International Journal of Knowledge Based Computer Systems
7 (1), June 2019, 19-24
http://www.publishingindia.com/ijkbcs/

Medical Health Posts Summarization Using Lesk Algorithm

Vinod L. Mane[1], Ashwini Abhale[2] and Sumit Khandelwal[3*]
[1]Assistant Professor, IT Dept., DYPCOE Akurdi, Pune, Maharashtra, India. E-mail: Vinod2789@gmail.com
[2]Assistant Professor, IT Dept., DYCOE, Akurdi, Pune, Maharashtra, India. E-mail: Ashwini.abhale@gmail.com
[3]Assistant Professor, Comp. Engg. Dept., MITAOE, Alandi, Pune, Maharashtra, India.
E-mail: Sumit3khandelwal@gmail.com
*Corresponding Author

Abstract: Today's world is all about information, mostly online. With the growth of internet many communication technologies emerged quickly as important information sources, such as emails, forums, social networking sites, etc. It is time and space consuming to handle such large amount of data. Text summarization is technique by which important portion of text are obtained. Traditionally these important portions are selected based on frequency of keywords, position of sentence, style of writing word, keywords in title, etc. Extractive text summarization is produced by concatenating several sentences taken exactly as they appear in the text. Sentences are selected based on some scoring techniques. In our approach, we are using simplified Lesk algorithm with some modification. Our approach is applicable to medical health posts. In this, sentences having important information from medical perspective are arranged in decreasing order of their weights. Based on given input percentage, relevant number of sentences is given as summary. We compared results with human expert summary. The proposed approach gives promising results.

Keywords: Health posts, Lesk algorithm, Summarization, UMLS.

I. Introduction

Summarization is nothing but converting original text into summarized text but without changing its meaning. Summary of particular text is important when content is important and we have very less time to read all the content [14]. In this era, World Wide Web is providing huge amount of information to people. This information is in every form like social, political, technical, economical, medical, etc. To efficiently use such vast amount of information text summarization is important. For many data handling techniques, especially natural language processing applications like information retrieval text summarization is playing significant role in getting important information from large amount of data [1].

There are two types of text summarization, abstractive text summarization and extractive text summarization [2]. Abstractive summaries are generated from author's point of view. Internal semantic representation of the text is built in abstractive summarization. This summarization technique generates a summary which is close to what a human would generate [14]. In this summary, keywords or sentences from original texts may be reused. On the other hand, Extractive summarization is based on selection of important sentences or phrases. Each sentence is evaluated based on predefined function and most important sentences are extracted in original form [14]. Several sentences from original text are used as it is to generate extractive summary.

Traditional approach to generate extractive summary is based on some hand tagged rules [3]. These rules include position of sentence in text, frequency of certain keywords in text, writing style of text, etc. But this approach is not promising as input texts are different.

Consider a scenario for an article about new drug. In this article title will be there also drug name, important methods are written in different format. So generating extractive summary will be easy using traditional approach of hand tagged rules. But if we consider text from forums where no title or different style of writing is there, above approach won't work efficiently. In such case semantic analysis of text helps to generate better results. The proposed approach uses simplified Lesk algorithm [1] with some modification and extracts important relevant sentences from text. To perform semantic analysis UMLS (Unified Medical Language System) [4] is used which gives meaning of keywords in terms of medical language.

The rest of the paper is organized as: Section II is about related work; Section III describes proposed approach; Section IV depicts experimental results along with comparison; Section V represents conclusion of the paper.

II. Related Work

Researchers are working on automated text summarization from around 1950s [15]. With the growth of internet many

communication technologies emerged quickly as important information sources, such as emails, forums, social networking sites, etc. It is time and space consuming to handle such large amount of data. Due to this information overloading researchers nowadays are interested in automatic text summarization.

