Role of Text Mining in Disease Prediction

Amit Kumar Tyagi[1] [0000-0003-2657-8700]
Vellore Institute of Technology, Chennai Campus, Chennai, 600127, Tamilnadu, India.
amitkrtyagi025@gmail.com

Sravanti Reddy[2]
Assistant Professor, VJIET, Hyderabad, Telangana, India
sravanthireddy.k8@gmail.com

Shabnam Kumari[3]
Research Assistant, AARIN, India.
shabnam.kt25@gmail.com

Abstract. Due to rapid creation of digital data by Internet of Things devices or smart devices, many new modern mining strategies/ techniques require to handle/ analyse this large amount of data. Note that more than 90 percent of today's data is in present (generated) unstructured or semi-structured data format (most of part of this data is being generated only in the past decade). The discovery of appropriate patterns and trends to analyse the text documents from this large big data (i.e., large volume of data) is a big issue. Text mining is a process of extracting interesting and non- trivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text (also other data format) and discover valuable information for future prediction and decision making process. Basically, there are two terms used in making or extracting some relevant information from a data-set, i.e., prediction modelling, and text mining. Predictive models are often used to detect crimes and identify suspects, after the crime has taken place/ to detect an email, how likely that it is spam. Similarly, text mining used in applications like digital libraries, academic research field, life science, social media, business intelligence, etc. Today's different text mining techniques are available for analysing the text patterns and their mining process, some of them are included here as: document classification (text classification, document standardization), information retrieval (keyword search/ querying and indexing), document clustering (phrase clustering), natural language processing (spelling correction, lemmatization, grammatical parsing, and word sense disambiguation), information extraction (relationship extraction/ link analysis), and web mining (web link analysis), etc.

This article discusses and analyse the text mining techniques and their applications in diverse fields of life. This work discusses about several use-cases, efficient algorithms like apriori algorithm, association rule mining, etc., which is used for frequent item set extraction (information retrieval and information extraction) and rule generation. Also, in result, generated several rules form a collected data-set to predict about a disease (as an example) will be discussed. In last, this work discusses detail descriptions about the terms classification, clustering, regression, association rule mining and outlier detection as a work-flow in analysing the data for producing a decision or making some prediction, also discussing some useful research gaps, challenges, issues (as its concluding remarks).

Keywords*: Clustering, Classification, Data mining tools, Disease prediction, Information extraction, Health care.*

1. Introduction - Text Mining

In the past decade, text mining or Text data mining has received attention from research community (due to recent development in Internet of things devices/ smart things). The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. In the text mining, we can analyse words, clusters of words used in documents, etc., or we could analyse documents and determine similarities between them or how they are related to other variables of interest in the data mining project. In general, text mining will "turn text into numbers" (meaningful indices), which can then be incorporated in other analyses such as predictive data mining projects, the application of unsupervised learning methods (clustering), etc. In simple terms, Text Mining defined as "the process of examining data to gather valuable information". Note that

Text mining, also known as text data mining which involves algorithms of data mining, machine learning, statistics, and natural language processing, attempts to extract high quality, useful information from unstructured formats. Text mining, which is often interchangeably used with "text analytics" is a means by which unstructured or qualitative data is processed for machine use. Text mining can help businesses listen to the right stories by extracting insights from a free text written by or about customers, combining it with existing feedback data, and identifying patterns and trends. Manual analysis alone is unable to capture this level of insight due to the sheer volume and complexity of the available data. We can say,

Text Mining = Statistical NLP + Data Mining Statistical

Here, Natural Language Processing is a set of algorithms for converting unstructured text into structured data objects, whereas data mining is the quantitative methods that analyse these data objects to discover knowledge. Note that text is main vehicle which contain a lot of information or knowledge in it.
Finally, text mining can be discussed in widely accepted definition as: "Text mining is the process of analysing data to capture key concepts and themes and uncover hidden relationships and trends without prior knowledge of the precise words or terms that authors have used to express those concepts". Some of the examples are:

- Risk management
- Knowledge Management
- Prevention of Cybercrime
- Customer Care Service
- Contextual Advertising
- Business Intelligence
- Spam Filtering
- Social Media Data Analysis

Text Mining techniques such as categorization, entity extraction, and sentiment analysis are made use of to extract the useful information and knowledge hidden in text content. Hence, some of the popular Text Mining applications include:

- Enterprise Business Intelligence/Data Mining, Competitive Intelligence
- E-Discovery, Records Management
- Risk management
- National Security/Intelligence
- Scientific discovery, especially Life Sciences
- Search/Information Access
- Social media monitoring

Text Mining and Data Analytics: As discussed above, text mining and text data mining are interchangeably terms. But, text analytics and data analytics both are different terms. In general, data analytics is being on data which may be present in structured or unstructured form, whereas text mining is being done only on unstructured data (on unlabelled data). There are five fundamental steps involved in text mining which are included here as:

- Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
- Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing (or data cleaning) allows us to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
- Convert all the relevant information extracted from unstructured data into structured formats.
- Analyse the patterns within the data via Management Information System (MIS).
- Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organisation.

A clear picture (explanation) of text mining process has been discussed in figure 1.

