A Comparative Review of Text Mining & Related Technologies

Roland Vasili
Dept. of Mathematics,
Informatics & Physics
Faculty of Natural Sciences
University of Gjirokastra
6001 Gjirokastra, Albania
rvasili@uogj.edu.al

Endri Xhina
Dept. of Informatics
Faculty of Natural Sciences
University of Tirana
1001 Tirana, Albania
endri.xhina@fshn.edu.al

Ilia Ninka
Dept. of Informatics
Faculty of Natural Sciences
University of Tirana
1001 Tirana, Albania
ilia.ninka@fshn.edu.al

Thomas Souliotis
Dept. of Informatics
University of Edinburgh
s1778881@sms.ed.ac.uk

Abstract

Text mining has become an established discipline in both research and business intelligence. It refers commonly to the method of extracting interesting information and knowledge from unstructured text. Society's future will be closely connected to handling large amount of data. Information may be available in various ways, either freely on the Web or on social networks. Text mining is a multi-disciplinary field in view of Data Mining, Computational Linguistics, Artificial Intelligence and Machine Learning, Statistics, Databases, Library and Information Sciences, and actually the new field of Big Data. Some of these disciplines will be compared based on the goals, data, algorithms, techniques and the tools they use, as well as the their outcome. All these subjects are similar, which is based on two fundamental facts: (1) all of them develop methods and procedures to process data, and (2) any data processing algorithm or procedure may belong to some or even all. The differences are in their perspectives. This difference in perspectives does not affect the procedures but it does affect the choice of them and, even more so, interpretation of concepts and results.

1. Introduction

The Text Mining field covers a wide research area and its methods can be applied in different contexts and for several purposes, depending on the needs of the specific task and the availability of data and expertise. To this aim, rather than being an exhaustive list of techniques and research directions, the Figure 1 shows that the text mining is a composite discipline that overlaps several branches of science. In the Figure 1 of [Tal16] Knowledge Data Discovery field from Fig. 4.1 in [Dea14] is added, that will help us understand that

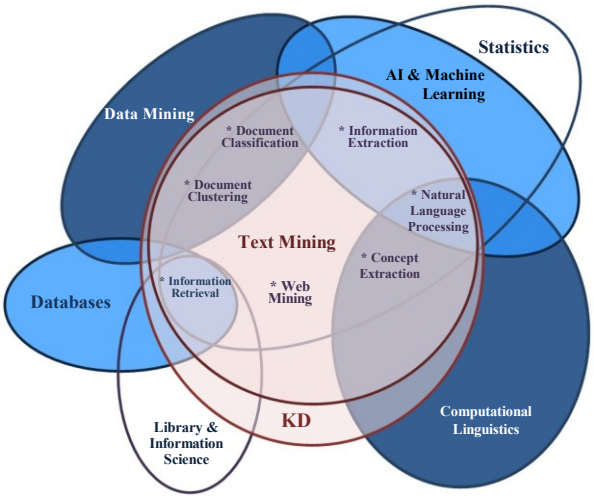

Figure 1: Multidisciplinary Nature of Text Mining (Composition of Fig.1 in [Tal16] and Fig. 4.1 in [Dea14])

the goal of all of these disciplines is knowledge discovery, and in this base they will be compared.

So, the goal of this paper is to outline the Text Mining landscape which in contrast to encompassed technologies like Data Mining, Natural Language Processing, Information Retrieval, Information Extraction, Artificial Intelligence and Machine Learning., it tries to depict the scale and potential scientific interaction with classic scientific areas, such as Statistics.

1.1 Definition of Text Mining

Text mining (TM), also called Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), is mainly used to define the procedure of extracting interesting and non-trivial data and knowledge from unstructured text ([Gök15]). There are many more definitions of text mining like the definition of the Oxford English Dictionary: "as the process or practice of examining large collections of written resources in order to generate new information, typically using specialized computer software". It widely covers a large set of related topics and algorithms for analyzing text, spanning various communities, including information retrieval, natural language processing, data mining, machine learning, many application domains web and biomedical sciences.

