoretical overview of text summarization tasks is briefly described.
- More importantly, the authors have described recent contributions and development in this area with the individual pros and cons of each module and the problems the research deals with.
- Efficient comparison is made between the existing systems in the market and it accuracy.

Disadvantages:
- The research proposal presented by the author is comparatively weak and dos not possess anecdotal evidence of the improvement in algorithms.
- The researchers have exemplified that the summarization task mainly including opinion summarization, have wide open issues which shall be researched in the upcoming years.

4.3. Online Product Review Summarization [6]

The author focuses on identifying important aspects from consumer reviews and, on that aspect, provides sentiment classification, and finally applies the ranking algorithm to determine the particular product ranking. This paper elaborates on pursuing Support Vector Machine as an avenue for machine learning to automatically predict review helpfulness. As proven by the results, their hypothesis of creating specialized machine learners to increase performance on partitioned datasets in an effective way to classify product reviews. The proposed system used hybrid classifiers SVM and Naïve Bayes with fuzzy logic. From the various classification methods used, the hybrid methodology in classifying the data was highly vigorous in nature. The accuracy of hybrid classifier method can be increased by increasing the number of classifiers. The authors used this technique to improve the system performance and reduce the error rating.

Advantages:
- The proposed system will give the improved aspect ranking system with few changes. The new reviews are easily updated.
- Positive and negative reviews are entering into the different windows.
- Comparatively it will take the less time for summarizing process and to give the final output.

Disadvantages:
- The author fails to provide any theoretical evidence of work in comparison chart with respect to other algorithms.
- The theoretical claim of solving the disadvantages of other systems is not supported by accuracy measuring graph to prove the statement.

4.4 Opinion mining from online hotel reviews –A text summarization approach [7]

In this research paper, the author discusses about the importance of providing automatically generated hotel summaries that could help the travelers in selecting hotels. This survey provides a unique multi-text summarization method for distinguishing the top most informative sentences obtained from the reviews of hotel. Previous studies on review summarization partially cover the importance of difference in opinion. The author has considered such factors and developed a new sentence importance metric. In order to determine the similarity of two sentences, the author proposes to use both the content and sentiment similarities. Furthermore, with an aim to differentiate the top sentences, and to partition sentences into k groups, the author uses k-medoids clustering algorithm. The medoids from these groups were then chosen as the final summarization results.

Advantages:
- This study considered all the factors such as the credibility of review authors, review time, review usefulness, and conflicting opinions. And in addition to the information (i.e., keywords and key phrases) extracted from the hotel reviews, is also considered.
- Sentence positions, sentence length, whether a sentence contained indicator phrases, and all the aforementioned factors were then combined to obtain the sentence importance score. In addition, the authors have considered to use text pre-processing methods, such as NGD and PMI, in order to handle different opinions among the sentences. Moreover, it is evident that the author used k-medoids clustering method with the intention to filter the inefficient and noisy sentences and further classify the top-k sentences from the user review summarization outcome.

Disadvantages:
- This study has several limitations as well. This survey consists a considerably small sample of dataset used for the experiments and all the participants in the dataset have similar background (i.e., graduate students). Using such a homogeneous group of participants may limit the generalizability of this study.

4.5. Graph-Based Text Summarization Using Modified TextRank [8]

This research paper describes a graph based text summarization method which captures the specifications of a text document. This method has been developed by using the concepts of modified TextRank techniques, and it is evaluated based on the concept of PageRank algorithms which was used for Web pages. The proposed work builds a graph with textual sentences as the nodes and likeness between any two sentences as the weight of the edge between them. This method shows improvement in results as compared with the traditionally used practices in finding weights of the correlated words in the sentences. The author cleverly expresses the accuracy of the method from the efficiency evaluation of proposed summarization method.

Advantages:

- The authors have provided a survey on various methods of abstractive summarization and extractive summarization and have also discussed the pros and cons along with their own proposed work.
- Better results of important sentences are obtained in similarity graph of sentences constructed using proposed modified cosine similarity.
- More information from the graph representation are used for better results and representation.

Disadvantages:
- There may be the possibility of selection of more than one alike sentences with high score for the summary.
- The research work does not overcome the existing problem of duplicated selection of sentences with same frequency. But the author suggests to work on data clustering algorithm along with their proposed work in future.