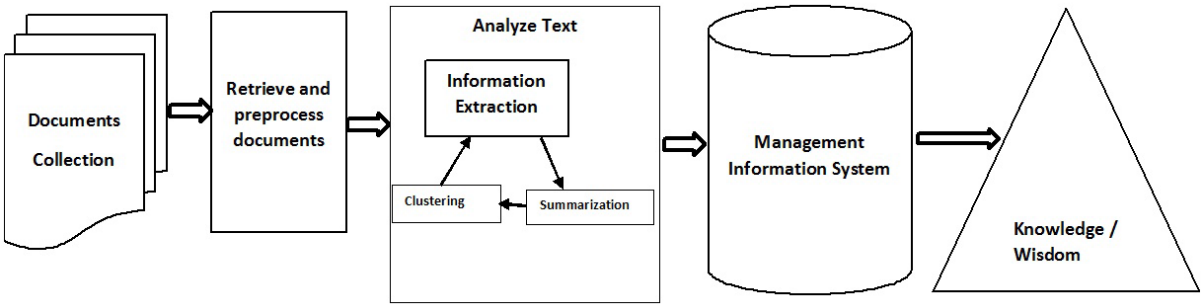

Figure 1: Text Mining Process [3]

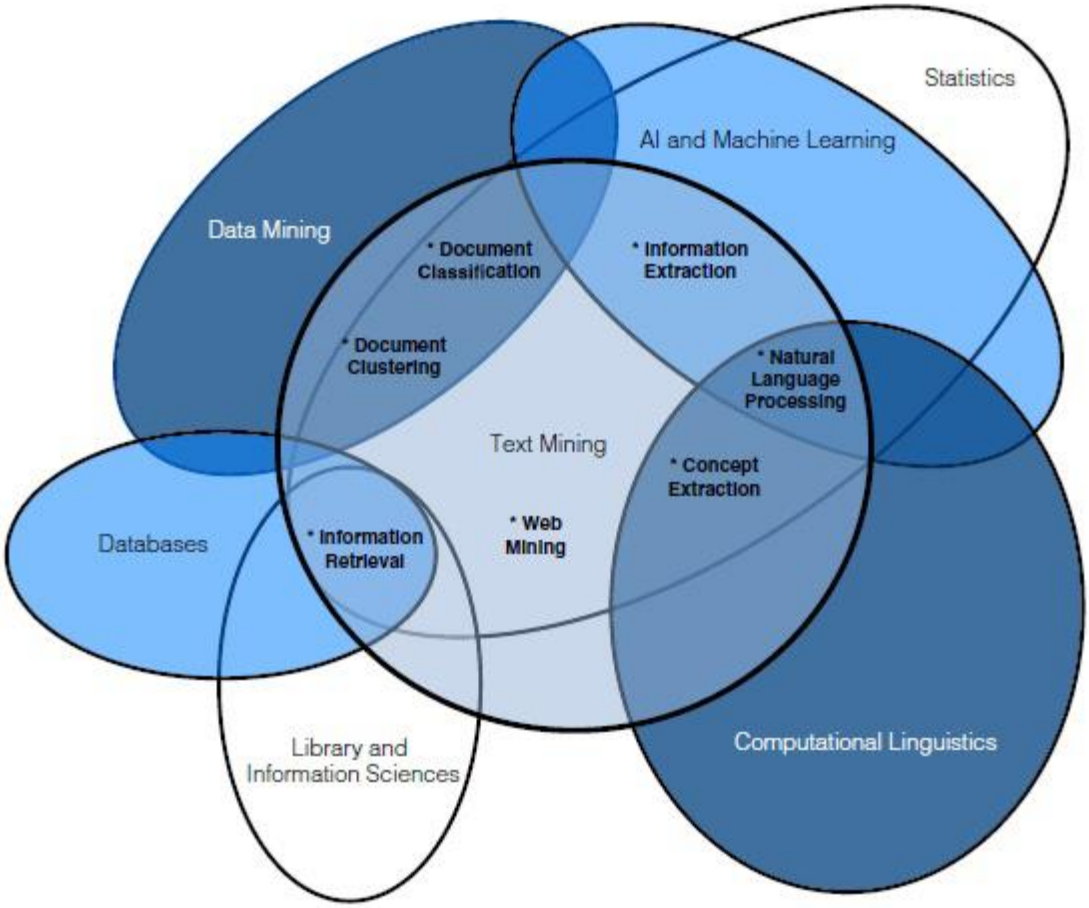

Figure 2: Venn diagram of text mining interaction with other fields given by S. M. Weiss et al. [1]

Now some of data analytics steps are:

- Data Requirement specifications: mainly depending on the objective or an experiment. Data might be numerical or categorial.
- Data Collection: Gather the information on focused factors which are recognized as data prerequisites. Data collection ensures that gathered data is accurate. Data is collected from various data sources and data fields.
- Data Processing: The gathered data must be processed or organised for data analysis.
- Data Cleaning: After processing, if the data is incomplete or containing redundant data it is removed.
- Data Analysis: Now, data is ready for analysis (inspecting, transforming). Correlation and regression analysis can be used to identify the relation among factors.
- Communication: The results of the data analysis are to be reported in a format as required by the clients to support their decisions.

Further, table 1 outlines differences between data mining and text mining and data analysis process in figure 3.

Table 1: Comparison between Data Mining and Text Mining

	Data Mining	**Text Mining**

Overview	A range of functions to search for patterns and relationships in structured data	A range of functions to turn unstructured textual data into structured information to enable data analysis
Data Type	Structured data from large datasets found in systems such databases, spreadsheets, ERP, CRM and accounting applications	Unstructured textual data found in emails, documents, presentations, videos, file shares, social media and the Internet.
Data Retrieval	Structured data is homogenous and organized making it easy to retrieve	Unstructured textual data comes in many different formats and content types located in a more diverse range of applications and systems.
Data Preparation	Structured data is formal and formatted facilitating the process of ingesting data into analytical models	Linguistic and statistical techniques – including NLP key-wording and meta-tagging – must be applied to turn unstructured into usable structured data.
Need for Taxonomy	There is no need to create a over-riding taxonomy for text mining	As the unstructured text comes in many different forms and formats, there needs to be an over-riding taxonomy for the data so that it can be organized into a common framework.

