CHAPTER 7

DATA PRE-PROCESSING IN TEXT MINING

Tuğçe AKSOY*, Serra ÇELİK, Sevinç GÜLSEÇEN*****

*Istanbul University, Informatics Department, Istanbul, Turkey
E-mail: aksoy0tugce@gmail.com

**Dr., Istanbul University, Informatics Department, Istanbul, Turkey
E-mail: serra.celik@istanbul.edu.tr

***Prof. Dr., Istanbul University, Informatics Department, Istanbul, Turkey
E-mail: gulsecen@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.07

Abstract

The fact that any kind of user has the ability to generate data with great ease at any time causes an increase in the importance of data mining. Considering the reality that the vast majority of the available data is composed of unstructured data and that the data in the text type is outnumbering, it proves the increasing interest in text mining and the abundance of studies in this field. However, in order to be able to examine an unstructured data type like text, which is quite different from machine language, it is necessary to make this data more structured and make the machine work. At this point, the data pre-processing step, which covers a large part of the entire text mining process, is of great importance. In this chapter, it is aimed to explain the text pre-processing phase on a basic level by supporting this using visuals. In doing so, it is primarily planned to focus on text mining and to explain in detail the characteristics of the data processed. In this context, it is aimed to explain the data pre-processing steps followed in order to overcome these difficulties by examining the difficulties created by the data in question. As a result, this chapter is a descriptive review of the data pre-processing phase in text mining, which covers some of the studies previously conducted on this subject.

Keywords: Text mining, Pre-processing, Linguistics

1. Introduction

This chapter is a compilation study on the basic level of data pre-processing steps in text mining. After a broad description of the concept of data mining, it is aimed to elaborate on text mining, to examine the pre-processing phase by explaining the characteristic features of the text data type and the difficulties it brings.

The concept, which emerged as Knowledge Discovery from a Database (KDD) and then evolved into data mining, involves the processing of existing raw data into usable knowledge. With the rapid development of technology, the production of huge amounts of data in seconds and the emergence of the concept of big data creates the need to accelerate the developments in the field of data mining. In the data mining process, which includes steps such as data selection, pre-processing, transformation, data mining and interpretation, the pre-processing step is the most important and the most time-consuming step. The fact that the amount of data produced every second increases greatly demonstrates the importance of the pre-processing step. The International Data Corporation (IDC) estimates that about 90% of all data is unstructured (Gantz & Reinsel, 2011). This again reveals the concept of pre-processing. Since the data pre-processing step is so important, it is aimed to introduce the subject at the beginning level and shed some general light on the subject. This study is a compilation study that combines the studies carried out in this field from past to present by making a literature review and explaining the pre-processing steps in detail with schemes.

The structure of the chapter is as follows: In the second section, the concept of data mining will be explained in detail and the basic concepts of data mining, which are classification, clustering and association will be explained and their connection with text mining will be elaborated. The importance of text mining, characteristics of data processed and why it is difficult to process will be explained. While doing this, questions such as what is text, a word and a paragraph are answered. In the third section, the text pre-processing step will be defined and each method used in this step will be examined individually and in detail. While doing this, basic examples will be presented with a better understanding of these methods with visuals. In the fourth section, the conclusion paragraph will be presented and a final evaluation will be made and the importance of the pre-processing process will be re-emphasized.

2. The Definitions of Data Mining and Text Mining

Structured data is data in a tabular form which has some relational rows (records) and columns (variables). The type of each variable is pre-defined - numerical, categorical, logical, date, image, and so on. It is the easiest data to work on. It is defined as a type of data in which

both structured and unstructured data coexist. Unstructured data is information that either does not have a predefined data model or is not organized in a pre-defined manner. It includes data types such as text, sound, image and video (Eberandu, 2016). It is the most difficult data to work with and needs to be transformed into structured form before it is processed. By its very nature, the concept of data mining can be defined in many different ways. The reason for this is that the process may vary according to the type of data and the needs of the person in many different fields and sectors. However, according to Han, Kamber, & Pei (2011), data mining means the process of extracting interesting patterns from large amounts of data and discovering information. This is because the main idea in data mining would not be efficient if huge amounts of raw data available worldwide are not processed. What is meant here is that the data becomes knowledge through the necessary steps and that this knowledge is exchanged by people and made available for a specific purpose. The data may be numerical, image, audio, video and text.

There are basically five steps in data mining (Fayyad, Piatetsky-Shapiro & Smith, 1996). These are the selection of data, pre-processing of data, transformation of data, data mining and interpretation. In the selection of the data, the data to be processed is determined and the information to be obtained from this data is predicted. In the data pre-processing phase, it is aimed to make the selected data suitable for performing the necessary mathematical operations and creating a model. In this step, deficiencies in the existing raw data are tried to be eliminated, the noise is removed and the data is made much more structured. At the end of this process, the attributes that are contained in the data are determined and by making use of these attributes, models are tried to be created by methods such as classification, clustering, association and regression performed at next step, which is data mining. Finally, there is the stage of interpreting the performance rates by calculating how well the models represent the data.

A more detailed review of the data mining methods before the data pre-processing step will help to understand how important the data pre-processing phase is for the next steps.

2.0.1. Classification

In general, the classification method is the process of categorizing the elements in the data. A dataset has certain variables e.g. inputs (predictors) and an output (class). Outputs are categorical or numerical results according to classicifaction algorithms such as Classification and Regression Trees (CART), and Random Forest. As a preliminary preparation of the classification method, the list of categories is determined so that the classification system and the elements of the data are split as sample data for each category (Jo, 2018).

