Text Mining Challenges and Applications, A Comprehensive Review

Dr. Muzammil Khan, Mr. Sarwar Shah Khan, and Dr. Yasser Alharbi

Department of Computer & Software Technology, University of Swat, Pakistan

College of Computer Science & Engineering, University of Hail, Saudi Arabia
Contact# 0092-334-9342615

Summary
Text Mining which is known as text analysis, is defined as the process to extract the proper text patterns from the unstructured text data, which are collected from different written resources. The unstructured text data almost free and found in many different locations such as newspapers, books, the internet, etc. Text mining enables companies to naturally process information and create significant experiences by applied different Artificial Intelligence (AI) Algorithms. Thus, this leads these companies to make appropriate decisions in a data-driven business. In this article, review the main challenges and assessed the applications of major text mining techniques. The applications of each technique are thoroughly evaluated with comprehensive analysis.

Key words:
Text Mining, Information Extraction, Summarization, TM Challenges, TM Application

1. Introduction

Text Mining (TM) is the breakthrough of the computer of new or previously unidentified information, which is automatically extracting information from different written resources (such as natural language text) [1, 2, 3]. Other words, Discover useful new and previously unknown “gems” of information in large text collections. Text mining, also known as text data mining [4], Intelligent Text Analysis [5], Text Analytics [6] or knowledge discovery from textual databases [7], generally refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents [8]. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases [9]. Text mining is loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, the text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, the text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial [10].

Structured data is data that resides in a fixed field within a record or file. This data is contained in the relational database and spreadsheets [11]. The unstructured data usually refers to information that does not reside in a traditional row-column database, and it is the opposite of structured data. Semi-Structured data is the data that is neither raw data nor typed data in a conventional database system [12]. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases or XML files, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents, HTML files, etc. [2]. Text mining is a new area of computer science research that tries to solve the issues that occur in the area of data mining, machine learning, information extraction, natural language processing, information retrieval, knowledge management and classification [13], clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling (*i.e.*, learning relations between named entities) building ontology [14].

In this review paper, the essential consideration on the technical challenges and comprehensive analysis of applications of all technologies of text mining which is highly beneficial for the new researchers.

2. Technologies of Foundation

There is a big difference between the human and computer languages but advances technologies which have commenced closing the gap. In natural language processing (NLP) field has developed technologies that teach computers the natural language. So they can easily understand, analyse, and even produced text [2]. Currently, in text mining searching for and model the hidden patterns and also handle the issues such as information extraction, classification, text representation and clustering [15]. Significantly, some technologies [2] are produced and also applied in the TM are described in details in [7] with their challenges and applications in the following sections.

Manuscript received December 5, 2019
Manuscript revised December 20, 2019
https://doi.org/10.22937/IJCSNS.2020.20.12.15

2.1 Challenges of Information Extraction (IE)

- Form the last several years, the majority of research focused on information extraction in English. The growing a large quantity of textual data in other languages results in shifting of the focus to non-English information extraction and language-independent "multilingual" information extraction technologies. Information extraction in languages other than English is, generally, more complicated, and the performance of non-English information extraction systems is normally lower. This is primarily because of the deficiency of core NLP factors and underlying lingual resources for various languages, but most of all due to the several lingual phenomena that are non-existent in English, which contains, "inter alia" [18]:
- Lack of whitespaces, which complicates word boundary disambiguation, "e.g., in Chinese" [18];
- Productive compounding, which complicates the morphological analysis, as in German, whereas around 10–15% words are combines whose decomposition is essential for higher-level Natural Language Processing [18].
- Complex proper name declension, which complicates named-entity normalisation. Typical for Slavic languages and exemplified in [20], which describes methods for lemmatisation and corresponding Polish person names.
- Zero anaphora, whose resolution is crucial in the context of CO task. The typical for Slavic, Romance and Japanese languages,[19,21];
- There are a lot of languages, and for all languages, the processing tools are not available [17].
- A lot of done on information extracting which is more complex conceptions such as events, opinions, sentiments, entities, and relationships [17].
- Disambiguating extracted mentions (Entities over time and tracking mentions) is common because the text is inherently ambiguous, must disambiguate and merge extracted data, understanding, correcting, incorporating user feedback, explanations, and maintaining extracted information like provenance [17].

2.1.1 Application of IE

The large numbers of applications of information extraction used in the wide range of domains. The structure and particular type of information to be extracted depending on application requirements. The information extraction example applications are described below [22]:

a. **Biomedical:** The researchers often require to sift through a huge amount of scientific articles to search discoveries associated with specific "genes, proteins or other biomedical entities". To assist this effort, simple search based on keywords matching may not be sufficient because the biomedical entities often have ambiguous names and synonyms, so it's very difficult to retrieve the accurate and relevant documents [30].

b. **Financial professionals:** The professionals often required to seek specific pieces of information from news reports to help their day-to-day decision making. For instance, "a finance company may need to know all the company takeovers that take place during a certain time span and the details of each acquisition". Automatically discovering the important information from documents needs the standard information extraction tools such as named entity recognition and relation extraction [22].