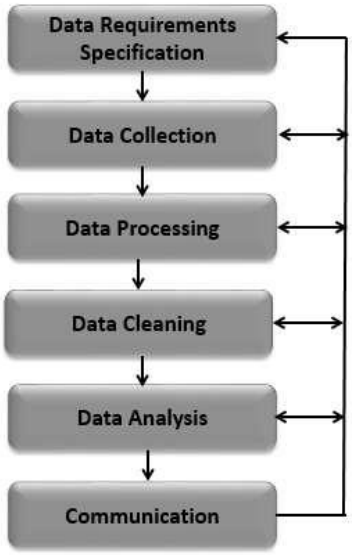

Figure 3: Data Analysis Process [2]

Hence, this section discusses the purpose of text mining in data mining, how the unstructured data is transformed to meaningful indices and the steps involved for text mining process. Also discussed about the popular text mining techniques and applications, difference between text mining and data mining and finally discussed about how text analytics are different from data analytics. Now, the remaining part of this article is organized as: Section 2 discusses about literature work related to text mining. Further, importance or scope of text mining in many applications is discussed in section 3. Section 4 discusses about medical care or healthcare application with introduction and importance in text mining. Further, section 5 discusses several existing algorithms or techniques available for text mining. Further, tools for text mining are discussed in section 6. Section 7 provides best tools explanation based on performance analysis of all existing tools. Further, section 8 discusses several issues and challenges raised in text mining process of unstructured data. Then, some research gaps are identified for future (as future work) in section 9. In last, this work is concluded with some future enhancement in section 10.

2. Literature Work

Unstructured text is very common in our daily life, i.e., produced in applications like e-healthcare. Most of the facts or information or data is produced/ generated in unstructured format available to a particular research or data mining project. Some of the work (with respect to data mining) has been discussed in next some paragraphs. In [3], authors described that gathering, extracting, pre-processing, text transformation, feature extraction, pattern selection, and evaluation steps are part of text mining process. In addition, different widely

used text mining techniques, i.e., clustering, categorization, decision tree categorization, and their application in diverse fields are surveyed. Further in [4], authors point out several issues with respect to text mining applications and its related techniques. They discussed that dealing with unstructured text is difficult as compared to structured or tabular data using traditional mining tools and techniques. They also discussed the use of text mining process in bioinformatics, business intelligence and national security system. Natural language processing and entity recognition techniques have reduced the issues that occur during text mining process. However, there exist issues which need attention. In table 2, we describe some text mining software using natural language processing and machine learning. Further in [5], authors explored MEDLINE biomedical database by integrating a framework for named entity recognition, classification of text, hypothesis generation and testing, relationship and synonym extraction, extract abbreviations. This new framework helps to eliminate unnecessary details and extract valuable information. Later in [6], authors analysed the text using text mining patterns and showed term-based approaches cannot analyse synonyms and polysemy properly. Moreover, a prototype model was designed for specification of patterns in terms of assigning weight according to their distribution. This approach helps to enhance the efficiency of text mining process. Recently in [7], authors presented a crime detection system using text mining tools and relation discovery algorithm was designed to correlate the term with abbreviation. In [8], authors presented a top down and bottom up approach for web-based text mining process. To combine the similar text documents, they apply k-mean clustering technique for bottom up partitioning. To find out the similarity within the document TF-IDF (Term Frequency- Inverse Document Frequency) algorithm has been used to find information regarding specific subjects. In [9], authors gave an overview of applications, tools and issues arise to mine the text. They discussed that documents may be structured, semi structured or unstructured and extracting useful information is a difficult task. They presented a generic framework for concept based mining which can be visualized as text refinement and knowledge distillation phases. The intermediate form of entity representation mining depends on specific domain. In [10], authors presented innovative and efficient pattern discovery techniques. They used the pattern evolving and discovering techniques to enhance the effectiveness of discovering relevant and appropriate information. They performed BM25 and vector support machine based filtering on router corpus volume 1 and text retrieval conference data to estimate the effectiveness of the suggested technique.

Table 2: List of Text Mining Software using NLP and Machine Learning.

S. No.	Types	Explanation
1	Amenity Analytics	Amenity Analytics' cloud-based text analytics tools allow companies in any industry to systematically extract actionable intelligence from any text (unstructured data) in real time. Amenity Analytics offers software products for text mining, sentiment analysis, and text analytics to automatically process structured data and unstructured data, using natural language processing, machine learning, artificial intelligence, and other technologies.
2	Linguamatics	Is a provider of text mining systems through software licensing and services, primarily for pharmaceutical and healthcare applications. The core natural language processing engine (I2E) uses a federated architecture to incorporate data from 3rd party resources. LabKey, Penn Medicine, Atrius Health and Mercy all use Linguamatics software to extract electronic health record data into data warehouses. Linguamatics software is used by 17 of the top 20 global pharmaceutical companies, the US Food and Drug Administration, as well as healthcare providers.
3	General Sentiment	General Sentiment's core system included a natural language processing engine to analyse text and entity management to track entities. The frontend consisted of depository servers and the Social Intelligence Platform, which hosted a number of applications that are useful for brand measurement. General Sentiment offered a software as a service (SaaS)-based solution delivered via the Amazon Cloud.iment tracked more than one billion entities and had some historical data dating back to 2004.