We have done an overview of the same in our previous work [5], [6], and [7]. Details are as follows:

Alok Pal, and D. Saha did summarization for different categories of texts like writer, soul, sports, etc. They used WordNet dictionary to perform syntactic analysis. Their approach is based on simplified Lesk algorithm. This algorithm along with WordNet dictionary is used to give weightage to each sentence in input text. They sorted the sentence in decreasing order of their weight. When percentage of summarization is provided as input, appropriate summary is generated from sorted sentences. They used precision, recall and f-measure as metrics to evaluate the results of summarization. Their approach doesn't give good results when input text contains more number of named entities. Therefore, better weighting should be done [1].

Rafael *et al.,* used sentence scoring technique to generate extractive text summary. They focused mainly to improve the quality of summarization. To perform summarization, they consider context of a text as their base. For better weighting, they combined word based, sentence based and graph based scoring methods together. They compared their results with summary generated by human experts by counting number of sentences common in both summaries. ROUGE is used as quantitative measure to evaluate summary [8].

Jayashree R *et al.,* used kannadawebdunia which is portal offering news from politics, sports, cinema, etc., to perform summarization. Their approach is based on keywords in text. By combining GSS coefficient and IDF methods along with TF, they extracted important keywords from portal. They gave weight for each keyword based on this combination. Weight of each sentence is calculated by summation of weights of all keywords in a sentence. Finally, when user gives input, most weighted sentences are generated as summary [9].

C. Lakshmi Devasana *et al.,* derived the structure of input text using rule reduction technique and developed a text analyzer. First tokens are created from input text then important features are identified and finally categorization and summarization is done. Rules are generated for noun phrase (NP), prepositional phrase (PP), possessives (POSS), verb phrase (VP). Text analyzer categorizes tokens as per these rules and summarizes them to formulate a sentence [10].

Rashmi Mishra *et al.,* classified summarization methods into categories like statistical, natural language processing, machine learning, and hybrid technique. Their research is concentrated on summarization of biomedical documents. They found more approaches towards generation of extractive summary. They showed that, though challenging, abstractive and graph based summaries are also increasing nowadays [11].

Alan R. Aronson gave metamap algorithm that mapsbiomedical text to concepts in the UMLS Metathesaurus. He describes different stages in which metamap works [12].

Selvani Deepthi Kavila, Radhika Y gave a comparative evaluation of statistical methods in extractive text summarization [14]. To achieve better performance, they used modified weighing method and modified sentence symmetric feature method. They added thematic weight and emphasize weights in conventional weighing method. In this three different algorithms for summarization are implemented and the performance is observed.

After comparison with different methods they show that modified weighing method is the best method with 80% efficiency. But main drawback of their work is this work is limited with extractive summarization only.

D. Y. Sakhare, Dr. Raj Kumar presented a hybrid technique of text summarization with the combination of dependency grammar and the sentence features [15]. They used DUC 2002 dataset for various compression ratios. They achieved F-measure of 80% for the compression ratio 50%.

Ahmad Ashari, Mardhani Riasetiawan used TextRank algorithms and Semantic Networks and Corpus Statistics to implement document summarizer system [16]. The TextRank algorithm includes various processes, namely tokenization sentence, creation of a graph, the edge value calculation algorithms using Semantic Networks and Corpus Statistics, vertex value calculation, sorting vertex value, and the creation of a summary. They calculated precision, recall, F-score values of summary and tested quality of output using ROUGE-N methods.

Murali Krishna V. V. Ravinuthala, Satyananda Reddy Chproposed a technique called thematic text graph which is nothing but directed as well as weighted graph representation of unstructured text [17]. Important keywords in the document are vertices of graph and edges are drawn between vertices based on theme of document. They carried out their experiment on standard data sets SemEval-2010 and DUC 2002 data sets. Their proposed keyword weighting model is more effective and generate better extractive summaries.

