Text Summarization using NLP Techniques

Poojitha Guntupalli
Erik Jonsson School of Engineering
University of Texas at Dallas
Richardson TX, USA
pxg190042@utdallas.edu

Ratna Mani Meghana Pisati
Erik Jonsson School of Engineering
University of Texas at Dallas
Richardson TX, USA
rxp190065@utdallas.edu

Shammi Akhil Lokam
Erik Jonsson School of Engineering
University of Texas at Dallas
Richardson TX, USA
sxl200002@utdallas.edu

Rohith Jallipalli
Erik Jonsson School of Engineering
University of Texas at Dallas
Richardson TX, USA
rxj200037@utdallas.edu

***Abstract*—Data is a valuable resource. As the data is increasing exponentially, we get challenges and opportunities in 4 dimensions that are volume, variety, velocity and veracity. Big Data directs to data which can't be prepared using basic traditional tools. It is also used to express huge datasets. Summarization alludes to the task of creating a short summary that catches the fundamental thoughts of an original text. It is a course of selecting a group of sentences from the input text data so that an extent of chosen sentences describe the original data without losing context and content. It means picking sentences that doesn't lose the semantic importance of the data in the input data. This is a course of creating brief, significant substance from various assets of data. The primary point of this paper is to perform text rundown utilizing NLP strategies like LSA and Text Rank on the amazon electronics reviews dataset and about presenting the methods to perform data summarization.**

Keywords—Summarization, Automatic text summarization, Text Rank, Data, NLP, LSA, CNN.

I. INTRODUCTION

A. Types of Text Summarization

Automatic text summarization is the task to produce a concise summary which gives important information in the original data. Automatic text summarization technologies are broadly used in the areas such as search engines. Summarization is a method which we try to understand the crucial points from huge document and shorten it to two to three paragraphs.

Here we have generally 2 sort of text summarisation. They are Abstraction and Extraction. Abstractive summarization is a method which creates a context in a general language by taking the relevance of the document and produces a summary which contains a representation of original text. Extractive summarization is a method where we extract several parts like phrases and statements from the original document to generate a summary. The vast majority of the text summarization techniques utilize extraction method because it's easy but abstraction method is a more broad and general approach that people use to summarize data.

B. NLP Techniques

Text Rank and Latent Semantic Analysis (LSA) are the techniques that are used in this report. Latent Semantic Analysis is a process that extracts and represents the meaning of words using statistical process. It converts a collection of unstructured texts to structured data. Text Rank is an algorithm that is similar to the execution of Page Rank Algorithm. The Page Rank algorithm finds the rank of web pages by taking into account the outlinks an inlinks to other web pages. At last the result includes most visited web pages by the users. Text Rank also uses the same method, it splits the given data into sentences and finds each sentences's rank depending on the term similarities. Lastly, we get sentences with their respective ranks. Sentences which have high ranks are used for the summary of the input data.

II. LITERATURE SURVEY

Paper[1] explains about summarization of a DB which has of two months of Convolution Neural Network World-view news. In these 2 techniques are taken to explain the data and performance of both evaluation metrics were calculated. These processes were used here are LSA and Relevance measure. The output results generated by both were also compared with some people. In spite of the distinctive reach for each technique followed to create the results, the exhibition is respected to be very possible.

Paper[2] in this presents methods which uses NLP techniques Latent Semantic Analysis. LSA has generated a fair score for the input text reviews. The performance measure or evaluation metric used in this report is RougeL which contains Precision, Recall and F measure. This evaluation measures are tested on various texts which contain reviews categorized on electronics. The new methods presented in the report are Topic and Cross. Latter method is generally performs better than all other processes and its performance doesn't change even if different methods for input matrix creation are used. This paper presents thoroughly about the technique described above.

Paper[3] this one produces a text by LSA. An advanced similarity matrix computing method is shown and used in this article. This created some good results. It also explains about summary length. The new text comparability calculation method has delivered incredible outcomes.

Paper[4] A graph based text-ranking model was introduced. Like the google's page rank calculation this is additionally a diagram based positioning algorithm. Being specific two chart strategies that are unaided are presented.

