Two - Level News Summarization with Sentiment Analysis

[1]Gayathri S [2]Emmanuel Donal [3]Roshan Sabu [4]Aleena Johnson [5]Ms Divya Mohan
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4]Albertian Institute of Science & Technology [5]APJ Abdul Kalam Technological University

Abstract

Text Summarization has always been an area of active interest in the academia. In recent times, even though several techniques have being developed for automatic text summarization, efficiency is still a concern. Given the increase in size and number of documents available online, an efficient automatic news summarizer is the need of the hour. In this paper, we propose a technique of text summarization which focuses on the problem of identifying the most important portions of the text and producing coherent summaries. People tend to read multiple news articles on a topic since a single article may not contain all important information. A summary of all the articles related to topic will save the time and energy. In this research, an extractive based approach is used to generate a two-level summary from online news articles. News topics covered include politics, sports health, science and movie reviews from, etc. The first-level summary generates the summary of each article and second level summary combines the first level summaries and generates the final summary. To understand the variation of these news articles, Sentiment Analysis is applied.
Keyword- Text Summarization, Extraction based Summarization, Sentiment Analysis

I. INTRODUCTION

By the wide exploration of internet, a huge collection of documents are available in the internet. Whenever people searches for a news topic, they are compelled to read all the documents related to their search. This takes large amount of time and it is very tough to summarize manually. There comes the value of automatic news summarization. Text summarization is the process by which it reduces the size of the document without loss of its meaning. Text summarization process done by a machine is entitled as automatic text summarization. Automatic Text Summarization, also known as Text Reduction, is a growing and interesting field in natural language processing. The core benefits of automatic text summarization are minimization of reading time and effort. It is possible that a user will miss some of the important information if he/she reads the articles from multiple news sources. If an automatic text summarizer is available then it will provide all the important information from these articles in one scan of all the documents. Text summarization methods can be categorized by the way it is done. There are multiple types of text summarization techniques [1], and are mentioned below.

A. Abstraction vs. Extraction

Text summarization is classified into extractive summarization and abstractive summarization. Extractive summarization extracts the important sentences or phrases from the original document. These highest ranked sentences are joined together to generate a summary without changing the original document. This application mainly targets journalists and analysts who want precise knowledge about a specific news topic.
Abstractive summarization understands the document and generate a summary close to the human made summary. It uses advanced natural language processing techniques to produce a summary.

B. Mono-lingual vs. Multi-lingual

Mono-lingual summarizer works only on one language, for instance English. In contrast, multilingual works with multiple languages, for instance English, Spanish, Japanese, and Hindi.

C. Single-document vs. Multi-document

Single-document summarizer uses only one document as an input and generate a summary out of it. Multi-document summarizer uses two or more documents of same/similar topic and generate a summary out of them.

D. Generic vs. Query-based

In some cases, user may find the entire information useful rather than a specific information. The summary of such information fit in to generic based summary. If a user needs a Specific information from the original document, then it is known as a query-based summary.

All rights reserved by *www.grdjournals.com*

E. Indicative vs. Informative

Indicative summaries highlight only the important information in few words from the document. These summaries are used to motivate the users to read the entire document. Whereas, Informative summaries can be treated as a replacement for the Original document as they provide concise information. In this research, a novel approach for a two-level text summarization from online news sources using extractive summarization with sentiment analysis is presented. Initially, important sentences from various news articles corresponding to a topic are extracted and summaries are generated individually. To explore further into the individual summaries, sentiment analysis is performed on them for a topic.

II. RELATED WORK

A. Extractive Summarization

The extractive based summarization selects the pivotal information from the original document. Extensive research has been done on extraction based summarization. Krishnaprasad et. al. implemented a single document extractive text summarizer for Malayalam language [1]. They used rank-based approach by providing a score to each word from each sentence based on their importance, selected top N ranked sentences, and generated a summary. The standard metric ROUGE is used to present the results. Feng et. al. built a single document extractive text summarizer by using corpuses of news stories . They used keyword-based approach where they searched using keywords to extract the related news stories about a topic stored in their corpus. Evaluations are done using ROUGE sets. Krzysztof et. al. developed a sentence-based extractive summarization for polish language. They used Term frequency – Inverse document frequency method and Polish news corpus to generate a summary. ROUGE evaluation was used to check the performance of their summaries.