4	Luminoso	Luminoso's software identifies and quantifies patterns and relationships in text-based data, including domain-specific or creative language. Rather than human-powered keyword searches of data, the software automates taxonomy creation around concepts, allowing related words and phrases to be dynamically generated and tracked. Commercial applications include analysing, prioritizing, and routing contact centre interactions. identifying consumer complaints before they begin to trend and tracking sentiment during product launches. The software natively analyses text in fourteen languages, as well as emoji.
5	Natural Language Toolkit (NLTK)	A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language.
6	OpenNLP	The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as language detection, tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing and coreference resolution. These tasks are usually required to build more advanced text processing services
7	Megaputer Intelligence	Derives actionable knowledge from large volumes of text and structured data, including natural language processing (NLP), machine learning, sentiment analysis, entity extraction, clustering, and categorization.
8	IBM SPSS	Provider of Modeler Premium (previously called IBM SPSS Modeler and IBM SPSS Text Analytics), which contains advanced NLP-based text analysis capabilities (multi-lingual sentiment, event and fact extraction), that can be used in conjunction with Predictive Modelling. Text Analytics for Surveys provides the ability to categorize survey responses using NLP-based capabilities for further analysis or reporting. SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners.

In [11], authors performed various experiments of classification using multi-word features on the text. They proposed a hand-crafted method to extract multi-word features from the data set. To classify and extract multi-word text they divide text into linear and nonlinear polynomial form in support of vector machine that improve the effectiveness of the extracted data.
Hence, this section tries to provide a light on “Literature work”, i.e., work which have been already done in the past decade by several researchers. Now next section will discuss about importance or scope of text mining is several real world applications.

3. Importance/ Scope of text mining: Explanation with an Application

In the past decade, roughly 80% has been produced by billions of smart devices (embedded with other smart objects). This data is being generated in form of unstructured, have no labelled. According to a study by the International Data Group (IDG), unstructured data is growing at an alarming rate of 62% per year. The same study also suggests that by 2022, close to 93% of all data in the digital world will be unstructured. These statistics can be alarming for enterprises that are already grappling with the issue of having to deal with loads of unstructured data. Hence, if we look at several real world’s applications then we find out that we require several new approaches/ models/ algorithms/ tools (a form of technology) to easily process this unstructured data and help organizations (industries) discover what's within it, i.e., with speed and accuracy. In result, text analysis or mining is used to extract useful information from this unstructured data. Text analysis (or text mining) is the process of analysing chunk of unstructured data to find out previously undiscovered information and insights that can be leveraged for informed decision making and other processes. The new age text analysis tools like

3RDi Search offer a host of text mining services like sentiment analysis, content classification, semantic search, content summarization, named entity recognition and more. Text analysis tools are based on a complex process that consists of several concepts, such as statistics, machine learning, natural language processing and more. For e-healthcare applications, text mining steps will be as figure 2.

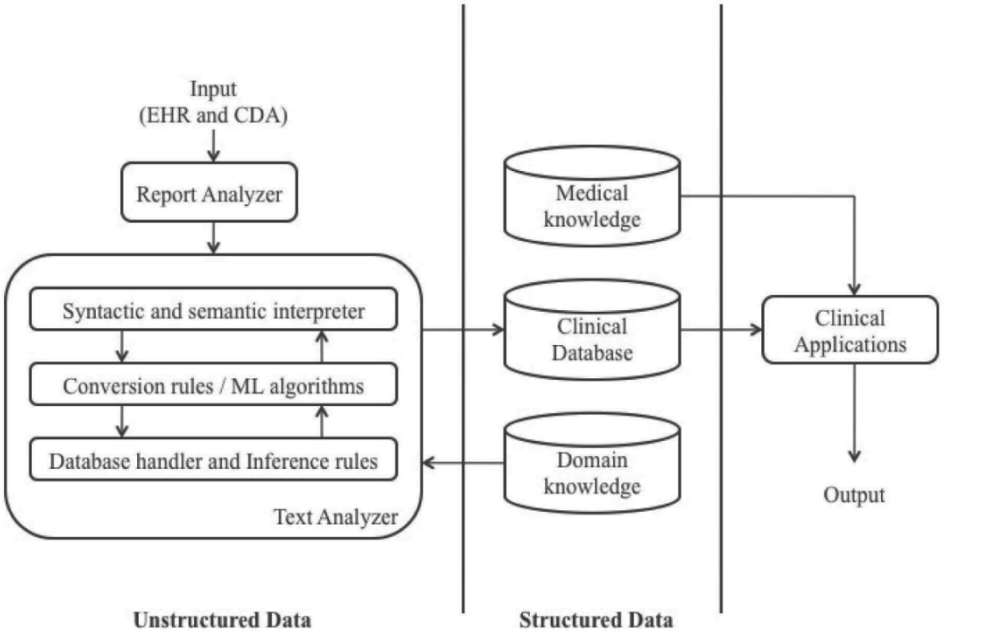

Figure 2: Process of Text Mining in Healthcare (unstructured and structured)

In figure 2, Text Mining steps are discussed as:

a) Extract information from unstructured data.
b) Extracted information converted into structured data.
c) Pattern identified from structured data.
d) Analyse the pattern.
e) Extract the valuable information.
f) Store in the database.

Hence this section discusses about importance or scope about text mining in various applications especially in healthcare (discussed in next section). Also, complete process for text mining in healthcare also discussed in this section. Now, next section will discuss about importance and discussion of text mining in healthcare application in detail.