III. THEORITICAL AND CONCEPTUAL STUDY OF TECHNIQUE

A. Need?

The necessity for summarization is increasing broadly every year. For instance input data consisting of large size

text document, it is preferable to read a short summary of the document instead of the entire text. As data is growing exponentially every minute, to understand or inspect the data, that much of time needs to be used. So, it is helpful if we provide core of the document. Trying to find the summary of the large document data takes a lot of time. Using these text summarization methods we get the core idea of the topic. This process saves a lot of time and work.Natural Language Processing provides different methods to find summaries of data. There are a lot of use cases of text summarization like news articles and product reviews.When users provide comments or reviews on products we can use this techniques for summarization. This will save time very much and it also gives the main point of what the user of the product wants to tell.We have used a dataset which contains Amazon online website reviews provided by different users of a product and find the summary of a product's reviews.

B. Concept

The main concept of this paper is to summarize a huge no of customer comments/reviews of some online service product and to give an outline about that. This would be verymuch helpful to get an idea about a product features without reading any comments. The summarization process can be done using methods like LSA(tf-idf and SVD based), and TextRank which uses graph based summarization method.

IV. EXPERIMENTAL STUDY

A. Dataset

The dataset consists of electronic products reviews from Amazon US Dataset. It contains the reviews of 30 electronic products from amazon. Approximately 40 reviews of a particular product are stored for each electronic product in separate file for the summarisation task. Each row in a file contains comments/review with columns as review_id, product_title, star_rating, vine, verified_purchase and review_body. The name of the file is product_id is named by its product_id.

B. Dataset contents

product_id.txt:

review_id- Its's an unique id given to a review.
product_title- It is in the text format. Tells the product title.
star_rating- Numerical value ranges from 1-5, tells rating of the product.
vine- It is a class label.
verified_purchase- It would be just 'Y'(Yes) or 'N'(No). It tells us whether the product is verified or not.
review_body- Contains the content of the actual review.

C. Latent Semantic Analysis(LSA)

LSA(Latent Semantic Analysis) is a technique for text summarisation. It uses a factorisation technique called Singular Value Decomposition. To use SVD method, the data needs to be in such a format that the sentences include all the highly important words. This can be accomplished in many ways. For our paper, we are using tf-idf similarity matrix.

Data pre-processing for Latent Semantic Analysis:

1. First, load the file as RDD which contains review_id and review_body.
2. The sentences in each review are split and are assigned an id called sentence_id. Therefore, sentence_id consists of sentence's index in that particular review along with the review_id.
3. Words are collected from the sentences using split, and WordNetLemmatizer is used for removing the stopwords. Words that have length less than five are removed.
4. At last, after all the pre-processing we have list contains all the words(wordList) and sentence_id.

There are three main steps in LSA:
Article[2]

1. Creation of Input Matrix using term frequency:
   Calculation of term-frequency is done using pre-processed data. After that the document frequency is found using term-frequency. IDF- Inverse Document frequency is calculated using document frequency and the total no of statements. tf-idf is the product of TF(term-frequency)and IDF(Inverse document frequency).
   This matrix contains information regarding the importance of all the words in a given document.

2. Singular Value Decomposition:
   The above matrix is a 2D matrix, SVD is used to decompose the matrix into two 2-D matrices.

   $$\boxed{I = USV^T}$$

   U - A Left Singular Matrix which contains the input matrix's rows as vector of extracted concepts

   S - A Diagonal Matrix containing the scalar values in non-increasing order.

   $\boxed{V^T}$- A Right Singular Matrix containing the input matrix as a vector of extracted concepts.

3. Keywords, Concepts and Sentences extraction:
   At last, we get the final summarization. The concepts or keywords are found using the Right Singular Matrix's rows and sentences are found using columns of the Right Singular Matrix.

D. Text Rank

TextRank is a graph-based synopsis method like the Google's-PageRank Algo. In the PageRank calculation rank is registered in light of the in joins and out joins in a webpage. The same technique is done as a component of Text-Rank calculation.

Data Pre-Processing for Text Rank:

1. Upload the record as RDD of review_id and review.

2. Divide the sentences in each file and allocate a sentence_id to each sentence. Here the sentence_id is the mix of review_id and the file of the sentence in that survey.
3. Presently we split the sentences into words and eliminate the stopwords utilizing WordNetLemmatizer. Words of length under five are eliminated so we can think about just the significant words.
4. At last, calculate the rank and the summary of each sentence and display the summary.

   similarity = (count of common words in both the ve rtices)/ (1 + log2(len(vertex1)) + log2(len(vertex2)))

TextRank Formula: TR(Vi) = (1-d) + d SUM,Vj in In(Vi) [[Wji * PR(Vj)]/ [SUM, Vk in Out(Vj) [Wjk]]]*

Further Steps Done:

1. Keyvalue pairs are generated Key is the sentence_Id and the words are the value.
2. A vertex should be created for each and every unique key value pair.
3. Edges are developed between these vertices. The weight of each edge is the closeness esteem between both the vertexes.