1.2 Text Mining Process

The text mining process (TM) can basically be summarized in three (3) steps below ([Kar05]):

- *Document Collection*
- *Document Preprocessing*
- *Text Mining Operations*

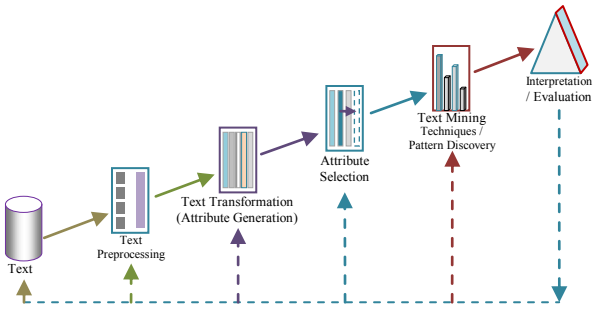

Figure 2: Text Mining Steps
(According to the process of Knowledge Discovery)

The steps of TM process regarding the output of results are shown in the Figure 2.
A TM system receives a collection of documents as input and then pre-processes each document by checking its format and set of characters. Next, these pre-processed documents go through the text analysis phase, by repeating the techniques until the required information is extracted. Figure 3 shows three techniques of text analysis, but other techniques,

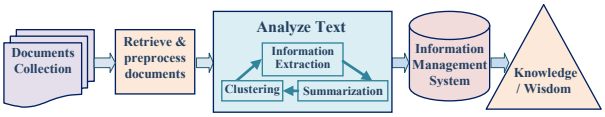

Figure 3: Text Mining Steps ([Lia12])

depending on the goal and the corporation, may also be used. Information derived from the extraction can be accessed by an information management system, producing valuable knowledge for the user of this system. Figure 4 analyzes the processing steps that a typical TM System follows.

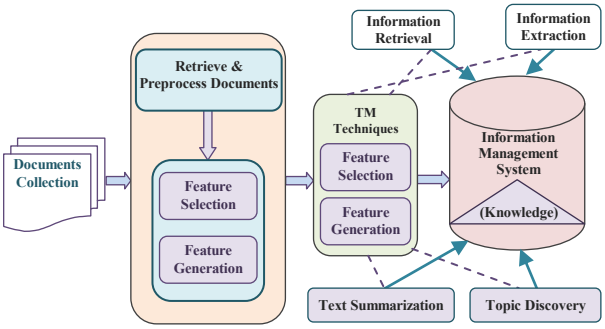

Figure 4: Text Mining System ([Lia12])

1.2.1 Document-Text Collection

The basic element of TM is the collection of documents of any text form. The number of texts in such collections may range from thousands to several millions.

Text collection can be static or dynamic. At the static approach, the original textbook total remains unchanged while at the dynamic, the textbook over time is classified into new or gets updated. Extremely large

collections and high-rate changing text collections are considered challenges and constitute the main object of Text Mining Systems. A peculiar example of a large dynamic collection of texts, used by millions around the world, is Pub Med (US National Library of Medicine 2018)[1]. It is an internet resource, which includes literature references related to biomedical and health sciences. It is worth pointing out that it includes over 25 million research reports in the biomedical field in which they are added, roughly 35,000 with 40,000 new items each month. In addition to that, unstructured data and free text are usually most of the data we encounter and this includes over 40 million articles in Wikipedia, 4.5 billion Web pages, about 500 million tweets a day, and over 1.5 trillion queries on Google in a year.

Therefore, to initiate the TM process, the user has to choose the desired collection of texts on which the procedure will be based on, and the variety of texts that will constitute the source of the data.

The following process involves the TM System, which has the ability (with the help of knowledge-discovery algorithms) to quickly and efficiently identify the patterns among a large number of natural texts.
But the realization of this requires the existence of elaborate text collections. For this reason, the most important TM process is the pre-processing phase of the texts under examination, and then, the successful implementation of the knowledge-discovery algorithms.

1.2.2 Text Preprocessing

Though this is considered to be the preliminary step to be conducted, before actually applying Text Mining algorithms/methods, it is a very important process. This routine itself is divided into a number of sub-methods which again have optional algorithms with their own set of advantages and disadvantages.

