Text Analytics on Big Data using Artificial Intelligence

Ain-ul-noor Fatima
National Textile University
anfatimakhan@gmail.com

Ayesha Muarraf
National Textile University
opinionated.gal4@gmail.com

Rehan Ashraf
National Textile University
rehan_ashraf94@yahoo.com

Waqar Ahmad
National Textile University
waqar@ntu.edu.pk

***Abstract*— Text analytics is the text substance analysis find mostly in emails, tweets, forums, text messages and other methods of communication links. This field is used in most industries for analyzing to help in many emails and also to analyze the customer comments questions and answers in different forums and sites. By using this technique, a conception analysis of a company, products, and brands by checking all positive or negative viewpoints can be done. The other name of this field is also called text mining and it is also a sub-branch of Natural Language Processing (NLP). Text analytics has a number of subdivisions such as Named Entity Recognition, Information Extraction, and Semantic web annotated domains representations etc. These days, text analytics is widely used in big data that's why we choose this field because many peoples used this data mining techniques for several purposes. And many of them achieve a high consideration e.g. machine learning that is used to display the semi-supervised improvement of the systems but the problems are that they show the many limitations that lead to the result that does not always make the best choice. Basic purpose to choose these subdivisions is to achieve the present and eventual applications of Text analytics.**

Keywords— Text Mining; Text Analytics; Big Data; Artificial Intelligence; Natural Language Processing;

I. INTRODUCTION

Natural language processing is a field of computer science, computational linguistics and Artificial Intelligence (AI) that is linked to the communication between the natural languages (human) and computer languages. Some of the authors use the term for this 'conversely'[1]. NLP is a sub-branch of artificial intelligence and in future, it will be leading to cognitive computing so most of the cognitive methods are granted as natural language assertion.

In this article, the major discussion is on Text Analytics. Text Analytics is the new name of data mining, text mining and natural language understanding. It is the method to change disorganized data into structured data. It can also perform manually but it is an insufficient method [2]. Due to manually annotation technique, we purposed text mining and NLP algorithms to find meaningful data in the big amount of text. A new name for text analytics achieves the success that is "Big Data" used for disorganized text, usually in the economic slightly than the educational area, most likely because disorganized free text accounts have 80% in the business area including social media, Wikipedia's and surveys etc. [3]. This topic is not covered in many academic papers but it is going to be changed in future. This is the analysis of new and already have unknown information and from different resources it automatically extracts information.

II. TEXT ANALYTICS

Text analytics is extended form of data mining and it basically tries to find textual patterns and connections from unorganized data instead of data that is stored in the database. The purpose of text analytics is to draw out information from the un-sorted, un-organized text. It is similar to data mining but in data mining first of all the data is processed and changed into structured data and then it is stored in the database[4]. Data mining deals with that structured and saved data. Text analytics deals with un-organized data such as the content of electronic mails, research papers, text documents, also articles of blogs on the internet, comments, and messages on social media. Text analytics is basically a branch or field which has been drawn using many concepts of data mining, machine learning and so on.

Text analytics has many application areas. For example, it can be used by companies to know the latest trends, likes, and dislikes of people. According to those trends, they can launch new products and there will be very less chance of that product to be a flop because that product will already be developed from studying psychology, likes and dislike of people through their social media profiles, comments, tweets, messages, etc.[5]. Text analytics field is basically very interesting but also complicated field because it is a combination of various fields as in linguistics, psychology, data mining, machine learning and many more. A text analytics app can be developed performing various steps. First of all, a tool or software collects a document and checks what format is it in, also identifies the character sets. Then it will start analyzing that text and will keep performing different processes on that text until the required information gets extracted. The tactic is to find patterns, similarities or connections using some set of rules of some kind of previous information and applying the rule of that text (Note. the rule explains the text). The text is divided into sentences and words [6]. Then further words are differentiated using different categories for example if it's a noun or adjective. The relationship or connections between those words Then the text is divided according to the noun phrase, etc. The recognizable (named) entities are identified. At the end, the relationships are found out by seeing what the noun or pronoun is referring to. The result of the process mentioned above provides

organized or structured information which can be further utilized by various means for example Q&A systems which are discussed in detail in this paper. In text mining process some methods can be used such as information extraction, topic detection, categorization, concept-linkage, etc.[4]. Figure 1 shows the framework of text analytics processes, the author explained each process one by one.