4. E-healthcare Application: Introduction and Importance

As we discussed in [20], some applications like defence, healthcare, agriculture, public transportation, etc., are growing at a rapid rate now days (in case of using internet of things). Such smart objects when are embed with machine, generate a lot of data in form (during communications; which is available in form of text) which need to mined using efficient text mining techniques. In healthcare, text mining techniques are used to extract useful information, i.e., to cure (or find solution for critical or harmful) diseases. Hence in summary, the major uses of a text mining tool in healthcare are for:

- Text Analytics: involves extracting useful information and patterns from text. Most tools provide this feature.
- Text Processing: involves transforming and manipulating unstructured text so that analysis methods can be applied to it.
- Classification/ Categorization: Many tools are used for classification and categorization of text/ documents.
- Sentiment Analysis: is used to identify subjective information from text. Many tools provide for sentiment analysis also called as Opinion Mining.
- Knowledge Discovery: deals with identification of useful information from huge amount of text. Most tools provide for Knowledge discovery and information retrieval features.

Moreover this, some approaches for Text Mining are:

- Using well-tested methods and understanding the results of text mining.
- "Black-box" approaches to text mining and extraction of concepts
- Text mining as document search.

Also, several other approaches are also being used in many applications. Hence in this article, this section discusses about importance of text data mining in an essential (useful) applications like healthcare. Now, next section will discuss about several existing text mining algorithms/ techniques available for healthcare applications.

5. Existing algorithms/ techniques for Text mining (i.e., for Classification, Clustering, Regression, Association Rule Mining and Outlier detection, etc.)

Mostly used techniques used in text mining techniques are Information Extraction (IE), Information Retrieval (IR), Categorisation, Clustering and Summarisation. All techniques have been discussed in table 1. Moreover this, there are some methods used in text mining which are included here as:

i. Unsupervised Learning Methods: Unsupervised learning methods are techniques trying to find hidden structure out of unlabelled data. They do not need any training phase, therefore can be applied to any text data without manual effort. Clustering and topic modelling are the two commonly used unsupervised learning algorithms used in the context of text data. Clustering is the task of segmenting a collection of documents into partitions where documents in the same group (cluster) are more similar to each other than those in other clusters. In topic modelling a probabilistic model is used to determine a soft clustering, in which every document has a probability distribution over all the clusters as opposed to hard clustering of documents. In topic models each topic can be represented as probability distributions over words and each document is expressed as probability distribution over topics. Thus, a topic is similar to a cluster and the membership of a document to a topic is probabilistic.

ii. Supervised Learning Methods: Supervised learning methods are machine learning techniques pertaining to infer a function or learn a classifier from the training data in order to perform predictions on unseen data. There is a broad range of supervised methods such as nearest neighbour classifiers, decision trees, rule-based classifiers and probabilistic classifiers.

iii. Probabilistic Methods for Text Mining: There are various probabilistic techniques including unsupervised topic models such as probabilistic Latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA), and supervised learning methods such as conditional random fields that can be used regularly in the context of text mining.

iv. Text Streams and Social Media Mining: There are many different applications on the web which generate tremendous amount of streams of text data. News stream applications and aggregators such as Reuters and Google news generate huge amount of text streams which provides an invaluable source of information to mine. Social networks, particularly Facebook and Twitter create large volumes of text data continuously. They provide a platform that allows users to freely express themselves in a wide range of topics. The dynamic nature of social networks makes the process of text mining difficult which needs special ability to handle poor and non-standard language.

v. Opinion Mining and Sentiment Analysis: With the advent of e-commerce and online shopping, a huge amount of text is created and continues to grow about different product reviews or users opinions. By mining such data we find important information and opinion about a topic which is significantly fundamental in advertising and online marketing.

vi. Biomedical Text Mining: Biomedical text mining refers to the task of text mining on text of biomedical sciences domains. The role of text mining in biomedical domain is two-fold, it enables the biomedical researchers to efficiently and effectively access and extract the knowledge out of the massive volumes of data and also facilitates and boosts up biomedical discovery by augmenting the mining of other biomedical data such as genome sequences and protein structures.

Some of used data modelling functions related to text mining are: association, classification, clustering, and regression. Hence, a text mining is being based on following process:

Sample document ->Transformation ->representation models ->learning ->specific models ->knowledge ->visualization

Table 3: Common Techniques used in Text Mining Analytic with Essential Information

Title	Objective	Characteristics	Tools
Information	Reconstructing a set of unstructured or semi-	Extract	Text Finder,

Extraction	structured textual documents into a structured database.	information from structured database	Clear Forest Text
	• The first step in the process of evaluation of unstructured data. • Involves tokenization and identification of named entities, key phrases and parts-of-speech. • Uses concept of pattern matching to find out predefined sequences if any within the data. • Identifies the relationship between entities and attributes.		
Categorization	Assigning one or more categories to an unstructured text document.	Document based categorization	Intelligent Miner
	• Works on an input-output principle wherein the system is given inputs regarding the pre-defined categories under which the data in the new documents is to be classified. • Consists of the following steps - processing, indexing, dimensional reduction and classification. • Uses the Nearest Neighbour classifier, Decision Tree, Naïve Bayesian classifier, and other statistical classification techniques.		
Clustering	Bringing together clusters of documents that have similar content.	Cluster collection of documents, clustering, classification and analysis of text document	Caroot, Rapid Miner
	• Generates multiple groups of documents known as clusters. • The content of documents in a specific cluster are very similar while that of documents in different clusters are not even remotely similar. • Differs from clustering as it brings together documents without the use of any pre-defined categories as reference. This technique works on semantics - the principle on which semantic search engines work. • K-means is a frequently used algorithm that brings great results.		
Visualization	Simplifying and enhancing the discovery of useful information with visual cues.		
	• Uses visual cues such as text flags to indicate individual documents or document categories and colours to indicate the density of a category, entity, phrase, etc. • Enables the user to zoom in/out or scale the document as required, without any loss of data.		