Most of the TM approaches are based on the idea that a text document can be described by the set of words contained in it i.e. bag-of-words representation. The preprocessing itself is made up of a sequence of steps ([Gup09]) (Figure 5). The steps are:

- *Text Structure Removal.*
- *Tokenization.*
- *Stopwords Removal.*
- *Filtering (Removing terms based on their length).*
- *Filtering (Removing terms based on their frequency).*
- *Part Of Speech Tagging (Syntactical and Semantical Analysis)*
- *Stemming.*
- *N- grams.*
- *Term weighting.*

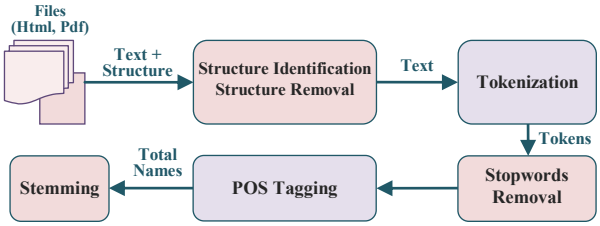

Figure 5: Text Preprocessing Steps

1.2.3 Text Representation

In order to apply TM techniques, the texts should be presented in a formatted form. We could say that the most familiar method of text representation is vector model. There are two main ways used for vector text representation:

- *Boolean Model.*
- *Term-Weight Model.*

2 Text Mining and Data Mining

Data Mining (DM) is a subfield of computer science which combines many techniques from statistics, data science, database theory and machine learning.

DM is simply the process of gathering information from huge databases that was previously incomprehensible and unknown and then using that information to make relevant business decisions. More simply, data mining is a set of various methods that are used in the process of knowledge discovery for distinguishing the relationships and patterns that were previously unknown. The final goal is the description of existing database data as well as forecasting and clarification of new data. We can therefore define data mining as a combination of various other fields like artificial intelligence, data room virtual base management, pattern recognition, visualization of data, machine learning, statistical studies and so on. The primary goal of data mining is to extract information from various sets of data in an attempt to transform it in

[1] https://www.ncbi.nlm.nih.gov/pubmed/: Accessed 1-4-2018

proper and meaningful structures for eventual use. It mainly includes procedures and tools of extracting patterns from the data set and relates exclusively to structured data. But in recent years, interest has also shifted to unstructured data (e.g. texts, images, paperwork, web pages, etc.) with the result of knowledge discovery from text (Text Mining). This shift is very important since most of the data nowadays are in unstructured textual form ([Gri08]). For example, a text file contains few structured elements such as author, title, date of creation etc. But it also contains large segments of unstructured text such as its summary and its contents. This requires both sophisticated linguistic and statistical techniques able to analyze unstructured text formats and techniques that combine each document with actionable metadata.

TM is an intense cognitive process through which the user interacts with a collection of texts using a set of analysis tools ([Seh04]). Similarly, as well as DM, TM aims at extracting useful information from data sources through recognition (identification) and examination of interesting patterns. Meanwhile, in the case of TM data sources are text collections, interesting motives are searched in unstructured textual data ([Nah02]).

Given the above definition, it is argued that the TM has its roots in the area of Knowledge Discovery (KD). Moreover, this is also used for the DM definition reference. Consequently, TM is similar to DM, mainly because in both cases, knowledge detection is based on processes of data preprocessing and pattern searching algorithms. However, this similarity may lead to overseeing their differences. Thus, the goal of the majority of the studies in those two areas, is to identify and analyze these differences.

3 Text Mining vs. Data Mining

The method of Knowledge Discovery from Data or Data Mining, namely finding useful patterns between data, is a very good solution for collecting and storing a huge volume of data. Though the scope of its implementation is extensive it is not a developing technology.

Instead, the knowledge discovery of textual data or Text Mining is a new method in the field of Knowledge Discovery, which is feasible because the information to be extracted refers to text.

The knowledge discovery from text resembles a lot to the classical method of knowledge discovery from data, since both are based on knowledge management. But, [Fra92] and [Raj97] concluded that the difference between these two domains is the type of data they use for Knowledge Discovery (KD). Thus, while DM uses data extraction techniques over structured data, TM does the same thing but for unstructured or semi-structured data, which is often referred to as textual data ([Gup09]).

