ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2022, 3(2), 107-120, https://doi.org/10.46656/access.2022.3.2(2)

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

BUSINESS DEMANDS FOR PROCESSING UNSTRUCTURED TEXTUAL DATA – TEXT MINING TECHNIQUES FOR COMPANIES TO IMPLEMENT

Denitsa Zhecheva[1], Nayden Nenkov[2*]

[1,2] Konstantin Preslavsky University of Shumen, Shumen, Bulgaria

e-mails: [1] mrszhecheva.denitsa@gmail.com, [2] n.nenkov@shu.bg*

Received: 22 February 2022 Accepted: 04 April 2022 Online Published: 17 April 2022

ABSTRACT

The rapid development of technology has caused a pervasive change in the way people and businesses live. Making sound business decisions is unthinkable without processing a large amount of data (publicly available and collected on the basis of problems) with high accuracy and quality. The importance of unstructured data acquires various sources is growing. Of particular value is the continuous flow of textual information that is generated every minute around the world in a different form (unstructured textual data). This is also the subject of this article. The aim of the article is to provide an analytical overview of the main methods of word processing that are applicable for pragmatic analysis of information flows from companies, such as: extraction, summarization, grouping and categorization of text. Some methodologies are based on NLP (Natural Language Processing), others on Bayesian logic and statistical theory and practice. From the review of various publications on the topic, conclusions are proposed for their practical applicability. This allows for an objective choice of appropriate tools for processing unstructured information and business intelligence. The results of the study can be successfully used to improve managerial decision-making, improve the quality of work of employees and reduce errors in overall marketing planning.

Keywords: unstructured textual data, business intelligence, NLP (Natural Language Processing), text mining, text extraction, text summarization, clustering, text categorization, text retrieval

JEL classification: *C15, C81, C82*
Paper type: Research article

Citation: Zhecheva, D., Nenkov, N. Business demands for processing unstructured textual data – text mining techniques for companies to implement. *Access to science, business, innovation in digital economy*, ACCESS Press, 3(2): 107-120. https://doi.org/10.46656/access.2022.3.2(2)

INTRODUCTION

In 1998, in one of its "Industry review" reports Merrill Lynch, a trader in the sale and purchase of corporate securities, indicates that: "unstructured data comprises the vast majority of data found in an organization, some estimates run as high as 80%." (Lynch, 1998). This trend retains as a result of the invention of new technology and innovations and other authors have measured that "unstructured data constitute 95% of big data" (Gandomi and Haider, 2015) and it is doubling its volume every three years (Cukier & Mayer-Schoenberg, 2013). The figures in 2020s are pointing out some big numbers as well. According to some statistics in 2020 people created 1.7 MB of data every single second of their time which means that we have created 2.5 quintillion data bytes in 2020 per day. If we have to link these statistics to unstructured data and restrict the numbers to emails and posts on the social media source "Tweeter" – every day 306.4 billion emails are sent and 500 million Tweets

are made (TechJury, 2021). It is clear that it is up to businesses to put in use this voluminous data, most of it unstructured, and extract the sensitive information from it also knows as business intelligence.

Not surprisingly, this need for information extraction and its conversion by businesses has been identified a long time ago. In 1865, Richard Millar Devens pointed the term "business intelligence" (Devens, 2016) with the same meaning as the world understands it today – a process nested in the company which is analyzing data and then working through it to deliver information that could be used for decision-making. More than 140 years later the business consulting companies across the globe are reporting the same – business intelligence is a top priority for decision-makers in companies (Gartner, 2009). Yes, old-dated term, but new rules are applied nowadays. These rules are shaped by the mentioned above voluminous unstructured data mass, captured today within "Big data". The majority of authors are defining big data as a collection of huge volume of data that has an architecture constructed on the "Roman census" method and the data accumulated is stored in unstructured format. The data collected is categorized as repetitive and nonrepetitive big data. It has been discovered by researchers that a micro part of the repetitive data is capturing some business value in comparison to the nonrepetitive data which has gathered the fuel of knowledge the business needs for better decision-making (Inmon, Linstend & Levins, 2019). This unstructured data that has a high percentage of business value is being stored in different formats and big part of it is held by unstructured text.

Corporate unstructured textual data could be found as we already illustrate above in emails and social media posts, but also in clients' reviews, web pages, blogs, news, CRM systems files, survey responses, etc. However, it is not limited to online activities only. There are numerous offline sources of unstructured textual data for business as well (such as marketing materials, contracts, offers, job performance evaluations, meeting summaries, formal correspondence, etc.). The sources are really diverse as well as the methods to process and analyze the generated textual data. In this regard a key contribution that this paper attempts to bring forth is the conceptualization of scientifically grounded collection and presentation of unstructured textual data and more specifically - the inspection of methods to be used to execute text analysis on newly generated organization-related data supported by real-life examples of application. This paper is organized as follows: (1) it provides highlights of the published hitherto methods for textual analysis of unstructured data and underlines the authors' point of view on them; (2) the paper then expands to a discussion which may assist and navigate decision makers in companies when choosing the right text mining method/s for business intelligence extraction suitable for their business model. Previous research has showcased and analyzed separate text mining methods while this paper tries to incorporate a larger body of up-to-date literature with practical business implications.