	• Places large sources of textual data into a visual hierarchy.		
Summarization	Automatically generating a summary/compressed version of the text with information that will be of the highest importance or relevance to the end user.	Reduce length by keeping its main points and overall meaning it is	Tropic Tracking Tool, Sentence Ext tool
	• Determines the most important points in a lengthy document that the user of the text analysis tool will find useful. • Involves 3 steps - Pre-processing, Processing, and Development. • The pre-processing step involves building a structured representation of the text. • The processing step involves application of algorithms to generate a summary of the text. • Uses semantics technology, similar to a semantic search engine, to retain the meaning of the text in the summary. • The development step is where the final text summary is obtained.		

Hence, table 2 discusses several essential features of text mining like techniques, characteristics, tools, etc. Such component or features have an important role in healthcare to cure patient's disease and predict useful information about patient's disease (or status of patient's health). Now, next section will discuss tools available for text data mining in detail. In continuation to this, next section will also discuss "how such tools are helpful in healthcare applications?"

6. Tools Available for Text Data Mining

Generally, text mining or text data mining is being Text Mining or knowledge discovery from text (KDT) (introduced by Fledman et al.), refers to the process of extracting high quality of information from text (i.e., structured such as RDBMS data, semi-structured such as XML and JSON, and unstructured text resources such as word documents, videos, and images). It widely covers a large set of related topics and algorithms for analysing text, spanning various communities, including information retrieval, natural language processing, data mining, machine learning many application domains web and biomedical sciences.
Text Mining Tools can be classified into three categories.

i. *Proprietary Text Mining Tools*: These tools are commercial text mining tools owned by a company. To use these tools purchase is required. Although demo/trial versions are available free of cost but have limited functionality. Note that 39 out of these 55 tools are proprietary tools.
ii. *Open Source Text Mining Tools*: These tools are available free of cost and also there source code and one can even contribute in their development. Note that 13 out of these 55 text mining tools are open source.
iii. *Online Text Mining Tools:* These tools can be run from the website itself. Only a web browser is required. These tools are generally simple and provide limited functionality. Note that 03 out of these 55 text mining tools are online web based tools.

Note that in text mining, the main tools are QDA Miner, WordStat and SimStat tools that allow users to explore, analyze and relate both structured (labelled) and unstructured (unlabelled) data. Moreover this, top free software for text analysis or text mining and text analytics are General architecture for text engineering, RapidMiner Text Mining Extension, Coding analysis Toolkit, KH Coder, Google Cloud Natural Language API, QDA miner Lite, Visual Text, TAMS, Pattern, Natural Language Toolkit, Datumbox, Apache Mahout, Carrot2, Textable, Apache OpenNLP, KNIME Text Processing, LingPipe, Gensim, tm-Text Mining Package, Aika, LPU, Apche Stanbol, Distributed Machine Learning Toolkit are some top free Text Analysis, Text Mining, Text analytics, Software, etc. Hence, now some important tools for detecting cancer disease/ healthcare are being discussed in table 4.

Table 4: Tools Used in Bio-Medical Text Mining for Cancer [12]

System	Brief introduction
ABNER	ABNER is a software tool for molecular biology text analysis. It uses linear-chain conditional random fields approach with orthographic and contextual features
GENIATagger	The GENIA tagger is specifically tuned for biomedical text such as MEDLINE abstracts. It is a useful pre-processing tool for information extraction from biomedical documents
LingPipe	LingPipe provides three generic, trainable chunkers to carry on named entity recognition. LingPipe can be used to identify biomedical entities such as genes, organisms, malignancies, and chemicals
Yapex	Yapex is a rule-based system named entity recognition system that utilizes lexical and syntactic analysis to identify protein names

Popular tool in cancer: Cancer Hallmarks Analytics Tool (CHAT) [13]

Explanation of CHAT: Now, to understand the molecular mechanisms involved in most cancers' improvement, considerable efforts are being invested in cancer studies. One way to organize existing knowledge on cancer is to utilize the widely accepted framework of the Hallmarks of Cancer. These hallmarks refer to the alterations in cell behaviour that characterize the cancer cell. Automatic text mining methodology and a tool (CHAT) capable of retrieving and organizing millions of cancer-related references from PubMed into the taxonomy. The efficiency and accuracy of the tool was evaluated by case studies. The correlations identified by the tool show that it offers a great potential to organize and correctly classify cancer-related literature. Furthermore, the tool can be useful, for example, in identifying hallmarks associated with extrinsic factors, biomarkers and therapeutics targets.
Hence, this section discusses about tools present in the text mining which are categorised into three types that are proprietary text mining tool, open source text mining tool and online text mining tool, also discussed about top popular tools in all fields. Now, next section will discuss about tools or algorithms based on their performance analysis of all existing algorithms.

7. Performance analysis of all existing algorithms and tools (based on popularity) available for Text Mining in Disease Prediction

After significant (e.g., frequent) words have been extracted from a set of input documents, and/or after singular value decomposition has been applied to extract salient semantic dimensions, typically the next and most important step is to use the extracted information in a data mining project.

i. Graphics (visual data mining methods): Depending on the purpose of the analyses, in some instances the extraction of semantic dimensions alone can be a useful outcome if it clarifies the underlying structure of what is contained in the input documents. For example, a study of new car owners' comments about their vehicles may uncover the salient dimensions in the minds of those drivers when they think about or consider their automobile (or how they "feel" about it). For marketing research purposes, that in itself can be a useful and significant result. We can use the graphics (e.g., 2D scatterplots or 3D scatterplots) to help us to visualize and identify the semantic space extracted from the input documents.
ii. Clustering and factoring: We can use cluster analysis methods to identify groups of documents (e.g., vehicle owners who described their new cars), to identify groups of similar input texts. This type of analysis also could be extremely useful in the context of market research studies, for example of new car owners. We can also use Factor Analysis and Principal Components and Classification Analysis (to factor analyse words or documents).
iii. Predictive data mining: Another possibility is to use the raw or transformed word counts as predictor variables in predictive data mining projects.