c. The feasibility to found an automatic set of services aimed at associating **weather forecasting** with event detection and IE applying social media streams [24].

d. **E-recruitment or job searching:** The large number of user are using the information extraction technique for e-recruitment, or job searching and the number is increasing day by day across the world [25].

e. **Electronic Medical Records (EMRs):** The medical Information recorded in (EMRs), clinical reports, and summaries have the possibility of revolutionising health-related research. Information extraction of EMR data can be used for disease registries, epidemiological studies, drug safety surveillance, clinical trials, and healthcare audits [17, 23].

f. The IE method applying to a corpus of **conference announcement** posted on conference web newsgroups [17, 26].

g. **World Wide Web**: The search engines have become an essential component of people's daily lives, and the search behaviours of users are easy to understand now [22].

h. Extracts **advertisements** information form **newspaper** [27] and extraction the news Item from Web Newspapers in text mining [28]

i. Newswire reports [17]

j. **Electronic mail [17, 29]:** The information extraction techniques are also used for email data. Email is one of the easiest and common way of communicating via text. The estimation is that an average computer user receives 35 to 45 emails every day. Several TM applications need to take emails as inputs, such as email filtering, email routing, email analysis, newsgroup analysis, and information extraction from email.

k. **Digital Libraries** (DL**):** The information extraction in digital libraries, "metadata" means is structured data, which helps the user discover and to process images and text documents [73]. With the "metadata" information, search engines can recover the needed documents more accurately. The scientists and librarians require to use substantially manual efforts and lots of time to produce the metadata for the text documents. "To relieve the hard labour, many attempts have been made towards the automatic generation of metadata based on information extraction techniques" [32].

l. **Personal Profile Extraction**: Person information management is a significant topic in both the industrial and the research community. A person can have different but related types of information: person profile such as homepage, affiliation, position, portrait, documents, and publications), contact information such as an address, telephone, fax number, and email. Nevertheless, the data is commonly hidden in heterogeneous and distributed web pages [33].

2.2 Challenges of Topic Tracking

The traditional methods of topic tracking specially used for event detection in the social media context, pose unique problems because of the distinguishing characteristics of textual data in social media such as follow [36].

Time Sensitivity: Unlike conventional textual data, the text in social media has real-time nature. Besides communication and sharing new ideas, users in social networks might post their views and feelings about the wide variety of recent events various times in the day [38]. "Users may want to communicate instantly with friends about What they are doing (Twitter) or What is on their mind (Facebook)".

Short Length: The majority of the social networks platforms limited the length of posts. For instance, "Twitter allows users to post tweets that are no longer than 140 characters". Unlike the standard text with a large number of words and their resulting statistics, short messages consist of few phrases or sentences. They can't provide enough information for effective similarity evaluate, the basis of various text processing techniques [35].

Unstructured Phrases: In contrast with well-written, structured, and edited news releases, social posts might include irregular, abbreviated words, polluted and informal content, a large number of meaningless messages, improper sentence structures, large amounts of spelling and grammatical errors, and mixed languages, which negatively affect the performance of the detection methods [36, 25].

2.2.1 Applications of Topic Tracking

There are several applications where topic tracking can be employed, such as:

- The topic tracking is commonly used in **industries.** Primarily, the industry can alert the companies anytime when the competitor is in the news. This method allows companies to know about the changes in the market or competitive products [2].
- Radio broadcasts [34]
- The **businesses** might want to track news on their products and own company [2].
- TV broadcasts [34]
- **Medical industry** [2].
- Newspaper and journal articles [67, 37]
- Education is the latest research area, which highly involves topic tracking [2].
- Newswires reports [34]
- Social networks such as Facebook, Instagram, Skype and Twitter [36]

2.3 Challenges of Summarization

- The main problem in document summarization lies in recognizing the most significant parts of the text and the lesser one [39].
- The major problem in summarization is, the computer system is capable of identifying the places, people and time, but it still very complicated to instruct the tools to examine semantics properly and to interpret meaning [3].
- The new technologies are developing for information communication with the high speed, a huge number of e-documents (electronic documents) are on-line available, and the users are facing difficulty to discover relevant information. Furthermore, internet technologies have provided huge collections of text on a variety of stories. So, many users get exhausted reading a huge amount of text document that they may ignore reading the important and interesting stories. Hence, the robust text summarization is currently highly needed in this generation [39].
- The automatic summary generation has a lot of challenging issues such as temporal dimension, redundancy, sentence ordering, co-reference, etc. that required specific attention. When summarizing multiple text documents, thereby making this task more complex [39].
- The text summarization is very important and useful in other technologies, such as information retrieval, text classification and Question Answering [41].

2.3.1 Applications of summarization

Email: Email-based text summarization is the kind of text summarization, were summarized the conversations of email. For communication, email is one of the effective ways because of its lack of cost, and speedily delivered [39].

Personalized summaries (PS): PS contains personal information about the user. The users have different requirements, so such "systems after determining the user's profile select the important content for generating the summary." In update summaries, it is considered that users have the basic information about the specific story and needs only the current updates regarding the story [39].