In order to apply the classification algorithms, the data must first be divided into training and test sets (Mitchell, 1997). The training set is used as sample data to create the classification capacity by using machine learning algorithms, and in the test set, the data elements are classified and the differences between the real and classified labels are observed (Jo, 2018).

Duda, Hart & Stark (2000) divide the classification into soft and hard classification in their studies. According to this distinction, the conditional probability of the class is calculated in the soft classification method and the class estimate is made according to the greatest probability. On the other hand, in the hard classification method, the classification boundaries are determined directly without estimating the probability of the class. In addition, Duda et al. (2000) have classified the classification method horizontally and hierarchically. In the horizontal classification method (Figure 2), the categories are predetermined as a single list, whereas in the hierarchical classification method (Figure 1), clustered categories exist in a number of categories and the categories have the characteristics of a tree model.

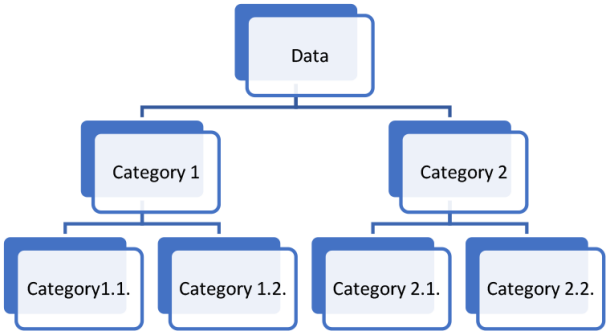


Figure 1: Hierarchical Classification Method

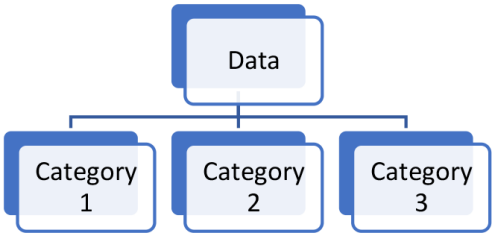


Figure 2: Horizontal Classification Method

2.0.2. Clustering

Clustering is the process of splitting a group of various data elements into subgroups, that is, clusters with similar properties. At this stage, first, unlabeled data elements are provided

and the similarity measures between them are calculated. The elements are subcategorized according to the similarities between them. The most important purpose of clustering is to maximize the similarity of elements in each cluster and minimize the similarity between clusters (Jo, 2006).

Duda et al. (2000) divide clustering into hard and soft clustering. In hard clustering, all elements can be collected in one cluster, whereas in soft clustering, each element can be clustered in more than one cluster (Figure 3).

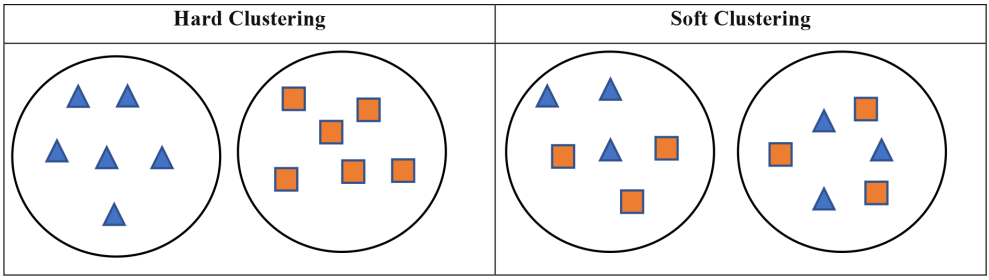


Figure 3: Differences Between Hard and Soft Clustering

Clustering is also divided into horizontal clustering and hierarchical clustering (Duda et al., 2000). In horizontal clustering, clusters are created as a single list while in the hierarchical clustering method, clusters are created in the tree model, and clusters can also have subsets.

The clustering method can automate the predetermination of categories, which is a prerequisite for the classification method. Horizontal or hierarchical clusters constitute the category list to be used in the classification method (Jo, 2006).

2.0.3. Association

Association is considered to be the extraction of the association between data elements in the if-then form (Jo, 2018). In other words, the association rule is a method aimed at revealing the related variables and determining the magnitude of the connection between them. It often involves identifying repetitive patterns and making predictions through them. It also provides great benefits in predictions in the fields such as purchasing, marketing and campaigning (Jo, 2018).

Confidence and support measures are frequently used when creating association rules. Support determines the rate at which a relationship is repeated in the entire data set, while confidence reveals the possibility of coexistence with two variables. If two variables are independent of each other, there is no association between them (Jo, 2018).

2.0.4. Regression

Regression is the process of estimating the output values of each data element. In other words, it is the process of calculating continuous data by examining the input data. Unlike the classification method which provides discrete values as outputs, regression gives continuous values as outputs (Jo, 2018). Regression gives numerical results based on a linear model such as shown in Figure 5. Regression works as a classification method if the dependent (output) attribute is categorical (as in the Logistic Regression model).

