Journal
of Danubian
Studies
and Research

# Survey of Text Mining Research Methods and Their Innovative Applicability

**Mihaela Chistol**[1]**, Mirela Danubianu**[2]

**Abstract**: Humans are social beings who feel a strong need for communication. From the earliest times, the exchange of information was based on primary skills such as sight and speech. Thus, at the beginning of the 20th century, a famous phrase was uttered that claims that "A picture is worth a thousand words". In the contemporary world, this phrase is no longer appropriate because with the discovery of the World Wide Web the textual revolution began. While digitalization continues at light speed, the need to process huge amounts of generated text resources is felt even more strongly. Therefore to solve the crisis of information overload, text mining is used, which is a new and interesting area of computer science research. This paper presents a methodological and conceptual theory of text mining along with the main methods behind it. Following an in-depth examination of the literature, the study shows the fundamental directions of text mining research such as classification, clustering, and information retrieval. In addition, the article presents state-of-the-art applications that implement the concept of text mining to solve problems in the real world.

**Keywords:** Text Mining Techniques; Classification; Clustering; Information Retrieval; Innovative Applications

## 1. Introduction

Communication is the key to the evolution of human society. The ability to understand each other allows us to work together in order to accomplish difficult tasks. As a result, humans hierarchically surpass any other species that live on Earth. Animals can also express their intention through vocal sounds. What made the difference between humans and animals? The answer is simple: the human's ability to articulate words, to express thoughts, feelings, and ideas.

Language is a powerful tool that people have improved over the centuries by inventing the alphabet, writing, spelling rules, and so on. Therefore, in the

[1] PhD student, Stefan cel Mare University, Suceava, Romania, Address: Strada Universității 13, Suceava 720229, Romania, Corresponding author: mihaela3milea@gmail.com.

[2] Associate Professor, PhD, Stefan cel Mare University, Suceava, Romania, Address: Strada Universității 13, Suceava 720229, Romania, E-mail: mdanub@eed.usv.ro.

contemporary world, we benefit from books, articles, libraries, data corpora, and the entire World Wide Web that guarantee access to any information at our fingertips. The proliferation of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, although the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant (Feldman & Sanger, 2007). Even for the people with a photographic memory, such as Kim Peek a savant who was the inspiration for the character Raymond Babbitt in the 1988 movie "Rain Man" (Kim Peek, 2021) who suffered from autism spectrum disorder but at the same time had the amazing ability to memorize the text from up to 8 books a day, would not be a feasible task.

In the information age, computers have the physical capability to store huge amounts of data. 80% of global data is in text form. The human brain process textual information using pattern recognition. But how do computers understand the information in text data? Recently, computers have improved their RAM and MHz properties, so they can perform complex analysis of a huge amount of textual material by using cutting-edge technologies, such as text mining.

Text mining is a new and emerging area of computer science research. This article tries to blend together the theory and practice of text mining methods. Following an in-depth examination of the literature, the study shows the fundamental directions of text mining research and outlines the main preprocessing techniques used by text mining systems.

## 2. Text Mining

### 2.1. Knowledge Discovery in Text (KDT)

Text mining is a modern technique for extracting knowledge from document collections through the identification and exploration of interesting patterns in the textual data of various types of documents – such as books, web pages, emails, reviews, reports or product descriptions.

The data sources and document formats can be diverse. Databases organize text sources as follows:

d. **Structured data**: This data is organized and follows a pattern, usually the information is arranged in rows and columns. The text contains precise facts and details such as addresses, biological parameters, or telephone numbers.

e. **Unstructured data**: This data is in a free form and does not have a predefined format. It may include messages or reviews collected from the web and social networks. Often, it contains media elements such as video and audio files.

f. **Semi-structured data**: It's name is suggestive and defines this data as a combination of unstructured and structured data formats. The most popular types of semi-structured files are JSON and XML, which although well defined, are not structured enough to meet the requirements of a relational database.

How can text mining process both structured and unstructured data as efficiently? Text mining "turns text into numbers" so as not to take into account the nature of data sources when applying the algorithm. Converting text into a structured, numerical format and applying analytical algorithms require knowing how to both use and combine techniques for handling text, ranging from individual words to documents to entire document databases (Miner, et al., 2012).