Text Summarization (TS): The TS and "Sentiment Analysis (SA)" together form "opinion mining," and they work together for producing such summarizes. In such summarizes opinions are initially observed and classified on the basis of subjectivity "whether the sentence is subjective or objective" and then on the basis of polarity "positive, negative or neutral" [16].

Survey summaries (SS): The SS has acquired a normal overview of a specific story or topic. These are generally lengthy as they contain the most significant facts, regarding persons, places or any other entities., Wikipedia articles, biographical summaries and Survey summaries, all these summaries come under this class [39].

The summarizing **news articles** as a way to select sentences in an extractive summarization [40].

2.4 Challenges of Classifications

- Feature vectors must acquire complex semantics of text.
- Binary or numeric characteristics obtained from word of phrase frequency must be noise-free.
- The structure of classifier like "Naïve Bayes" uses high dominance of model rather than hidden text characteristics thereby suppressing performance of the classifier.
- Information retrieval systems experience diverse nature of texts with highly variable content, quality and length.
- The performance of the machine learning model will be degraded if an ill-sampled data is presented to it while training and some classes are not observed by it.
- During the training phase of the machine learning model, the attained knowledge often escaped from given real data and hence resulting in deterioration of performance.
- The classification of text sometimes goes more subjective due to the presence of unknown classes and outliers.
- The volume of training data plays a great role in learning a model. Training data must be labelled and big enough to cover all the upcoming classes.
- Human labelers expressively bias the training data which may yield a wrong training of the model.
- In-text classification issue consists of a huge number of closely related classes, for instance: "Google directory contains around two billion categories in a deep hierarchy hence, making it difficult to correctly classify the test data through machine learning".
- In the case of large volumes of training data, stemming and lower-casing may decline the performance of statistical ML methods. For instance: words like "oxygenate" and "oxygenation" yields "oxygen" as an outcome of stemming, thereby thrashing the real semantics of text.

2.4.1 Applications of Text Classification

News article classification: Classifying a huge amount of unclassified archival documents such as academic papers, legal records and newspaper articles. For instance, newspaper articles can be classified as "features", "sports" or "news" [42].

Automatic email filtering: A lot of machine learning classification techniques are recently used to successfully detect and filter spam emails [43].

Webpage classification: The Webpages classification is the process of classifying web documents into predefined categories based on their content [44].

Word sense disambiguation (WSD): There were identified a range of linguistic phenomena such as "preferential selection or domain information" that is relevant in resolving the ambiguity of words [45].

2.5 Challenges of Clustering

Sparse Feature Vector: The number of words being very less, the feature vector produced from the short text is normally sparse in nature. The sparsity of the feature vector is the main issue in clustering short text data, and resolving this issue is a challenging job.

Synonymy's: There is **a** couple or more words having a similar meaning. For example; words Beautiful, Attractive, Pretty, Lovely, Stunning have the same meaning. So it is also a challenging task to decide in which cluster such words would be placed especially in the case when such words are found in short texts.

The identifying of distance measurement: The distance measurement (numerical attributes) which is used as a standard equations like "eucledian", "manhattan", and "maximum" distance metrics. All the three are special cases of "Minkowski distance". But identifying the measurement for categorical attributes is very complicated [52].

The number of clusters: Identifying the number of clusters is a complicated task if the number of class labels is not known in advance. Carefully analyze the number of clusters is necessary to generate actual results. Else, it is found that heterogeneous tuples may combine or similar type's tuples may be broken into many. This could be catastrophic if the method used is hierarchical. Because in the hierarchical method if a tuple gets wrongly combined in a cluster that action cannot be undone. While there is no perfect way to examine the number of Clusters. There are few statistics which helps to analyze the process such as the Cubic Clustering Criterion (CCC), the Approximate Overall R-Squared and the Pseudo-F statistic [52].

Lack of class labels: For real datasets "relational in nature as they have tuples and attributes" the distribution of data has to be done to understand where the class labels are? [52]

Database Structure: The real-life data may not always contain clearly identifiable clusters. Also, the order in which the tuples are arranged may affect the results when a method is executed if the distance measures used is not perfect. With a structureless data (for e.g. Having large of missing values), even identification of an appropriate number of clusters will not yield good results [52].

Database different Types of attributes: The databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other kinds such as "nominal", "ordinal", "binary" etc. So these attributes have to be converted to categorical type to make computation simple [52].

Selecting the initial clusters: For partitional technique, most of the methods mention k initial clusters to be randomly selected. The careful and comprehensive examination of data is needed for the same. Also, if the initial clusters are not properly selected, then after a few iterations, it is found that clustering may even be left empty [52].

2.5.1 Applications of Clustering

Medical Field: In medical imaging, clustering analysis is very significant used to differentiate among different kinds of tissues and blood (PET Scans). The clusters also applied in the analysis of antimicrobial activity to analyze the patterns of antibiotic resistance [46].

Business and Marketing: Partitioning the general population of consumers into market segments and to better understand the relationship among various groups of customers can be done with the help of clustering and the analysis will be used by many of the market researchers [47].

