A Comprehensive Study of Text Mining Approach

Abhishek Kaushik† and Sudhanshu Naithani††

Kiel University of Applied Sciences† Kurukshetra University ††

Summary
Text mining or knowledge discovery is that sub process of data mining, which is widely being used to discover hidden patterns and significant information from the huge amount of unstructured written material. The proliferation of clouds, research and technologies are responsible for the creation of vast volumes of data. This kind of data cannot be used until or unless specific information or pattern is discovered. For this text mining uses techniques of different fields like machine learning, visualization, case-based reasoning, text analysis, database technology statistics, knowledge management, natural language processing and information retrieval. Text mining is largely growing field of computer science simultaneously to big data and artificial intelligence. This paper contains the review of text mining techniques, tools and various applications.

Key words:
Text Mining, Data Mining, Natural Language Processing, Machine Learning, Visualization, Text Analysis.

1. Introduction

Today's world can be described as the digital world as we are being dependent on the digital / electronic form of data. This is environment friendly because we are using very less amount of paper. But again this dependency results in very large amount of data. Even any small activity of human produces electronic data. For example, when any person buys a ticket online, his details are stored in the database. Today approx 80% of electronic data is in the form of text [10]. This huge data is not only unclassified and unstructured (or semi-structured) but also contain useful data, useless data, scientific data and business specific data, etc. According to a survey, 33% of companies are working with very high volume of data i.e. approx 500TB or more. In this scenario, to extract interesting and previously hidden data pattern process of text mining is used. Commonly, data are stored in the form of text [15]. And then following three steps [16] takes place:-

a) Data is preprocessed.
b) A technique of text mining is applied.
c) Here the results of bare analyzed.

Text mining and data mining are similar, except data mining works on structured data while text mining works on semi-structured and unstructured data [10]. Data mining is responsible for extraction of implicit, unknown and potential data and text mining is responsible for explicitly stated data in the given text [2]. On the other hand potential information extraction is common to both [2]. Some other terms such as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT) are also used in place of text mining sometimes [10]. Figure 1 shows the general process of text mining.

2. State of the Art

Simon Balbi [5] described text mining on elementary forms. Raymond J Mooney [6] gave a framework called DISCOTEX (Discovery from Text EXtraction). Martin Krallinger [7] introduced Prodisen corpus using information extraction and text classification techniques for the automatic identification and construction of text-based protein and gene description records. Jadhav Bhushan [9] proposed a way for searching research papers by using clustering. Jacob Kogan [11] reported clustering scheme that applies singular value decomposition (SVD) based algorithm. Jae-Hong Eom [12] provided an intelligent machine learning based text mining system for mining biological information called PubMiner. Yu-Seop Kim [13] tried to find optimal dimensionality in data-driven models, Latent Semantic Analysis (LSA) model and Probabilistic Latent Semantic Analysis (PLSA) model.

3. Technology Premise of Text Mining

There are several technology premises for mining the text. Some of them are listed below and later on to describe one by one.

a) Summarization
b) Information extraction
c) Categorization
d) Visualization
e) Clustering
f) Topic tracking
g) Question answering
h) Sentiment analysis

a) Summarization: - Summarization is one of the most important techniques. In simple words summarization is a process of making summary of any document containing large amount of information while theme or main idea of

Manuscript received February 5, 2016
Manuscript revised February 20, 2016

document is maintained. It helps the user to understand whether a particular document is useful for him or not [10]. Compression is also related to summarization, but is not in human readable form [2]. For an example of summarization, 'abstract' part of a research paper describes the main idea of research conducted by author(s).

b) Information Extraction: - Information Extraction utilizes relations within the text. It uses pattern matching for it [17]. For example, in RDBMS stored data is available in the form of tables. When data is in unstructured form, information cannot be extracted with ease (no fixed reference). In IE natural language document is converted into structured one and then the knowledge is extracted [16]. IE process extracts small chunks/fragments (people, place, time, date, address) of text by matching patterns. This technique provides impressive results when it is applied to very vast volume of text. By using machine learning an IE system can be constructed, but it cannot provide fully accurate results [6]. That's why errors can be detected in the output of automatically extracted database. In DISCOTEX [6] framework author first construct a database by using learned information extraction system and then utilize mining to discover knowledge which can be again used to fortify the accuracy of IE. IE [17] has several types:-

