Movie Review Summarization and Sentiment Analysis using RapidMiner

Alaa F. Alsaqer and Sreela Sasi
Computer and Information Science Department
Gannon University
Erie, PA, United States
afalsaqer@windowslive.com, sasi001@gannon.edu

***Abstract*— Automatic text summarization is one of the important challenges of natural language tasks. It will help the readers save time to get the important information from a lengthy document automatically. Sentiment Analysis is the process of identifying and categorizing opinions expressed in a piece of text computationally to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral. This research focuses on improving the summarization accuracy with sentiment analysis of movie review posts using RapidMiner operators. The first model of summarization is built using the Aylien Text Analysis extension. The proposed second model is built using the Text Processing extension. For both these methods the sentiment analysis is done using the same Aylien Text Analysis extension for evaluating the summarization results. An accuracy of 90% is achieved for sentiment analysis using the first model and 96% for the second model.**

Keywords*— *Text Mining; Multi-Document Summarization; Automatic Summarization; Sentiment Analysis; Opinion Mining;

I. INTRODUCTION

Text Mining, also called text analysis, is a process of extracting meaningful information from a text document by using text-processing techniques [1]. It plays a significant role in a variety of applications such as language detection, sentiment analysis, spam filtering, hand writing analysis, text summarization, categorization, and classification [2]. Although, it has been evolving rapidly there are still some more challenges and hence needs more research. The different metric used for measuring summarization are efficiency based on time, accuracy, the number or the size of documents, and other features.

Text documents are considered as unstructured data that require pre-processing of the text before applying the sentiment analysis and summarization processes on it. This will help end-users or business managers to make appropriate decisions for purchasing new products or getting new services. There are many reviews about products and services available online. Therefore, collecting and analyzing the huge collection of reviews from the web require the use of text summarization and sentiment analysis techniques. There are three steps for applying text summarization or sentiment analysis. This includes retrieving the html documents from the web, convert them to text, and extract textual features. Then summarize this dataset into a limited number of sentences. The summary can be verified by comparing them with the sentiment analysis obtained on these original documents.

The remaining content of this research is organized as follows: In Section II, the Related Research in literature are presented. In Section III, the Methodology for text summarization and sentiment analysis are given. Section IV presents the Simulation and Results. This is followed by Conclusion in Section V, and Future Work in Section VI.

II. RELATED RESEARCH

A. Summarization

Text Summarization is an important challenge of natural language understanding. It concentrates on reducing the size of the text document to a fewer number of important sentences. The aim is to produce a condensed representation of an input text that captures the core meaning of the original document.

The research review paper in [3] presented three major types of automatic text summarization systems based on the output format of summary, the input as number of documents to be summarized, and the purpose of summary. The three types of text summarization mentioned in this review are abstraction/extraction based summary, the single document/multi-document summary, and the summary based on generic/query-based approaches.

Text summarization could be done by using various methods and features. Textual features include term frequency, location, cue method, title/headline word, sentence length, similarity, proper noun, and proximity. These features play a significant role in selecting key sentences to create a summary. Extractive text summarization techniques including Term Frequency method, Cluster based method, Graph Theoretic approach, Machine Learning approach, LSA method, Text summarization with Neural Networks, Automatic text summarization based on fuzzy logic, and Query based extractive text summarization are discussed in this research review [4].

Research in [5] used unsupervised method to accomplish text summarization automatically by applying deep learning on Bilingual documents. The Bilingual documents include two languages English and Hindi. Their model used deep learning

978-1-5090-6590-5/17/$31.00 ©2017 IEEE

to explore the textual features of sentences to obtain a summary based on sentences.