A. Text Analytics Application Area

Security: By monitoring and analyzing text on websites and blogs, threats and unethical stuff can be found out which can be used for security purpose.
Marketing: By analyzing textual data we can identify which type of products people like and which type of products won't be profitable.
Processing emails and messages: We can filter the huge amount of emails based on some textual criteria (e.g. some words)[7]. It is very useful in some scenarios like if in an organization we want to divert messages or emails to concerned or appropriate departments.

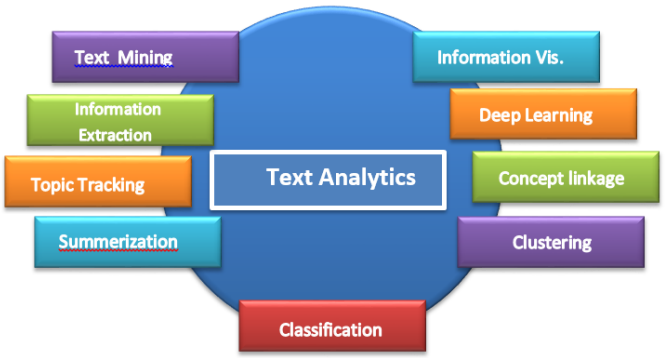

Figure1: Text Analytics Framework

III. Literature Review

In this section, a literature review has been covered. We studied five papers on text analytics. In Table1 we have rated/evaluated the papers we studied.

	Paper Title	Proposed Technique	Performance of paper	Data set	Limitations of papers
1	Mining Knowledge from Text Using Information Extraction [8]	Information Extraction Technique	Good	Nil	Only covers IE portion of Text Analytics
2	Big Data Analytics: Challenges and Applications for Text, Audio, Video and Social Media Data [9]	Text Mining	Not so good	Nil	Paper is not written so well, even grammar mistakes are there.
3	A Comprehensive study of Text Mining Approach [10]	Text Mining	Average	Nil	The material is helpful but the paper is not well written.
4	Text Analytics: the Convergence of big data and Artificial Intelligence[11]	Text Analytics in Big Data	Very Good	Nil	This paper is very helpful but fewer examples are given.
5	Text Mining with Information Extraction [7]	Information Extraction	Good	Nil	This paper only covers information extraction.

Table1: Comparison of Different Authors' Work

IV. Information Extraction

Information extraction is the widely studied area of NLP. This is the focused area because information collection is the big study in data mining [7]. It works on finding the accurate information in the free text. For example in database records, it is the form of structured articles. Records may contain real-world entities having attributes given in the text. In many organizations, more than 75 % data exists in a disorganized way. Extraction of organized information from disorganized data is a basic task in Artificial intelligence. Information extraction used text mining framework DiscoTEX (Discovery from Text Extraction)[7]. It plays an important role by recovering a collection of text documents to extracted items in data mining modules. The collection of documents (with their filled templates) requires extraction rules that can be tested on novel documents. For example, a detailed job posting document is to be extracted with information extraction methods. A detailed document is given below:
Title: Web Designer Address: Islamabad
The single person is answerable for the implementation and design process of web interfacing modules and back-end development responsibilities. A selected candidate should have experience includes: HTML, HTML5, javascript, CSS, Oracle, Java, C/C++, ODBC, PHP, ASP, and VBScript. Platforms used: Visual studio, Polaris, Windows. BS-SE or relevant and 3 years' experience is required in the relevant field. So the document format will be like in table 1:

1	Title	Web Designer
2	Location	Islamabad
3	Languages	C/C++, Java, HTML, HTML5. Java Script, PHP, ASP, VBScript
4	Platforms	Visual Studio, Polaris, Windows
5	Degree Requires	BS-SE
6	Experience	3 years
7	Areas	Database

Table2: Filled Job Template

Information extraction classifies key phrases and communication within the text. This is done by looking at the predefined arrangement in a text. This method is known as pattern matching that is placed on the 'Regular Expression'. The important form of information extraction is "Named Entity Recognition". NER concern with the predefined

categories, for example, the name of people, organization, companies, locations, cities, temporal expressions, percentages, business values, stock exchange values, genetics, quantities etc. NER explore to establish and classify elements in text into predefined categories. Many tools used for this purpose are as follows:

a. Apache openNLP

It is a library in a machine learning toolkit for the working of natural language processing.it supports all the NLP task for example sentence segmentation, part of speech tagging, parsing etc. [3]

b. Stanford named entity recognizer

It is a Java implementation of NER. It labels sequence of words in the text that will be the name of things. Ling Pipe: It is software for refining text using computational linguistics. It is also used for finding the name of organizations, the name of peoples in news. It automatically analyzes the Twitter search results into sections and advises right spellings of queries.

Here, the simple diagram shows the overview of text mining process. Data mining suppose that the mined information is already a relational database. Figure 2 shows the text mining process that takes the input of raw/plain text and appearance from NLP operations and gives an output that is the structured/organized data sufficient for ordinary purpose analysis software's. [12] Figure 2 shows the text mining process.

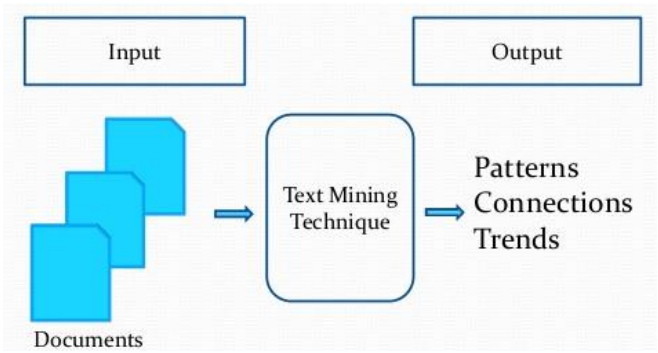

Figure 2: Text Mining Process

In many applications, electronic information is in the form of NLP documents except that the structured database. Information extraction addresses the problem of converting a collection of text documents into the structured database. The database is composed of information extraction module can be given to the KDD (knowledge Discovery Database)[13] module for more information. After all of this framework, get the outcomes.

V. TOPIC DETECTION

There are always some "Keywords" in articles which are brief and give the gist of substance to the reader. A large article or topic can be summed up in a brief way by discovering those keywords in that topic or article.[14] Now that internet has a huge amount of data, these keywords can be used for searching purpose. These keywords are used in many other techniques as well as searching such as text mining and topic detection. It is a very time-taking process of identification of these keywords in a document so it is preferable to use a tool for that.

This is where topic detection tool comes in handy. Google and Youtube provide this topic detection or topic tracking tool. This tool basically keeps a record of user's history, types of the files viewed by the user and recommends similar files to the user.[11] Google also allows the user to select some "Keywords" and whenever some news relating to those keywords becomes available, the user is notified.

VI. SUMMARIZATION

A summary is the main theme in the unstructured text. It has two ways to describe, Informative summary and indicative summary. The purpose of the informative summary is to replace the original document. And indicative summary supports decisions for example "can I read the original document yes or no"? Due to the unstructured text on the web, it is more important to use some methods for where summarization will be done automatically. The basic purpose of this method is to take more text as input and to alter into a small text its mean so-called summery of unstructured data [15]. A summary of text should be descriptive and should keep the meaning of the original text.