Knowledge discovery from data is implemented in the databases where the data is structured and described by a unique structure where each instance of a problem is determined by a specific and fixed set of features ([Kan09]).

Instead, in the case of knowledge discovery from text, the data is semi-structured or unstructured and cannot be described by any set of fixed features ([Liu11]). For this reason, the method tries to bring the text in the appropriate form for the direct application of its computing applications.

In the case of knowledge discovery from texts, there are two approaches regarding the representation of the text. In the first approach, the presence of a feature (word) in a text is taken into consideration. Thus, when a new instance of the problem occurs, what is controlled is the presence of instances of the features (words) in different classes of the problem. The class in which most words are present is the desired class.

In the second approach, for each feature we hold the frequency of its appearance in a text. Thus a new instance class derives from the frequency of the presence of text words in different classes of the problem. The class in which the most displayed and the most frequent word of the text is the desired class.

In addition to the data type, [Dör99] separated these fields from the complexity of the steps that followed for knowledge discovery. The general steps followed by DM are:

(1) *identifying the data collection,*
(2) *preparation and features selection and*
(3) *distribution analysis.*

Even though TM does not deviate from these steps, the selection of features is different, since it is not practical to be responsible for the examination of the features and decide which of them should be used.

The other point where they differ is in distribution analysis, where multi-dimensional vectors are to be treated. This implies that there must be special versions and implementations of DM algorithms. However, these differences do not prevent [Hea99] from declaring that TM is an extension of DM. It is not clear to what extent this statement may be true as there are no studies that agree or disagree with it. Some basic

differences between TM and DM are also presented in [Ber09] work, which are seen in Table 1. In Table 2 we show some additional features :

Table 1 : Differences between TM and DM [Ber09]

Text Mining	Data Mining
Relies on unstructured or semi-structured data	Relies on fielded (structured) data
Term extraction takes place based on semantic algorithm	Involves numerically based statistical analysis
Documents containing overlapping concepts can be organized together	Allows for temporal analysis
Documents containing overlapping concepts can be placed together partially	Clustering based on coding
	Involves co-occurrence matrices and histograms

Table 2 : TM vs. DM

Base for C/sion	Data Mining	Text Mining
Concept	Data mining is a spectrum of different approaches, which searches for patterns and relationships of data.	Text mining is a process required to turn unstructured text document into valuable structured information.
Retrieval of data	With standard data mining techniques reveals business patterns in numerical data.	With standard text mining methods discovers a lexical & syntactic feature in the text.
Type of Data	Discovery of knowledge from structured data, which are homogeneous and easy to access.	Discovery of text from unstructured data which are heterogeneous, more diverse.

However, [Fan06] considers Text Mining as an interdisciplinary field based on other disciplines, such as Data Mining, Information Retrieval, Computational Statistics, Computer Science and Linguistics.

4 Text Mining vs. NLP

Natural language processing (NLP) is a subfield of computer science (CS), artificial intelligence (AI), and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding ([Nav18]), that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

TM refers to a subset of data mining concerned with discovering knowledge from various sources; especially, unstructured texts, which are still considered the greatest easily accessible source of knowledge. In TM, the main problem arises when trying to extract explicit and implicit ideas and semantic links among different ideas using NLP methods. The objective is to obtain a full understanding of vast amounts of text data. Many of the text mining algorithms extensively make use of NLP techniques, such as part of speech tagging (POS tagging), syntactic parsing and other types of linguistic analysis ([Kao07]). TM is greatly connected to NLP, but it is also related to processes in statistics, machine learning, information extraction, information management etc. During its procedure of finding out hidden secrets, TM has a very important part in upcoming applications of NLP field, like Text Understanding ([Sal18]).

Text Mining deals with the text itself, while NLP deals with the underlying/latent metadata.
Answering questions like - frequency counts of words, length of the sentence, presence/absence of certain words etc. is actually text mining.