The paper contains generic theories which are not tool-specific and additionally - specific tools established and pointed by other authors as complementing examples.

MATERIALS AND METHODOLOGY

1. Conceptualization of unstructured data

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

The main path standing in front of every business when a decision to deal up with unstructured textual data is taken is starting with the collection of it through a specific source/s and usually the options are close to myriad. We must take into account that the extracted data could be either subject to individual analysis with its only outcomes or it could then be added to the previously collected and structured data and become part of an integral data analysis which could give an overall view of the business operations and processes. The decision to pass up the separate interpretation depends only on the specific pre-set main purpose for executing the whole data processing.

After collection of data, a transformation from unstructured to structured data follows which would activate one's ability to make sharp analysis for the business. We have to mark that for some of the reviewed authors this transformation can be skipped which will lead to working directly with the data and analyzing it in its unstructured format. This approach is also mentioned by W. Inmon and A. Nesavich: "One approach is to look at and gather the textual data in the unstructured environment. When there, the textual data is analyzed and manipulated in the unstructured environment. The unstructured environment appears to be a natural place to do textual analysis because, after all, the text resides in the unstructured environment." (Inmon & Nesavich, 2007). This approach assumes the application of a different analysis pattern each time we collect unstructured data. This is surely a time consuming process and one will not be able to use the positives of an encapsulated business analysis described above. The benefits of the transformation-based approach which reconstructs unstructured data to one with a structure and then works with it through data analytics tools are even more affirmative for the business because the process of transformation could be established as continuously ongoing transformation and this way to become a channelized business unit in the company operating with a good level of autonomy and resource efficiency.

By reviewing the approach that is transforming the unstructured data to structured, we can delve into the works of many scholars and their research in the field of text analysis. The main idea standing behind the transformation process is to find a context in the unstructured textual data. Part of the companies have found the need of this contextualization long time ago. The first attempts have been made through mechanism called "NLP". NLP, meaning Natural Language Processing, is a technology that contributes the human-like understanding of text in different languages by a computer. It is referred to as a part of Artificial Intelligence field of study. Natural language processing systems were developed as early as back in 1970s such as "SHRDLU" in Massachusetts Institute of Technology (MIT) (Stephens, 2002). Over the time the NLP mechanism has been repeatedly reconsidered and improved because of its limitations like the fact that NLP does not consider the emphasis on words. Or the fact that it is difficult for NLP to find the logic behind words in a specific language as every language has expanded through the years (Inmon, Linstend and Levins, 2019). These are some of the main reasons for NLP not to be used as a single approach for analyzing unstructured textual data. Consequently, other tools and processes are in need to be developed as add-on to NLP and to extract context for business intelligence and deliver value to companies (Hänig & Schierle, 2010). Also,

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2022, 3(2), 107-120, https://doi.org/10.46656/access.2022.3.2(2)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

owning to the dramatic changes that occurred with the more diverse communication channels and sources, the evolution of contextualization reached to text mining.

We can conclude that the improvement of the combination between NLP and artificial intelligence techniques gives a good effect on unstructured textual data processing. It decreases the level of complexity to analyze and also raises the level of utilization of the collected data. Another positive consequence can be also traced in the above two – the more data businesses are processing; the more accurate decision making will become.