B. Sentiment Analysis

Microblogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these microblogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on microblogs. One challenge is to build technology to detect and summarize an overall sentiment. We look at one such popular microblog called Twitter and build models for classifying "tweets" into positive, negative and neutral sentiment. Agarwal et. al. presented a novel approach by analyzing sentiment of tweets using polarity based approach where tweets were classified into positive, negative or neutral sentiment .They used unigram, feature and tree kernel-based approach for classifying the tweets and achieved 71.35% accuracy using Support Vector Machine (SVM). Mirani and Sasi identified sentiments of ISIS related tweets with their exact locations by using polarity based approach. They used SVM, Random Forest, Bagging, Decision Trees and Maximum Entropy Algorithms and achieved more than 90% average accuracy.In this research, an individual summary is extracted from news articles from multiple sources to achieve the best results. Sentiment analysis is applied on these individual summaries to identify how reliable are these news sources.

C. Abstractive Text Summarization using Sentiment Infusion

Text Summarization is condensing of text such that, redundant data are removed and important information is extracted and represented in the shortest way possible. With the explosion of the abundant data present on social media, it has become important to analyze this text for seeking information and use it for the advantage of various applications and people. From past few years, this task of automatic summarization has stirred the interest among communities of Natural Language Processing and Text Mining, especially when it comes to opinion summarization.[3] Opinions play a pivotal role in decision making in the society. Other's opinions and suggestions are the base for an individual or a company while making decisions. In this paper, we propose a graph based technique that generates summaries of redundant opinions and uses sentiment analysis to combine the statements. The summaries thus generated are abstraction based summaries and are well formed to convey the gist of the text.

III. METHODOLOGY

Extractive summarization is used for finding the important sentences from the news articles. These news articles' topics comprises politics, sports, health, science and movie review. All these prime news topics are considered to analyze and understand the results in this research. There are many research papers available for each news topic. Goal of extractive text summarization is selecting the most relevant sentences of the text. Summarization system consists of 3 major steps, Pre-processing, Extraction of feature terms and algorithm for ranking the sentence based on the optimized feature weights.

All rights reserved by *www.grdjournals.com*

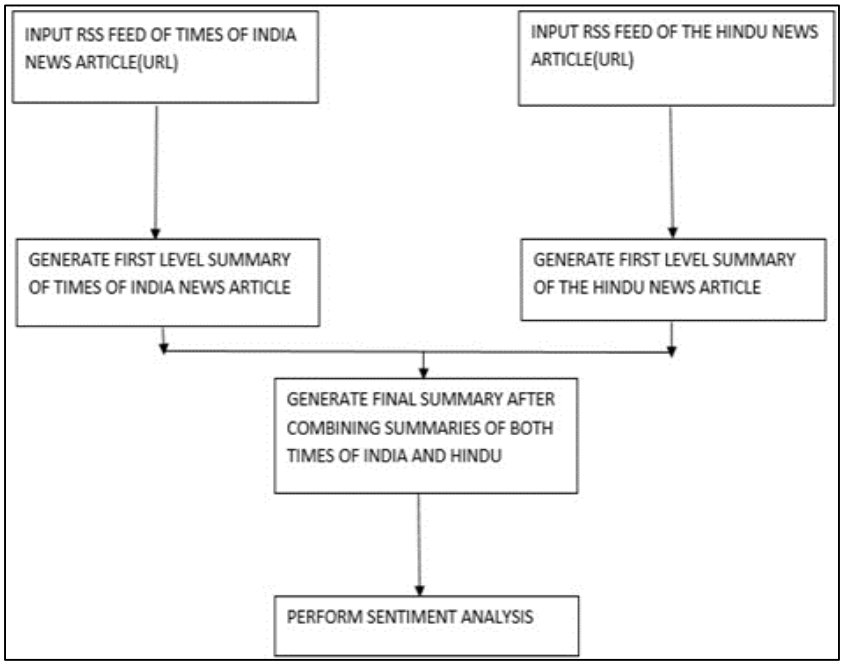

Fig. 1: Architecture diagram of text summarization

The RSS feed (URL) of the news articles from Times of India and The Hindu newspapers are fetched and provided as the input for the first level summary. Here we apply extraction based method which identifies the important sentence and arrange them in the order of their importance. A Second level summary is generated after combining the first level summaries of news articles which was previously generated. In order to identify whether there is any different view point of newspapers on the same news topic sentiment analysis is applied.