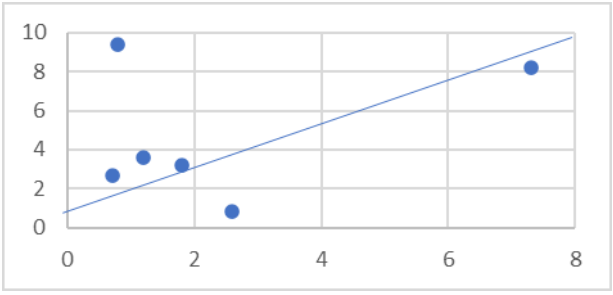


Figure 4: Regression Analysis

2.1. Text Mining

In general, text mining is the process of extracting quality information from the text by making use of numerical methods and techniques. In different studies, it is also defined as the process of extracting implicit information from text data (Feldman & Sager, 2007). It is a more specific sub-branch of data mining. In text mining, it is aimed to obtain information that is not yet known from very large unstructured text data. According to Kalra and Aggarval (2018), information retrieval is actually text mining. Another view takes a more technical approach to text mining and defines text mining as a set of statistical and computer science techniques developed specifically for analyzing text data (Zanini & Dhawan, 2015). Examined in detail, it is understood that text mining is not actually a new concept and it is an extension of data mining. Therefore, many algorithms used in data mining can also be used in text data, and therefore in text mining. The only difference is that while data mining deals with structured, quantitative data, text mining deals with unstructured or semi-structured data. In fact, the goal is to extract meaningful numerical indexes from the text that the computer can understand. While doing this, statistical methods are used extensively. With text mining, the information contained in the text can be categorized and clustered to obtain results such as word frequency distribution, and distributions. Then association and predictive analyses can be applied to words. For this reason, calculating classification, clustering and

association rules of the text is the basic function of text mining (Jo, 2018). This is the reason for the detailed descriptions and definitions made in the previous section. If the concepts of classification, clustering and association are understood, the importance of data pre-processing phase in text mining will be more easily realized.

The process in text mining is no different from data mining. Only the types of data being processed differ.



Figure 5: The Text Mining Process (Gaikwad, Chaugule & Patil, 2014)

Looking at the history of text mining, the National Center for Text Mining (NaCTeM) is the first public institution established worldwide (Zanini & Dhawan, 2015). It was established by the UK's Joint Information Systems Commission (JISC). The first activities of text mining were observed in the 1980s, and at first text mining was only dealing with data in databases and data warehouses. Nowadays, with the developing technology, there has been an intense interest in this field - unstructured data has reached a very high rate of 90%, and there are many types of text data such as text messages, e-mails, social media activities, blog contents and internet searches that are of interest to text mining.

Text mining is used in classifying, organizing and summarizing documents, estimating and developing contextual suggestion systems. Recently, text mining has been increasingly used in epidemiology, economics and education, and applied research to gain insights into the market and consumer, particularly in relation to businesses (Zanini & Dhawan, 2015). The information obtained from this research is very useful in decision-making processes.

The types of text mining can be divided into seven groups according to their function:

1) Document Classification
2) Document Clustering
3) Information Retrieval
4) Web Mining
5) Information Extraction
6) Natural Language Processing
7) Term Extraction

2.1.1. Difficulties

The actual source of text data is, in fact, language, and it is not possible, at least for the time being, to establish a universal model in the text mining process, as each language has its own characteristics. It is very difficult to create a generally valid model that can be adapted to the different uses of a single language, let alone between languages. The fact that each language has different syntax, and the ability of a single language to produce a wide range of syntaxes within and especially in daily use, makes this process even more challenging. In addition to syntax, there is great ambiguity in language. A word can have different meanings depending on the context of the sentence in which it is used or different words can be used for the same meaning. Adopting this ambiguity and diversity of language into a mathematical model needs a complex structure and intensive language knowledge. It is even more challenging to process data because of the fact that the majority of the data that text mining deals with includes text messages, social media content, and e-mails which are also generated without much attention to grammatical rules such as spelling and punctuation. In order to carry out studies such as classification and clustering on text data, it is necessary to determine the attributes of the data elements, and even if the data is a single paragraph text, almost every word will come as a separate attribute, and processing these attributes causes a great deal of time and space consumption. The ambiguity of the language can be examined in four categories (Sheeba & Vivekanandan, 2012):

i) Homophony

Those are words that have the same spelling but have different meanings depending on the context of the sentence.

Example:

I left my phone.

My phone is on the left side of the table.

ii) Synonymy

Those words have different spelling but have the same meaning.

Example:

The small child was sleeping.

This kid is so smart.

iii) Polysemy

A word has different but interrelated meanings depending on the context of the sentence.

Example:

The woman's face was beautiful.

You have to face with the consequences.

He sat facing the door.

iv) Hyponymy

It means semantic inclusion between lexical units.

Example:

Dog - Animal

In this example, the dog has a hyponymy relationship with the animal species.

All these ambiguities make it difficult to conduct semantic analyses in the text mining process, to correctly label the words according to the word type and many other processes.

2.1.2. What is text?

Text is an unstructured data type consisting of arrays called words (Salton, 1998). According to Jo (2018), the text is a collection of sentences or paragraphs written in natural language. In the first definition, the concept is discussed in a technical context, while in the second definition, there is a more linguistic perspective. If both definitions are considered, the text consists of elements of very different dimensions. Words are the basic units of the text and the words come together in accordance with grammatical rules to form sentences and sentences form logical paragraphs (Jo, 2018). The emphasis here is important. Although it is stated in the previous sections of the chapter that the language can vary in syntactic terms, this diversity takes place within the framework of certain rules. Therefore, the words that make up sentences have to comply with certain rules. It is somewhat easier to produce models within the framework of these rules. However, there are no grammatical rule limitations when it comes to paragraphs. The important thing is that the combined sentences follow each other with a certain logic and sub-context (Jones and Manu, 1999). This requires significant semantic analysis and semantics is one of the most challenging concepts in the field of natural language processing.