Text summarization is the beneficial branch of the text analytics. It is also the branch of natural language generation. It helps to find the document is fulfills the user requirements or not. Text summarization develops and snips the document in time that the user will read the first paragraph[4]. The basic idea of the summarization is to minimize the length and detail of document including main points and meanings. For example, if we take the data from Wikipedia on any topic. Wikipedia contains many details of the topics or articles. We read and summarize the data in some sentences by using any summarization technique. Another example is to take the one paragraph and text will be summarized using some technique.

"Hi All, as of today MetaStock has several new functions. The most important new feature is the ability to display forward heat rate charts. Also, notice that the interface looks different. This reflects and accommodates the new features[19]. If you have any question regarding this new version of MetaStock, please contact Bella Santuri". The summary of this paragraph will be:

"MetaStock used several different functions and interface. Or MetaStock has a new interface and new features". A tool that is used by the text analytics is "sentence extraction". It is used to find the sentences from an article. This tool may also use for searching heading in documents and subtopics in order to find the major points of documents. Summarization technique can be organized in two different methods, Shallow analysis, and deeper analysis.

a. Shallow analysis: prescribed the linguistic level of illustration and try to select the important part of the text.

b. Deeper analysis: conclude the correct level of representation of original text.[11]

Figure 3 is the detail description of summarization process where documents are summarized in a short detail.

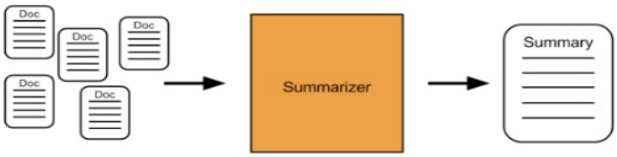

Figure 3: Summarization Process

VII. Categorization and Classification

Categorization is basically (as the word suggests) a process of putting the similar things under one category or class. In other words, things are sorted into categories for our convenience. Categorization uses a strategy to find out the main topics a document is covering. After this, the relationships are found out which can be done using synonyms and some other methods. At the end, documents are ranked on basis of how much content they have on a particular topic. [6]

VIII. Information Visualization

Information visualization provides a new method in which textual sources in an organized manner (hierarchy), providing browsing capabilities which allows the reader to figure out and study related topics without putting so much effort in searching. It comes handy when there's a large number of documents and reader wants to study related topics in those documents.[16] This process is categorized into three parts:

i. Data preparation where the thing is which data is used for this process and what data it will be.
ii. Data extraction and analysis for visualization data from original data.
iii. Mapping algorithms are important to map visualization.

This is a conceptual diagram of information visualization where we take data from different sources and find there meaning and take the analysis of the given data. Then the graphical representation of analyzed data is available for users. Figure 4 displays the overall description of concepts related to information visualization.

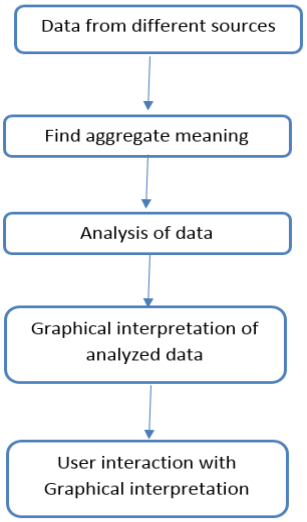

Figure 4: visualization process steps

Tag clouds and word clouds are common examples of putting similar words together. Figure 5 displays the word cloud that is used for visualization.