World Wide Web: In social networks, the clustering is significantly used to recognize communities within large groups of people. Clustering may be used to create a more relevant set of search results compared to normal search engines like Google [48].

Image Processing: The clustering is very importantly used in image segmentation to separate a digital image into distinct regions for object recognition or edge detection [49].

Social Science: Clustering is used in crime analysis to discover the areas where the larger incidences of specific kinds of crime. By identifying these "distinct areas" or "hot spots" where a similar crime has happened over the specific time period, in this way effectively manage law enforcement resources [50].

Education: In education, the clustering analysis is applied to identify the different groups of students or schools with similar properties.

Climatology: Clustering algorithms are applied in the analysis of weather and climate to identify discrete groups of "atmospheric and oceanic" structures and evolutions that happen more often than would be expected [51].

Bioinformatics: In bioinformatics, the clustering analysis is used to identify **a** gene, protein network and gene expression data.

Customer Recommendation: In customer recommendation system objects are customers and attributes are products purchased by customers. This system is helpful to better understand the buying behaviour of consumers. So it can help to analysis which product is highly purchased in which region, age group etc.

Compression of Data: Cluster analysis is useful in data compression. The information which is present in the data set is abstracted in clusters.

2.6 Challenges of Text visualization

To developed and design the proper visualization techniques are very complicated task, which as five major issues.

Usability: The development of InfoVis has been driven by real-world applications and user requirements. Typically, the users are mostly involved with the "visualization system or toolkit" to achieve his analysis tasks. To facilitate the visualization, designers design an effective visualization toolkit /system, the scientists have designed the set of advanced empirical evaluation techniques and design study techniques, as well as several design theories [70].

Visual Scalability: The Visual scalability is defined as the capability of visualization tools to effectively display large data sets in terms of either the number or the dimension of individual data elements. Scalability is a fundamental issue for Information Visualization, especially with the boom in big data analytics [71].

Integrated analysis of heterogeneous data: The heterogeneous data are data from multiple sources and in varying formats. Integration and analysis of heterogeneous data are one of the most significant issues for versatile applications [72].

In-situ visualization: In-situ visualization incrementally produced visual presentations when new data arrive. The effective way to analyze and understand "streaming data". The streaming data is defined as "data with a regular rate of flow through hardware". Typical examples include log data such as search sensor logs and logs, periodically updated social media data (e.g., tweets), and stock data. Due to the rapid rate of incoming data, and the large size of the data stream model, analysis of such streaming data poses a big problem in visualization [70].

Errors and uncertainty: Real-world data sets often contain errors and/or uncertainties, for instance, noisy and inconsistent social media data provided by users every day, imprecise data from sensors, or imperfect object recognition in video streams. For instance, data transformation, data filtering, or data sampling may generate errors and inconsistencies into the visualization, which is another major source of uncertainty [69].

2.6.1 Application of Text visualization

- The **information visualization** is used by the government for identifying terrorist activities or to retrieve information about criminal and crimes that may have been previously thought unconnected. It provides them with a map of all possible relationships among suspicious activities [53].
- **Visualizing Social Networks**
- **Social network detection**

2.7 Challenges Question Answering

There are various challenges are present which plays a vital role in question answering system such as

Question Classes: Various schemes are considered in answering questions. Based on the class, a question falls, a specific scheme can be used in answering it, and such a scheme may not work for another question of a different class. Hence, a complete understanding of what class a question falls is required to be able to answer questions correctly [55].

Question Processing: In "natural language", the same question may be posed in various ways. The question may be asked "assertively" or "interrogatively". The need to understand the semantics comes up, that is knowing what the query focus is before trying to discover a solution to the query. The whole act involved in discovering what category a question belongs to is known as "Question Processing" [54].

Data Sources for QA: The answers to questions asked or posted to a question answering system are sourced from a knowledge base. The "base" or "source" must be relevant and exhaustive. It may be a collection of text documents, the web, or a database from where we can get the answers [56]

Answer Extraction: The class of query asked informs the type of answer that will be extracted from any source a QA is using. So the required to understand and generate the expectation of the user from the question provided is the aim in "Answer Extraction" [57]

Answer Formulation: Basic extraction can be sufficient for certain queries. For some questions, the solutions are extracted in parts from various bases which are then merged to answer asked questions [54]

Real-time Question Answering: Real-time question answering needs instantaneous answers (replies) to questions posed in "Natural Language". In cases where instantaneous replies are needed, the required to answer questions regardless of it complexity in seconds comes up,therefore the required for architectures that can generate valid answers in given time constraint [58].