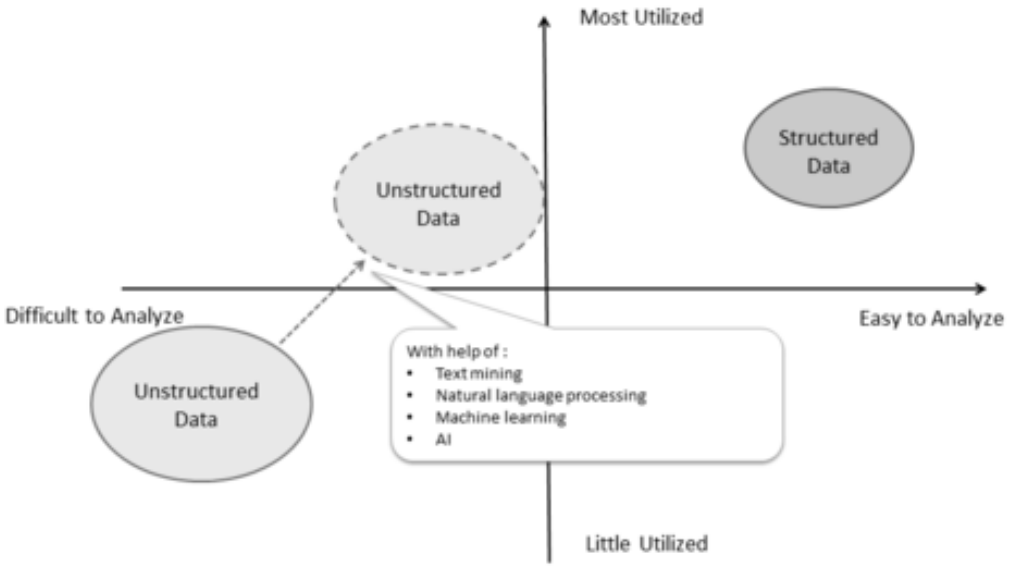

Figure 1. The use of unstructured data becomes easier

Source: https://blog.yokogawa.com/blog/find-value-in-unstructured-plant-data

2. Conceptualization of text mining techniques

The vast majority of authors in the field of text mining classified the same five main types of text mining techniques: (1) text extraction, (2) text summarization, (3) clustering, (4) text retrieval and (5) text categorization.

2.1. Text Extraction

While focusing on the first type of text mining technique Jiang opined that text extraction is the method for "finding structured information from unstructured textual data" (Jiang, 2012). More specifically, the text extraction process is connected to the scanning through the text and finding relevant words and expressions. We can distinguish between three main types of text extraction through the reviewed studies: entity recognition, relation extraction and keyword extraction which can be used properly in almost every textual data analysis by businesses.

Entity recognition (entity extraction) could be found in scientific literature also as "named entity recognition" and it refers to the processing of unstructured text and finding a set of episodes of words and phrases which are belonging to a specific predefined Named Entities categories (Mikheev, Moens & Grover,

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

1999). These categories could be locations, names of people, names of organizations, specific dates, etc. There are entirely automated extraction approaches that have been developed through the years and are designed with the help of machine learning (fast and effective alternatives for businesses). A good example is "LODifier". It uses the nature of named entity recognition, the essence of deep semantic analysis, also Semantic Web vocabularies and word sense disambiguation for its main purpose – the extraction of named entities and links between them in text. It has high practical potential due to the options for conversion into a RDF (Resource Description Framework) representation of the collected data and the linkage to DBpedia and WordNet (Augenstein, Padó & Rudolph, 2012).

The second type of text extraction is relation extraction. The central conception embedded in this term is extracting semantic relationships from an unstructured textual data. There is a simple reason why human cannot annotate semantic information from the data – because of the volume and different structure variations of the text which will make it difficult for humans to process. Thus at all times computer annotates the semantic information through supervised and semi-supervised approaches (Bach & Badaskar, 2007).

The third type of text extraction is keyword extraction. It refers to a powerful approach for keywords extraction in order to capture the 'key' components of a specific text (Firoozeh et al., 2020). Keywords are subject to one or more words which are giving a glimpse of the content of the text. The general perception in most of the research papers in this field of study is that keyword extraction is part of the bigger body of literature named 'text extraction' but there are also authors who are referring this term to 'information retrieval (IR)'. Their argument is based on the fact that keywords are used regularly within information retrieval systems as queries determinants because of their ability to be defined, shared and remembered easily (Rose et al., 2010). As with the other types of text extraction, the artificial intelligence methods are also implemented to improve and upgrade the classic process of keywords extraction. This approach has firstly been introduced by Turney who gives a detailed description of the supervised machine learning tasks embedded in the keyword extraction process (Turney, 2000).

After reviewing the main types of text extraction, we can arrive at the conclusion that none of all text extraction types can be defined as a unified domain of studies. The main reason standing behind this conclusion is connected to the core of the unstructured textual data – it is never repeatable and it is unique on its own as a representation of the unique thoughts of humans and different variations of expressing them. Other logic is connected to the diverse sources from which the data could be extracted. As we already described in this paper, the structure of every source is one of a kind and we cannot count to lay out a specific pattern from any kind of structure.

2.2. Text Summarization

The second type of data mining main technique mentioned is text summarization. The term is defined as method operated by a computer which generates summaries of a textual data. It is subject of research in both natural language processing and artificial intelligence domains (Impelsys, 2021) which is an expected fact given the already discussed specificities of data processing in the beginning of this chapter. Although it is a

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

widely used technique nowadays, it was first researched in 1950s from Lunch (1958) and Baxendale (1958) (Bhartiya & Singh, 2014). "Text summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs from the original document and concatenating them into shorter form. Abstractive summarization aims to interpret and examine the source text and creates a concise summary that usually contain compressed sentences or may contain some novel sentences not present in the original source text." (Bhide, 2016). Despite the fact that there are multiple studies on this topic, the research community stands for three core characteristics of the text summarization only. A summary as result to the text summarization process should be short, it must contain important information and it can be extracted by single or multiple text/s.