i- Entity Extraction (associating nouns with entities)
ii- Concept Extraction (noun and phrases)
iii- Token Extraction (characters without separator)
iv- Term Extraction (token with specific semantic purpose)
v- Atomic Fact Extraction (subject with actions)
vi- Complex Fact Extraction

c) Categorization: - Categorization is a supervised learning technique which places the document according to content. Document categorization is largely used in libraries [2]. Document classification or text categorization has several application such as call center routing, automatic metadata extraction, word sense disambiguation, e-mail forwarding and spam detection, organizing and maintaining large catalogues of Web resources, news articles categorization etc. For text categorization many machine learning techniques has been used to evolve rules (which helps to assign particular document to particular category) automatically [2].

Fig.1. General Process of Text Mining.

There are two ways for document classification which are SAS Content Categorization and SAS Text Miner. Former one can parse, analyze and extract content while latter one depends on frequency of occurrences.

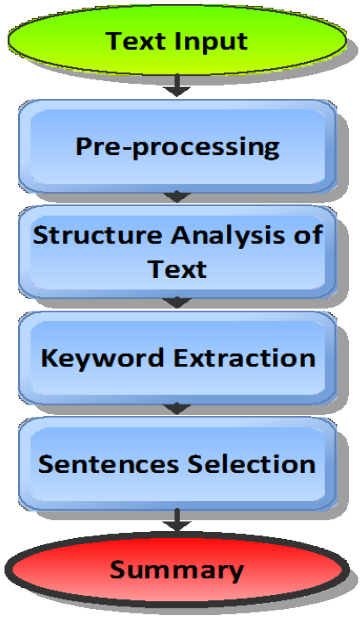

Fig.2. Process of Summarization.

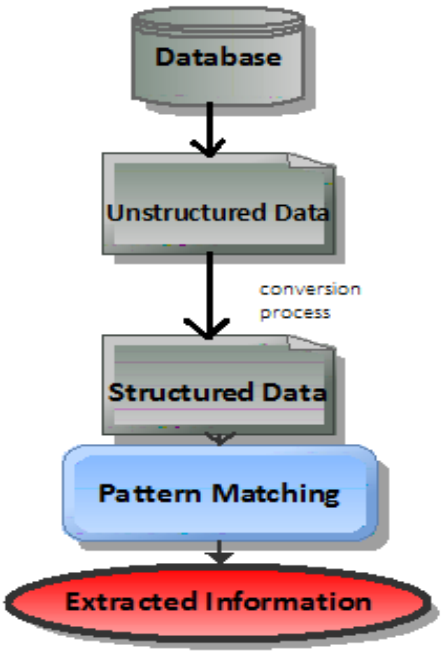

Fig.3. Steps of Information Extraction.

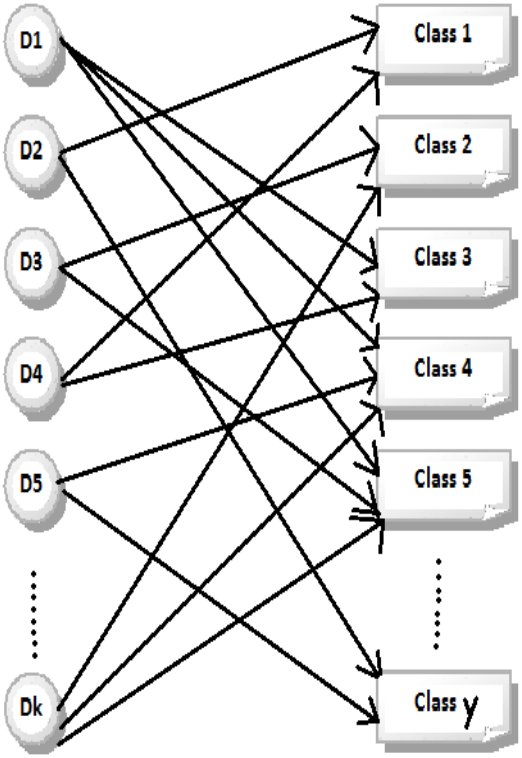

Fig.4. Classification process with multiple classes per document.(Here D stands for document & total documents are k and total classes are y. Each document has multiple classes.)

d) Visualization: - Visualization which dates back to 2nd century is a computer graphic effect to represent information and revealing relationships. According to David McCandless [22] (author, data journalist, and information designer):-

"By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you're lost in Inform, an information map is kind of useful."