B. Extractive Vs. Abstractive

The approaches for summarization of documents can be divided into extractive or abstractive at a high-level standpoint. Most successful summarization systems utilize extractive approaches in which a few sentences from the original text are selected to produce a condensed version. In contrast, abstractive summarization attempts to produce a new text summary in which the original text does not appear as a part of the summary [4]. In Extractive Summarization, the sentences extracted from the original documents are coherently combined to get a shorter version of the original text without substantially changing the meaning. In Abstractive Summarization, new sentences are generated by analyzing the semantic content of the original documents using natural language generation techniques. Both types of summarization have been extensively studied for factual and edited text documents. The evaluative text summarization has focused on sentiment analysis and information extraction most of the times [6]. A recent research presented an abstractive summarization approach by using a model that consists of Wordnet dictionary and Simplified Lesk algorithm to obtain a sentence-level summarization [7].

Research in [8] utilized the Fuzzy Logic approach with WordNet to do extractive summarization based on semantic features. They applied a combination of three summarization methods to obtain three different summaries. These are used to generate the final summary to get better results. The three methods used to accomplish that task are fuzzy logic based method, bushy path method, and wordnet synonyms method.

C. Single document Vs. Multi-documents

Text summarization could be done on single document to reduce the size of the text content by generating a short text summary. In addition, text summarization could be done for multiple documents to obtain a summary of each document or all the documents in a single summary. The research in [7] used single document summarization. Research in [8] used a multi-document summarization.

D. Generic Vs. Query-based approach

The text summarization could be categorized based on the purpose of summarization to generic approach or query-based approach. Generic approach focused on summarized text document irrespective of its topic. In query-based approach, the summary of documents on a topic or a query from a user using the keywords on the same domain is done [9].

E. Sentiment Analysis

Sentiment analysis is considered as opinion mining. Since there is a large collection of opinions on the web, some form of summary of opinions is needed [10]. The research in sentiment analysis has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In fact, there are many different techniques available for finding sentiment analysis including Support Victor Machine (SVM), Naïve Bayes, and Artificial Neural Network (ANN) [11].

In addition, another research presented a comprehensive study of Text mining approaches [2]. This study includes summarization, information extraction, categorization, visualization, clustering, sentiment analysis, and many other topics and methods of text mining.

F. Extrinsic Vs. Intrinsic evaluation approach

Text summarization evaluation was divided into two categories the extrinsic and intrinsic. In the extrinsic evaluation, the summary will be checked based on how it will affect on achieving tasks such as relevance assessment, reading comprehension, etc. The intrinsic evaluation, evaluate a summary based on itself like focusing on its information [12].

As text summarization is based on sentence extraction, there is a possibility of selecting different content at different times from the same source document. Because of this situation, researches in [7] and [8] performed a quantitative evaluation approach by using human-made summaries to evaluate the results obtained from automatic summarization. However, people are subjective because a summary of particular text document could be deferred from one person to another. Also, a particular person could make different summary in different times from the same content of text. Therefore, qualitative evaluation approach may be needed as this quantitative evaluation approach is not the only way of evaluating summarization. That could be done by getting a sentiment analysis of a text document automatically and then comparing it with the results generated by automatic summarization technique.

The current research focuses on improving the performance of mining the review posts using text summarization and sentiment analysis. An extractive summarization technique is applied on multiple documents in this research. The sentences obtained from summarization are verified with the sentiment analysis result done on the original document. The Architecture model is explained in Section III.

III. THE ARCHITECTURE MODEL

The Architecture Model for Multi-Document Summarization and Sentiment Analysis is shown in Fig. 1. The html files are retrieved from the web and used it as data input. The pre-processing consists of converting the unstructured data to structured data, changing the upper cases to lower cases, and getting the roots of words using English dictionary. Then the sentiment analysis and summarization models will be applied in parallel. In the sentiment analysis phase, the opinion of each review will be determined. In the summarization phase, the text of each review will be summarized to a limited number of

sentences. Finally, both results will be used for evaluation by using the extrinsic evaluation approach. The Simulation results are explained in Section IV.