NLP on the other hand allows you to answer questions like; - What is the sentiment? - What are the keywords? (using POS tagging & parsers) - What category of content it falls under? - Which are the entities in the sentence? etc.

Text mining is the process of mining text in the context of data mining, when we consider as data just text. Mining is about extracting useful information from the available data. Information could be patterns in text or matching structures but the semantics in the text is not considered. The goal is not about making the system understand what does the text convey, rather about providing information to the user based on a certain step by step process.

Natural language is what humans use for communication. Processing such data is NLP where the

data could be speech or text. Thus, the main goal is understanding what is the semantic meaning conveyed in it. Therefore, we can understand why we care about grammatical part of speeches and the lexical relations among them.

Speech recognition systems could be a part of NLP, but it has nothing to do with TM. It may seem like NLP is a more general, significant concept, because it uses TM, however, it's actually the other way round. TM uses NLP, because it makes sense to mine the data when we understand the data semantically.

Table 3 shows the top 5 Comparison between Text Mining vs. Natural Language Processing:

Table 3 : TM vs. NLP Comparison[2]

Basis of Comp/son	Text mining	NLP
Goal	Extract high-quality information from unstructured & structured text. Information could be patterned in text or matching structure but the semantics in the text is not considered.	Trying to understand what is conveyed in natural language by human- may text or speech. Semantic and grammatical structures are analyzed.
Tools	• Text processing languages like Perl • Statistical models • ML models	• Advanced ML models • Deep Neural Networks • Toolkits like NLTK in Python
Scope	• Data sources are documented collections • Extracting representative features for natural language documents • Input for a corpus-based computational linguistic	• Data source can be any form of natural human communication method like text, speech, signboard etc • Extracting semantic meaning and grammatical structure from the input • Making all level of interaction with machines more natural for human
Outcome	Explanation of text using statistical indicators like • Frequency of words • Patterns of words • Correlation within words	Understanding what conveyed through text or speech like • Conveyed sentiment • The semantic meaning of the text so that it can be translated to other languages • Grammatical structure
System Accuracy	Performance measure is direct and relatively simple. Here we have clearly measurable mathematical concepts. Measures can be automated	Highly difficult to measure system accuracy for machines. Human intervention is needed most of the time. For example, consider an NLP system, which translates from English to Hindi. Automate the measure of how accurately system doing translation is difficult.

To conclude[2], both TM and NLP try to extract information from unstructured data. TM is concentrated on text documents and mostly depends on a statistical and probabilistic model to derive a representation of documents. NLP tries to get semantic meaning from all means of human natural communication like text, speech or even an image. NLP has potential to revolutionize the way humans interact with machines e.g. AWS Echo and Google Home.

5 Text Mining vs. Web Search

Text Mining is different from the concept referred as Web Search. In addition to the differences between TM and DM, explained above, [Gup09] tries to establish boundaries between TM and web search.

The main part of web search is the web engine. A web engine has three main parts: (1) Crawler: Gathers the contents of all web pages (using a program called a crawler or spider), (2) Indexer: Organizes the contents of the pages in a way that allows efficient retrieval (indexing), and (3) Ranker: Takes in a query, determines which pages match, and shows the results (ranking and display of results).

The difference from TM lies in the fact that internet users are searching for something that exists, which has been found and was previously written by a person, while TM is aimed at detecting previously unknown information ([Gök15]). So, the problem is to separate

[2] https://www.educba.com/important-text-mining-vs-natural-language-processing/: Accessed 5-9-2018

the material that is not related to your needs and keep the essentials in order to find the information you need.

6 Text Mining vs. Information Retrieval

Information retrieval ([Rij79]) is used to search documents or information in documents. Generally, it is a subject of information science and computer science. Its main uses are for access to books and journals from universities and public libraries and the most notable application is as web search engines. With the great improvement of the web, a huge amount of information is available online for the daily user. A user will try to retrieve relevant information from web search engines with a question or a query. Information retrieval helps the process to return a set of documents that meets the requirements of the user's query.