Visualized text is easy to understand as compared to numerical data because anyone can easily scan trends from a well drawn picture [21].

Visual text mining or information visualization is a process where the vast volume of text is processed and provides browsing capabilities [10]. Information visualization is currently being used in social network analysis (SNA). It is also used by governments in case of national security and by investigation agencies to investigate any mishappening. The visualization process can be divided into three parts [10]:-

i) - Data preparation (decide which and what data is to be used for visualization process and form data space)

ii) - Data analysis and extraction (analyze and extract data for visualization of original data and form visualization data space)

iii) – Visualization mapping (use mapping algorithm to map visualization, data space to visualization target)

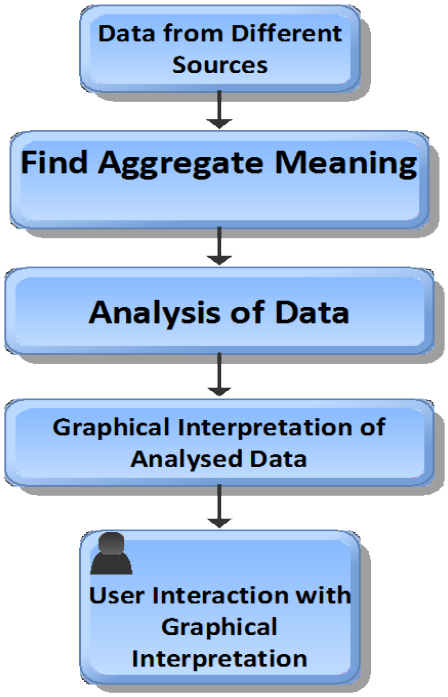

Fig.5. General steps in the process of Visualization.

e) Clustering: - Text Clustering is a document's textual similarity based unsupervised technique (does not require training data) which is used by data analysts to divide the text into mutually exclusive groups. Text mining has very less application of clustering. Clustering can be divided into 2 parts- hierarchical clustering and partitional clustering.

Hierarchical clustering outputs a single clustered tree and again subdivided into a bottom up hierarchical clustering and top down hierarchical clustering.

In partitional clustering document set is divided into k disjoint point sets [16]. For example k-means algorithm.

Fig. 6 shows generalization of Clustering process. Here are no of different kinds of documents which are grouped on the basis of textual similarity (In this case no of documents are 12 and we used the same color to show the textual similarity)

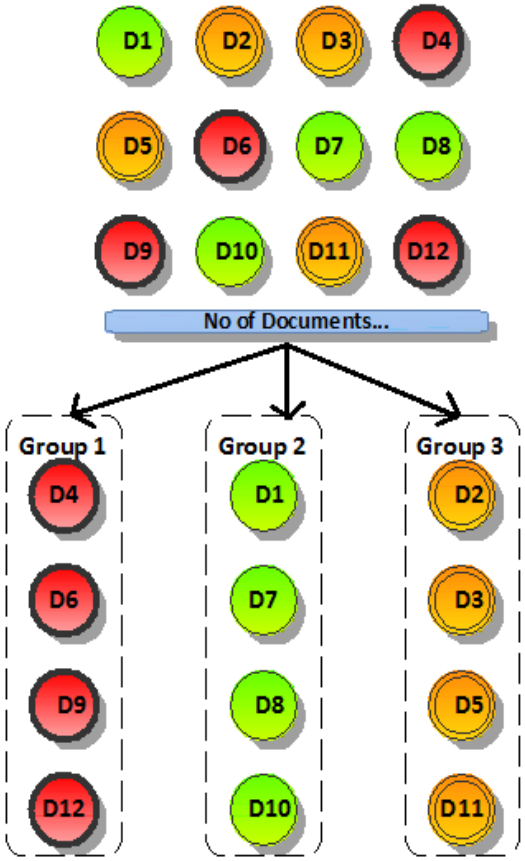

Fig.6. Generalization of Clustering.