A. Pre-Processing

This step involves Sentence segmentation, Sentence tokenization, Stop word Removal.

1) Sentence Segmentation

It is the process of decomposing the given text document into its constituent sentences along with its word count. In English, sentence is segmented by identifying the boundary of sentence which ends with full stop (.) , question mark (?), exclamatory mark(!).

2) Tokenization

It is the process of splitting the sentences into words by identifying the spaces, comma and special symbols between the words. So list of sentences and words are maintained for further processing.

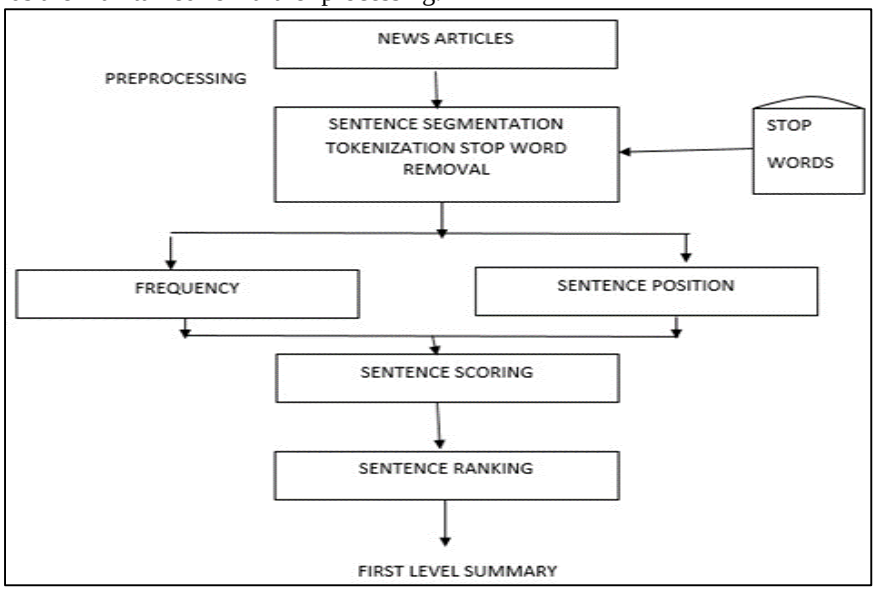

Fig. 2: Steps in Preprocessing

All rights reserved by *www.grdjournals.com*

3) Stop Word Removal

Stop words are common words that carry less important meaning than keywords .This words should be eliminated otherwise sentence containing them can influence summary generated.

B. Feature Extraction

After an input document is tokenized, it is split into a collection of sentences. The sentences are ranked based on two important features: Frequency, Sentence Position value.

1) Frequency

Frequency is the number of times a word occurs in a document. If a word's frequency in a document is high, then it can be said that this word has a significant effect on the content of the document. The total frequency value of a sentence is calculated by sum up the frequency of every word in the document.

2) Sentence Position Value

Position of the sentence in the text, decides its importance. Sentences in the beginning defines the theme of the document whereas end sentences conclude or summarize the document. The positional value of a sentence is computed by assigning the highest value to the first sentence and the lowest value to the last sentence of the document.

C. Sentence Scoring

The final score is a Linear Combination of frequency, Sentence positional value, weights of Cue Words and Similarity with the title of the document.

D. Sentence Ranking

After scoring of each sentence, sentences are arranged in descending order of their score value i.e the sentence whose score value is highest is in top position and the sentence whose score value is lowest is in bottom position.

E. First Level Summary Generation

In this research, the extraction-based method is applied to generate a summary of online news articles. The steps involved in First-level summary is shown in Fig. 3. Single news article may contain multiple paragraphs. These paragraphs are converted to sentences and then from sentences to words.