III. Proposed Work

We are focusing on text obtained from health forum posts. These posts are summarized according to given percentage of summarization. We have given weights to sentences based on their contents. If sentences contain name of drug, disease or symptoms such sentences are having higher weights than others. E.g. The sentence "I am taking xanax since four months for my anxiety" is having higher weight than the sentence "I've been down this road before, but this pain is worst". The earlier sentence contains words 'xanax' and 'anxiety' which is having meaning 'drug' and 'disease' respectively. While

the second sentence contains the word 'pain', this is having meaning 'symptom'. Thus earlier sentence is given more weight than second sentence. To identify name of drug, disease or symptom we used dictionary obtained from UMLS (Unified Medical Language System). This dictionary informs whether a particular word is drug, disease or symptom. Once the weight of each sentence is obtained, they are arranged in decreasing order of their weights. Then, as per given percentage number of sentences from original text are displayed as summary. In order to preserve flow, the sentences are displayed in original sequence as they appear in the text. The original Lesk algorithm [1] has been modified for our application as shown in Fig. 1.

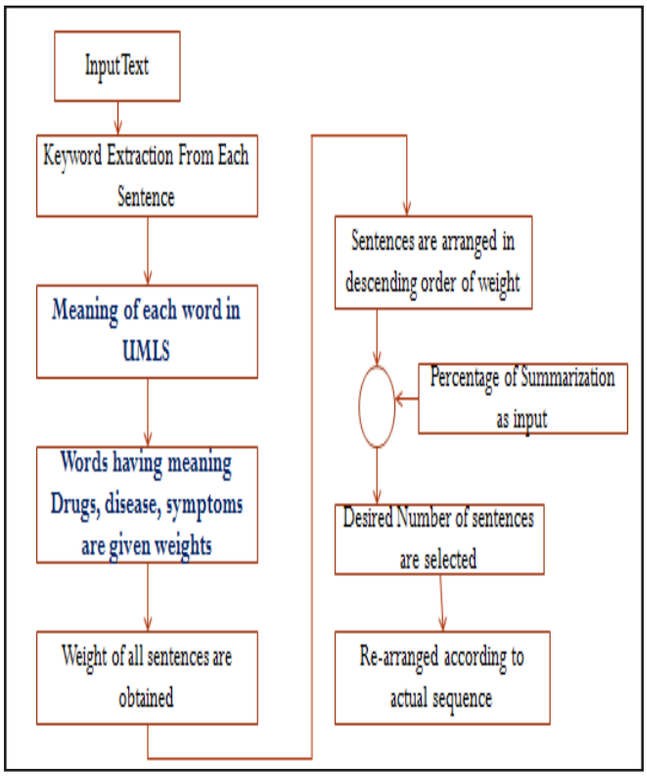

Fig. 1: Modified Lesk Algorithm

IV. Experimental Results

The objective of our proposed algorithm is to improve the quality of text summarization which contains medical keywords i.e., keywords having meaning in medical terminology e.g., drug, disease, symptom, etc., To test our objective, summary generation process has been modeled based on the weights of the keywords obtained using Modified Lesk Algorithm. This section describes the experiments carried out to test the accuracy of keyword weighting based on modified Lesk algorithm and the quality of generated summaries. Our System is implemented using java programming language. We used UMLS (Unified Medical Language System) dictionary to identify the meaning of keyword. Our dataset for experimentation is obtained from healthborads.com [18] which is social networking platform to discuss about health related issues.

The algorithm is tested on different input posts.

We have done rigorous experimentation by considering different input files having a varying number of health posts from different users. Details are described below:

We began experimentation by considering an input file having 500 posts, then we considered 1000 posts, later on we progressed with 1500, 2000 and 2500 posts respectively. Each post is nothing but something written by each user for his/her disease or drug. When we are considering 500 posts, it means 500 users have written their view on particular drug/disease.

Also, we varied the percentage of summarization each time for the inputs from 25%, 50%, 75% and 85% respectively. Results obtained for 500, 1000 and 2500 posts are as illustrated in Fig. 2, 3 and 4.