The text includes variables such as the size, author and title of the text as well as the paragraphs that form it , and is considered short text if the text consists of a single paragraph,

medium-length text if it consists of a single group of paragraphs, and long text if it consists of multiple groups of paragraphs (Jo, 2018).

2.1.3. What is a sentence?

A sentence is a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses (https://www.lexico.com/en/definition/sentence). A sentence can only consist of a single complement, that is, it can only have the subject and predicate, or it can consist of more than one complement. Sentences begin with uppercase letters in most languages and end with punctuation, such as dots, question marks, exclamation points, and triple dots. All of these elements are objects of the pre-processing step and will be explained in more detail in the next section. In addition, the words in the sentence are separated by spaces in many languages, and in fact, this feature of languages is similar to the tokenization process (Manning, Raghavan, & Schutze, 2009).

2.1.4. What is a word?

The word is considered as the basic unit of a text. Since the text data processed in text mining will form a very long string as a whole, in the pre-processing phase, the text is segmented into a word list and the other processes are performed through this list. The reason for choosing the word as the smallest unit is that the smallest unit that can be evaluated as meaningful and an attribute is also the word and, as mentioned earlier, the main purpose of text mining is to extract meaningful information from unstructured data.

In text mining, there is a term called stop-words, which means unnecessary words. Stop-words are the words that are necessary in grammatical terms and that are very common in the text but do not make any sense alone. Due to the fact that they are frequently included in the text due to grammatical rules, they cause wrong measurement results in making necessary statistical measurements and therefore stop-words are extracted from the text in the pre-processing phase. This will be explained in more detail in the next section.

3. A Detailed Description of Pre-processing Steps

3.1. What is data pre-processing?

The process of data pre-processing in text mining consists of the steps of converting the original text data into a raw data structure that distinguishes important textual features between text categories (Srividhya & Anitha, 2010). There are many methods of data pre-processing in text mining and they all try to make documents or document collections

structured in some way. As many different methods emerged during the effort to make the text structured, these methods evolved over time. Data pre-processing is an important part of the natural language processing system as well as text mining because at this stage, characters, words and sentences are determined and the characteristics of these elements are specified and transferred to the next stage, then they are used in information retrieval or machine translation systems. The raw data, which is formed after the data that is planned to be processed is collected, goes through the visualization stage in order to understand its structure. Correlation matrices can be used to perform this process, especially in text mining, and thus the similarity between the attributes that make up the text can be examined and more reliable results can be obtained in the following steps. After the data is better understood through visualization, the actual data pre-processing step is initiated and data cleaning is performed. In the context of text mining, clearing of data involves steps such as removing stop-words, punctuations and special characters (Kalra & Aggarwal, 2018) and will be explained in more detail in this section. Performing the data cleaning process earlier will reduce the size of the data to be dealt with in the next steps, thus achieving more optimal results in terms of time and space. This is because not only are stop-words eliminated in the data pre-processing phase, but also words in multiple forms are reduced to a single form (Kadhim, 2018). Obviously, due to the detailed processing of the pre-processing step, the phase covers about 50% to 80% of the entire text mining process. Then, in order to carry out other text mining steps following the data pre-processing step, the attributes created as a tabular form are labeled and the task of each item in the text is determined.

For the reasons already mentioned, an efficient pre-processing step should effectively represent the text in terms of both space (storage) and time (information retrieval) requirements**,** as well as good retrieval performance (precision and recall) (Giagole, Patil & Chaudhari, 2013). It is understood that the purpose of the data pre-processing phase is to present the text as an attribute vector by separating each text into individual words and to establish a relationship between the obtained attributes and the text.

Finally, to summarize why the pre-processing phase is important in text mining: First, it reduces the file size of the text data because stop-words correspond to approximately 20% to 30% of the total number of words in the text, and stemming reduces the index size by almost 40% to 50% (Gurusamy & Kannan, 2014). Secondly, it is important to make information retrieval systems work more effectively. Because stop-words are useless in searches and text mining, they can cause confusion in information retrieval systems, and stemming is used to match similar words in the text file (Gurusamy & Kannan, 2014).

Feldman & Sanger (2006) divide the pre-processing phase into two types of methods: task-oriented pre-processing methods and other methods. In the task-oriented pre-processing method, there is usually a problem that needs to be solved, such as extracting titles or authors from a PDF file, and structured text descriptions are made by creating tasks and sub-tasks through this problem (Feldman & Sanger, 2006). Other pre-processing methods consist of formal methods that are created to analyze complex structures and that can also be used on texts produced in natural language, and they include classification schemes, probabilistic models and rule-based system approaches (Feldman & Sanger, 2006). Although the two methods differ from each other, the aim of both is to store the most meaningful information as an attribute and exclude unnecessary elements, and both methods deal with unstructured or semi-structured data.