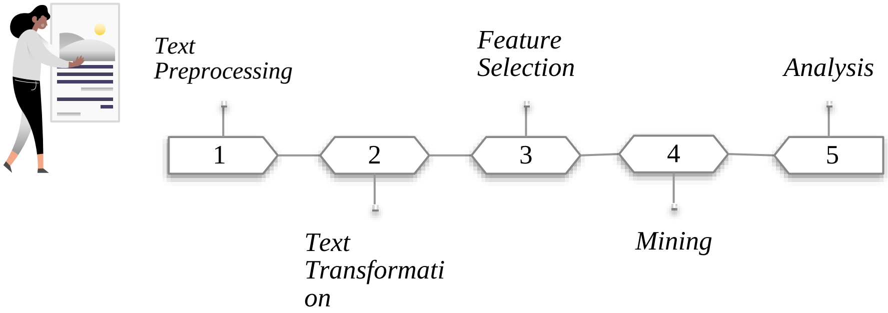

**Figure 6. Overview of KDT Process**

Knowledge Discovery in Text (KDT) is the process that explores large datasets in order to identify useful and relevant patterns within them. This process is also known as Text Data Mining (TDM) because it can be seen as a data mining process that explores text data.

KDT is a multi-step process that involves text preprocessing, text transformation, feature selection, and the application of the mining algorithm. The last stage consists of a performance analysis that is done by determining a series of statistical indices in order to discover richer knowledge. The performed steps depend on the application's purpose but, most systems follow the stages presented in Figure 1.

**Tokenization** is a preprocessing method which breaks a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The main goal of this step is the investigation of the words in a sentence (Kowsari, et al., 2019).

**Stop Words** is a technique that removes the words that do not bring a significant amount of information for the text mining process. For example, the word *the* is removed because it is common and does not have a positive impact on the process.

**Stemming** is the process of reducing the word to a base form. Text stemming modifies words to obtain variant word forms using different linguistic processes such as affixation[1] (Singh & Gupta, 2016). For example, the stem of the word "learning"

[1] The process of adding a new part (a prefix or suffix) to the beginning or end of a word, so that the word's meaning is changed.

is “learn”.

**Capitalization** is used to turn the words from the data sources into lower case. This is a common approach that helps to improve the performance of the text mining process that handles large documents with different capitalizations. This technique projects all words in text and document into the same feature space, but it causes a significant problem for the interpretation of some words e.g., “US” (United States of America) to “us” (pronoun) (Singh & Gupta, 2016).

**N-Gram** is a technique applied after the preprocessing stage. When computing the n-grams, the data is transformed into a vector representation suitable for input into mining algorithms. Among the researchers’ most preferred schemes are Term Frequency Inverse document frequency (TF-IDF), Bag-of-Words (BoW), and Word2Vec.

### 2.2. Core Text Mining Techniques

The text mining research field is constantly evolving. Since it is more often used in various fields, for example, natural language processing or web mining, researchers had to develop new methods that can extract useful information effectively and meet the needs of modern society.

Text mining can be divided into seven practice areas, based on the unique characteristics of each area (Miner, et al., 2012).

8. **Information retrieval** (IR): Information retrieval is a process whose main goal is to search and retrieve the relevant patterns from large datasets.

9. **Clustering**: Described abstractly, clustering is the process that identifies the essential information in text corpora and gathering these into relevant groups called “clusters”.

10. **Classification**: Classification is one of the most popular text mining methods because it is suitable for solving various problems. Classification, also known as text categorization (TC), assigns a known set of categories to unlabeled instances using a trained model.

As with many other artificial intelligence (AI) tasks, there are two main approaches to text categorization. The first is the knowledge engineering approach in which the expert’s knowledge about the categories is directly encoded into the system either declaratively or in the form of procedural classification rules. The other is the machine learning (ML) approach in which a general inductive process builds a classifier by learning from a set of pre-classified examples. In the document management domain, the knowledge engineering systems usually outperform the ML systems (Feldman & Sanger, 2007).

11.**Web mining**: The World Wide Web is the largest and most widely known repository of hypertext. Hypertext documents contain text and generally embed hyperlinks to other documents distributed across the Web (Chakrabarti, 2002). Web mining is the process of analyzing and exploring the web pages in order to identify meaningful patterns. It uses methods borrowed from statistics and machine learning, to extract the knowledge from structured and unstructured data.

12.**Information extraction** (IE): This is a text mining technique that focuses on the extraction of structured data from unstructured text. The training of an information extraction system requires a solid understanding of specialized algorithms.

13.**Natural language processing**: Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI, concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP combines computational linguistics, rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data (Natural Language Processing (NLP), 2020).