f) Topic Tracking: - As the name suggests topic tracking is a process of following any specific topic based on requirement or interest. It predicts possible topic of interest for use on the basis of his topic interest history [10]. The Google search engine is utilizing it very effectively. For example, when a user search an item and after some time he goes for another item, then he gets ads related first item on different websites opened as a result of second search. If a student searches for some universities and after some time he open social networking site (Facebook). In this case he will get sponsored pages and ads of other related universities on his wall. On the other hand it also has limitations. If a user subscribes for "artificial intelligence" then he will get new information on it bat very less of them will have an actual use of the user. Topic tracking has very large use in the field of medical, research and education.

g) Question Answering: - Natural language queries or question answering is responsible to decide a way to find a more suitable answer for a particular question. For example, when user asks Google any question or search something, then Google gives him best matched answers or links for that question by searching keywords of question in the database. It can use more than one mining techniques at a time [10] for example it can use IE and question categorization to extract entities and to assign the question respectively. Q&A can be used in several web applications, medical field as well as for education.

h) Sentiment Analysis: - Sentiment Analysis which is also known as opinion mining is configured of user's emotion, mostly into several classes which are positive, negative, neutral and mixed [17]. It is mainly used to get people's view or attitude towards anything which includes services and products. For an example, Amazon's website provides space for user's comment on their all the selling products. By this peoples can express their attitude towards any specific item which is again helpful for other buyers because they can read reviews from previous buyers. From company's (Amazon) perspective, it helps them to improve their product's quality by making required modifications in it. Now with the growing popularity and reach of social networking sites any organization can get a large quantity of data (reviews) related to their products. This enables them to analyze real time opinion of customers very quickly [17]. So finally Sentiment Analysis is a process of automatic extraction of features by mode of notions of others about specific product, services or experience [1] which can be more productive with other analytics/mining techniques.

4. Tools of Text Mining

There are some tools on the internet which can be used for mining the text. These tools follow step by step process for the purpose of mining. For an example to analyze event data [24] by text mining tools first structure of event data is converted to country-event-month ID, name of the event and short description of the event. After these these steps are followed:-

1. Import event data into mining tool
2. Linguistic processing
3. Factor analysis
4. Cluster analysis

Here are some tools for text mining which can be found on the internet:-

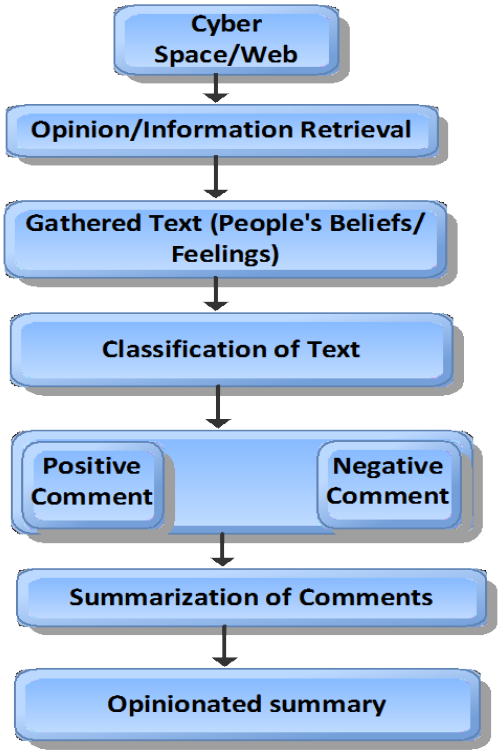

Fig.7. Architecture of Sentiment Analysis/Opinion Mining

a) Text Analyst: natural language text analysis software (Megaputer Intelligence): - Megaputer is one of manufacturer of mining related tools. Its mining system applies lots of analytic functions to investigate text. To manufacture this tool 20 years of research is used. The main advantage of this system is it can distill the semantic network of a text completely autonomously without prior development of a subject-specific dictionary by a human expert [19].

b) Intelligent Miner for Text (IBM Software): - IBM is well known worldwide. Its intelligent miner is a combination of analysis tools. Its language identification tool, feature extraction tool, Summarizer tool, topic categorization tool and clustering tool perform tasks according to their name [19].

c) Text Finder (Paracel Inc.): - Paracel's text finder is very fast and accurate system. It can easily search, filter and categorize large volume of text. It is very useful for time critical problems. It can filter 40M byte text/second which is equal to 50000 pages [19].