Feldman & Sanger (2006) grouped the data pre-processing step according to their functions in a different way and divided the process into three subprocesses as shown in Figure 6:

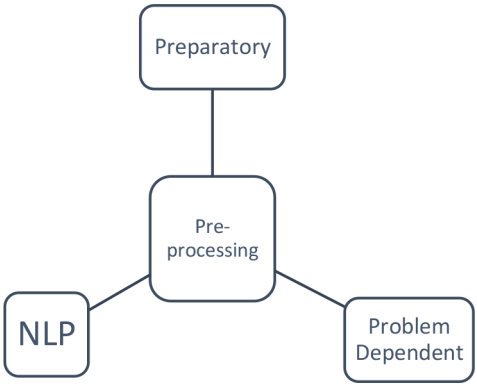


Figure 6: Feldman and Sanger (2006) Data Pre-processing Stages

At the preparatory stage, the raw data is first transformed into the appropriate structured form for the subsequent linguistic processing, and in doing so, divides the text into paragraphs, columns or tables. Steps such as determining words, POS tagging, syntactic parsing and morphological analysis are involved in the natural language processing. The output of this process is generally meaningless to the end user and is used as input to the next stage, the problem dependent stage. In the problem dependent stage**,** a final semantic representation of the data is created. In text mining, this process is usually completed with classification and information retrieval methods.

If we take a closer look at the linguistic steps at the stage of natural language processing, it is observed that linguistic methods are given different priority orders and even some of the studies include the methods that other studies do not mention. For example, while Brants (2003) argues about the natural language processing taking place during the information retrieval process, he lists methods such as stemming, POS tagging, compound recognition, de-compounding, chunking, and word sense disambiguation. However, Lourdusamy and Abraham (2018) list the pre-processing steps as tokenization, stop-word filtering, POS tagging, stemming, document indexing, grammatical parsing, text summarization, TF-IDF and chunking in their study. Kalra & Aggarwal (2018), on the other hand, believe that creating the vector space model of the attributes obtained in the list is another pre-processing step. In their work, Srividhya & Anitha (2010) include document indexing, and listing the attributes of the texts and measuring the term weighting of them in the pre-processing phase. TF-IDF is not described as a separate pre-processing step for them, but as a method that is frequently used in calculating term weights.

Although data pre-processing steps differ in the studies, it should be remembered that it is necessary to choose these steps according to the content of the text to be processed and the priority order of these steps should be determined according to the text. In line with this context, the following table summarizes the data pre-processing steps as a general review of all studies:



Figure 7: Text Pre-processing Steps

3.2. Text pre-processing steps

3.2.1. Tokenization

The first step in the text pre-processing phase is tokenization. Since we work on words - which are considered to be the smallest meaningful units in the text - they must be tokenized first. In general, the text is separated into words by spaces or punctuation marks and is organized in a list at this step. According to Jo (2018), the tokenization step is a prerequisite for stop-word removal and stemming. As the sub-stages of the tokenization step, capital letters can also be converted to small letters and special characters, symbols and numbers can be removed. However, converting the first letter of the words to lower case may affect the

chance of getting healthy results in the text mining process where there is a purpose such as selecting proper names. Such steps should be carried out taking into account similar situations. In one study, the removal of punctuation marks was considered as one of the basic steps in tokenization (Kalra & Aggarwal, 2018). Feldman & Sanger (2006) argue that this step is not limited only to identifying words, but in general, parsing the main text into paragraphs, sentences, phrases, words, and even morphemes that are accepted as the smallest meaningful units in linguistics, but the most commonly preferred parsing is on a sentence and word level. However, the main problem in parsing sentences is the determination of the beginning and end of sentences accurately. This is because the full stop, often used at the end of the sentence, can be used not only for this function, but also for abbreviations of titles in languages such as English. The full stop at the end of ordinal numbers in Turkish also causes great problems and inaccurate tokenization. Srividhya & Anitha (2010) divided the tokenization step into three phases: The first is to convert the text into a word count, i.e. to create a bag of words. Then, the necessary cleaning and filtering operations should be performed, that is, spaces, special characters and symbols should be removed and finally the text is converted into an attribute list, that is, separated into words, terms or properties. It is understood from these stages that in this study, the cleaning and filtering stage is accepted as a basic step in tokenization.

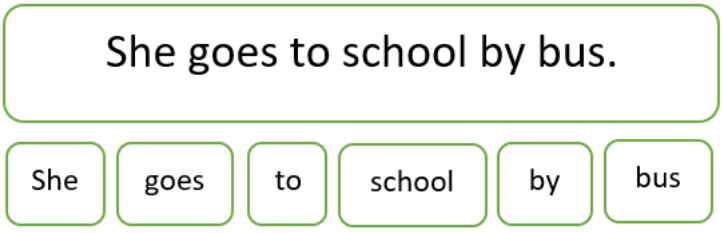


Figure 8: Tokenization

3.2.2. Stemming

In the stemming phase, the attributes in the list created in the previous step are separated from their prefixes and suffixes and converted to a stem, which is the nominative version of words. Although there are some differences between the meanings of the derivative and nominative words, they have relative meanings, and therefore, in order to avoid having to deal with a lot of data and improve performance, only one of the words with the same root goes on to the next steps and the derivative ones are excluded. However, in order to carry out this step, it is necessary to have an intensive linguistic knowledge and to develop language-

specific methods. Because each language has different grammar rules and each prefix or suffix is added to the words in accordance with these rules, the removal process should be performed according to those. Otherwise, the root words that do not actually exist can be obtained as outputs and may negatively affect the results of the study. This step is usually applied to nouns, verbs and adjectives (Kowalski & Maybury, 2000).