Based on the last characteristic mentioned there are single-document summarization and multiple-document summarization methods. Those which are based on providing a summary from a single document listed by two researchers from Carnegie Mellon University (Das & Martins, 2007) are mainly based on: (1) Naïve-Bayes methods which are capable of learning from the data (Edmundson, 1969), (2) Decision trees models which are studying the usefulness of single features in the text (Lin, 1999), (3) Hidden Markov models which are using sequential model for uncovering dependencies between sentences (Conroy & O'leary, 2001), (4) Log-Linear models used by some authors to show that this model extracts summaries better than Naïve-Bayes model (Osborne, 2002), (5) Neural networks algorithms which use also third party datasets to extract a summary (Svore, Vanderwende & Burges, 2007).

Das and Martin managed to describe in details the field of multiple-document summarization (Das & Martins, 2007) that has been settled by researchers in Columbia University and their system called SUMMONS – the first multiple-document summarizer (McKeown & Radev, 1995). Next big contribution in this field of study is a text summarization processed with the power of graph-based method where the summary content is introduced as nodes and edges of a graph (Mani & Bloedorn, 1997). One year later the topic-driven summarization was introduced and also the term maximal marginal relevance was set (Carbonell & Goldstein, 1998). Another approach has been invented in 1999 and described in details few years later. It is called cluster centroids where all documents are represented as bags-of-words (Radev *et al.*, 2004). One more approach for multi-document summarization is brought to our notice and it solves the problem of summarization of multi-language document/s (Evans, McKeown & Klanvas, 2005).

2.3. Clustering

Clustering is defined as unsupervised grouping of specific objects (patterns) into groups (clusters) based on their characteristics (Tsai, Wu & Tsai, 2002). The data which is subject to the data mining process in cluster analysis is affected by certain clustering algorithms to be transformed from raw data to clusters of specific data at the end of the process. A variety of models can be used on the raw data and some authors define seven of them as main applications of the cluster modeling: (1) hierarchical algorithms, (2) partitions algorithms, (3) density-based algorithms, (4) grid-based algorithms, (5) model-based algorithms, (6) graph-based algorithms and (7) combinational algorithms (Patel & Thakral, 2016). Each one of these algorithms has advantages and

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

disadvantages and it is good for businesses to choose wisely which model to operate during their collection of textual unstructured data depending on its characteristics and the results set to be achieved through the process of analysis.

Hierarchical models in clustering perform a hierarchical grouping breakdown on the raw data. There are two types of hierarchical algorithms – divisive and agglomerative algorithms. The divisive ones, also known as top-down methods, are defined by some authors as three-step algorithms. First, a splitting process is subdivisioning the raw data into two sub clusters. In the next step, a local evaluation of the results of the preliminary detached divisions is made. The final step consists of a resulting dendrogram – a result from the process of the determination of the node levels (Roux, 2018). The other type of hierarchical clustering represented by agglomerative algorithms, also known as bottom-up methods, is approach in which each observation is presented as a single cluster. Then, pairs of clusters are shaped as the algorithm moves up to the top of the hierarchy (Sasirekha & Baby, 2013). Nowadays there are cultivated ready-to-use packages (libraries) available using the hierarchical clustering methods which can be imported directly to a variety of advanced integrated development environments. An example for such package is the C++ library 'Fastcluster' which is also "featuring memory-saving routines for hierarchical clustering of vector data and improves both asymptotic time complexity and practical performance comparing the existing implementations in standard software" (Müllner, 2013).

Partitioning algorithms divide the raw data into multiple subsets (partitions): k-means algorithms, CLARA algorithms, etc. While in density-based clustering algorithms, there are clusters made of data objects that could be found in regions with a high density of data objects. These clusters are disconnected by regions with low density of data objects which are mainly defined as outliers or noise (Kriegel et al., 2011). There are clustering-based solutions dealing with stream data using density-based techniques. An example here is 'D-Stream'. The algorithm behind this framework can produce and correct clusters in real time (Chen & Tu, 2007).

The other clustering algorithms are also providing practical support to the process of data analytics: grid-based algorithms have multi-dimensional grid architecture and they are convenient because of their quick processing time; model-based algorithms seek to enforce data to some mathematical model and this way they are defined as both statistical and artificial intelligence approaches; graph-based algorithms come in handy when the need of visualization of objects that are related to each other has arisen; combinatorial clustering "is performed by minimizing an objective function under a condition to satisfy a constraint." (Kumagai et al., 2020).