Multilingual (or cross-lingual) question answering: "Cross-lingual QA" or "Multilingual QA" involves retrieving answers from sources various from the language the query was expressed. Different English Question Answering systems data sources exist, but some other languages still lack such resources, e.g. "Urdu, Hindi"[59]

Interactive QA: Additional information about asked questions from users can help guide the question-answering process. Therefore the require for an interactive system that is not boring in the sense that it relates back and clear doubts in case it discovers the question ambiguous [54]

Advanced reasoning for QA: In advanced reasoning, the QASs does more than producing what it discovers in the dataset. It does more by learning facts and using reasoning to produce new facts which can be applied to better answer posed questions [54]

Information clustering for QA: Retrieving precise information for simple questions has distorted to a tough and resource expensive act due to excessive information growth in the web. Clustering decreases the search space and by so doing reduce the workload of methods [54]

User profiling for QA: The need to tailor QASs replies around the users that is asking the question is another issue. The intention of the user can be known by analyzing the user's earlier queries. To do this, there is the required to develop the user's profile [60]

2.7.1 Applications of QUESTION ANSWERING

Web applications: The companies can mostly applied Q&A methods internally for staff who are searching the answers for the common questions.

Education: The question answering techniques are also used in education [2].

Medical: The Q&A techniques are very useful in the area of medical, where the people are frequently asked questions [2].

Banks, insurance and financial markets etc. [2]

2.8 Challenges of Association rule mining

There is a lot of challenges of ARM.

- In a single level or multiple levels of association rules; the most significant problem is concerned with accurate data source in an appropriate data format. Which encoding technique should be applied to convert the transaction tables is a major problem because these encode tables are applied to support the concept hierarchy of multiple levels [67].
- The problem to develop/design techniques for multiple-level association rules to decrease the number of iteration and to achieve time efficiency. The time efficiency can be achieved by reduction of database scans at each level. The redundancy of association rules is a major problem in "association rule discovery".
- The techniques which have lower CPU overhead and decrease the I/O overhead associated with previous techniques are desirable.

- The ARM is to the finding of the usefulness of association patterns the task of decision making is found to be incorporated flawlessly within the “association mining process”.
- Single access to data: In data streams, data are coming continuously with fast speed and in large volume. As a consequence, in many cases, it is impractical to store all data in persistent media and in other cases, it is too expensive to “randomly” access data multiple times. The main problem is to find frequent itemsets, while the data can only be assessed once [66].
- Unbounded data: The characteristic of data streams is that data are unbounded. In comparison, storage that can be applied to find or maintain frequent itemsets is limited. The problem is to use limited storage to find frequent dynamic itemsets from unbounded data [66].
- Real-time response: The data stream applications are commonly time-critical, there are requirements on response time. For some restricted scenarios, techniques that are slower than the coming data rate are useless. The issue is efficiently mine frequent itemsets in real-time [66].

2.8.1 Applications of Association rule mining

Market Basket Analysis (MBA): MBA is the most common application of ARM that finds the relations between the items obtained by the customers [61,65].
Intelligent transportation system (ITS): ITS is innovative information technology, processor technology integration and switch technology that is used to the entire transportation system. ITS is build within the accurate, versatile role, actual time and well-organized integrated transportation controlling system [61].
Web Log Data: The tremendous use of the internet has completed the automatic knowledge extraction from web log files [62]
Identification of Frequent Disease Data: The mining is a process of describing or extracting interesting information, “information” or “patterns” from data in a huge database [63].
Computer-Aided Diagnostic System (CAD) of Breast Cancer: With the extensive application of computer and knowledge, the quantity of data generated by many disciplines has enlarged quickly. In order to mine valuable knowledge from that data, DM or TM techniques are recycled. By using data removal methods, hopeful results have been obtained in the treatment, diseases diagnosis, image examination, drug growth, organ transplantation, scientific study etc. [64].
Recommender frameworks: “Recommender systems are designed for offering products to the potential customers”. Collaborative Filtering is known as a common way in Recommender framework, which offers recommendation made by similar users in the case of entering time and previous transactions [68].
Web utilization mining, interruption recognition, Continuous generation, and bioinformatics.

Table 1. Comprehensive analysis of Text mining Application

Name Of Applications	IE	TT	SUM	CLASS	CLUS	CL	InfoV	Q&A	RM	NLP
MEDICAL										
FAQ's	✓			✓		✓		✓		✓
Drug design	✓				✓	✓				✓
New treatment		✓				✓			✓	✓
Biomedical Research	✓				✓				✓	✓
Patient Record / electronic medical records (EMRs),	✓	✓			✓			✓		✓
Computer Aided Diagnostic (CAD) System for Breast Cancer									✓	✓
BUSINESS										
Competitive Examination		✓	✓							✓
Media impact / Examination		✓								✓
Current Awareness		✓								✓
Intellectual property infringement	✓	✓			✓					✓
Customer Recommendation System					✓				✓	✓
Customer support for FAQ's	✓		✓	✓	✓				✓	✓
Social network detection							✓			✓
Content personalization										✓
Business Marketing		✓			✓				✓	✓
GOVERNMENT										