d) Callable Personal Librarian (CPL): - CPL is for users' expert of C language. CPL system can manage structured data, hypertext and multimedia application. CPL's API constructs searching and retrieving application. It is basically famous for discovering concepts on the fly without any additional effort [19].

e) Vantage Point: - It is very useful for competitive intelligence and technology management. Here search is conducted at starting on database by using a search engine. Then raw data merge into one document for preprocessing. It can extract phrases and words very easily by using NLP. It is most effective with several thousand records [19].

f) Fulcrum Search Server: - It extracts information from online sources like e-commerce, customer care, online technical support etc. Its searching process is high and reliable. It can provide very good results from natural language queries [19].

g) Isaac and Amberfish: - These two are produced of Etymon, which develops tools for constructing info architecture. Search is open source which supports Boolean queries and field based searching. With it an interface for web search is also provided. Its commercial version is called Amberfish [19].

h) ISIS: - It is product of the United Nations Education, Scientific and Cultural Organization (UNESCO). This organization is responsible for justice, human rights and fundamental freedoms without distinction of race, sex, language or religion.
ISIS is combination of lots of software for same data formats. For it very strong retrieval engine and more powerful formatting language is utilized. It also accommodates for international info format [19]. It is free for non commercial use.

i) AE1: - AE1 is the first search engine manufactured by Answer Logic which interprets documents and questions by using NLP techniques [19]. It first analyzes text by using language processor and then stores results as concepts in IdeaMap. There is a document manager which manages all the functions regarding documents. To get output language processor matches the concept of question with stored concepts in IdeaMap and provide best matched concepts [19].

j) WordSmith Tools: - This tool is released by Oxford University Press. It analyzes behavior of words in the text. it has some functions [19] described below:-

1. **Wordlist** enables user to see all word-cluster present in text in alphabetical order.
2. **Concord** helps to see any word or phrase in the text.
3. **Key Word** is able to find keywords in a text.

k) Harvest: - It is designed by Internet Research Task Force Research Group. It is a combination of tools to work on the internet and can gather, extract, organize and replicate info. It is popular for its quality to work in different formats on different machines [19].

5. Applications in Text Mining

Text mining applications are being used in all those fields where we have to get most valuable and useful data from an enormous amount data like banks, IT sector, research, energy, media, political analysis, healthcare etc.
Here are the main applications of text mining:-

a) Competitive Intelligence: - Competitive Intelligence is a process of collecting all possible information about market trends and other competitors so that by analyzing this data specific patterns and current requirements can be developed which will further contribute in the company's strategies [10].

b) Detection of Junk Emails: - Text mining is also utilized to detect unwanted junk e-mails automatically. These emails can be classified according to predefined frequency terms [16].

c) Management of Human Resources: - Text mining also can be used to manage human resources. For example, analyzing staff's opinion is the best use of mining in an office. Other than that storing new CV's, monitoring company's progress and monitoring satisfaction levels of employees are also important tasks which can be done by mining techniques [10].

d) Customer Relationship Management: - It's the duty of CRM to provide quick response to any client's query or message. By using text analysis these messages/queries are diverted to the appropriate person or service for further process [10].

e) Multilingual Applications of Natural Language Processing: - Utilization of text mining techniques to analyze different web pages in a different language is a classic example of multilingualism [10]. There are other applications also like speech recognition system.

f) Classification of NEWS as Text: - Normally people like to see headlines in a newspaper which involves naming of any person, country or organization [16]. Manually doing this job is hectic and time consuming process. So text mining techniques are used to perform this task.

g) Classification of Scientific Documents: - In the past, scientists had to go through a number of articles to find articles in their interest. With the passage of time documents have grown at very high rate so making classification more complex ever. Replacement of keyword queries with the structured automated process by topic scoring engine reduced research time and improved results quality. By using regression analysis, classification rules are generated to cope up with this problem [23].

h) Sentiment Classification: - The purpose of social media has created many chances for people to publicly voice their beliefs, simply when they are employed to deliver an opinion hit a vital problem. Sentiment Analysis is a case of natural language processing which could mark the mood of the people about any specific product by analysis and classifying it as positive, negative or neutral. Sentiment Analysis is a process of automatic extraction of features by mode of notions of others about specific product, services or experience. The Sentiment Analysis tool is to function on a series of expressions for a given item based on the quality and features. Sentiment analysis is also called Opinion mining due to the significant volume of opinion [1].