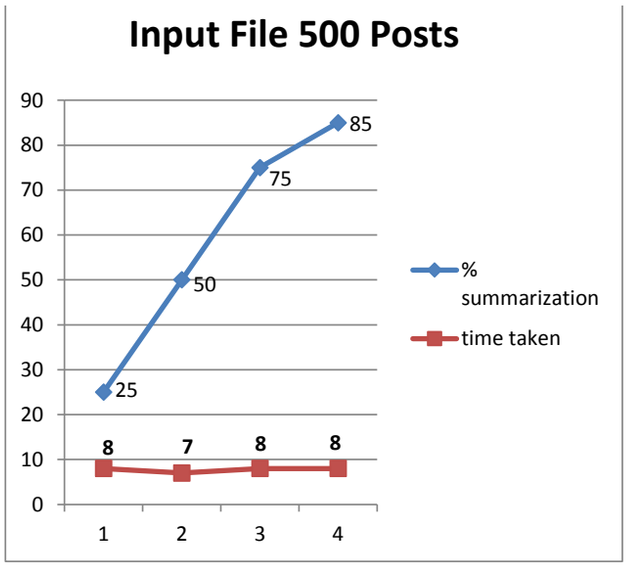

Fig. 2: Summary Percentage Vs. Time Taken (500 Posts)

Fig. 2 shows time (in seconds) taken to generate summary along with percentage. When we give 50% that means output will be 250 posts out of 500.

Same experiment is carried out with different number of posts.

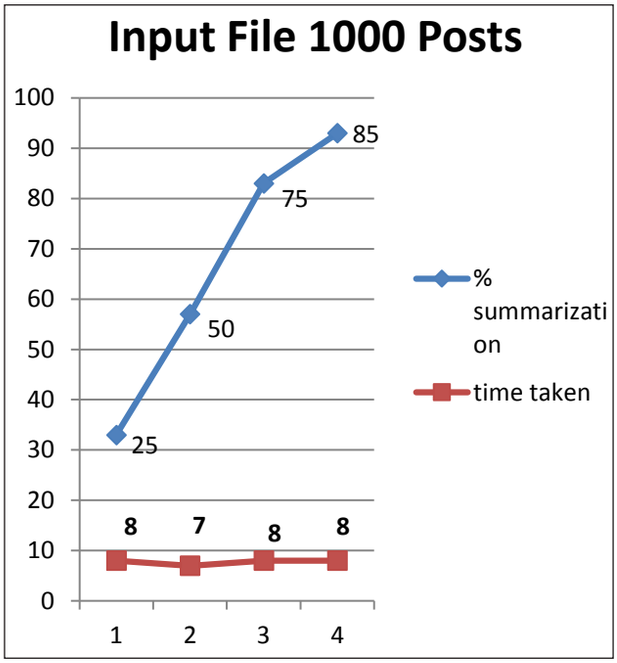

Fig. 3: Summary Percentage Vs. Time Taken (1000 Posts)

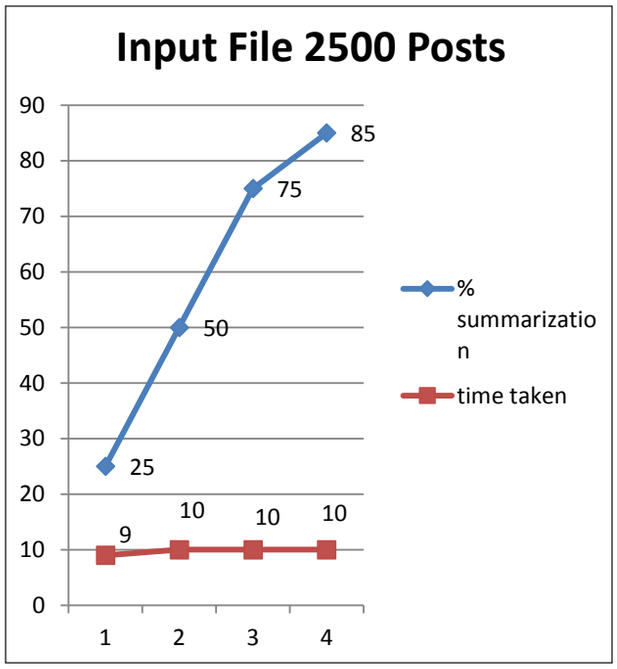

Fig. 4: Summary Percentage Vs. Time Taken (2500 Posts)