Lourdusamy & Abraham (2018) divide the problems that may occur in the stemming step: over stemming and under stemming. In the case of over stemming, two separate words having two different roots are stemmed to the same root while in the case of under stemming, two different words having the same root are stemmed to two different roots.

Example:

Membership (Input) -> Member - ship = Member (Output) [Name]

Discovered (Input) -> Discover - ed = Discover (Output) [Verb]

Historical (Input) -> Historic - al = Historic (Output) [Adjective]

Different methods are also mentioned in the stemming step and are divided into linguistic/dictionary-based stemming and Porter-style stemming (Brants, 2003). Accuracy of the linguistic/dictionary-based stemming is much higher in methods, but cost is also proportionally increased and its usage area is narrower, because language-specific methods need to be developed. In the Porter-style method, lower accuracy rates are obtained, but in the same direction, the cost is lower and healthy results can be achieved in the field of information retrieval with these methods.

3.2.3. Stop-word Removal

Stop-words are very high frequency words that do not contain any content information (Zhai & Massung, 2016). They consist of grammatical words of the language. They are necessary for the formation of sentences that meet the rules of grammar and do not have any meaning by themselves. Since the main purpose of text mining is to extract meaningful information from the text being processed, stop-words should be removed. These may be words in the language such as conjunctions, prepositions and pronouns. For example; but, however, because, with, like, it, this and these. Srividhya & Anitha (2010) argue that the step of removing stop-words should be done immediately after the tokenization step because the more the elimination of unnecessary data is carried out in the first stage, the higher the performance of the model, in other words, it provides more optimum results in terms of time and space. A list of pre-formed stop-words specific to a language is loaded into the system

and in case of matching with the existing words in the text, these words are removed (Kowalski & Maybury, 2000).

3.2.4. POS Tagging

POS tagging is the process of labeling the words in the text that are present as input according to their tasks in the sentence. POS tagging is a very important step for identifying neighboring words by labeling language-specific elements of sentences such as nouns, verbs, adjectives, prepositions, conjugations, adverbs, and analyzing syntactic structure and observing the relationship between words (Lourdusamy & Abraham, 2018). However, there are also studies that consider this step as an additional process (Jo, 2018). But there are different opinions objecting those studies, too. POS tagging plays a very important role in many natural language processing areas such as speech recognition, machine translation, information retrieval and information extraction (Singh, 2018). POS tagging is generally examined in two categories: rule-based approaches and statistical approaches. Rule-based approaches require an advanced linguistic expertise and a comprehensive collection that requires labor and cost. In addition, since it is necessary to create separate corpora specific to each language, it is not possible to have a universal characteristic. Although the problem of universality has not been solved, a transformation-based approach has been proposed as an alternative to this approach and is intended to automatically learn from the corpora. On the other hand, statistical methods benefit from Decision Trees and the Hidden Markov Model and are not specific to a particular language, they are universal. The data obtained after POS tagging can be used in a different function as stemmed and labeled words can represent separate dimensions in the vector space model (Brants, 2003). Thus, the model can give much more detailed results about the data. These labels provide information about the semantic properties of the text (Feldman & Sanger, 2006).

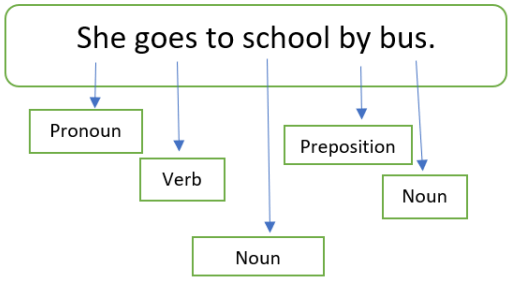


Figure 9: POS Tagging (Lourdusamy and Abraham, 2018)

3.2.5. Parsing and Chunking

Parsing and chunking steps are often given together in studies. Chunking separates the words in the sentence into basic phrases. Examples of these are noun phrases, adjective phrases and verb phrases. The phrases obtained by chunking form one dimension of the vector space model (Brants, 2003). Unlike parsing, chunking is used to create a hierarchical structure between elements of a sentence. Parsers are the comprehensive linguistic analysis of the text. These parsers are called syntactical parsing and divided into two categories depending on grammatical formalism: constituency and the dependency parser (Singh, 2018). The constituency parser separates the phrases in a sentence according to a hierarchical order and visualizes the relationship between the phrases. The root of the tree starts with "S", which represents the Sentence. Each sentence has a noun phrase and verb phrase, and they are referred to as "NP" and "VP", respectively (Zhai & Massung, 2016). In other branches, the labels of the words that were implemented in the previous step according to the content of the sentence are included in this tree. Shallow parsing, which is another method of parsing, is preferred when speed is important and forms a general model by separating only the noun and verb phrases without analyzing all the phrases in the sentence. This decomposition method is generally used to examine semantic relationships of phrases in sentences and to create functional classification labels. Dependency parser examines a sentence by reviewing the dependency of the words in pairs. Each dependency represents a linguistic function. These parsers see language as a set of relationships between words and create a graph for each sentence.