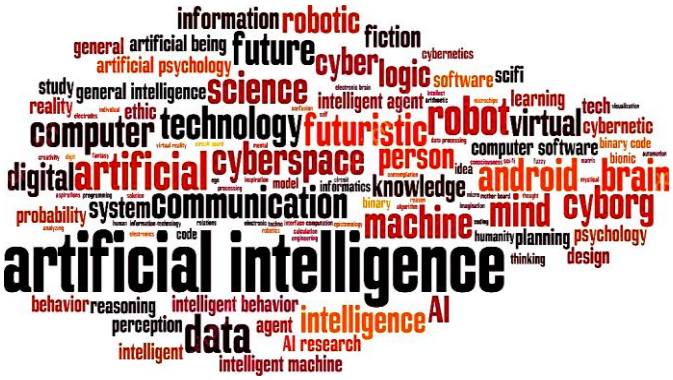

Figure 5: Example of word cloud

IX. Concept Linkage

Concept linkage method is used to create relationships or connections between documents that cover similar topics or concepts. This method is very useful in searching. In research, concept linkage is very helpful. Nowadays so much research has been done on the same topic of a field and before getting to any conclusion it is absolutely necessary to study the previous research done on that concept or topic. This is where this concept-linkage becomes very useful [10].

X. Question Answering

It is a computer science technique under the field of natural language processing to find the best answer of a given question that is relating to automatically answer the questions asked by the humans in natural languages[12]. This technique includes many sub-techniques describe here, for the question understanding from information extraction, categorization, and

question topology, for the right selection and generation of the answer.

Different systems used for this technique. Some of them are:

a. OpenEphyra

It is an open source QA system which was originally derived from Ephyra that is developed by Nico Schlaefer and his participation in the TREC QA competition.[12] Because of some reasons this system has been abandon so some alternative is replaced just like Yoda QA.

b. Yoda QA

It is usually a basic QA system. It is an open domain information that cannot be complete from big datasets. [14]

QA system includes some examples using natural language collection documents are:

i. A collection of Wikipedia pages.
ii. Organization internal documents and web pages
iii. A set of World Wide Web pages.

XI. CLUSTERING

Document or text clustering is different from categorization and classification because categorization process is done for supervised data, using predefined topics the process is done.

There aren't predefined topics in clustering. Document or text clustering is an app of cluster analysis on textual data. Clustering is basically a process of grouping related or similar documents. This clustering can be useful using search engines because a search engine returns a huge amount of pages against a query. A clustering tool can be helpful as it can automatically group the returned pages into categories and meaningful information. Carrot 2 is open-source software which can be used for this process.[17]

XII. DEEP LEARNING

Deep learning is a branch or sub-field of machine learning which is related to algorithms which are inspired by the structure and behave of a human brain. Deep learning is the use of artificial neural networks. Deep learning is using brain simulations to make algorithms better and revolutionize the world of artificial intelligence and machine learning[7]. Deep learning is done using supervised data. The performance of deep learning is directly proportional to the amount of data available. Today, a huge amount of data is available on the internet on almost everything.

Deep learning revolves around Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). RNN revolves around the idea of using sequential information. The output is always dependent on previously provided information. RNN uses memory to remember previous information [16].CNN consists of different layers. Each layer performs a different function on data, providing the output. CNN can be used for topic detection, sentiment analysis, and many more purposes.

An example of usage of deep learning in text analytics is Word2vec. Word2vec is a tool inspired by deep learning which takes the text as input and gives output in form of numeric vectors which represent each word. The vectors show the semantic similarity between words.

Stuff that still needs to be figured out in text analytics: The data is increasing with each passing day and some of the data is not sorted or organized like the data of social media. Even after the cleansing process, still, information is not in a form to be easily used for getting reliable predictions. Another problem is that many social networking groups and companies do not provide all the data publically. We need the data for text analytics.