2.4. Text Classification

Text classification, also known as text categorization, can be defined as the process of classifying textual data based on predefined categories, topics, etc. Text classification is a central approach in NLP, working together with machine learning methods. The majority of researchers in the field of text mining have come to the conclusion that text classification could be used in broader area of applications such as: topic analysis (understanding the topic of a specific text), spam detection (the detection of spam messages/e-messages

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

throughout the context of the source), intent detection, language detection (the detection of a language used in a given text), etc. We can notice that special attention is paid to sentiment analysis through multiple studies.

Sentiment analysis, also known as opinion mining, is defined as the process of extracting and analyzing human's sentiments towards specific domains such as products, services, etc. (Birjali, Kasri and Beni-Hssane, 2021). Lately, a popular approach to work with sentiment analysis is the Long Short-Term Memory (LSTM) approach. LSTM is proposed Reiter and Huber (1997) and its core is to polish the architecture of recurrent neural network and more specifically - to handle the problem with long-term dependencies by the classic RNNs examined by Hochreiter in 1991 and Bengio in 1994. LSTMs are designed to remember information for long periods of time and with a different structure of the repeating module from the classic RNNs.

Some authors propose an innovative approach for analyzing textual data by using text categorization and the Latent Categorization Method (LCM). This combination was used for categorizing and automatically naming "IS research topics in 14,510 abstracts from 65 Information Systems journals". This method managed to form categories by a wide array of text (Larsen et al., 2008).

2.5. Text Retrieval

Text retrieval (TR) is a process of finding an accurate textual data with an unstructured nature compatible with predefined criteria by using a query. It is a widely used technique to accommodate the need of finding ad hoc information in relevant document from a collection of documents. (Zhai & Massung, 2016). We can indicate two processes involved in the series of actions to perform text retrieval: (1) indexing which is using metadata to frame a computable representation of the content, (2) retrieval which matches queries to documents. This pair of query document indexing and retrieval is built on lexical approach. The currently used worldwide search engines are based on same lexical approach in combination with bag-of-words model (Tamine & Goeuriot, 2021).

RESULTS AND DISCUSSION

The world of business has come a long way in terms of unstructured text processing. It is important to note that the employers know very well how diverse the sources for unstructured textual data extraction are. This made them 'hunt', explore and then implement partially new architectures for their sources of data that provide all the information needed for better data mining. For example, part of companies offering services through internet managed to switch their booking process to online booking forms, instead of phone call booking requests. This lead also to the reorganization of many units and processes in companies such as: reorganizing office spaces; changes in cost-per-employee index; directing more attention to UX design; searching for better IT solutions to increase the swiftness of the new booking process and many more. All in the name of getting more accurate information from their customers. Others, product-offering companies, are trying to differentiate special places in their websites where customers can leave reviews. Even more, they are encouraging the using of hashtags in some social media platforms so the process of extracting information by their comment become easier. The main idea standing behind these new approaches may be different for each

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

business (to improve an existing product/service; for better sales planning; to easily develop a new product/service; to expand to another market/s, etc.) but they all end on same result - the level of accuracy of customer knowledge management (CKM) increases. CKM, "which refers to the management of knowledge from customers, i.e. knowledge resident in customers" (Gibbert, Leibold & Probst, 2002) is thoroughly examined by many authors but only a few of them studied the extraction of knowledge from customers through data-driven approaches. These particular studies have found that the effectivity of the CKM raises through the usage of data mining techniques (Zhan, Tan & Huo, 2019).

Despite the above mentioned different ways in which companies choose to organize their work practices the text information received is mainly submitted in the form of free text in a non-unified way. This unstructured raw textual data is yet extensively used even in emergency department records of hospitals for example. And NLP and machine learning techniques again are the main contributors in the process of analyzing these narratives. This way the text mining process becomes subject to patterns recognition in specialists' reports by identification of specific configurations found in patients' information. Besides emergency departments, in most businesses - the better knowledge and understanding of text mining methods companies have, the larger the potential for improvements of overall performance may take place; thus - better implementation and usage of these methods can also positively impact business results of organizations. Substantially, text mining techniques can yield significant information about important insights that companies from different industries can use to adjust their marketing strategies. Different techniques and their application are presented below:

Text summarization, as one of main text mining techniques, could solve a huge amount of problems in the companies of almost every sector. It saves time in first place and this is of major importance as time is the most valuable resource for every business, but also for its customers. A good sector-specific example where text summarization usage is recommendable is travel reviews and travel shopping websites. They have a huge number of reviews and it is a true nightmare for the user to go through all the reviews to pick the right future travel property. Text summarization is a useful technique in this particular situation – it can summarize the listed travel properties and this way instead of wasting user's time it will give him a comfortable and quick option to pick a property through summarized details like district, property type, facilities, review score, cleanness review, etc.

On the other hand, text summarization could decrease the error rate in decision-making based on the omission of important information and even more – this way higher quality of work will be established and the employees' performance can be more successful. It is discovered that error rates in companies with common working practices vary from 10 to 30 errors per hundred opportunities. Alternatively, best performing companies using data-driven management methods have 5 to 10 errors per hundred opportunities (GoodSign, 2015). And again, it takes time for the management to process information compiled in the data analysis process and to deliver insights of it to the respective subordinates in the organizational structure.