Homeland security: finding terrorist networks	✓	✓			✓	✓	✓			✓
Law enforcement: crime prevention	✓	✓			✓	✓	✓			✓
Intelligence analysis	✓								✓	✓
Security Application	✓			✓	✓					✓
Banks, insurance and financial markets	✓			✓				✓	✓	✓
EDUCATION										
Research on a topic		✓	✓	✓						✓
Citation analysis	✓				✓		✓			✓
FAQ's	✓			✓	✓			✓		✓
Education Data Mining		✓			✓			✓		✓
Social Science / Twitter		✓			✓					✓
Climatology					✓					✓
Computer Science		✓			✓					✓
Spatial Data Analysis					✓					✓
Survey Summarizing			✓							✓
Sentiment / opinion mining			✓							✓
Web Search	✓			✓	✓			✓	✓	✓
Conference Announcement	✓	✓								✓
Digital Libraries	✓									✓
OTHER										
TV & Radio Broadcast	✓	✓								✓
Data visualization			✓				✓		✓	✓
Human Resource Management			✓							✓
Electronic mail	✓		✓	✓						✓
Person Profile Extraction	✓		✓							✓
Weather forecasting	✓									✓
Word sense disambiguation (WSD)				✓						✓
Advertisements information form newspaper	✓	✓		✓						✓
Compression of Data					✓					✓
Newswires reports		✓								✓
E-recruitment, / Job Searching	✓									✓
Journal Articles		✓	✓							✓
Questioning in Natural Language								✓	✓	✓
Multilingual	✓	✓	✓	✓	✓			✓	✓	✓

Color	Denoted Name	Full Name
🟥	IE	Information Extraction
🟧	TT	Topic Tracking/Detection
🟨	SUM	Text Summarization
🟩	CLASS	Classification (Categorization)
🟩	CLUS	Clustering
🟦	CL	Concept Linkage
🟦	InfoV	Information Visualization
🟦	Q&A	Question Answering
🟪	ARM	Association Rule Mining
🟧	NLP	Natural Processing Language

3. Conclusion

Text Mining is considered as AI technology that aims to extract, analyse, and process a tremendous amount of unstructured text data to be used invaluable business insights companies. Internet is an example of Written resources that generate daily an enormous amount of unstructured text which needs to be transferred into readable and understandable information to the machines. This review provided a brief introduction about the Text Mining, terminologies and its process for both the text mining and information extraction with a variety of its techniques and applications to understand the basic concepts. In addition to this, review paper has explored in detail the variety of related research areas with their used applications, and challenges. Text analysis is a new trend of artificial intelligence that grows rapidly with many of improvement in its applications and technology.

4. References

[1] Witten, I. H. Text Mining **(2004)**.

[2] Weiguo, Fan, Wallace Linda, Rich Stephanie, and Zhang Zhongju. "Tapping into the Power of Text Mining." *Journal of ACM, Blacksburg* **(2005)**.

[3] Petr Knoth, Phil Gooch"An Introduction to Text Mining Research Papers" Mendeley 22 September **2015**

[4] MUZAMMIL KHAN, A., B. MUSHTAQ RAZA, and C. NASIR RASHID. "Appropriate length of text line with special relationship to eye blink to reduce maximum focus loss." *ICOMP 2010: proceedings of the 2010 international conference on internet computing (Las Vegas NV, July 12-15, 2010).* (**2010)**

[5] Feldman, Ronen, and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)." In KDD, vol. 95, pp. 112-117. **(1995)**.

[6] Antonio Moreno, Teóflo Redondo "Text Analytics: the convergence of Big Data and Artifcial Intelligence" Article in International Journal of Interactive Multimedia and Artificial Intelligence March **(2016).**

[7] Sukanya, M., and S. Biruntha. "Techniques on text mining." 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). IEEE, **(2012)**.

[8] Akilan, A. "Text mining: Challenges and future directions." 2015 2nd International Conference on Electronics and Communication Systems (ICECS). IEEE, **(2015)**.

[9] Azeroual, Otmane. "Text and Data Quality Mining in CRIS." Information 10.12: 374**(2019).**

[10] Perivasamy, S. K. T. "An Efficient Clustering Algorithm for Text Mining Using Greedy Approach." International Journal of Advanced Research in Computer Science & Technology (IJARCTS) 12 **(2014)**.

[11] Lindell, Jim. Analytics and Big Data for Accountants. John Wiley & Sons, **(2018)**.

[12] Zhang, Kuo, et al. "A Semantics Enabled Intelligent Semi-structured Document Processor." International Conference on Trustworthy Computing and Services. Springer, Berlin, Heidelberg, **(2013)**.

[13] Miloš Radovanović, Mirjana Ivanović "TEXT MINING:APPROACHES AND APPLICATIONS" Novi Sad J. Math.Vol. 38, No. 3, 227-234 **(2008)**

[14] Islam, Shaziya, and Manpreet Kaur. "Knowledge-Based Text Mining in Getting Perfect Preferences in Job Finding." Recent Findings in Intelligent Computing Techniques. Springer, Singapore. 43-54 **(2018)**.

[15] Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining." Ldv Forum. Vol. 20. No. 1. 2005.

[16] Khan, Sarwar Shah, et al. "CHALLENGES IN OPINION MINING, COMPREHENSIVE."

[17] Ahmad, Peerzada Hamid, and Shilpa Dang. "A Comparative Study on Text mining Techniques." International Journal of Science and Research, ISSN: 2319-7064.(2014)

[18] Piskorski, Jakub, and Roman Yangarber. "Information extraction: Past, present and future." Multi-source, multilingual information extraction and summarization. Springer, Berlin, Heidelberg, 2013. 23-49.