XIII. TOOLS

There are many tools available on the internet for text analysis. These tools follow some steps for analysis purposes. For example, to analyze some event data by text analytics tools firstly event data structure is changed into country-event-month ID, event name and some short detail of the event. For this event analysis follow these steps:

i. Change event data into mining methods
ii. Linguistics processing
iii. Factor analysis
iv. Cluster analysis

Here are some tools for text mining process:

a. Megaputer Intelligence:

The mining systems apply to many analytical functions to find the text. 20 years of research is used to mining this tool. The basic purpose of this tool is that it can extract the semantic network of text is perfectly individually without the proceeding development of a subject-specific dictionary by human experts[5].

b. Intelligent Miner in text IBM Software:

Intelligent minor is a combination of analysis tools. It used the language identification tool, summarization tools, features extraction tools, topic categorization tool and clustering tools perform many tasks [13].

c. Paracel's Text Finder:

It is very fast and accurate system. This tool can easily filter, search and categorize text. Useful for time-critical problems[11].

d. Search Server Fulcrum:

This tool extract information from many online sources e.g. customer care, online support material etc. it provides very high and reliable searching process and gives good results from Natural Language Queries[8].

e. Harvest:

This tool is designed for internet research task group. Many combinations of tools are available on the internet and can collect, extract and organized the information. Its quality of work is good [8].

XIV. APPLICATIONS OF TEXT ANALYTICS

a. IBM's Watson

IBM has a broad scope in Artificial Intelligence and is considered as a father of Artificial intelligence. IBM developed many systems in AI. Some of them are checkers players, Deep Blue and many more. Recently work of IBM is on advanced Question & Answer and cognitive computing in which the "Synapse" and "Watson" are good examples[18].

IBM Watson is a question answering computer system accomplished of answer the questions in NLP developed by IBM. This system is specially developed for the answer question on a quiz show. It is a platform for to build different technologies in NLP and machine learning to represent a large amount of disorganized data. Watson represents information of documents in a database then make patterns of different relationships conclude from different concepts in documents than develop a question answering system to query the details of the database[11]. It used supervised learning techniques. It also used the NLP to understand the concept of grammar and find all possible meanings and resolve what is actually being asked. Then presents the results based on the material and information provided.

b. IPsoft's Amelia

Ipsoft is known for developing expert systems that perform domestic servant type tasks rather than people performing those tasks. Amelia is one of their products. Amelia understands the semantics of the language and solves queries related to business. It starts with learning the manual just like a human. It has the capability of learning through the experience and the interaction or communication with customers. That's what makes it human-like. Amelia learns through the semi-supervised machine learning[3].

Amelia has both episodic and semantic memories. Episodic one remembers events and experiences in sequence. Semantic memory, as the name suggests, it has information about facts. In other words, semantic memory has the knowledge about the world. Amelia can easily and quickly understand the context and retrieve the information from its semantic memory. It also interacts with the customer emotionally. In other words, we can say that shows empathy and compassion to the customers which lead to the better user experience. This software is used to help desks and contact centers [3].

XV. FUTURE WORK

Text analytics technologies are used in many industries nowadays. For example, business, finance, marketing, healthcare, media analysis. It extracts data from not only traditional data sources but also extracts data from online resources and social media. For example, the public going towards the largest generator of text contents e.g. WhatsApp, Facebook, Instagram, etc.

Many data is generated beyond feedback channels continues to develop at an aggressive rate providing business with the assets information of their customers, looking at the customers experiences and business success, organization are going to join text analytics to gain the unstructured text help in open survey questions, social media post, and other sources of feedback data. It helps to business listen to right stories by extract vision from a free text written by or about customers than combine this data with the existing feedback data and finds trends and patterns.

This field is covered currently in many areas such as customer experience or social listening. Scientific experimentation and technical innovation are also covered are this area. Multi-lingual analytics is a branch of AI and also in machine translation. Customers experience, market research, customer vision, digital analytics are enhanced through text analytics. Lexical and semantic networks, systaltic rule system will continue their work in the areas of emotion analytics, affective states compounded of speech, text and images and facial expression analysis.

Super textual communications such as emoji on social media or messaging needs extract their vision to meaningful analytics. Semantic search, knowledge graphs, speech analytics and the capability to compose articles such as email text messages and translations from text, data, rules etc. All these fields, that need to be covered or already has been covered in text analytics.