Our results illustrate that for different percentage summarization that is 25%, 50%, 75% and 85% the processing time taken is the same for 500 posts as well as 1000 posts, and it is maximum 8 sec. From 1000 posts when we increased to 2500 posts processing time required is still optimum and is maximum 10 sec.

Evaluation of our proposed system has been done by comparing system generated summaries with those generated by 2 human experts. We have given some random number posts to human expert and asked them to generate summary of those posts. And same posts we used as input for our system. We compared both the results and we found that our system give almost 80% similar result as that of human expert's summary.

Posts about drug, disease are given as input for summarization. Same input is given to human experts to generate summaries. Parameters considered for comparison are precision, recall and f-measure [13]. Results obtained are promising.

TABLE I: HUMAN EXPERT 1 VS. SYSTEM SUMMARY

Posts	Correct	Missed	Wrong	Precision	Recall	F-measure
Post1	5	1	1	0.833	0.833	0.833
Post2	7	0	1	1.0	0.87	0.933
Post3	8	2	4	0.8	0.66	0.72
Post4	3	1	1	0.75	0.75	0.75
Post5	7	1	0	0.875	1.0	0.933

Table I shows sample result generated for 5 different posts after comparing with expert summary 1. Summarization percentage chosen was 50%. Here, post 1 contains 14 sentences and summarization percentage is 50. As shown in above Table, 5 sentences were present in summaries of human as well as system (correct), 1 sentence is not present in system summary (missed) and 1 sentence is present in system summary but it should not be not there (wrong).

Fig. 5 shows precision and recall values while Fig. 6 shows F-measure values when compared with expert summary 1.