[19] Iida, Ryu, and Massimo Poesio. "A cross-lingual ILP solution to zero anaphora resolution." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[20] Piskorski, J.,Wieloch, K., Sydow, M.:Onknowledge-poor methods for person name matching and lemmatization for highly inflectional languages. Inf. Retr. 12(3), 275–299 (2009)

[21] Zavarella, V.,Tanev, H.,Piskorski, J.:Event extraction for Italianusing acascade of finite-state grammars. In: Proceedings of FSMNLP 2008, Ispra (2008)

[22] Jing Jiang "INFORMATION EXTRACTION FROM TEXT" Singapore Management University C.C. Aggarwal and C.X. Zhai(eds.),Mining Text Data, DOI 10.1007/978-1-4614-3223-4_2 © Springer Science+Business Media, LLC 2012.

[23] Ford, Elizabeth, et al. "Extracting information from the text of electronic medical records to improve case detection: a systematic review." Journal of the American Medical Informatics Association 23.5 (2016): 1007-1015.

[24] Rossi, C., et al. "Early detection and information extraction for weather-induced floods using social media streams." International journal of disaster risk reduction 30 (2018): 145-157.

[25] Awan, Ahmed, et al. "A New Approach to Information Extraction in User-Centric E-Recruitment Systems." Applied Sciences 9.14 (2019): 2852.

[26] Korde, Vandana. "Information extraction for personalised services based on conference alerts." International Journal of Data Mining, Modelling and Management 8.1 (2016): 93-105.

[27] FabrizioSebastiani.2002. Machine learning in automated text categorization. ACM computing surveys (CSUR) 34,1(2002),1–47.

[28] M Khan, AU Rahman, A Ahmad, S Khan A Content-based Technique for Linking Dual Language News Articles in an Archive. Journal of Information Science

[29] Laclavík, Michal, et al. "Email analysis and information extraction for enterprise benefit." Computing and informatics 30.1 (2012): 57-87.

[30] Gong, Lejun. "Application of biomedical text mining." Artificial Intelligence: Emerging Trends and Applications (2018): 417.

[31] Khan, Muzammil, and Arif Ur Rahman. "A Systematic Approach Towards Web Preservation." Information Technology and Libraries 38.1 (2019): 71-90.

[32] Muzammil Khan, Arif Ur Rahman, M. Daud Awan." Exploring the Digital World of Newspaper Archives", A Science and Technology Journal, Portugal, Vol.32 No.6, pp. 430-449

[33] Emami, H., H. Shirazi, and A. Abdollahzadeh. "A Semantic Approach to Person Profile extraction in Farsi Text." The Journal of Information Systems and Telecommunication (JIST), pp: 232 243 (2017).

[34] JingQiu, LeJian Liao, XiuJie Dong, (2008), ― Topic Detection and Tracking for Chinese News Web Pages‖, International conference on Advanced Language Processing and Wen Information Technology, IEEE

[35] Huang, Jiajia, et al. "A probabilistic method for emerging topic tracking in microblog stream." World Wide Web 20.2 (2017): 325-350.

[36] Zarrinkalam, Fattane, and Ebrahim Bagheri. "Event identification in social networks." Encyclopedia with Semantic Computing and Robotic Intelligence 1.01 (2017): 1630002.

[37] Anup Kumar Kolya, Asif Ekbal, Sivaji Bandyopadhyay, (2009), ―A Simple Approach for Monolingual Event Tracking System in Bengali‖, 8th International Symposium on Natural Language Processing, IEEE

[38] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter", Comput. Intell., vol. 31, no. 1, pp. 132164, 2015.

[39] Gambhir, Mahak, and Vishal Gupta. "Recent automatic text summarization techniques: a survey." Artificial Intelligence Review 47.1 (2017): 1-66.

[40] Nazari, N., and M. A. Mahdavi. "A survey on Automatic Text Summarization." Journal of AI and Data Mining 7.1 (2019): 121-135.

[41] Perea-Ortega, José M., et al. "Application of text summarization techniques to the geographical information retrieval task." Expert systems with applications 40.8 (2013): 2966-2974.

[42] Ramdass, Dennis, and Shreyes Seshasai. "Document classification for newspaper articles." Document classification for newspaper articles (2009).

[43] Dada, Emmanuel Gbenga, et al. "Machine learning for email spam filtering: review, approaches and open research problems." Heliyon 5.6 (2019): e01802.

[44] Patel, Keyur J., and Ketan J. Sarvakar. "Web page classification using data mining." International Journal of Advanced Research in Computer and Communication Engineering 2.7 (2013): 2513-2519.

[45] Morariu, D., R. Cretulescu, and Macarie Breazu. "Word Sense Disambiguation for Text Mining." The third international conference in Romania of "Information Science and Information Literacy". 2012.

[46] Alashwal, Hany, et al. "The Application of Unsupervised Clustering Methods to Alzheimer's Disease." Frontiers in computational neuroscience 13 (2019).