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

As prime text mining technique, clustering is applicable for a variety of business practices in conjunction with its main role in Functional data analysis (FDA). FDA is part of statistical analyzation of data represented as curves, images and all types of graphical representations that trace a process or indicator over time. For companies operating in the field of sports trading, financial trading, all types of companies offering business predictive models to their corporate clients (from sales and new market entry to gauging buyers' propensity to purchase and internationalization opportunities), etc., clustering is an essential analytical tool.

All mentioned in chapter "Conceptualization of text mining techniques" types of text classification techniques (sentiment analysis, topic analysis, spam detection, intention detection, etc.) are unquestionably a dominant tool for companies from all sectors to get the sentiment of a text generated through web site reviews, social media comments, customer surveys, etc. The result received, mainly as simple as it is – positive, neutral or negative, can provide a better vision of public opinion within marketing campaigns, services and products. Sentiment analysis for example, could give a more detailed view and form sub-sections of the main analysis through techniques such as finding the sentiment of a separated sentences or providing a more complex result (sentiment numeric score) which can indicate positive and negative parts of the text. The last example is also known as a good practice for businesses to incorporate in their tactical operations that can allow for maximum utilization of the benefits that sentiment analysis offers – a combination with aspect mining. This approach identifies different parts of the text and as result - the analysts can get the sentiment for the pricing, for the customer service, etc.

The examples described above form a small part of the benefits that put in superior position among competitors the businesses chosen the data-driven approach powered by AI for its decision-making. And, even more, when a combination of text mining techniques is adequately made, they result in integral data analysis which effect on the predefined areas for improvement is impressive. A good example is healthcare industry and more specifically – healthcare professionals relying highly upon feature extraction, clustering and classification techniques applied to clinical datasets for designing pattern-specific treatments and decreasing the number of harmful reactions by medicines (Aarushi and Arunava, 2021). These two researchers used the studies of Viveka and Kalaavathi, GraciaJacob and Ramani to be even more specific: the clusters can lead us to a specific type of drugs which may provoke a harmful reaction (Viveka & Kalaavathi, 2016); the process of classification helps in the assigning of parts of patient's data to predetermined classes; the pre-existing data sets may be cleaned up from predictors of pathologies for better and more swiftly data exploration and analysis (GraciaJacob & Ramani, 2012). This way the life of their patients is improving but also there is a serious reduction in hospitalization costs (Jain & Ghosh, 2021).

Another successful combination of text mining techniques was discovered by Sharma and Panigrahi related to banking sector. Their review of literature about data mining techniques able to be used for a practical detection of financial accounting frauds shows that Bayesian belief network and decision trees were among the most widely used ones for solving the defined problem (Sharma & Panigrahi, 2012). As decision tree and Naïve Bayes are already mentioned conceptualization chapter above (instead of Bayesian belief network) it is

objective to state that nowadays Naïve Bayes is better known techniques and this is for a reason. As other authors concluded Naïve Bayes is a limited form of Bayesian belief network but once used as a classifier it shows a better performance in this specific task (Friedman, Geiger & Goldszmidt, 1997).

The right combining of text mining techniques could have a great quantity of business advantages and profit gain. From better office locations analysis for telecommunication companies and better warranty analysis for product manufacturers to more precise property-prices trend analysis for real estate agencies and present-day-talked-about-topic analysis for television producers – there is always a room for improvement but with the power of text analytics it may be possible for businesses to succeed in doing it.

Along with the core techniques described in details in this chapter it is recommendable for companies to stay informed about all new hybrid tools that are specifically designed to incorporate all functions contained in the text processing – extraction, analyzing, etc. An example for such tool is "VisualUrText" which is a tool that aims to combine the process of extraction, processing and analyzing the unstructured business related text data and to visualize at the end a cleaned text. There is a list of possible form which are capturing the final result such as Document Term Matrix (DTM), Frequency Graph, Network Analysis Graph, Word Cloud and Dendrogram (Zainol, Jaymes & Nohuddin, 2018).

CONCLUSION

One of the main challenges standing in front of every business nowadays is to refine the ongoing process of complex data analysis where the main part of the data is in unstructured format. This improvement on company level is accompanied by choosing modern, automatic text mining method to do the main work. Thus, a proficiency in knowing best methods combinations, best modern practices and the most suitable text analytics tools can give every company the power to dominate in its field of expertise.

Based on the literature reviewed for this paper another direction for further discussion arises. We can notice that almost every text mining algorithm is working with NLP in English or with only one language. Although it is considered that the language of science is English and it is also the most used language for business relations all over the world, it will be best to direct the future development power to more multilingual techniques for text analytics. This way a larger part of businesses could benefit from the text mining methods.