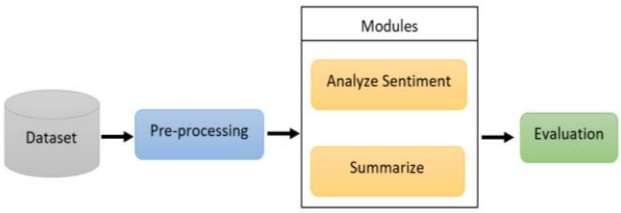


Fig. 1. Architecture of Multi-Document Summarization and Sentiment Analysis

IV. SIMULATION AND RESULTS

Simulation is done using Aylien extension and Text processes extension provided by RapidMiner. These extensions are capable of dealing with multiple documents. One hundred documents are used for simulation. After applying summarization and sentiment analysis approaches on text documents, the accuracy of performance has been measured.

A. Data Component

The dataset used for simulation is the 'Sentiment polarity datasets v2.0' [13]. It includes 27889 collections of movie-review documents, which are unprocessed html files. Each file contains more than 26 sentences of review posts. This dataset is also useful for training and testing the sentiment analysis because according to the dataset website, it has movie-review documents labeled with respect to their overall sentiment polarity (positive or negative). The current research implemented 100 text files from the same dataset for summarization and for sentiment analysis.

B. Implementation Component

RapidMiner version 7.3 with Aylien text analysis and Text processing extensions is used for implementation. It includes various operators that could be applied on different types of data including text documents. It has two views named as 'design' and 'result'. In the design view, it allows the user to build a model visually by dragging and dropping operators into the workspace and connect them to each other. Also, it lets the user to execute his own code by using a specific extension. In the result view, there are advanced ways to display the results using charts and statistics table with useful details [14], [15]. In addition, Rapid Miner is one of the big data analytics tools available [16] and [17].

C. Model Component

The implementation consists of two phases. The first phase is for reading the html documents and applying sentiment analysis process on these documents. The second phase focuses on the summarization process in order to get the five (by choice) key sentences out of each document. This was found by repeated trials.

In the first phase, there are three elements for the model. As a main process, the 'Process Documents From Files' operator is used to load all the text documents from the folder by editing its parameters for setting the class name and direction for each category. Inside the 'Process Documents From Files' operator, the 'Transform Cases' operator is added to transform letters from capital to small. Then the 'stem' operator is used to make words return to its root. These processes are shown in Fig. 2 and 3.

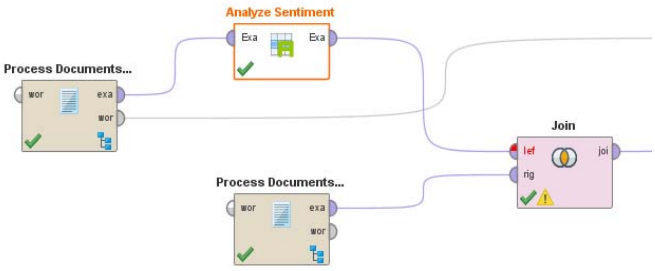


Fig. 2. Main process of 'Process Documents from Files'

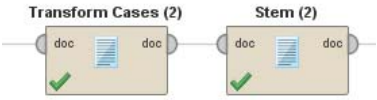


Fig. 3. The sub-processes of the first phase

The output of the 'Process Documents from Files' operator is an example set containing the requested feature values. Then the 'Analyze Sentiment' operator is used to analyze the semantic of the text. The result from this phase is presented in Fig. 4. The output from this model of sentiment analysis is verified with the output from the text summarization model.

Row No.	label	polarity_confidence	subjectivity_...	text	metadata_file	metadata_p...	metadata_d...
1	M	0.996	0	review for hitc...	0002.html	C:\Users\Ala...	Oct 11, 2001 ...
2	M	0.987	0	review for bey...	0003.html	C:\Users\Ala...	Oct 11, 2001 ...
3	M	0.718	0	review for an...	0004.html	C:\Users\Ala...	Oct 11, 2001 ...
4	M	0.995	0	review for an...	0005.html	C:\Users\Ala...	Oct 11, 2001 ...
5	M	0.987	0	review for rad...	0006.html	C:\Users\Ala...	Oct 11, 2001 ...
6	M	0.985	0	review for nig...	0007.html	C:\Users\Ala...	Oct 11, 2001 ...
7	M	0.990	0	review for nig...	0008.html	C:\Users\Ala...	Oct 11, 2001 ...
8	M	0.985	0	review for leth...	0009.html	C:\Users\Ala...	Oct 11, 2001 ...
9	M	0.978	0	review for leth...	0010.html	C:\Users\Ala...	Oct 11, 2001 ...