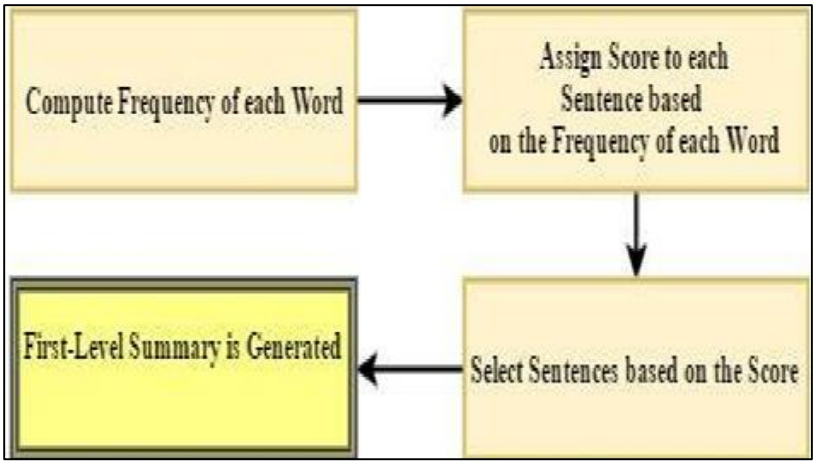

Fig. 3: Process for First-level Summary Generation

Step 1: the frequency of each word is calculated and its threshold is set. In simple terms, how many times each word has appeared in the sentence is calculated. This word frequency is calculated by number of times the word appeared in a sentence divided by total number of words in a sentence. The words that have a frequency lower than the set threshold or higher than the set threshold will be ignored. This is done because some words contain less value, and may not be removed during the stop words process. For example, if words such as "block" or "ban" are used one time or nine times respectively in a sentence then these words will be eliminated.

Step 2: each sentence will be assigned a score based on the frequency of words. This process will be repeated until all the sentences of news article are given a score. In simple words, the score is a priority number that will be allocated to each sentence. Finally, the sentences with a top priority number will be selected and the First-level Summary is generated.

F. Second Level Summary Generation

People are likely to read more than one news article on a similar topic and each article may contain 30 – 40 sentences. This situation takes more time to read multiple news articles. In order to save time and effort, a Second-level Summary will help the user to get

All rights reserved by *www.grdjournals.com*

a better idea of the content provided in various related news articles from different news channels. In this research, a Second-level Summary is generated from all the important sentences of the First-level summaries on a news topic using the same extractive-based approach. Two/ Three First level Summaries on a topic are used to generate the Second-level Summary.

G. Sentiment Analysis

Many researches are available on sentiment analysis of social media; but very less research has been done summarizing news articles on a topic from various newspapers. The polarity of the sentiments can be positive or negative. [4]

IV. SIMULATION

In this research, the open source Java language is used, as it is easy to use and provides a plethora of packages for better statistical analysis and visualization. The first step for generation of summary is to fetch the URLs of news articles and this was done using RSS feed. Following a website's RSS feed gives you the opportunity to stay up-to-date on everything that website publishes. The standard orange RSS logo pictured below is a dead giveaway. Click on it, and it will take you to the website's RSS feed. From there, you can get the URL for feed source settings. Right click on the website's page, and choose Page Source. In the new window that appears, use the "find" feature (Ctrl + F on a PC or Command + F on a Mac), and type in RSS. You'll find the feed's URL between the quotes after href=.The Pre-processing of these news articles is done using regular expression and Tokenization.

Sentiment analysis can be performed by using stanford NLP library which is available in java. A sentiment analysis code gives the sentiment of a sentence based on a dictionary of words which tag words as positive and negative and give them a score between -10 and 10.

Stop word: it is the list of most common words which don't have any sentiments.

V. CONCLUSION AND FUTURE WORK

This paper presents a novel two-level Extraction based News Summarization with Sentiment Analysis. The two-level summary provides only the important content from various related online news articles on a news topic in one place. Sentiment Analysis provides the view of various news channels.

As a future work, combining this Extraction-based method with Abstraction-based method would enhance the news summarization results. Building this research in the form of an Android and IOS applications would greatly help many people who use cell phones to read the news.

REFERENCES

[1] Krishnaprasad, A. Sooryanarayanan and A. Ramanujan, "Malayalam text summarization: An extractive approach," 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, 2016, pp. 1-4.
[2] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R,"Sentiment Analysis of Twitter Data," in Proc of ACL HLT Conf, 2011
[3] ATSSI: ABSTRACTIVE TEXT SUMMARIZATION USING SENTIMENT INFUSION Author Rupal Bhargava Yashvardhan Sharma Gargi Sharma
[4] Sentiment analysis algorithms and applications: A survey Walaa Medhat , Ahmed Hassa ,Hoda Korashy

All rights reserved by *www.grdjournals.com*