Formulae used for calculating similarity-

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|)+log(|S_j|)}$$

4. In light of the edges made we produce an Adjacency list for the vertices made.

We announce number of emphasis made and iteratively process the text position of every vertex by utilizing the formulae:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

At last the Vertex positions are arranged from high to low and the sentences with high position are picked.

Commitment of every vertex is registered by utilizing rank: 0.15 + 0.85 * rank

E. Evaluation Metrics

For finding the efficiency of the summarisation techniques, ROUGE Article[5] is used. The abbreviation for ROUGE is Review Oriented Understudy for Gisting Evaluation. It assesses the output summary with the actual summary of the given input document data.

Types of ROGUE metrics are:

1. ROUGE-N: This metric is an n-gram based comparison of words in summaries, which means it compares n consecutive words, for example ROUGE-1 has n=1 and ROUGE-2 has n=2.
2. ROUGE-L: This metric compares the longest common subsequence of summary in in the system along with longest common subsequence of summary the resultant output.
3. ROUGE-W: This metric is the similar to ROUGE-L metric but longest common subsequence is weighted.
4. ROUGE-S: This metric is based on skip bigram.
5. ROUGE-SU: This metric uses both skip-bigram and unigram.

For our implementation of the technique, ROUGE-L method is used for evaluating the summaries found.

V. EXPERIMENTAL RESULTS

The procedures described above were applied to a dataset including reviews for 30 products on Amazon's electronics website, and the results were evaluated. Figures 1(LSA) and 2(Textrank) demonstrate the summary obtained by the discussed methodologies on the Coby 8 GB 1.8-Inch Video MP3 Player. Figure3,4 demonstrate the summaries and ranks of two particular products.

The precision recall are calculated for a two particular products as shown in figure 5,6.Rouge metrics is used to evaluate the performance of our implementations. We examined the rouge-1, rouge-2, and rouge-L f-stat scores on a specific product, as shown in Figure 7, to see which one delivers better results. The best results were obtained by rouge-2, as shown below(from the peak shown in the figure).

There are no reference summaries available due to dataset limitations. As a result, they are created using software packages, such as pytldr for LSA. Figure 8 shows the graph of LSA scores.

In Fig-1, the data is summarized for each product i.e., each file, the concepts, keywords and sentences are displayed which refers the data summarization.

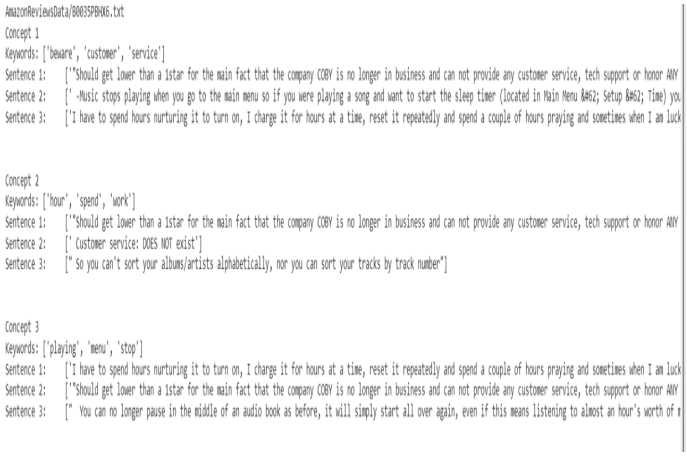

Fig-1.1

```
Concept 4
Keywords: ['file', 'audio', 'sorted']
Sentence 1:    ['I have to spend hours nurturing it to turn on, I charge it for hours at a time, reset it repeatedly and spend a couple of hours praying and sometimes when I an
Sentence 2:    [' -Music stops playing when you go to the main menu so if you were playing a song and want to start the sleep timer (located in Main Menu &#62; Setup &#62; Time
Sentence 3:    ['"Should get lower than a 1star for the main fact that the company COBY is no longer in business and can not provide any customer service, tech support or honor


Concept 5
Keywords: ['sorted', 'language', 'german']
Sentence 1:    [' And with pictures, it makes them all so that the picture is upright, rather than filling the whole screen sideways, even though it rotates music videos sidewa
Sentence 2:    ['"everything worked fine when the unit was plugged into the computer via USB but once I got the music all loaded up and disconnected the USB charger from comput
Sentence 3:    ['"Very nice design and size , problem is the software that is just too basic and will give you trouble organizing music when added to the device , If you dont n


Average score of AmazonReviewsData/B0035PBHX6.txt is 0.004539496501556231
```