Fig. 4. The results from the first phase

The main process of reading text documents and the sub-processes including text summarization are shown in Fig. 5. The main process of the second phase includes the 'Process Documents from Files' operator. This contains sub-processes

for pre-processing using 'Transform Cases' and 'Stem' operators, and the 'Summarize' operator for summarization.

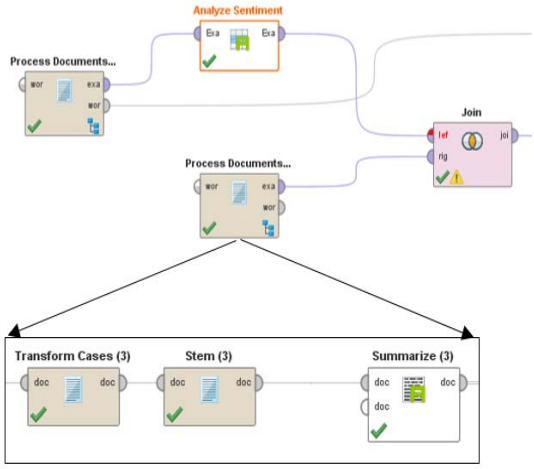


Fig. 3. The main processes and sub-processes of the second phase

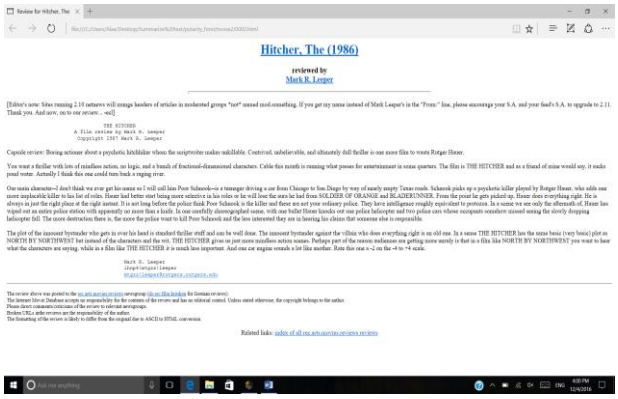
In the second phase, the summarization model is applied to extract key sentences from the documents. The text documents are read by using 'Process Document from Files' operator, followed by the 'Summarize' operator. In this summarization process, a number of key sentences will be extracted from the documents based on extractive summarization. During this stage, user can determine the number of sentences that shall be presented in the result. The number of sentences could be one sentence to ten sentences. By repeating this experiment, it is found that five sentences will give better accuracy for this data set. The result is shown in Fig. 6.

sentence_1	sentence_0	sentence_2	sentence_4	sentence_3
You want a th...	Review for Hit...	Schnook pick...	Mark R. Leeper	It is not long ...
Bruce (Jeff G...	Review for Be...	A series of bi...	Unless state...	When the fina...
We saw the fi...	Review for An...	Unfortunately,...	Please direct ...	I hope that if t...
His client is L...	Review for An...	ANGEL HEA...	Mark R. Leeper	I doubt that R...
Finally, Allen ...	Review for R...	While HANNA...	Broken URLs...	It's funny, in r...
End of bad n...	Review for Ni...	I was also im...	Related links:...	Bad acting, st...

Fig. 6. The results from the second phase

The user needs to read only five sentences instead of reading the whole document. All the reviews had more than 25 sentences in the current dataset. The summarization process reduced the text document down to an average of 86%. This helps the user to save time by reading only the important information from a long document and to determine whether the review is positive or negative or neutral.