XVI. CONCLUSION

Text analytics has a huge history and is a constantly evolving field. Especially when it comes to big data, a huge amount of textual information is produced with each passing day. This textual information contains knowledge about the world. This knowledge can be used for various purposes e.g. marketing purposes, academic purpose and so on. Cognitive products have already been developed using this knowledge like Watson and Amelia. These products interact with humans. They help in performing human-like tasks through cognition. Textual data from all over the world is increasing everyday especially the data from social media. All that needs to be done is to somehow clear the noise and irrelevant data and process textual information using that textual data. Future seems very bright but still, there are open areas for research in textual analysis and there is stuff (problems) that needs to be figured out.

XVII. REFERENCES

[1] C. Suh-lee, J. Jo, and Y. Kim, "Text Mining for Security Threat Detection Discovering Hidden Information in Unstructured Log Messages," 2016.

[2] D. Miao, "A Recommendation System Based on Text Mining," 2017.

[3] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text Mining : Techniques , Applications, and Issues," vol. 7, no. 11, pp. 414–418, 2016.

[4] C. A. S. J. Gulo and R. P. M. R. Thiago, "Text Mining Scientific Articles using the R Language," pp. 60–69,

2015.
[5] R. A. Sinoara, “Text mining and semantics : a systematic mapping study,” 2017.
[6] R. E. Thomas, “Co-Clustering with Side Information for Text Mining.”,Communication and Electronics Systems (ICCES), International Conference,IEEE, 21-22 Oct. 2016.
[7] R. J. Mooney and U. Y. Nahm, “Text Mining with Information Extraction,” no. September 2003, pp. 1–16, 2005.
[8] R. J. Mooney and R. Bunescu, “Mining Knowledge from Text Using Information Extraction,” vol. 7, no. 1, pp. 3–10.
[9] A. P. F. O. R. T. Ext, A. Udio, and V. Ideo, “BIG DATA ANALYTICS: CHALLENGES AND APPLICATIONS FOR TEXT, AUDIO, VIDEO, AND SOCIAL MEDIA DATA,” vol. 5, no. 1, pp. 41–51, 2016.
[10] A. Kaushik and S. Naithani, “A Comprehensive Study of Text Mining Approach,” vol. 16, no. 2, pp. 69–76, 2016.
[11] A. Moreno and T. Redondo, “Text Analytics : the convergence of Big Data and Artificial Intelligence,” pp. 57–64.
[12] S. Fong, S. Zhou, and L. Moutinho, “Text Analytics for Predicting Question Acceptance Rates,” vol. 424, no. August 2015.
[13] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, “The Comparison of Text Mining With Naive Bayes Classifier , Nearest Neighbor , and Decision Tree to Detect Indonesian Swear Words on Twitter.” Conference: 2017 5th International Conference on Cyber and IT Service Management (CITSM),2017.
[14] T. Matsumoto, W. Sunayama, Y. Hatanaka, and K. Ogohara, “Data Analysis Support by Combining Data Mining and Text Mining,” 2017.
[15] S. M. Road, “Techniques on Text Mining,” no. 978, pp. 269–271, 2012.
[16] Y. Zhang, M. Chen, and L. Liu, “A Review on Text Mining.” Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference, 2015.
[17] R. E. Thomas, “Text Mining Improved Clustering Technique using Metadata for Text Mining.”, 2016.
[18] S. Wang *et al.*, “Text mining for identifying topics in the literature about adolescent substance use and depression,” *BMC Public Health*, pp. 4–11, 2016.
[19] A. Paul, A. Ahmad, M. M. Rathore and S. Jabbar, "Smartbuddy: defining human behaviors using big data analytics in social internet of things," in IEEE Wireless Communications, vol. 23, no. 5, pp. 68-74, October 2016.