Author Contributions: Conceptualization, D.Z.; methodology, N.N.; formal analysis, D.Z.; investigation, D.Z.; project administration, N.N.; data curation, D.Z.; resources, D.Z..; supervision, N.N.; validation, D.Z. and N.N.; writing—original draft preparation, D.Z. and N.N.; writing—review and editing, N.N.
All authors have read and agreed to the published version of the manuscript.

Data Availability Statement:

The data presented in this study are available on request from the corresponding author.

Conflict of interests

The authors declare no conflict of interest.

References

Augenstein, I., Padó, S., Rudolph, S. (2012). LODifier: Generating Linked Data from Unstructured Text. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds) The Semantic Web: Research and Applications. ESWC 2012. Lecture Notes in Computer Science, vol 7295. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30284-8_21

Bach, N. and Badaskar, S. (2007) 'A review of relation extraction', *Literature review for Language and Statistics II*, 2, pp. 1–15.

Bhartiya, D., & Singh, A. (2014). A Semantic Approach to Summarization. ArXiv, abs/1406.1203.

Bhide, M. (2016). Single or Multi-document Summarization Techniques. *International Journal of Computer Science Trends and Technology (IJCST)*, 4(3), pp.375-379. Available at: http://www.ijcstjournal.org/volume-4/issue-3/IJCST-V4I3P63.pdf.

Birjali, M., Kasri, M. and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, vol.226, doi: https://doi.org/10.1016/j.knosys.2021.107134.

Carbonell, J. and Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336. https://doi.org/10.1145/290941.291025

Chen, Y. and Tu, L. (2007). Density-based clustering for real-time stream data. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142. https://doi.org/10.1145/1281192.1281210

Conroy, J. M. and O'leary, D. P. (2001) Text summarization via hidden Markov models. SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 406–407. https://doi.org/10.1145/383952.384042

Das, D. and Martins, A. (2007) A survey on automatic text summarization.

Edmundson, H. P. (1969) New Methods in Automatic Extracting. *Journal of the ACM*. vol. 16, Issue 2, pp. 264-285. https://doi.org/10.1145/321510.321519

Evans, D. K., McKeown, K., Klanvas, J. L. (2005) Similarity-based multilingual multi-document summarization.

Firoozeh, N., Nazarenko, A., Alizon, F., Daille, B. (2020) Keyword extraction: Issues and methods. *Natural Language Engineering*. 26(3), pp. 259–291. doi:10.1017/S1351324919000457.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian Network Classifiers. *Machine Learning*, 29(2), pp. 131–163. doi: 10.1023/A:1007465528199.

Gibbert, M., Leibold, M. and Probst, G. (2002) Five Styles of Customer Knowledge Management, and How Smart Companies Use Them To Create Value. *European Management Journal*, 20(5), pp. 459–469. https://doi.org/10.1016/S0263-2373(02)00101-9.

Gracia Jacob, S. and Ramani, G. (2012) Data Mining in Clinical Data Sets: A Review. *International Journal of Applied Information Systems*, vol.4, Issue 6, pp. 15–26. doi: 10.5120/ijais12-450774.

Hänig C., Schierle M., T. D. (2010) Comparison of structured vs. unstructured data for industrial quality analysis. Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I, pp.432-438. WCECS 2010, October 20-22, San Francisco, USA. ISSN: 2078-0966 (Online)

Impelsys (2021) *An overview of Text Summarization in Natural Language Processing*. Available at 12.02.2022: https://www.impelsys.com/an-overview-of-text-summarization-in-natural-language-processing/.

Inmon, W. H., Linstend, D. and Levins, M. (2019) *Data Architecture. A primer for the Data Scientist*. Academic Press. eISBN: 9780128169179, pISBN: 9780128169162

Jain, A., & Ghosh, A. (2021). Novel Insights into Data Mining to Improve the Specificity of Pharmacovigilance and Prevent Adverse Drug Reactions in Psychiatric Patients. *Asia Pacific Journal of Health Management*, *16*(3), 130-136. https://doi.org/10.24083/apjhm.v16i3.985

Jiang, J. (2012). Information Extraction from Text. In: Aggarwal, C., Zhai, C. (eds) Mining Text Data. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_2

Kriegel, H. *et al.* (2011) Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), pp. 231–240.

Kumagai, M., Komatsu, K., Takano, F., Araki, T., Sato, M. and H. Kobayashi. (2020). Combinatorial Clustering Based on an Externally-Defined One-Hot Constraint. *2020 Eighth International Symposium on Computing and Networking (CANDAR)*, pp. 59-68, doi: 10.1109/CANDAR51075.2020.00015.

Larsen, K., Monarchi, D., Hovorka, D., Bailey, C. (2008) Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems. *Decision Support Systems*, 45(4), pp. 884–896. https://doi.org/10.1016/j.dss.2008.02.009.

