Text Summarization of Amazon Customer Reviews using NLP

Himik Nainwal
Dept. of Computer Science
Graphic Era Hill University
Dehradun, India
mailtohimik@gmail.com

Ashish Garg
Dept. of Computer Science
Graphic Era Deemed University
Dehradun, India
geuashishgarg@gmail.com

Aytijha Chakraborty
Dept. of Computer Science
Graphic Era Hill University
Dehradun, India
ac.aytijha@gmail.com

Divyanshu Bathla
Dept. of Computer Science
Graphic Era Hill University
Dehradun, India
bathlad33@gmail.com

***Abstract*— The advancement in technology and the ease to buy products from e-commerce websites have changed the traditional way of shopping. To ensure the quality of products customers mostly believe in reviews of products, but a large amount of online data and the increasing number of users make it challenging for humans to manually summarize product reviews. This study aims to investigate the use of machine learning, natural language processing (NLP), and deep learning techniques for text summarization specifically for product reviews. The methodology used in this study involved applying text summarization techniques to condense product review text into a more manageable and informative format. The major findings of this study suggest that text summarization is a useful tool for reducing reading time and increasing the amount of important information from product reviews. The implications of this study suggest that text summarization can be a valuable tool for efficiently processing and understanding large amounts of online product review data.**

Keywords—Text Summarization, NLP, Machine learning, Deep learning, Word cloud

I. INTRODUCTION

A tremendous amount of data is accessible on digital platforms, it is important to have a tool that can help us get the desired data rapidly. Text Summarization is one such tool that condenses a large amount of information into a concise form through the process of selection of important information as it is a tedious and time-consuming task for individuals to manually select the gist of the elaborated text.

With the rapid growth in users and their feedback on a product, it is difficult for an individual to read a lot of reviews and grasp most of them [1]. As sometimes a piece of wrong information or less information can influence an individual to change his decisions. Hence, the accuracy of the summary of reviews matters a lot. An image is worth more than a thousand of words [2]. Here the main idea is to produce an image consisting of the most frequent words of combined texts to get the overall meaning or theme. In the past many years, different techniques have been used to summarize text where word cloud model and long short-term memory (LSTM) are one of the techniques that have been used widely to generate promising results [3]-[4]. The text summarization can also be improved with the help of sentiment analysis where the important reviews can be selected on the basis of sentiments [5] and learning the sentiment of reviews can also help in increasing the overall customer service [6]. Many researchers believes that deep learning can also be used for text summarization. Deep learning techniques provides promising results and can reduce train loss so that valuable data can be stored [7].

The aim of this paper is to describe such an algorithm that can easily help us identify the accurate summary of features and reviews of a product from Amazon.

II. LITERATURE REVIEW

Boost in technology and the availability of a large range of products on e-commerce websites have shifted the customer interest from offline to the online market but before buying customers prefer to read the customer review to ensure the quality of products. But reading reviews is time-consuming. Most humans are habitual of reading short summaries rather than reading big paragraphs. The traditional way of reviewing comments to understand the good and bad features of the product before purchasing on Amazon can be tedious and time-consuming. An alternative approach is to use images that provide a concise and accurate summary of all reviews, either positive or negative, through text summarization techniques and can significantly reduce the amount of time required for reading and comprehension.

Nisha et al. [8] proposed a hybrid model to summarize hotel reviews where the Support vector machine (SVM,) Genetic algorithm, and Naïve Bayes are used as classifiers and the result shows that hybrid classifier models overpowered single classifiers models. Gupta et al. [9] has studied many methods for sentiment analysis and text summarizing. The system learns and analyzes the attitudes and emotions present in the text using a technique called sentiment analysis.

Liu et al. [10] proposes frameworks for abstractive and extractive models and demonstrate that Bidirectional Encoder Representations from Transformers (BERT) and two stages fine-tuning approach which can enhance the generated summary quality and can be useful for text summarization. Liu et al. [11] proposed a new iterative refinement algorithm, and the outcome of the proposed algorithm is comparable to state-of-the-art methods.

Uddin et al. [12] proposed an aspect and statistical-based algorithm that summarizes the developer's opinion about Application Programming Interface (API) so that developers can choose the API with more accuracy in less time for development work. Bai et al. [13] proposed a pre-trained recommendation model and joint summarization model to predict review rate and experimental results show that the proposed model is better than various existing baseline models. Subha et al. [14] introduced three sequences of steps to predict the sequence of words and to automate text summarization and the steps are word segmentation, morphological reduction, and conference resolution.