The concept of information retrieval is really old. The first time that someone mentioned in a paper the ability of a computer to retrieve relevant pieces of information was in 1945 in the article "As We May Think" by Vannevar Bush [Sin01]. Since then, many other techniques have been shown until the last two decades in which web search engines have boosted the need of a large-scale information retrieval system.
There are different mathematical models for the information retrieval. Common models are set-theoretic, algebraic and probabilistic models. Set-theoretic models represent documents as sets of words or phrases. Algebraic models convert documents and words in vectors, matrices and tuples. Probabilistic models treat the information retrieval as a probabilistic inference.

It is important to differentiate between TM and Information Retrieval (IR). We can say that TM represents subsequent evolution (transformation) of IR.

In retrieving information, the search is conducted only for texts that already contain the answers to questions rather than search for new knowledge ([Hea99] & [Seh04]). In general, IR's goal is to extract all documents that are closer to the answer of a question. Thus, it is the activity of obtaining information resources (usually documents) relevant to an information need from a collection of information resources ([Fal95], [Man08]). Searches can be based either on metadata or on full-text indexing. Therefore, IR mostly focuses on facilitating information access rather than analyzing information and finding hidden patterns, which is the main purpose of text mining. IR does not care a lot about processing or transforming text, whereas text mining can be considered as going beyond information access to further aid users to analyze and understand information and ease the decision making. Table 4 illustrates some of the differences between TM and IR:

Table 4 : Differences between TM and IR

Basis of Comp/son	TM	IR
Goal	Extract high-quality information from unstructured & structured text. Information could be patterned in text or matching structure but the semantics in the text is not considered.	Finding answers and information that already exist in a system Creating answers and new information by analysis and inference – based on query
Scope	• Data sources are documented collections • Extracting representative features for natural language documents • Input for a corpus-based computational linguistic	• Unstructured information (text, images, sound, though spoken, image, video, email, Web, multimedia, ...) • Structured information ((DBMS), Data analysis systems, Expert systems)
Outcome	Explanation of text using statistical indicators like • Frequency of words • Patterns of words • Correlation within words	Text retrieval deals with computerized retrieval of machine-readable text, Speech retrieval deals with speech, Cross-language retrieval uses a query in one language & finds documents in other languages , Q-A IR systems retrieve answers from a body of text, Image retrieval finds images on a theme. IR dealing with any kind of other entity or object

The most important distinction between TM and IR is the output of each process. In the IR process the result consists of documents, some of which may be

clustered, ordered or scored but at the end to get the information we have to read the documents. In contrast the results of TM process can be features, patterns, connections, profiles or trends, and to find the information we need, we don't necessary have to read the documents.

7 Text Mining and Statistics

7.1 What is Statistics?

Statistics consists of a set of mathematical methods related to the collection, organization and analyzation of data. These techniques (and more) are used so as to extract some useful outcomes depending on our needs, while all the potential techniques used are categorized in two main categories the descriptive and the inferential.

In the descriptive statistics the initial data are used only for processing reasons and producing some useful conclusions based on them. However, no potential forecasts are made based on this data and no results are really inferred other than some simple outcomes only for the current data. These predictions are actually part of the second big category, the inferential statistics, where useful estimations are made for future events based on the current data.

7.2 Statistics: The Science of Learning from Data

Statistics is another broad subject which deals with the study of data, that is widely applied and plays a very important role in all areas of science. Statistics provides the methodology for making conclusions from data. It gives different methods to gather data, analyze them and interpret results and is widely used by scientists, researchers, and mathematicians in solving problems.

Though statistics provides the methods for data collection and analysis, it helps to obtain information from numerical and categorical data. Categorical data refers to unique data, e.g. blood group of a person, marital status, etc.

Statistics is highly significant in data related studies because it helps in,

- *Deciding the type of data required to address a given problem*
- *Organizing and summarizing data*
- *Analysis to be done to draw conclusions from data*
- *Assessing the effectiveness of results and to evaluate uncertainties*

The methods provided by statistics include,

- *Design for planning and conducting research*
- *Descriptions which implies exploring and summarizing data*
- *Making predictions and inference using the phenomena represented by data*.