4.6. User Review Classification and Star Rating Prediction By Sentimental Analysis And Machine Learning Classifiers [9]

Here, in this research work, the authors have proposed a unique approach of classification by performing abstract level sentimental analysis of user review by N-Gram classification and POS tagging. The derived classification obtained from the proposed methodology is then used as an entity for developing machine learning algorithm. This research work highlights upon the proposed technique with promising outcomes and high accuracy by analyzing the data with the help of two algorithms: Maxent Model and Naïve Bayes classifier after analyzing few algorithms including SVM and Random Forest. A proper methodology to classify the user reviews with greatest precision is developed in this paper. Sentiment Classification of User Reviews splits the sentence using N-Gram method and POS Tags in order to bifurcate positive and Negative meaning of the sentence. And then performs star rating prediction, accurately. This model has much relevant scope of developing a topic modelling system from the existing work done on this research as the precision achieved here should be utilized anyhow.

Advantages:
- In this research, the authors present a fine blend of machine learning algorithms and natural language processing methods with adequate amount of proofs.
- This model predicts the star rating of the user reviews of restaurant and performs sentimental analysis which evaluates positive/negative word count used in the review which helps in understanding the real motive of the review quite easily.
- The accuracy of the proposed algorithm and their approach has proven to be the most accurate amongst other classifiers and text mining algorithms.
- Users get benefit with the true and authenticate reviews as per the star ratings predicted against each review. In this way, a user can know the worthiness of that review, while the sentimental classifying of the reviews benefits the user in understanding the real meaning of the review.

Disadvantages:
- The proposed work is currently genre dependent. This algorithm is not tested on user reviews of different genres.
- There is still some more scope to add text summarization methods and natural language processing techniques for acute understanding of the user review.

V. CONCLUSION

After an effective study of many research and study papers, we have developed a cumulated understanding of the nuances of text summarization as well as sentiment analysis and how they can be effectively used in the prime aim of understanding the user reviews. The study of the current algorithms suggests that there are lot of further scope in developing an enhanced system that can comparatively improve the understanding of user reviews. In order to contribute towards the field, we decided to implement few of the algorithms on our datasets to effectively understand the working. Based on the study, we propose to develop a methodology that consists a mixture of extractive text summarization and sentiment classifiers using NLP techniques. We shall try to overcome the problems faced in few of the existing works and develop a novel approach.

REFERNCES

[1] Medium. 2020. *A Quick Introduction To Text Summarization In Machine Learning*. [online] Available at: https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

[2] Analytics Vidhya. 2020. *Introduction To Text Summarization Using The Textrank Algorithm*. [online] Available at: https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python

[3] Sharma, H., 2020. *Text Summarization In Python With Spacy Library - Present Slide*. [online] Present Slide. Available at: https://www.presentslide.in/2019/08/text-summarization-python-spacy-library.html

[4] Ghosh, D. (2019). A Sentiment-Based Hotel Review Summarization. Advances In Intelligent Systems And Computing, 39-44. doi:10.1007/978-981-13-7403-6_5

[5] Pecar, S. (2018). Towards Opinion Summarization of Customer Reviews. Proceedings Of ACL 2018, Student Research Workshop. doi:10.18653/v1/p18-3001

[6] Online product review summarization - IEEE Conference Publication. (2020). Ieeexplore.ieee.org. Retrieved 29 March 2020, from https://ieeexplore.ieee.org/document/8276030

[7] Hu, Ya-Han, Yen-Liang Chen, and Hui-Ling Chou. 2017. "Opinion Mining From Online Hotel Reviews – A Text Summarization Approach". Information Processing & Management 53 (2): 436-449. Elsevier BV. doi:10.1016/j.ipm.2016.12.002.

[8] Mallick, C., Das, A., Dutta, M., Das, A., & Sarkar, A. (2018). Graph-Based Text Summarization Using Modified TextRank. Soft Computing In Data Analytics, 137-146. doi:10.1007/978-981-13-0514-6_14

[9] Shah, A., Kothari, K., Thakkar, U., & Khara, S. (2019). User Review Classification and Star Rating Prediction by Sentimental Analysis and Machine Learning Classifiers. Information And Communication Technology For Sustainable Development, 279-288. doi:10.1007/978-981-13-7166-0_27