Fabbri et al. [15] proposed an end-to-end model which includes an extractive summarization model and an SDS model. Experimental results show that the proposed model provides a comparative result as compared to other existing models. Gamzu et al. [16] proposed a model which extracts helpful sentences with their sentiments i.e., negative, and positive regarding the product, and the result of the experiment outperforms several baselines.

Hou et al. [17] proposed a model which takes multi-domain input such as emotion, usage condition, and product affordance for atomization the process of summarization. Jiang et al. [18] proposed a summary Generation framework to deal with unsupervised and supervised scenarios and to identify important sentences they designed two pre-processes models of re-ranking and selection. Yumo et al. [19] propose a framework of course- to find which is robust across query and domain type and estimate whether the text segment contains relevant information or not. Huang et al. [20] proposed the concept of changing the negative sentimental changes into positive sentimental changes and to evaluate the model Bilingual Evaluation Understudy (BLEU), perplexity and subjective human assessments are used, and the experimental result gives satisfactory results.

Hariprasad et al. [21] proposed a classification model that uses a multi-kernel SVM (MKSVM) classifier with hyper-heuristic salp swarm optimization (HHSSO) to tune parameters. The proposed model is tested on a benchmark dataset and experimental results show that the proposed model provides better accuracy for big data classification. Saini et al. [22] proposed a multi-view-based framework for summarizing scientific documents that don't use any label data for computation and for analyzing the performance, the proposed framework is compared with a single view framework on SciSummNet 2019 dataset and results are calculated using the ROUGE measure.

Çebi et al. [23] implement three neural network-based models that are Attention Based Seq2Seq Neural Network, Pointer Generator Seq2Seq Neural Network, and Reinforcement Learning with Seq2Seq Neural for text summarization. The accuracy of these models is measured using ROUGE-1, ROUGE-2, and ROUGE-L scores. Naithani et al. [24] proposed steps to optimize the technique of text mining and used aspect-based classification techniques for sentiment analysis on social media comments about covid-19.

Hailu et al. [25] proposed a framework to automate text summarization and can successfully determine the salient top n sentence of the source and the result of the experiment shows promising results. Belwal et al. [26] proposed a model which includes a semantic measure and topic modeling with the vector space model to find a summary of a given text and the result of the proposed model is similar to human-generated summaries.

In the past many years, different methodologies and techniques have been used. Joint summarization models, Word cloud, and neural network-based hybrid models are widely used to summarize big data, and BERT and sentiment analysis can also be used to further enhance the accuracy. Along with that various frameworks have been proposed by different researchers to automate the summarization task which provides considerable results.

III. Proposed Model

We have proposed an algorithm that can easily help us to identify the exact/accurate summary of features and reviews of a product from Amazon. Figure 1 shows the flow diagram of the proposed model.

Algorithm

```
1.  function categorization (reviews, ratings)
2.          length ← (length of reviews)
3.          repeat 4 - 7 until length
4.          if ratings >3.5
5.          then pos [] ← (reviews[i])
6.          else neg[] ← (reviews[i]
7.          Update pos and neg
8.  end function
9.  function image_generate(data[])
10.         data ←del_punctuations(data)
11.         data ←del_stop_words(data)
12.         data ← lowercase(data)
13.         data ← stemming(data)
14.         data ← data_selection(data)
15.         data ← tokenization(data)
16.         image ← wordcloud(data)
17.         return image
18. end function
```

In this algorithm, at first, dataset will be prepared by taking reviews for a product and dataset will be categorized by checking whether the ratings are greater than or less than 3.5 (can be changed). Categorizing all reviews into positive and negative will help us in getting the exact meaning of words shown on the image.

After categorizing process of filtering have been done to remove stop words. Stop words is a group of words which is always present in the text data and aren't that important but decreases the accuracy. This group contains is, was, for, of, it, an, etc. As filtering is a process, we use stop words to remove words that are not of concern.

After removal of stop words, stemming will be performed which is one of the most common data pre-processing operations and applied in almost all-Natural

Language Processing. Stemming includes removal a part of a word or to reduce it to its stem or root. Some algorithms are there which help us to cut off those words.

After stop words removal and stemming, tokenization has been performed which includes converting each word into token and after than converting all tokens into lowercase.

Graph-based approaches to Text Summarization would be the Text Rank algorithm, developed by Google. It investigates the links among a set of pages that connect each other and use those relationships to determine the importance of pages.