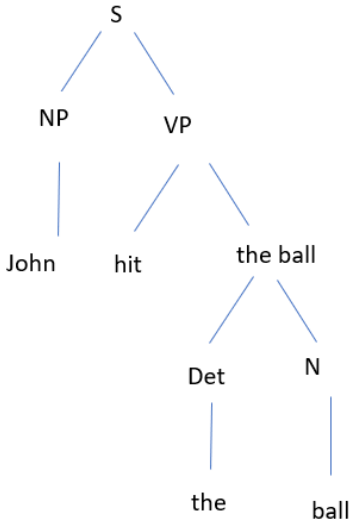


Figure 10: Parsing and Chunking (Zhai & Massung, 2016)

3.2.6. Word Sense Disambiguation

Word sense disambiguation is the process of distinguishing the correct meaning of the word within the context. When used in information retrieval, words are replaced by meaning in the vector space model (Brants, 2003). It is a step that needs to be done in sentences containing words that have different meaning according to the context. The problem of monophony mentioned in the previous section is the greatest example of this.

3.2.7. Dimensionality Reduction

Just as the removal of stop-words serves to remove words with very high frequency from the data, dimensionally reduction is intended to remove words with very low frequency. Document frequency is an important concept in this process and represents the number of documents in which a term exists. The most preferred method for dimensionally reduction is document frequency threshold. Words that are present in less than "m" documents are not considered as an attribute and are thus removed. "m" is a pre-determined threshold. This method is based on the idea that words that do not appear in a text at a certain frequency do not have an informative feature.

3.2.8. Term Weighting

Term weighting means calculating the weight of each word and assigning it an importance value. Each word in the text has different levels of importance (Salton & Buckley, 1988). Determining the term weights of the words before the stop-words removal and dimensionally reduction steps make those steps perform more effectively. There are three main factors that influence the importance of terms in a text: Term Frequency, IDF (Inverse Document Frequency) and Document Length Normalization (Karbasi & Bughanem, 2006). Therefore, Term Frequency and TF-IDF are the most commonly used term weighting methods. Term weights of the words are calculated according to their frequency of occurrence in the text. IDF is calculated according to the frequency of words found in all documents in the document database. The TF-IDF model is very popular in text classification and almost all models used for this process are variations of TF-IDF (Chrisholm & Kolda, 1998).

If all existing documents are called "D",

"w" is a word,

"d" is a document,

"w_d" is weight,

and the formula is:

$$w_d = f_{w,d} * \log(|D|/f_{W,D})$$

$f_{w,d}$ in the formula represents the frequency of "w" word in the "d" document. |D| is the size of the dataset and $f_{W,D}$ represents the frequency of "W" words in the "D" documents. The result of the TF / IDF measurement is a vector in which various terms exist together with their weights.

3.2.9. Document Indexing

Document indexing is defined as the process of converting a text into a list of words (Kowalski & Maybury, 2000). As mentioned in the first section, there are different opinions about the steps to be followed in the data pre-processing phase and this difference is mostly observed in the document indexing step. While document indexing is a term that expresses the data pre-processing phase in some studies, in others, it is considered as a framework covering the steps of dimensional reduction, term weighting and forming a vector space model. Therefore, in some studies, it is argued that document indexing has the same function as tokenization and in other studies, it is defined as a process in which keywords representing the documents in the best manner are selected and these keywords are appointed as the weight of these documents in the vector space model (Srivighya & Anitha, 2010). In this respect, document indexing is a step closely related to the weighting and dimensionally reduction.

3.2.10. Vector Space Model

The list of words obtained after tokenization is not yet suitable for numerical processing. Therefore, these words need to be converted to a numerical value. And each word should be converted to a term vector. The term vector facilitates further processing by giving numerical values to each word in the text. There are three ways to convert words into term vectors: Term Frequency, Term Occurrence and TF-IDF. The TF-IDF model is the most commonly preferred among them and it gives more weight to important terms and less weight to less important words. Vector values are observed between 0 and 1. In this context, 0 indicates that the word has no significance in the context of the text, while 1 indicates that the term is relevant to the text.

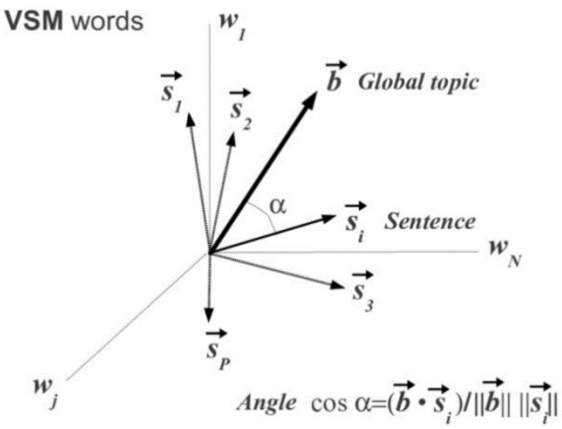


Figure 11: Vector Space Model (Moreno, 2012)

4. Conclusion and Evaluation

The aim of this chapter is to explain the data processing steps performed in text mining, and to convey the background information necessary to understand the subject and to support them with visuals. Firstly, explanations about the data (which is the main object of all these operations) are made and then the concept of data mining is examined. In this stage, classification, clustering, association and regression steps are mentioned among those models that are frequently used in data mining, what information is needed in these processes and what preliminary steps are needed to obtain healthy results are explained. Then, the concept of text mining, which is a sub-category of data mining, is tried to be defined and the fields that text mining is used are explained and the difficulties posed by the text data are mentioned. The reasons for the difficulty of expressing natural language with numerical algorithms are tried to be explained. Then, the text data involved in the text mining process was examined in more depth and the answers to the questions such as "what is text, a sentence and a word?" were sought. After providing all the necessary preliminary information, the data pre-processing steps were explained in detail and supported with visuals by emphasizing how important the data pre-processing stage plays in the whole text mining process.