Fig-1.2

In fig-2, the top ranks of the processed sentences of each product are displayed.

```
Top ranked reviews of file AmazonReviewsData/B0035PBHX6.txt
Rank= 1.42    Sentence= ['"This works well, but does not keep all the features of previous models, like music speed, brightness adjust, and back-light timeout time']
Rank= 1.29    Sentence= ['" It still plays music well, but I don't like all the problems and glitches that came with this new version"]
Rank= 1.15    Sentence= ['" I can't use the genres, the all songs, or a few other options I would normally use on the music button"]
Rank= 1.09    Sentence= ['Great MP3 player have had aprox 3 years and its still going strong!!!!!!']
Rank= 1.04    Sentence= [' So, I got myself another Coby MP3 player!!!! So glad I did!!!!! I love it and use about every day!!!"']
```

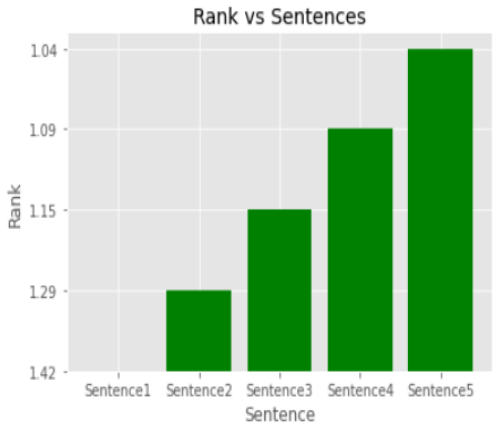

Fig-2

```
Top ranked reviews of file AmazonReviewsData/B0085QVKN6.txt
Rank= 1.4     Sentence= ['"So far a GREAT little speaker, I have been using it for about a month at work and it sounds great, every one has been asking me about it']
Rank= 1.32    Sentence= ['I now own two of these! Great little speaker and love the included stereo cable for quick hookup for non-Bluetooth devices']
Rank= 1.32    Sentence= ['Incredible sound from this little speaker! Battery life is pretty good']
Rank= 1.24    Sentence= ['Love these little speakers! They sound is really loud especially when you daisy chain them']
Rank= 1.24    Sentence= ['"I love this speakers! The sound is very good :-) And I love the size of the speakers, it is perfect to take on trips']
```

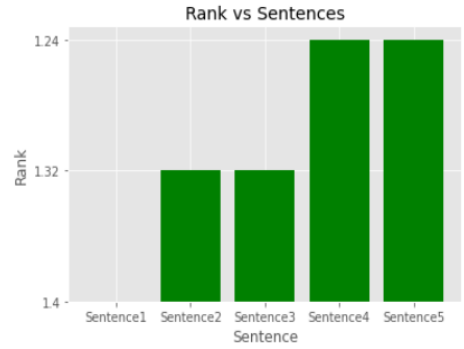

Fig-3

```
Top ranked reviews of file AmazonReviewsData/B0085TBBD8.txt
Rank= 1.38    Sentence= [' Their overall sound quality is decent for a pair of cheaper headphones ( if your concerned about sound quality then spend the extra money for something better)']
Rank= 1.38    Sentence= ['"these are perfect for running! I\'ve had several pairs and I\'ll keep buying these! They stay in even when you\'re sweaty and the sound is great too']
Rank= 1.32    Sentence= ['"These are the best headphones!! Great sound!! Sometimes they last couple years,, sometimes not but the most comfortable and great bass sound!!!"']
Rank= 1.27    Sentence= ['These earphones are the only earphones that we have found that stay over the ear on smaller ears! The sound quality is great too!']
Rank= 1.26    Sentence= ['"I bought three pairs because of how the ear buds fit and feel and the sound quality is great']
```