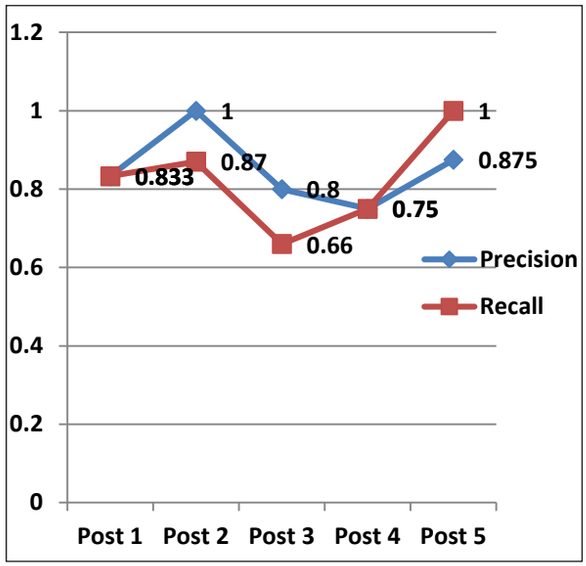

Fig. 5: Comparison with Human Expert Summary 1

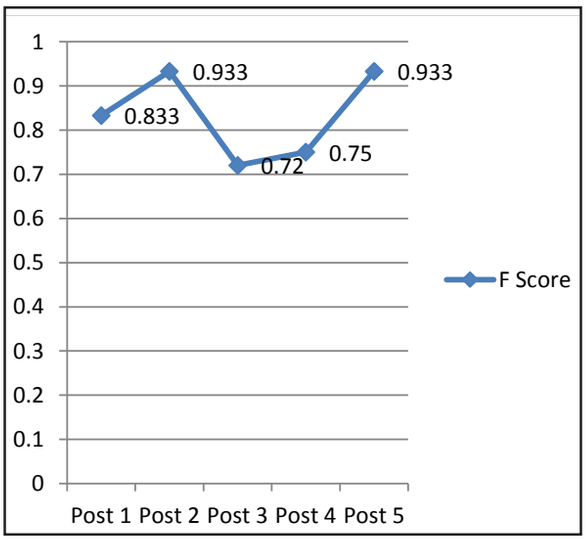

Fig. 6: F-measure for Comparison with Human Expert Summary 1

TABLE II: HUMAN EXPERT 2 VS SYSTEM SUMMARY

Posts	Correct	Missed	Wrong	Precision	Recall	F-measure
Post1	5	1	1	0.833	0.833	0.833
Post2	6	1	1	0.85	0.85	0.85
Post3	9	1	2	0.9	0.81	0.85
Post4	2	2	2	0.5	0.5	0.5
Post5	5	3	3	0.625	0.625	0.625

Table II shows sample result generated for 5 different posts after comparing with expert summary 2. Summarization percentage chosen was 50%.

Fig. 7 shows precision and recall values while Fig. 8 shows F-measure values when compared with expert summary 2.

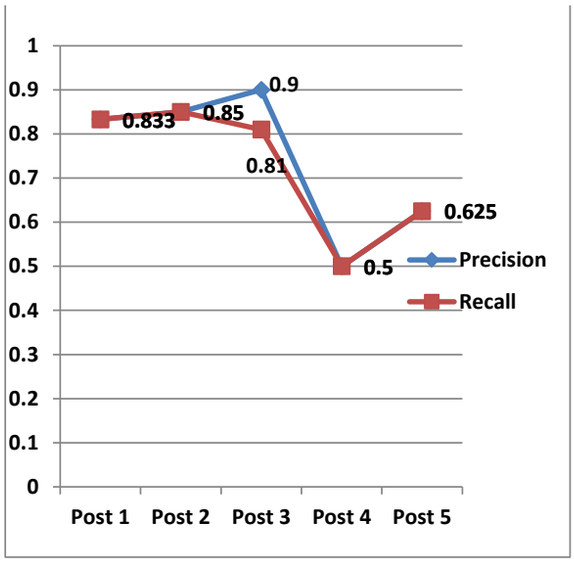

Fig. 7: Comparison with Human Expert Summary 2

When we compared all the posts with varying number of percentage, we found that our system gives good results with 80% similarity to that of human generated summaries.

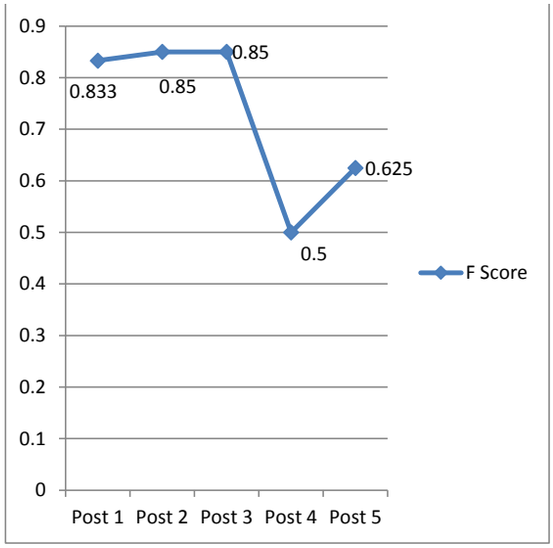

Fig. 8: F-measure for Comparison with Human Expert Summary 2

V. CONCLUSION AND FUTURE SCOPE

From the high values of F-score obtained in our experimentation we conclude that our approach is a promising one towards generated health summaries.

This approach can be used to summarize documents from any field. To extract semantic information from a sentence, only a semantic dictionary is needed. Using dictionary of economic keywords this approach can be used to summarize economical documents, news. Also, this approach helps to categorize text into different categories like sports, education, medical, economic, political, etc. We didn't consider spelling mistakes, typo errors in posts. Also, some authors are asked questions in their posts, from summary point of view these may not be necessary.