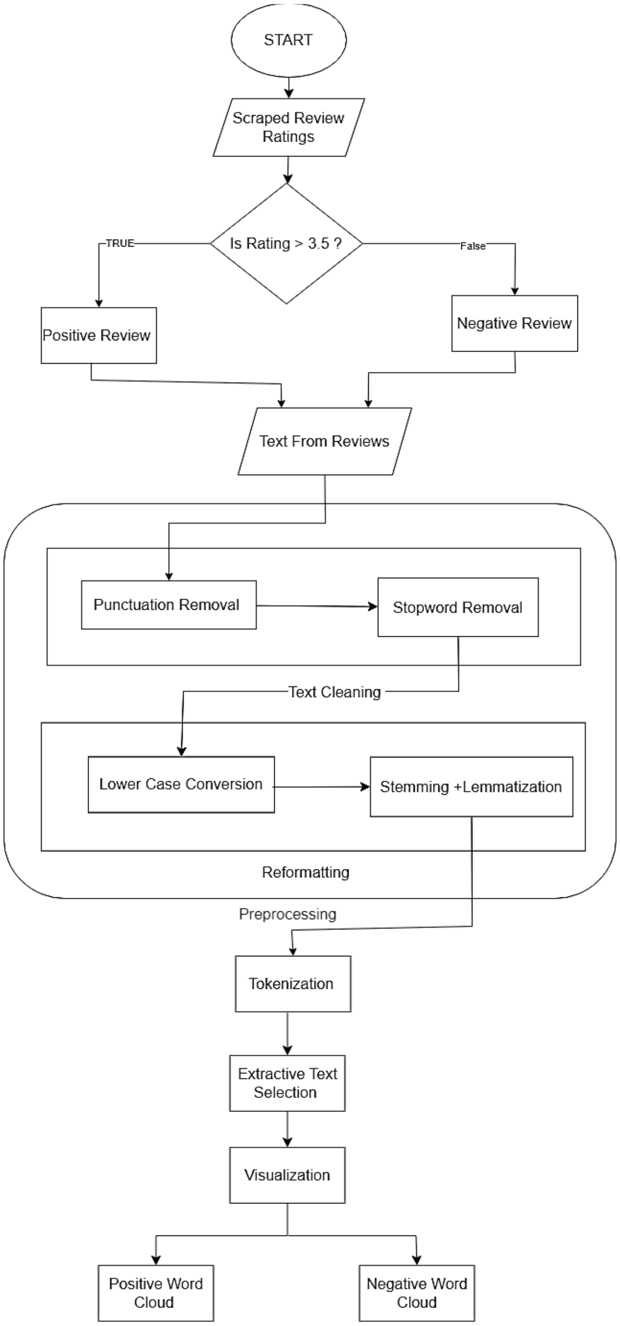

Fig. 1. Flowchart of Proposed Model

For example: Let there be 3 reviews- 1, 2, and 3, and let them be connected as shown in Fig 2.

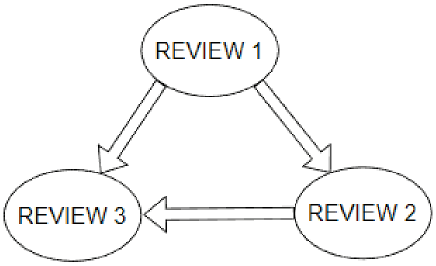

Fig. 2. Text Rank Approach

Here, 2 reviews lead to review 3, 1 review leads to review 2, and 0-pages lead to Page 1. So, the order of importance would be review 3 > review 2 > review 1.

In the case of text summarization, the sentences are considered pages/nodes and any overlapping words are considered links. Accordingly, the sentences are ranked in order of importance.

Some of the words in the word cloud are larger than others. This is because the size of a word in the word cloud is proportional to the frequency of the word inside the corpus.

This model can excel for businesses in finding the customer's main feedback and identifying what employees think about their organization.

IV. Implementation and Results

The dataset used is an updated version of the Amazon review dataset that was released in 2014. It includes reviews, product metadata, and links, as well as additional features such as an increased number of reviews (233.1 million compared to 142.8 million in the previous version), newer reviews (ranging from May 1996 to Oct 2018), transaction metadata for each review, product images taken after receipt, more detailed metadata of the product landing page, and an expansion of product categories with the inclusion of 5 new categories.