Hence, this section discusses about performance analysis of existing tools in text mining for disease prediction. Mostly classification and prediction analysis are used in bio-medical area. Now, next section will discuss several issues and challenges, arising in text mining process (in prediction of diseases).

8. Issues and Challenges faced in Text Mining (in Prediction of diseases)

Many issues occur during the text data mining process and effect the efficiency and effectiveness of decision making. Complexities can arise at the intermediate stage of text mining. In pre-processing stage various rules and regulations are defined to standardize the text that makes text mining process efficient. Before applying pattern analysis on the document there is a need to convert unstructured data into intermediate form but at this stage mining process has its own complications. Sometime real theme or data mislay its importance due to the

modification in the text sequence [14]. Another major issue is a multilingual text refinement dependency that creates problems. Only few tools are available that support multiple languages [15]. Various algorithms and techniques are used independently to support multilingual text. Because numerous important documents persist outside the text mining process because various tools do not support them. These issues create a lot of problems in knowledge discovery and decision making process. Infect real benefit is difficult to attain by using the existing text mining techniques and tools because its rarely support multilingual documents [16].
Integration of domain knowledge is an important area as it performs specific operations on specified corpus and attains desired outcomes. In this situations domain knowledge from which document corpus to be extracted need to integrate with the computing abilities from which information have to be attained. According to the requirements of the field, experts are needed to work collaboratively from diverse domains to extract more effective, precise and accurate results [14], [17]. The use of synonyms, polysems and antonyms in the documents create problems (abstruseness) for the text mining tools that take both in the same context. It is difficult to categorize the documents when collection of document is large and generated from diverse fields having the same domain. Abbreviations gives changed meaning in different situation is also a big issue [18]. Varying concepts of granularity change the context of text according to the condition and domain knowledge. There is need to describe rules according to the field that will be used as a standard in the area and can be embedded in text mining tools as a plug-in. It entails lots of effort and time to develop and deploy plug-ins in all fields separately. To develop plug-ins in depth and proper knowledge about the specific domain will be required [16], [19]. Natural languages have lots of complications in itself that create problem in text refinement methods and the identification of entity relationship. Words having same spelling but give diverse meaning, for example, fly and fly. Text mining tools considered both as similar while one is verb and other is noun. Grammatical rules according to the nature and context is still an open issue in the field of text mining [19].

Issues and Considerations for "Numericizing" Text

- Large numbers of small documents vs. small numbers of large documents. Examples of scenarios using large numbers of small or moderate sized documents were given earlier (e.g., analysing warranty or insurance claims, diagnostic interviews, etc.). On the other hand, if your intent is to extract "concepts" from only a few documents that are very large (e.g., two lengthy books), then statistical analyses are generally less powerful because the "number of cases" (documents) in this case is very small while the "number of variables" (extracted words) is very large.
- Excluding certain characters, short words, numbers, etc. Excluding numbers, certain characters, or sequences of characters, or words that are shorter or longer than a certain number of letters can be done before the indexing of the input documents starts. We may also want to exclude "rare words," defined as those that only occur in a small percentage of the processed documents.
- Include lists, exclude lists (stop-words). Specific list of words to be indexed can be defined; this is useful when we want to search explicitly for particular words, and classify the input documents based on the frequencies with which those words occur. Also, "stop-words," i.e., terms that are to be excluded from the indexing can be defined. Typically, a default list of English stop words includes "the", "a", "of", "since," etc., i.e., words that are used in the respective language very frequently, but communicate very little unique information about the contents of the document.
- Synonyms and phrases. Synonyms, such as "sick" or "ill", or words that are used in particular phrases where they denote unique meaning can be combined for indexing. For example, "Microsoft Windows" might be such a phrase, which is a specific reference to the computer operating system, but has nothing to do with the common use of the term "Windows" as it might, for example, be used in descriptions of home improvement projects.
- Stemming algorithms. An important pre-processing step before indexing of input documents begins is the stemming of words. The term "stemming" refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of verbs are identified and indexed (counted) as the same word. For example, stemming will ensure that both "traveling" and "travelled" will be recognized by the text mining program as the same word.
- Support for different languages. Stemming, synonyms, the letters that are permitted in words, etc. are highly language dependent operations. Therefore, support for different languages is important.

Hence, this section discusses about complex issues a raised during text mining process, and a major issue is a multilingual text refinement dependency. Also discussed why to remove stop words and synonyms from input data and what is the role of stemming algorithms? Now, next section will discuss several research gaps identified in text mining for near future.