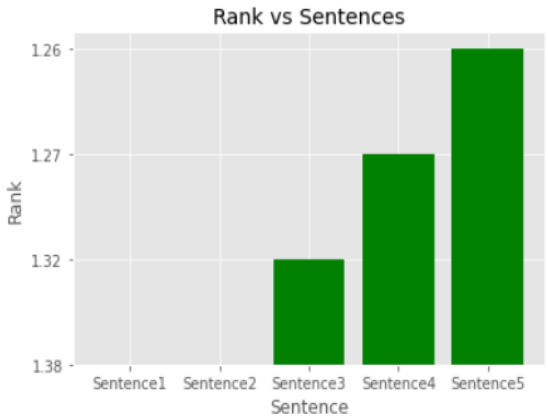

Fig-4

```
The calculated precision  =  0.050892534653011585
The calculated Recall =  0.3032319952774498
The calculated F Measure =  0.07078129215820075
The calculated rfScores= [0.05089253 0.303232   0.07078129]
```

Fig-5

```
The calculated precision  =  0.04160983941635799
The calculated Recall =  0.4160563252813139
The calculated F Measure =  0.06185166224963845
The calculated rfScores= [0.04160984 0.41605633 0.06185166]
```

Fig-6

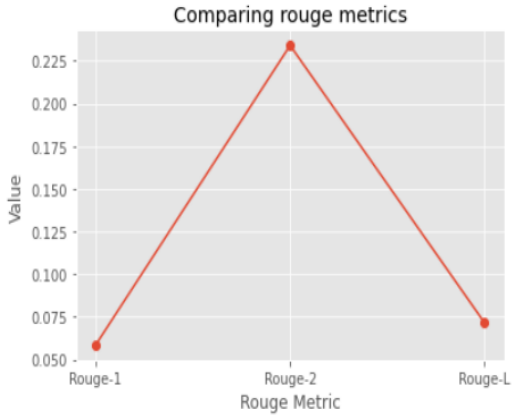

Fig-7

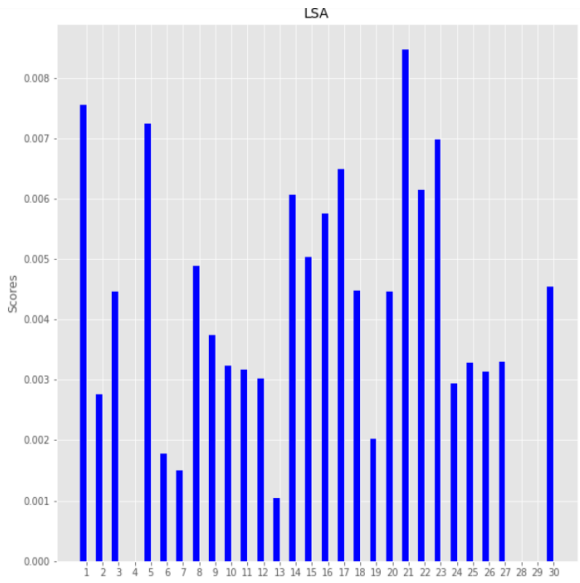

Fig-8

VI. Conclusion and Future Work

In this project, We concentrated on two extractive synopsis methods: LSA and Text Rank which we applied on the Amazon electronics dataset. To find the efficiency of the techniques we used evaluation metric Rouge. In practice, text summary is created by using summaries written by human. We have implemented the code for LSA and textrank and displayed the summaries, ranks of each sentence, precision, recall, F score of each product. We have used packages to produce reference summaries, so the results of our project are on the lower side. As a future work we can compare the summaries obtained from both the techniques and also run the methods on a standardised dataset which would allow to evaluate the NLP techniques in a more efficient way.

References

[1] Y. Gong, X. Liu: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of the 24 th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States 2001, pp. 19-25

[2] M. G. Ozsoy, I. Cicekli and F. N. Alpaslan:Text Summarization of Turkish Texts using Latent Semantic Analysis. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–876, Beijing, August 2010

[3] Josef Steinberger and Karel Ježek. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation(2004), In Proc. ISIM ’04

[4] Mihalcea and P. Tarau. TextRank - bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, 2004.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Lin, Chin-Yew and E.H. Hovy 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27 - June 1, 2003.

[6] https://ieeexplore.ieee.org/document/5197422

[7] Michael Armbrust, Reynold S. Xin, Spark SQL: Relational Data Processing in Spark

[8] “Text Summarization Evaluation – BLEU vs ROUGE”

[9] Praveen Dubey, “Understand Text Summarization and create your own summarizer in python.”

[10] Pranay, Aman and Ayush, “Text Summarization inn Python: Extractive vs Abstractive techniques revisited”.

[11] Sciforce, “Towards Automatic Text Summarization: Extractive Methods”.