The dataset obtained from the source was utilized with the aim of obtaining reviews and ratings for a specific product. Two data frames were then created, one containing positive reviews and the other containing negative reviews, through the process of filtering using a rating point of approximately 3.5. Both datasets were subsequently utilized to generate word clouds, with the positive and negative reviews being differentiated by being processed independently of one another. Text from the reviews was also subject to preprocessing, including the removal of punctuation and stop words. Punctuation was removed through the use of regular expressions, while stop words were imported from the NLTK corpus and the language was set to English. An alternative approach, using spaCy, was also considered. To ensure uniformity and better accuracy, all words were converted to lowercase.

After these steps, normalization techniques such as PorterStemmer() and WordNetLemmatizer() were applied for the purpose of stemming and lemmatization, respectively.

After completing the reformatting steps, in preprocessing steps, which included tokenization and the use of the NLTK library, the text rank algorithm was employed to extract the relevant text. Subsequently, utilizing the word cloud method, an image was generated.

After performing all steps, a word cloud image is generated, as shown in Fig 3.

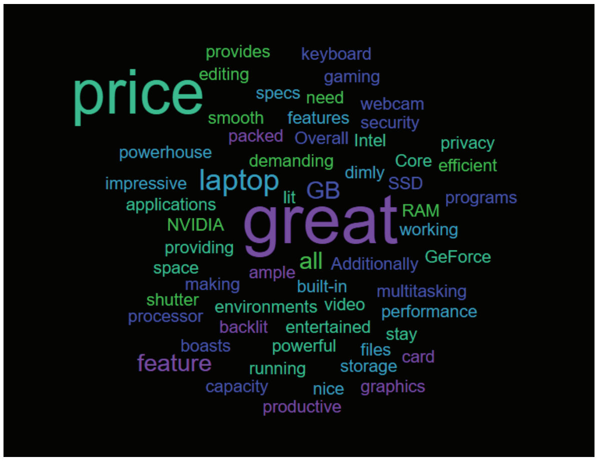

Fig. 3. An image formed by a Word Cloud

As it can be observed from Fig 3 see the larger words show the higher frequency of the words used in reviews. By looking at these types of images in the reviews section of Amazon's website can help a customer easily visualize which features are there positive and negative.

V. Conclusion

The proposed model generates a considerable image which helps extract the summary of reviews conveniently and further helps to save time to extract useful information from the reviews. This can help customers easily to decide whether to buy or not without wasting that much time on reading reviews. In machine learning every problem has a false positive part, which can decrease the accuracy of the model and to decrease the false positive, sentiment analysis can be used along with the proposed model which is left as future work.

VI. Refrences

[1] H. S. Choi and S. Leon, "An empirical investigation of online review helpfulness: A big data perspective," Decision Support Systems, vol. 139. Dec. 2020. doi: 10.1016/j.dss.2020.113403.

[2] L. T. Ferdous, C. Singh, and T. Rana, "A Picture Is Worth a Thousand Words: Audit Efficiency and Risk Management through Data Visualization," Handbook of Big Data and Analytics in Accounting and Auditing, pp. 17–39, 2023. doi: 10.1007/978-981-19-4460-4_2.

[3] S. Djamasbi, M. Siegel, and T. Tullis, "Can Fixation on Main Images Predict Visual Appeal of Homepages?," In Proc. of the 2014 47th Hawaii International Conference on System Sciences. IEEE, Jan. 2014. doi: 10.1109/hicss.2014.54.S.

[4] Djamasbi, M. Siegel, and T. Tullis, "Can Fixation on Main Images Predict Visual Appeal of Homepages?" In Proc. of the 2014 47th Hawaii International Conference on System Sciences. IEEE, Jan. 2014. doi: 10.1109/hicss.2014.54.

[5] J. Shah, M. Sagathiya, K. Redij, and V. Hole, "Natural Language Processing based Abstractive Text Summarization of Reviews," In Proc. of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, Jul. 2020. doi: 10.1109/icesc48915.2020.9155759.

[6] C.-F. Tsai, K. Chen, Y.-H. Hu, and W.-K. Chen, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," Tourism Management, vol. 80, Oct. 2020. doi: 10.1016/j.tourman.2020.104122.

[7] A. K. Mohammad Masum, S. Abujar, M. A. Islam Talukder, A. K. M. S. Azad Rabby, and S. A. Hossain, "Abstractive method of text summarization with sequence-to-sequence RNNs," In Proc. of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, Jul. 2019. doi: 10.1109/icccnt45670.2019.8944620.