Conclusion

Text mining is one of the fastest growing fields today. With the passage of time its importance is only going to increase because rate of data production is very high. Automatic text mining has a long way to go because it is not in the position to challenge the human's capabilities. From last few years text mining (sentiment analysis) is largely being used to predict the results of elections at national and state level which is most significant development in the field recently. On account of growing interaction of text mining to some other fields, especially with machine learning, visualization and natural language processing, it is possible to design more effective and useful text mining system. Text mining is also being used by industry and it is generating the sheer amount of knowledge which cannot even consume by humans. In this paper we tried to present an overview of text mining approach with its techniques, tools and applications.

References

[1] Abhishek Kaushik and Sudhanshu Naithani, "A Study on Sentiment Analysis: Methods and Tools" International Journal of Science and Research (ISSN (Online): 2319-7064, Volume 4 Issue 12, December 2015).

[2] Ian H. Witten et al, "Text Mining".

[3] Patricia Cerrito and John C. Cerrito, "Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs".

[4] Bill Hollingsworth, Ian Lewin and Dan Tidhar, "Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a Prerequisite for E-Science Text Mining".

[5] Simona Balbi, Sergio Bolasco and Rosanna Verde, "Text Mining on Elementary Forms in Complex Lexical Structures" JADT 2002 : 6es Journees Internationales d'Analyse statistique des Donnees Textuelles.

[6] Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction" Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.

[7] Martin Krallinger and Rainer Malik, "Text Mining and Protein Annotations: the Construction and Use of Protein Description Sentences" Genome Informatics 17(2): 121{130 (2006)

[8] Marti Hearst, "Text Mining Tools: Instruments for Scientific Discovery" IMA Text Mining Workshop April 17, 2000.

[9] Jadhav Bhushan G, Warke Pushkar U, Kuchekar Shivaji P and Kadam Nikhil V, "Searching Research Papers Using Clustering and Text Mining" International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014).

[10] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications" Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[11] Jacob Kogan, Charles Nicholas and Vladimir Volkovich, "Text Mining with Hybrid Clustering Schemes".

[12] Jae-Hong Eom and Byoung-Tak Zhang, "PubMiner: Machine Learning-Based Text Mining System for Biomedical Information Mining" AIMSA 2004, LNAI 3192, pp. 216–225, 2004.

[13] Yu-Seop Kim, Jeong-Ho Chang and Byoung-Tak Zhang, "An Empirical Study on Dimensionality Optimization in Text Mining for Linguistic Knowledge Acquisition".

[14] Miloš Radovanović and Mirjana Ivanović, "Text Mining: Approaches and Applications".

[15] Ah-Hwee Tan, "Text Mining: The State of the Art and the Challenges"

[16] Divya Nasa, "Text Mining Techniques- A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.

[17] Goutam Chakraborty, Murali Pagolu and Satish Garla, "Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS".

[18] Rashmi Agrawal and Mridula Batra, "A Detailed Study on Text Mining Techniques" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[19] Jan van Gemert, "Text Mining Tools on the Internet: An Overview".

[20] Micah J. Crowsey, Amanda R. Ramstad, David H. Gutierrez, Gregory W. Paladino, and K. P. White Jr., Member, IEEE "An Evaluation of Unstructured Text Mining Software".

[21] Zhao Kaidi, "Data Visualization"

[22] White Paper on "Turning Big Data Into Big Insights: The Rise of Visualization-based Data Discovery Tools" sponsored by intel, March 2013.

[23] Bernd Drewes, "Some Industrial Applications of Text Mining".

[24] Theoni Stathopoulou, "Using Text Mining Tools for Event Data Analysis".

Abhishek Kaushik is currently working in Siemens, Germany as a Master thesis student. He is in the final phase of completing his Masters degree in Information Technology from Kiel University of Applied Sciences. Before starting his Masters he received his Bachelor's of Technology in Computer Science Engineering from Kurukshetra University in 2012.

Sudhanshu Naithani has received his Bachelor's of Technology in Computer Science Engineering from Kurukshetra University in 2015. He is currently working as a research assistant under Assistant Professor Ravinder Madan at Manav Bharti University, Solan.