14.**Concept extraction**: Is the process of concept extraction from textual data.

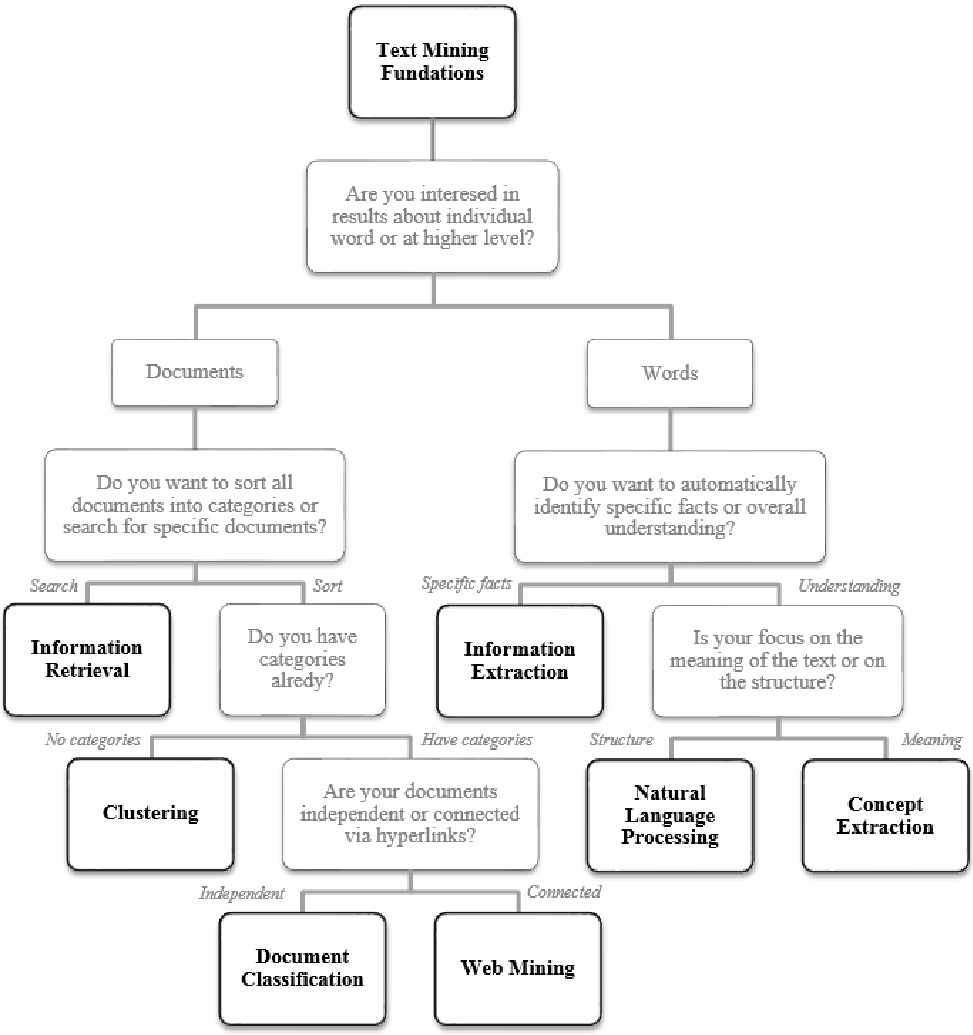
Text Mining Fundations
Are you interesed in results about individual word or at higher level?
Documents
Words
Do you want to sort all documents into categories or search for specific documents?
Do you want to automatically identify specific facts or overall understanding?
Search
Sort
Specific facts
Understanding
Information Retrieval
Do you have categories alredy?
Information Extraction
Is your focus on the meaning of the text or on the structure?
No categories
Have categories
Structure
Meaning
Clustering
Are your documents independent or connected via hyperlinks?
Natural Language Processing
Concept Extraction
Independent
Connected
Document Classification
Web Mining

**Figure 7. A Decision Tree for Finding the Right Text Mining Practice Area (Miner, et al., 2012).**

Text mining techniques employ different algorithms and tools of data mining, statistics, machine learning, and computational linguistics.

**Table 5. Text Mining Algorithms and the Corresponding Practice Area**

| Algorithm | Practice Area |
|---|---|
| Naïve Bayes | Document classification |
| Conditional random fields | Information extraction |
| Hidden Markov models | Information extraction |
| k-means | Clustering |
| Singular value decomposition (SVD) | Document classification, clustering |
| Logistic regression | Document classification |
| Decision trees | Document classification |
| Neural network | Document classification |
| k-nearest neighbors | Document classification |
| Regression | Classification |

## 3. Text Mining Applications

People keep valuable information in text format. Access to this information is the reason why text mining has become an important area of research that is attracting more and more investment from the business environment.