[8] N. Yadav, R. Kumar, B. Gour, and A. U. Khan, "Extraction-Based Text Summarization and Sentiment Analysis of Online Reviews Using Hybrid Classification Method," In Proc. of the 2019 Sixteenth International Conference on Wireless and Optical Communication Networks (WOCN). IEEE, Dec. 2019. doi: 10.1109/wocn45266.2019.8995164.

[9] P. Gupta, R. Tiwari, and N. Robert, "Sentiment analysis and text summarization of online reviews: A survey," In Proc. of the 2016 International Conference on Communication and Signal Processing (ICCSP). IEEE, Apr. 2016. doi: 10.1109/iccsp.2016.7754131.

[10] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders." arXiv, 2019. doi: 10.48550/ARXIV.1908.08345.

[11] Y. Liu, I. Titov, and M. Lapata, "Single Document Summarization as Tree Induction," In Proceedings of the 2019 Conference of the North. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1173.

[12] G. Uddin and F. Khomh, "Automatic summarization of API reviews," In Proc. of the 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, Oct. 2017. doi: 10.1109/ase.2017.8115629.

[13] Y. Bai, Y. Li, and L. Wang, "A Joint Summarization and Pre-Trained Model for Review-Based Recommendation," Information, vol. 12, no. 6, May 24, 2021. doi: 10.3390/info12060223.

[14] R. S. Shini and V. D. A. Kumar, "Recurrent Neural Network based Text Summarization Techniques by Word Sequence Generation," In Proc. of the 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE, Jan. 20, 2021. doi: 10.1109/icict50816.2021.9358764.

[15] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model." arXiv, 2019. doi: 10.48550/ARXIV.1906.01749.

[16] Gamzu, H. Gonen, G. Kutiel, R. Levy, and E. Agichtein, "Identifying Helpful Sentences in Product Reviews." arXiv, 2021. doi: 10.48550/ARXIV.2104.09792.

[17] T. Hou, B. Yannou, Y. Leroy, and E. Poirson, "Mining customer product reviews for product development: A summarization process," Expert Systems with Applications, vol. 132, pp. 141–150, Oct. 2019. doi: 10.1016/j.eswa.2019.04.069.

[18] W. Jiang, J. Chen, X. Ding, J. Wu, J. He, and G. Wang, "Review Summary Generation in Online Systems: Frameworks for Supervised and Unsupervised Scenarios," ACM Transactions on the Web, vol. 15, no. 3. Association for Computing Machinery (ACM), pp. 1–33, 2021. doi: 10.1145/3448015.

[19] Y. Xu and M. Lapata, "Coarse-to-Fine Query Focused Multi-Document Summarization," In Proc. of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.296.

[20] Y.-F. Huang and Y.-H. Li, "Sentiment Translation Model for Expressing Positive Sentimental Statements," In Proc. of the 2019 International Conference on Machine Learning and Data Engineering (iCMLDE). IEEE, Dec. 2019. doi: 10.1109/icmlde49015.2019.00025

[21] M. S. Ali and D. Hariprasad, "Hyper-heuristic salp swarm optimization of multi-kernel support vector machines for big data classification," International Journal of Information Technology, 2023. doi: 10.1007/s41870-022-01141-2.

[22] N. Saini, S. M. Reddy, S. Saha, J. G. Moreno, and A. Doucet, "Multi-view multi-objective clustering-based framework for scientific

document summarization using citation context," Applied Intelligence, 2023. doi: 10.1007/s10489-022-04166-z.

[23] Y. Yüksel and Y. Çebi, "TR-SUM: An Automatic Text Summarization Tool for Turkish," Engineering Cyber-Physical Systems and Critical Infrastructures, pp. 271–284, 2023. doi: 10.1007/978-3-031-09753-9_21.

[24] K. Naithani and Y. P. Raiwani, "Novel ABC: Aspect Based Classification of Sentiments Using Text Mining for COVID-19 Comments," Communications in Computer and Information Science, pp. 203–219, 2022. doi: 10.1007/978-3-031-24352-3_17.

[25] T. T. Hailu, J. Yu, and T. G. Fantaye, "A Framework for Word Embedding Based Automatic Text Summarization and Evaluation," Information, vol. 11, no. 2, 2020. doi: 10.3390/info11020078.

[26] R. C. Belwal, S. Rai, and A. Gupta, "Text summarization using topic-based vector space model and semantic measure," Information Processing and Management, vol. 58, no. 3, 2021. doi: 10.1016/j.ipm. 2021.102536.