So, Statistics is essentially a part of the process of TM. It is the science of learning from data. Also, it provides tools and techniques for dealing with large amounts of data. Statistics includes a number of processes, like:

- *The planning behind data collections*
- *Data management*
- *Drawing inferences from numerical data facts*

7.3 Text Mining vs. Statistics

Scientific literature suffers from lack of articles on comparisons such as TM and Statistics, even on DM and Statistics. So, since TM is a subfield of DM, we will base our comparison to DM and will check if it is valid for TM then will present it in the comparative Table 5[3] bellow.

In practice, comparing Statistics means comparing what is defined in terms of a set of tools, namely those being taught in graduate programs, i.e. Probability theory, Real anlysis, Measure Theory, Asymptotics, Decision theory, Markov chains, Martingales, Ergotic theory, etc. The field of Statistics seems to be defined as the set of problems that can be successfully addressed with these and related topics ([Fri98]).

For this reason, our comparison will not be a thorough one, based on multiple literature resources, but a little more simplified. Yet this analysis will still based on some scientific criteria. Our sources will be multiple web sources, but mainly three scientific articles: [Fri98] by Jerome H. Friedman of Stanford University, that explains the connection between Statistics and DM, [Sap00] by G. Saporta, that focuses on how DM could be used in official statistics and [Has14] by Hassani, Saporta, and Silva, that presents a thorough review of published work to date on the application of data mining in official statistics, and on identification of the techniques that have been explored.

---

[3] https://www.educba.com/data-mining-vs-statistics/: Accessed 5-9-2018

Table 5 : TM vs. Statistics Comparison Table

Text Mining	Statistics
Explorative: It digs out data first, builds model to discover novel patterns & make theories.	*Confirmative*: It provides theory first and then tests it using various statistical tools.
Data used is Numeric or Non numeric.	Data used is Numeric.
Inductive Process (Generation of new theory from data)	Deductive Process (Does not involve making any predictions)
Data collection is less important.	Data collection is more important. ([Sap00])
Involves Data Cleaning.	Clean data is used to apply statistical method.
Needs less user interaction to validate model hence, easy to automate.	Needs user interaction to validate model hence, difficult to automate.
Suitable for large data sets	Suitable for smaller data sets
It's an algorithm which learns from data without using any programming rule.	Formalization of relationship in data in the form of mathematical equation.
Use heuristics think (rules used to form judgments and make decisions)	Does not have scope for heuristic think.
Classification, Clustering, Summarization, Estimation, Association Rules, Topic Modeling, Visualization	Descriptive Statistical, Inferential Statistical
Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Certain Scientific Applications etc.	Demography, Actuarial Science, Operation research, Biostatistics, Quality Control etc. ([Has14])

8 Conclusion

In this article we attempted to briefly describe the differences of Text Mining with other related disciplines, while making a concise presentation.

In summary, it is noted that TM and all these sciences (even statistics) may seem indistinguishable due to its close connection. It is clear, however, that statistics is actually a tool or method for all these sciences, while most of them spread over a wide domain where a statistical method is an essential component. Text Mining has developed recently with big data and will continue to grow in the following years as data growth seems to be never-ending. This also applies to the other disciplines, which means that the data driving the algorithms, methods and decisions need to be high-quality. Nonetheless, all disciplinary fields described briefly in this review, cover the major areas of working with data and problems on various areas related to this data. The emerging picture reveals a blend of theory and practice that reflects each discipline rather than a unified system. Hopefully, a productive merging of TM approaches through increased cross-disciplinary research can develop and advance not only TM but all these fields. The rate of change in the text mining field is so rapid that the information is likely to be measurably different in the following years.