Fig. 7 shows the original text document for review before and after summarization. The original document is in html format. It contains more than 25 sentences. The review text also includes the information related



sentence_1	sentence_0	sentence_2	sentence_4	sentence_3
You want a th...	Review for Hit...	Schnook pick...	Mark R. Leeper	It is not long ...

Fig. 7. The original text document of review before and after summarization

to a reviewer, date of review, and the movie title, etc. The result after summarization consists of only top five sentences.

Fig. 8 shows the final process for combining both the results. Fig. 9 presents the results with the polarity for the summarized sentences. Fig. 10 shows two example documents used for summarization from column 'sentence_3'. It is clear that there is a correspondence between the summarization results and the polarity obtained from sentiment analysis.

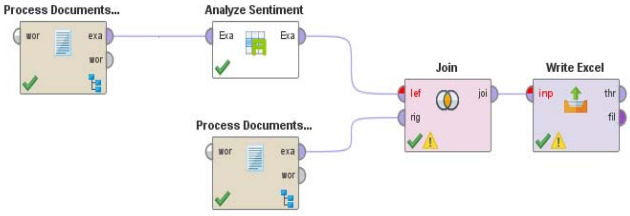


Fig. 8. The final step using Join process

polarity	subjectivity	sentence_1	sentence_0	sentence_2	sentence_4	sentence_3
positive	unknown	You want a th...	Review for Hit...	Schnook pick...	Mark R. Leeper	It is not long ...
negative	unknown	Bruce (Jeff G...	Review for Be...	A series of bi...	Unless state...	When the fina...
positive	unknown	We saw the fi...	Review for An...	Unfortunately,...	Please direct ...	I hope that if t...
positive	unknown	His client is L...	Review for An...	ANGEL HEA...	Mark R. Leeper	I doubt that R...
positive	unknown	Finally, Allen ...	Review for R...	While HANNA...	Broken URLs...	It's funny, in r...
negative	unknown	End of bad n...	Review for Ni...	I was also im...	Related links:...	Bad acting, st...

Sentiment analysis results + summarization results

Fig. 9. The final results after using Join process.

polarity	sentence_3
positive	It's funny, in retrospect, that Orson Welles name comes up so often in the
negative	Bad acting, stilted dialogue,

Fig. 10. Examples of results after joining both example sets results.

For instance, the sentiment analysis is negative when the text includes some negative words such as 'Bad acting, stilted dialog.' It will be positive if the text includes some positive words such as 'It's funny'.

In Fig. 11, the accuracy of sentiment analysis is presented using the confusion matrix for sentiment analysis. It showed an accuracy of performance as 70.94%. The polarity of sentences in the original document is obtained using the 'Analyze Sentiment' operator. This is compared with the polarity of sentences in the original document compiled manually. This is done to check the capability of the 'Analyze Sentiment' operator provided by RapidMiner.

confusion matrix	true.positive	true.negative	class percision
pred.positive	56	17	76.71232877
pred.negative	17	27	61.36363636
class recall	76.7123288	61.36363636	
Accuracy	70.9401709	70.94%	

Fig. 11. Confusion matrix for sentiment analysis for accuracy

Fig. 12 presents the results of sentiment analysis using Pie chart. It shows the percentage of positive, negative and neutral views as 55%, 44%, 1% respectively.

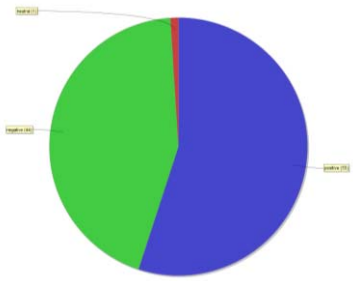


Fig. 12. Pie chart showing the percentage of positive, negative and neutral views

A new model is built using RapidMiner environment and is shown in Fig. 13. The summary is done by ranking sentences with weight values. The same dataset of text documents used for the first model is taken as an input. In the pre-processing step, every text document is tokenized to sentences, cases are transformed to lower case, and stop words are removed. After pre-processing, TF-IDF value for each token (sentence) is calculated. Then, tokens are ranked by using the 'Prune' operator in RapidMiner. The token weights are sorted in descending order by using the 'Principle Component Analysis (PCA)' operator. The top 5 sentences are selected and displayed as output. The summarized sentences have the same polarity as the original movie reviews. These