Lin, C.-Y. (1999) Training a selection function for extraction. in. CIKM '99: Proceedings of the eighth international conference on Information and knowledge management, pp. 55–62. https://doi.org/10.1145/319950.319957.

Mani, I. and Bloedorn, E. (1997) Multi-document summarization by graph search and matching. *arXiv preprint cmp-lg/9712004*.

McKeown, K. and Radev, D. R. (1995) 'Generating summaries of multiple news articles. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 74–82. https://doi.org/10.1145/215206.215334

Mikheev, A., Moens, M. and Grover, C. (1999) Named Entity Recognition without Gazetteers. in EACL 1999, *9th Conference of the European Chapter of the Association for Computational Linguistics*, June 8-12, 1999, University of Bergen, Bergen, Norway. pp. 1-8. https://doi.org/10.3115/977035.977037

Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, *53*(9), 1–18. https://doi.org/10.18637/jss.v053.i09

Osborne, M. (2002) Using maximum entropy for sentence extraction. *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pp. 1–8. Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics. doi: 10.3115/1118162.1118163

Patel, K. M. A. and Thakral, P. (2016) The best clustering algorithms in data mining. *2016 International Conference on Communication and Signal Processing (ICCSP)*, pp. 2042–2046. doi: 10.1109/ICCSP.2016.7754534.

Radev, D. R. , Jing H., Sty M. *(2004). Centroid-based summarization of multiple documents. Information Processing and Management, vol. 40, no. 6, pp. 919–938.*

Rose, S., Engel, D., Cramer, N., Cowley, W. (2010) 'Automatic keyword extraction from individual documents', *Text mining: applications and theory*, 1, pp. 1–20. Editor(s):Michael W. Berry, Jacob Kogan. https://doi.org/10.1002/9780470689646.ch1

Roux, M. (2018) A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, Springer Verlag, 2018, 35 (2), pp.345-366. doi:10.1007/s00357-0189259-9. hal-02085844

Sasirekha, K. and Baby, P. (2013) Agglomerative hierarchical clustering algorithm- A Review. *International Journal of Scientific and Research Publications*, (IJSRP), Volume 3, Issue 3, March 2013 Edition.

Sharma, A. and Panigrahi, P. (2012) A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications*, 39(1), pp. 37–47.

Stephens, K. R. (2002) What has the Loebner Contest told us about conversant systems. p. 2005.

Svore, K., Vanderwende, L. and Burges, C. (2007) Enhancing single-document summarization by combining RankNet and third-party sources. *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 448–457. Prague, Czech Republic. Association for Computational Linguistics.

Tamine, L. and Goeuriot, L. (2021) Semantic Information Retrieval on Medical Texts: Research Challenges, Survey, and Open Issues. *ACM Computing Surveys (CSUR)*, 54(7), pp. 1–38. https://doi.org/10.1145/3462476

Tsai, C.-F., Wu, H.-C. and Tsai, C.-W. (2002) A new data clustering approach for data mining in large databases. *Proceedings International Symposium on Parallel Architectures, Algorithms and Networks. I-SPAN'02*, pp. 315–320. doi: 10.1109/ISPAN.2002.1004300.

Turney, P. D. (2000) Learning algorithms for keyphrase extraction, Information retrieval, 2(4), pp. 303–336. https://doi.org/10.1023/A:1009976227802

Viveka, S. and Kalaavathi, B. (2016) Review on clinical data mining with psychiatric adverse drug reaction. *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, pp. 1–3. doi: 10.1109/STARTUP.2016.7583945.

Zainol, Z., Jaymes, M. T. H. and Nohuddin, P. N. E. (2018) Visualurtext: a text analytics tool for unstructured textual data. *Journal of Physics: Conference Series*. IOP Publishing, p. 12011.

Zhai, C. and Massung, S. (2016) *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool.

Zhan, Y., Tan, K. H. and Huo, B. (2019) Bridging customer knowledge to innovative product development: a data mining approach. *International Journal of Production Research*, 57(20), pp. 6335–6350. doi: 10.1080/00207543.2019.1566662.

About the authors

Denitsa ZHECHEVA

PhD student in computer science with specialization in artificial intelligence (AI) in Konstantin Preslavsky University of Shumen and with track record in Project Management.
Professional interests include but not limited to: artificial intelligence, business intelligence.

ORCID ID: https://orcid.org/0000-0002-4996-1369

Nayden NENKOV

PhD, Professor of the Computer Science in College in Dobrich, Konstantin Preslavsky University of Shumen, Bulgaria.
Professional interests include but not limited to: artificial intelligence, business intelligence, expert system, machine learning, neural networks, robotics, algorithms.

ORCID ID: https://orcid.org/0000-0002-1895-2662