Text processing is an emerging technique that provides unlimited applicability. Whether you work in marketing, education, biotechnology, or customer support, you can take advantage of text mining to make your job easier. For a short time, remember all repetitive and tedious tasks you have to deal with on a regular basis. Now imagine all the things you could achieve if you no longer had to deal with these responsibilities. Being free of manual tasks allows you to focus your energy on planning development strategies.

The current text mining systems, developed by academic researchers or corporate programmers- are built to solve a concrete problem in order to satisfy the industry's needs. Among the most significant applications that address important text mining issues are:

f. **Sentiment Analysis**: It is a tool designed to analyze human feelings and opinions from text data collected from blog posts, emails, comments, reviews, and so on.

What are my customers saying about me? Customer feedback is a very useful source of information on customer satisfaction. For example, it is useful for organizations to be able to extract the body of main "themes" and affective responses associated with their products from customer feedback and reviews or from public blogs that are relevant to the respective products or services (Li & Wu, 2010).

g. **Healthcare**: Text mining is a popular and commonly used technique in the biomedical field. This is due to the efficiency with which text mining implements the automation methods for extracting valuable information from medical reports.

h. **Fraud Detection**: Text data analysis must be performed on correct data, that does not contain inaccurate information. To protect themselves from fraud, companies that hold most of their data in the text format, such as insurance and finance, have developed fraud detection methods and now are able to process claims safely.

i. **Spam Filtering**: Email filtering is a technique we all benefit from, and which was developed from the need to prevent contaminating computer systems with malicious activities.

j. **Automatic Text Translation**: A common application of text processing relies upon the accurate interpretation of actionable content in an input text or corpus of text (or spoken command) to create a facility of automatic translation of text documents from one language to another (Miner, et al., 2012).

**Table 2. Text mining use cases and related areas.**

| Topic | Practice Area |
|---|---|
| Feature selection | Classification |
| Sentiment analysis | Classification |
| eDiscovery | Classification |
| Keyword search | Information retrieval |
| Document clustering | Clustering |
| Document similarity | Clustering |
| Web crawling | Web mining |
| Link analytics | Web mining |
| Part of speech tagging | Natural language processing |
| Question answering | Natural language processing |
| Link extraction | Information extraction |
| Synonym identification | Concept extraction |

## 4. Summary

Text mining is a vast and complex research field, and often its documentation is heavy and extremely theoretical. Thus, the young researchers feel confused and get hurt in the fight for the necessary information. This article is dedicated to them and presents a methodological and conceptual theory of text mining along with the main methods behind it. Following an in-depth examination of the literature, the study shows the fundamental directions of text mining research such as classification, clustering, information retrieval and presents state-of-the-art applications that implement the concept of text mining to solve problems in the real world.

## 5. Acknowledgement

The work of the first author is supported by the project ANTREPRENORDOC, in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847.

The work of the second author was carried out in the framework of the research project DREAM (Dynamics of the Resources and technological Advance in harvesting Marine renewable energy), supported by the Romanian Executive Agency for Higher Education, Research, Development and Innovation Funding – UEFISCDI, grant number PN-III-P4-ID- PCE-2020-0008.

## References

Chakrabarti, S. (2002). *Mining the Web Discovering knowledge from hypertext data.* Bombay: Indian Institute of Technology.

Feldman, R. & Sanger, J. (2007). *The text mining handbook. Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Press.

*Kim Peek* (2021). *Wikipedia*. https://en.wikipedia.org/wiki/Kim_Peek.

Kowsari, K.; Meimandi, K. J.; Heidarysafa, M.; Mendu, S.; Barnes, L. & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*.

Li, N. & Wu, D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis. Support Syst.*, pp. 354-368.

Miner, G.; Delen, D.; Elder, J.; Fast, A.; Hill, T. & Nisbet, R. & Balakrishnan, K. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Waltham: Academic Press is an imprint of Elsevier.

*** (2020). *Natural Language Processing (NLP)*. IBM Cloud Education: https://www.ibm.com/cloud/learn/natural-language-processing.

Singh, J. & Gupta, V. (2016). Text Stemming: Approaches, Applications, and Challenges. *ACM Computing Surveys*, pp. 1-46.