It is understood from all the studies examined that the data pre-processing stage has a very important role in the text mining process. The data pre-processing step, which has a portion of about 50% to 80% of the entire text mining process, is of great importance in structuring the unstructured text data and creating the necessary models by means of algorithms. However, in this step, the methods chosen according to the type and purpose of

the text and the order of priority of these methods may vary. Although these differences exist, each method is of great importance in the process in which they are involved, and each step is a continuation or prerequisite of another. Therefore, before starting the whole text mining process, the available data should be examined very well, the content of the data should be known, the objectives should be determined by establishing solid foundations and necessary pre-processing steps should be followed in this direction. When all these conditions are fulfilled, it is foreseen that healthier text mining results will be obtained.

References

Brants, T. (2003, Ocak). *Natural Language Processing in Information Retrieval.* Conference: Computational Linguistics in the Netherlands.

Chrisholm, E. & Kolda, T.F. (1998). New Term Weighting Formulas for The Vector Space Method in Information Retrieval, Technical Report, Oak Ridge National Laboratory.

Duda, R.O., Hart, P.E. & Stark, D.G. (2000). Pattern Classification. Access Address: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.320.4607&rep=rep1&type=pdf

Eberandu, A.C. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Emerging Trends & Technology in Computer Science, 38(1), 46-50. DOI: 10.14445/22312803/IJCTT-V38P109

Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge, NY: Cambridge University Press.

Gantz, J. & Reinsel, D. (2011). Extracting Value from Chaos, IDC Iview.

Gaikwad, S.V., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. International Journal of Computer Applications, 85(17), 42-45.

Giagole, P.C., Patil, L.H. & Chaudhari, P.M. (2013). Pre-processing Techniques in Text Categorization. International Journal of Computer Applications. Access Address: https://pdfs.semanticscholar.org/ff34/7657082e70347a916548a9fe567ab791162a.pdf

Gurusamy, V. & Kannan, S. (2014). Pre-processing Techniques for Text Mining. Date: 18 February 2018, https://www.researchgate.net/publication/273127322_Pre-processing_Techniques_for_Text_Mining

Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques. USA: Elsevier Inc.

Jo, T. (2018). Text Mining: Concepts, Implementation and Big Data Challenge. Poland: Polish Academy of Science.

Jo, T. (2006). The Implementation of Dynamic Document Organization Using the Integration of Text Clustering and Text Categorization. University of Ottawa. http://dx.doi.org/10.20381/ruor-19708

Jones, K.S., & Manu, I. (Ed.). (1999). Automatic Summarizing: Factors and Directions in Advanced Automatic Summarization (pp.1-12). Cambridge, MA: MIT Press.

Kadhim, A.I. (2018). An Evaluation of Pre-processing Techniques for Text Classification. International Journal of Computer Science and Information Security, 16(6).

Kalra, V. & Aggarwal, R. (2018). Importance of Text Data Pre-processing & Implementation in RapidMiner. Proceedings of The First International Conference on Information Technology and Knowledge Management, (pp. 71-75). DOI: 10.15439/2018KMK6

Karbasi, S. & Boughanem, M. (2006). Document Length Normalization Using Effective Level of Term Frequency in Large Collections. Advances in Information Retrieval, Lecture Notes in Computer Science, 3936/2006, 72-83.

Kowalski, G.J. & Maybury, M.T. (2000). Information Storage and Retrieval Systems: Theory and Implementation. Boston: Kluwer Academic.

Lourdusamy, R. & Abraham, S. (2018). A Survey on Text Pre-processing Techniques and Tools. International Journal of Computer Sciences and Engineering, 6(3).

Manning, C.D., Raghavan, P., Schutze, H. (2009). Introduction to Information Retrieval. Cambridge, NY: Cambridge University Press.

Mitchell, T. (1997). Machine Learning. McGraw, NY: Hill Companies.

Moreno, J. (2012). Artex is AnotheR TEXt summarizer. CoRR, abs/1210.3312

Salton, G. (1998). Automatic Text Pre-processing: Transformation, Analysis and Retrieval of Information by Computer. Tokyo: Addison Weseley Publishing Company.

Salton, G. & Buckley, C. (1988). Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24(5), 513-523.

Sheeba, J. & Vivekanandan, K. (2012). Improved Unsupervised Framework for Solving Synonym, Homonym, Hyponym & Polysemy Problems from Extracted Keywords and Identify Topics in Meeting Transcripts. International Journal of Computer Science, Engineering and Applications (IJCSEA), 2(5), 85-92.

Singh, S. (2018). Natural Language Processing for Information Retrieval. arXiv:1807.02383 [cs.CL]

Srividhya, V. & Anitha, R. (2010). Evaluating Pre-processing Techniques in Text Categorization. International Journal of Computer Science and Applications, 2010.

Zanini, N. & Dhawan, V. (2015). Text Mining: An Introduction to Theory and Some Applications. Research Matters: A Cambridge Assessment Publication, 19, 38-44.

Zhai, C., Massung, Z. & Özsu, M.T. (Ed.). (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool Publishers.