9. Research Gaps Mitigated for Future as Opportunities

In this work, we provide a detailed introduction to the field of text mining in disease prediction (healthcare). We provided an overview of some of the most fundamental algorithms and techniques which are extensively used in the text domain. This paper also overviewed some of important text mining approaches in the biomedical domain. Even though, it is impossible to describe all different methods and algorithms thoroughly regarding the limits of this article, it should give a rough overview of current progresses in the field of text mining. Text mining is essential to scientific research given the very high volume of scientific literature being produced every year []. These large archives of online scientific articles are growing significantly as a great deal of new articles is added in a daily basis. While this growth has enabled researchers to easily access more scientific information, it has also made it quite difficult for them to identify articles more pertinent to their interests. Thus, processing and mining this massive amount of text is of great interest to researchers. Some interesting facts about used techniques for text mining are:

- Natural Language Processing and Machine Learning: Most of the tools employ Natural Language Processing (21%) and/or Machine Learning techniques (21%) for mining text.
- Statistical Methods: as used for data mining are also applied for text mining. In fact most of the tools use statistical methods (11%) in conjunction with other methods.
- Artificial Intelligence (9%): techniques such as neural networks are also employed in many text mining tools.
- Classification techniques (8%): are also used to categorize text and documents. These classification techniques must be able to handle unstructured data.
- Linguistic Learning (5%), Semantic Analysis (five percent), and Predictive Modelling (seven percent) techniques are also employed for mining text.

Hence, this section discusses about how text mining tools are used for disease prediction. And discussed how efficiently text mining tools applied in all the fields. Most popular tools employ NLP. Now, next section will summarize this work in brief.

10. Summary

Text is the main essential vehicle which carries a lot of information (present in unstructured form), so we require text mining to determine useful patterns and relationships (including reason behind that). On another hand, most of the people confused with text mining and data mining. As we have discussed a complete difference in section 1 and table 1. Now, some important with respect to text mining and data mining can be included as:

- Unstructured versus Structured data: Data mining systems essentially analyse figures that may be described as homogeneous and universal. They extract, transform and load data into a data warehouse. Business analysts use data mining software applications to present analysed data in easily understandable forms, such as graphs. Currencies, dates, names, might have to be managed, but they are easy to link to data and do not require any deep understanding of their context. Text mining tools have to face major technical challenges such as heterogeneous document formats (text documents, emails, social media posts, verbatim text, etc.), as well as multilingual texts and abbreviations and slang typical of SMS language.
- Deployment time: Data mining is focused on data-dependent activities such as accounting, purchasing, supply chain, CRM, etc. The required data is easy to access and homogeneous. Once algorithms are defined, the solution can be quickly deployed. The complexity of the data processed make text mining projects longer to deploy. Text mining counts several intermediary linguistic stages of analysis before it can enrich content (language guessing, tokenization, segmentation, morpho-syntactic analysis, disambiguation, cross references, etc). Next, relevant terms extraction and metadata association steps tackle structuring the unstructured content to nurture domain-specific applications. Moreover, projects may involve some heterogeneous languages, formats or domains. Finally, few companies have their own taxonomy. However, this is mandatory for starting a text mining project and it can take a few months to be developed.
- Technology perception: Data mining has been considered a proven, robust and industrial technology for many decades. Text mining was historically thought of as complex, domain-specific, language-specific, sensitive, experimental, etc. In other words, text mining was not understood well enough to have management support and therefore, was never valued as a ‘must-have’. However, with the advent of digitalization, the rise of social networks and increased connectivity, companies are now more concerned about their online reputation and are looking for ways to increase loyalty with customers in a world of increasing choice. As a result, sentiment analysis is the new focus of text mining. Companies have realized that information is a strategic asset made of text and that text mining is no longer a luxury, but a necessity.

Hence, we have seen text mining is in growing phase and attracting more and more researchers yesterday around the world. Text mining is the future to extract useful data from this high volume of unstructured data. Hence, researchers, scientist are invited to do more and innovative research in this respective area and provide effective solution to the society.

Bibliography

[1] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, "Text mining: predictive methods for analyzing unstructured information", Springer Science and Business Media, 2010.

[2] Hilfiker, J. N., Sun, J., & Hong, N. "Data analysis", In *Springer Series in Optical Sciences*. https://doi.org/10.1007/978-3-319-75377-5_3

[3] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[4] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012.

[5] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and ˇ M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.

[6] B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013.

[7] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," Information Sciences, vol. 275, pp. 314–347, 2014.

[8] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," International Journal of Computer Applications, pp. 159–171, 2013.

[9] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications, tools and issues-an overview," International Journal of Computer Applications, vol. 80, no. 4, 2013

[10] P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on. IEEE, 2015, pp. 634–638.

[11] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan, vol. 51, 2007, p. 45.

[12] FeiZhua,ChengZhanga,et.al, "Biomedical text mining and its applications in cancer research", journal of biomedical informatics, volume 46,issue 2,pages 200-211

[13] Simon Baker, Imran Ali, Ilona Silins, Sampo Pyysalo, et.al, "Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer", Bioinformatics, Volume 33, Issue 24, 15 December 2017, Pages 3973–3981.

[14] A. Henriksson, J. Zhao, H. Dalianis, and H. Bostrom, "Ensembles of ¨ randomized trees using diverse distributed representations of clinical events," BMC Medical Informatics and Decision Making, vol. 16, no. 2, p. 69, 2016.

[15] H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," International Journal of Computer Applications, vol. 75, no. 16, 2013.

[16] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, "Automatic extraction of synonymy information: -extended abstract," OTT06, vol. 1, p. 55, 2007.

[17] B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," IJSEAT, vol. 3, no. 3, pp. 69–71, 2015.

[18] A. Kaklauskas, M. Seniut, D. Amaratunga, I. Lill, A. Safonov, N. Vatin, J. Cerkauskas, I. Jackute, A. Kuzminske, and L. Peciure, "Text analytics for android project," Procedia Economics and Finance, vol. 18, pp. 610–617, 2014.

[19] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, vol. 3, 2013, pp. 3–5.

[20] Tyagi, Amit Kumar, Building a Smart and Sustainable Environment using Internet of Things (February 22, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019.