REFERENCES

1. A. R. Pal, and D. Saha, "An approach to automatic text summarization using wordnet," in *IEEE Int. Advance Computing Conference (IACC),* Gurgaon, India, 2014.
2. E. Lloret, and M. Palomar, "Text summarization in progress: A literature review," *Artificial Intelligence Review*, vol. 37, no. 1, pp. 1-41, January 2012.
3. C. Y. Lin, and E. Hovy, "Identify topics by position," in *Proc. of the 5th Conf. on Applied Natural Language Processing*, Washington DC, pp. 283-290, 1997.

4. "Unified medical language system (umls)," Nlm.nih.gov, [Online]. Available: www.nlm.nih.gov/research/umls/
5. V. L. Mane, S. S. Panicker, and V. B. Patil, "Knowledge discovery from user health posts," in *IEEE 9th Int. Conf. on Intelligent Systems and Control (ISCO),* Coimbatore, India, 2014.
6. V. L. Mane, S. S. Panicker, and V. B. Patil, "Summarization and sentiment analysis from user health posts," in *IEEE Int. Conf. on Pervasive Computing (ICPC),* Pune, India, 2015.
7. V. L. Mane, S. S. Panicker, and V. B. Patil, "Knowledge discovery from various algorithms: A survey," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 7477-7479, 2014.
8. R. Ferreira, F. Freitas, L. de S. Cabral, R. D. Lins, R. Lima, G. Franca, S. J. Simske, and L. Favaro, "A context based text summarization system," in *IEEE 11th IAPR Int. Workshop on Document Analysis Systems*, pp. 66-70, 2014.
9. R. Jayashree, K. S. Murthy, and B. S. Anami, "Categorized text document summarization in the Kannada language by sentence ranking," in *12th Int. Conf. on Intelligent Systems Design and Applications (ISDA),* IEEE, Kochi, India, 2012.
10. C. L. Devasena, and M. Hemalatha, "Automatic text categorization and summarization using rule reduction," in *IEEE Int. Conf. on Advances in Engineering, Science and Management (ICAESM-2012)*, pp. 594-598, IEEE, 2012.
11. R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. D. Fiol, "Text summarization in the biomedical domain: A systematic review of recent research," *Journal of Biomedical Informatics*, Elsevier, vol. 52, pp. 457-467, 2014.
12. A. R. Aronson, "Effective mapping of biomedical text to the UMLS metathesaurus: The metamap program," in *Proc. of AMIA Symp.*, pp. 17-21, 2001.
13. J. Han, M. Kamber, and J. Pie, *Data Mining Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers, 2012.
14. S. D. Kavila, and Y. Radhika, "Extractive text summarization using modified weighing and sentence symmetric feature methods," *International Journal of Modern Education and Computer Science*, vol. 7, no. 10, pp. 33-39, 2015, doi: 10.5815/ijmecs.2015.10.05.
15. D. Y. Sakhare, and R. Kumar, "Syntactic and sentence feature based hybrid approach for text summarization," *International Journal of Information Technology and Computer Science*, vol. 3, pp. 38-46, 2014, doi: 10.5815/ijitcs.2014.03.05.
16. A. Ashari, and M. Riasetiawan, "Document summarization using textrank and semantic network," *International Journal of Intelligent Systems and Applications(IJISA)*, vol. 9, no. 11, pp. 26-33, 2017, doi: 10.5815/ijisa.2017.11.04.
17. M. K. V. V. Ravinuthala, and S. Reddy Ch., "Thematic text graph: A text representation technique for keyword weighting in extractive summarization system," *International Journal of Information Engineering and Electronic Business(IJIEEB),* vol. 8, no. 4, pp. 18-25, 2016, doi: 10.5815/ijieeb.2016.04.03.
18. "Ambien discussions (experiences, side effects, dosages, etc...)," Healthboards.com, [Online]. Available: https://www.healthboards.com/drugtalk/ambien/index.htm