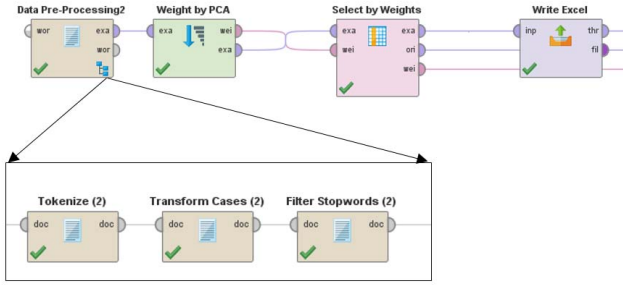


Fig. 13. The main and sub-processes of the new Summarizing model

steps are presented on Fig. 14.

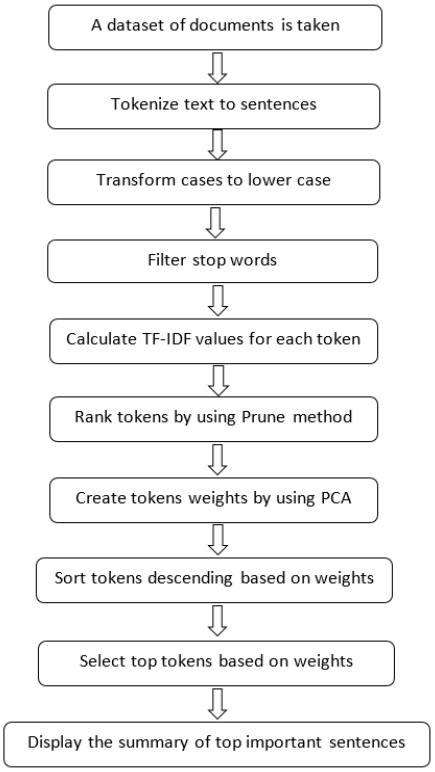


Fig. 14. The diagram of the Summarizing model

Table 1 shows the implementation results of the summarization models based on sentiment analysis on text documents. The documents set consists of 100 text document in html format. The new model has an accuracy of 96% compared with 90% for the Aylien summarization model provided by the RapidMiner. This result clearly indicates that this new approach of summarization model using TF-IDF,

'Prune' and 'PCA' operators is a better approach for summarization.

TABLE I.

COMPARISON OF ACCURACY (SUMMARIZATION MODELS)

The Performance of Summarization Models		
No.	***Summarization model***	***Accuracy (SA)***
1	Summarization (Current method)	96%
2	Summarization (Aylien method)	90%

Additionally, the current research results achieved better accuracy compared with the results of research in [5] and [7]. In [5] the summarization model using Deep learning showed 85% accurate results. Also, the summarization approach in [7] used Wordnet dictionary and Simplified Lesk algorithm, and tested on 50 different types of text documents from different five categories with 80% accuracy. By comparing their results with the current research, their research focused only on summarizing one document every time but the current research focuses on summarizing multiple documents at the same time. The reduction in terms of number of sentences is 86% of the original text and the accuracy is 96%. The current approach tested on 100 different text documents from various categories for more than 20 movies.

V. CONCLUSION

This research presented a method for multiple document summarization and sentiment analysis using RapidMiner. The first model of summarization is built using the Aylien Text Analysis extension. The proposed second model is built using the Text Processing extension. For both these methods, the sentiment analysis is done using the same Aylien Text Analysis extension for evaluating the summarization results. An accuracy of 90% is achieved for sentiment analysis using the first model and 96% for the second model.

VI. FUTURE WORK

For future work, this model could be applied to other types of text documents such as data from Twitter or comments posted on Facebook about products, services, etc. This may help end-users to make right decisions or companies to improve their services.