References

[Ber09] C. Berkouwer. Master Thesis: *The Reflection of Foresight in Defense Policy Making : A Comparative Study of the United Kingdom and the United States*, March 2009

[Dea14] J. Dean. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners: pp 56*. John Wiley and Sons, Inc., 2014

[Dör99] J. Dörre, P. Gerstl, R. Seiffert. *Finding Text Mining: Nuggets in Mountains of Textual Data. KDD '99 Proceedings of the Fifth ACM SIGKDD Intern. Conference on K. Discovery and Data Mining: 398–401*, August 1999

[Fan06] W. Fan, L. Wallace, S. Rich, & Z. Zhang *Tapping the Power of Text Mining. Communications of the ACM, 49 (9): 76–82*, September 2006

[Fal95] C. Faloutsos, D. W Oard. *A survey of information retrieval and filtering methods. Technical Report. University of Maryland at College Park*, MD, USA 1995

[Fra92] W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus. *Knowledge Discovery in Databases : An Overview. AI Magazine, 13 (3): 57–70*, September 1992

[Fri98] J. H. Friedman. *Data Mining and Statistics: What's the Connection? Computing Science and Statistics Vol. 29 (1): 3-9*, Ed. D. Scott 1998

[Gri08] S. Grimes. *Unstructured data and the 80 percent rule. Clarabridge Bridgepoints newsletter 23, column, "Experts Corner: Seth Grimes."*, August 2008

[Gök15] A. Gök, A. Waterworth, P. Shapira. *Use of web mining in studying innovation. Scientometrics 102 (1): 653–671*, Jan. 2015

[Gup09] V. Gupta, G. Lehal. *A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence, 1(1): 60–76*, Academy Publisher, August 2009

[Has14] H. Hassani, G. Saporta, E. S. Silva. *Data Mining and Official Statistics: The Past, the Present and the Future. Big Data Vol. 2 (1): 34-43*, March 2014

[Hea99] M. A. Hearst. *Untangling Text Data Mining. Proceedings of ACL '99: the 37th Annual Meeting of the Association for computational Linguistics*, University of Maryland, June 1999 (invited paper)

[Kan09] Y. Kano, W. A. Baumgartner, L. McCrohon, S. Ananiadou, K. B. Cohen, L. Hunter, T. Tsujii. *Data mining: concept and techniques. Oxford Journal of Bioinformatics, Volume 25, Issue 15: 1997-1998*, August 2009

[Kao07] A. Kao, S. R. Poteet. *Natural language processing and text mining*. Springer, 2007

[Kar05] H. Karanikas, Th. Mavroudakis. *Text Mining Software Survey. RANLP Text Mining Workshop No 1: 39-48*, September 2005

[Lia12] S. H. Liao, P. H. Chu, P. Y. Hsiao. *Data Mining Techniques & Applications - A Decade Review from 2000 to 2011. Expert Systems with Applications, Vol. 39 (12): 11303–11311*, Elsevier Ltd., September 2012

[Liu11] F. Liu, X. Lu. *Survey on text clustering algorithm. Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS), China, 901-904*, 2011

[Man08] C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York 2008

[Nah02] U. Y. Nahm, J. R. Mooney. *Text Mining with Information Extraction. Technical Report SS-02-06*, Department of Computer Sciences, University of Texas, March 2002

[Nav18] Roberto Navigli. *Natural Language Understanding: Instructions for (Present and Future) Use. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence: 5697-5702*, Early Career, July 2018

[Raj97] M. Rajman, R. Besançon. *Text mining: Natural language techniques and text mining applications. Data Mining and Reverse Engineering: Searching for Semantics: IFIP TC2 WG2.6 IFIP 7th Conference on Database Semantics (DS-7): 50-66*, January 1997

[Rij79] C. J. Van Rijsbergen. *Information Retrieval, London: Butterworths, 2nd edition*, November 1979

[Sal18] S.A. Salloum, A.Q. AlHamad, M. Al-Emran, K. Shaalan. *A Survey of Arabic Text Mining. Studies in Computational Intelligence, vol 740: 417-431*, Springer, Cham, January 2018

[Sap00] G. Saporta. *Data Mining and Official Statistics. Quinta Conferenza Nationale di Statistica*, ISTAT, Roma, November 2000

[Seh04] A.K. Sehgal. *Text Mining: The Search for Novelty in Text. Ph.D. Comprehensive Examination Report, Dept. of Computer Science*, The University of Iowa, April 2004

[Sin01] A. Singhal. *Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43*, December 2001

[Tal16] R. Talib, M. Kashif, Sh. Ayesha, F. Fatima. *Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science & Applications Vol. 7(11): 414-418*, November 2016