[47] Ližbetinová, Lenka, et al. "Application of cluster analysis in marketing communications in small and medium-sized enterprises: An empirical study in the Slovak Republic." Sustainability 11.8 (2019): 2302.

[48] Belk, Marios, et al. "Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques." Journal of Systems and Software 86.12 (2013): 2995-3012.

[49] Khan, Sarwar Shah, et al. "Hyperspectral image classification using nearest regularized subspace with Manhattan distance." *Journal of Applied Remote Sensing* 14.3 (2019): 032604.

[50] Alkhaibari, Adel Ali, and Ping-Tsai Chung. "Cluster analysis for reducing city crime rates." 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT). IEEE, 2017.

[51] Straus, David M. "Clustering Techniques in Climate Analysis." Oxford Research Encyclopedia of Climate Science. 2019.

[52] Parul Agarwal, M. Afshar Alam, Ranjit Biswas "Issues, Challenges and Tools of Clustering Algorithms" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011 ISSN (Online): 1694-0814 www.IJCSI.org

[53] Khan, Muzammil, and Sarwar Shah Khan. "Data and information visualization methods, and interactive mechanisms: A survey." International Journal of Computer Applications 34.1 (2011): 1-14.

[54] Ojokoh, Bolanle, and Emmanuel Adebisi. "A Review of Question Answering Systems." Journal of Web Engineering 17.8 (2018): 717-758.

[55] Stupina,A.A., Shigina,A.A., Shigin,A. O., Karaseva, M. V., and Korpacheva, L. N. 2016. Question Answering system. IOP Conf. Ser.: Mater. Sci. Eng. 155 012024

[56] Jurafsky, D., and Martin, J.H. 2015. Question Answering. In: Computational Linguistics and speech recognition. Speech and Language Processing. Colorado.

[57] Xianfeng, Y., and Pengfei, L. 2016. Question Recommendation and Answer Extraction in Question Answering Community. International Journal of Database Theory and Application, 9 (1):35–44

[58] Agichtein, E., and Savenkov, D. 2016. CRQA: Crowd-Powered Real-Time Automatic Question Answering System. HCOMP

[59] Foster, G. F., and Plamondon, L. 2003. Quantum, a French/English CrossLanguage Question Answering System. CLEF.

[60] Bergeron, J., Schmidt, A., Khoury, R., and Lamontagne, L. 2016. Building User Interest Profiles Using DBpedia in a Question Answering System. AAAI Publications, The Twenty-Ninth International Flairs Conference.

[61] Shaukat, Kamran, Sana Zaheer, and Iqra Nawaz. "Association rule mining: an application perspective." International Journal of Computer Science and Innovation 2015.1 (2015): 29-38.

[62] Yan Hai, Xiu-li Li, "A General Temporal Association Rule Frequent Itemsets Mining Algorithm", IJACT, Vol. 3, No. 11, pp. 63 ~ 71, 2011.

[63] Y Wang, H. Y Wang and D. W Zhang, et al, "Research on Frequent Itemsets Mining Algorithm based on Relational Database", Journal of Software, vol. 8, no. 8, pp. 1843-1850, 2013.

[64] J. Y Li, J. P Wang and H. X Pei, "Data Cleaning of Medical Data for Knowledge Mining", Journal of Networks, vol. 8, no. 11, pp. 2663-2670, 2013.

[65] Hlaing, Moe Moe. "ECLAT based market basket analysis for electronic showroom." (2019).

[66] Rashid, Md Mamunur, Iqbal Gondal, and Joarder Kamruzzaman. "Mining associated sensor patterns for data stream of wireless sensor networks." Proceedings of the 8th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks. 2013.

[67] Yadav, R. U. C. H. I. K. A., K. A. N. W. A. L. Garg, and P. R. I. Y. A. N. K. A. Khurana. "Issues and Challenges associated with Association Rules Mining Algorithms." (2014).

[68] Varzaneh, Hossein Hatami, et al. "Recommendation systems based on association rule mining for a target object by evolutionary algorithms." Emerging Science Journal 2.2 (2018): 100-107.

[69] Wu, Y., Yuan, G.-X., Ma, K.-L.: Visualizing flow of uncertainty through analytical processes. IEEE Trans. Vis. Comput. Graph. 18(12), 2526–2535 (2012)

[70] Khan, Sarwar Shah, et al. "Visual Features Comparison of Smartphone and Tablet in Visual Mobile Data Mining Framework." *IJCSNS* 20.5 (2020): 44.

[71] Khan, Muzammil, Sarwar Shah Khan, and M. Daud Awan. "Comparative Exploration of Features for Data Mining Results by Legend Navigation Interactive Technique." *International Journal of Database Theory and Application* 9.9 (2016): 49-58.

[72] Khan, Muzammil, et al. "EVALUATING INTERACTIVE VISUALIZATION TECHNIQUES ON SMALL TOUCH SCREEN DEVICES." International Journal of Grid and Distributed Computing (IJGDC) 12.02 (2019): 31-48.

[73] Khan, Muzammil, et al. "The role of news title for linking during preservation process in digital archives." *Library Hi Tech* (2020).