ACKNOWLEDGMENT

The author, Alaa Alsaqer, would like to express her deepest gratitude to the Ministry of Higher Education at Saudi Arabia represented by SACM in Washington DC, the United States, and King Faisal University in Alhasa, Saudi Arabia for supporting.

REFERENCES

[1] Kotu, Vijay, and Bala Deshpande. *Predictive Analytics And Data Mining: Concepts And Practice With Rapidminer*. 1st ed. Morgan Kaufmann Publishers, 2015. Print.

[2] Kaushik and S. Naithani, "A Comprehensive Study of Text Mining Approach," in *IJCSNS*, 2016. [Online]. Available: http://paper.ijcsns.org/07_book/201602/20160212.pdf. Accessed: Dec. 11, 2016.

[3] N. Bhatia and A. Jaiswal, "Automatic text summarization and it's methods - a review," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 65-72.

[4] D. Gaikwad and C. Mahender, "A Review Paper on Text Summarization", *in International Journal of Advanced Research in Computer and Communication Engineering*, IJARCCE, 2016.

[5] S. P. Singh, A. Kumar, A. Mangal and S. Singhal, "Bilingual automatic text summarization using unsupervised deep learning," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 1195-1200.

[6] L. Wang, "SUMMARIZATION AND SENTIMENT ANALYSIS FOR UNDERSTANDING SOCIALLY-GENERATED CONTENT," in *Cornell University, http://www.ccs.neu.edu/*, 2016. [Online]. Available: http://www.ccs.neu.edu/home/luwang/thesis/thesis.pdf. Accessed: Dec. 15, 2016.

[7] A. R. Pal and D. Saha, "An Approach to Automatic Text Summarization using WordNet," 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, 2014, pp. 1169-1173. [Online].

[8] J. Yadav and Y. K. Meena, "Use of fuzzy logic and wordnet for improving performance of extractive automatic text summarization," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 2071-2077.

[9] Y. Kumar, O. Goh, H. Basiron, N. Choon and P. Suppiah, *A Review on Automatic Text Summarization Approaches*, 1st ed. Journal of Computer Science, 2016.

[10] N. Sharef, H. Zin, and S. Nadali, "Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data," in *Science Publications - Journal of Computer Science*, 2016. [Online]. Available: http://thescipub.com/PDF/jcssp.2016.153.168.pdf. Accessed: Dec. 11, 2016.

[11] A. Kaushik, A. Kaushik, and S. Naithani, "A Study on Sentiment Analysis: Methods and Tools," in *International Journal of Science and Research (IJSR)*, 2014. [Online]. Available: https://www.ijsr.net/archive/v4i12/NOV151832.pdf. Accessed: Dec. 15, 2016.

[12] I. MANI, *Summarization Evaluation: An Overview*, 1st ed. DARPA's Translingual Information Detection, Extraction, and Summarization (TIDES) research program, 2002.

[13] J. Khairnar1, "Sentiment Analysis Based Mining and Summarizing Using SVM-MapReduce," in *(IJCSIT) International Journal of Computer Science and Information Technologies*, 2014. [Online]. Available: http://www.ijcsit.com/docs/Volume%205/vol5issue03/ijcsit20140503314.pdf. Accessed: Dec. 12, 2016.

[14] M. Hofmann and R. Klinkenberg, Eds., *RapidMiner: Data mining use cases and business analytics applications*. United States: Taylor and Francis, 2016.

[15] RapidMiner GmbH. (2016). RapidMiner 7, Operator Reference Manual. [Online]. Available: https://rapidminer.com/wp-content/uploads/2014/10/RapidMiner-5-Operator-Reference.pdf. Accessed on: Apr. 26, 2016.

[16] G. Piatetsky, "KDnuggets Annual Software Poll:RapidMiner and R vie for first place", *Kdnuggets.com*, 2013. [Online]. Available:

http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html. [Accessed: 27- Apr- 2017].

[17] Forrester Research, Inc., "The Forrester Wave™: Predictive Analytics And Machine Learning Solutions, Q1 2017", forrester.com, Cambridge, MA, USA, 2017.