Offering a Combined Approach to The analysis of News database, using the software Rapid Mine Case study: News agency in Persian

A.Saberi[1]*, S.Ahmadiyan[2]

1-Master of Electronic Commerce, Khaje Nasir Toosi University of Technology
2-Master of Information Systems Management, Imam Hossein University

ABSTRACT - One of the most widely used approaches in the management of Web-based systems, use the online agency and Data related to them. Online web agency, like the heart of a community and is one of the most critical information resources in a society. The importance of this topic in the news military organizations will be doubled. So having a right way and a unique conceptual approach to analyze and categorize the resources, Can provide numerous benefits and assistance to decision-makers within the organization, especially military organizations in carrying out decisions.

In this study, firstly we contribute to the study and understanding of the concepts of data mining methods. Secondly, analysis of different methods of text processing must be done from different perspectives So as to identify the positive strategies and strengths of a variety of methods and The Finally, the analysis of news database, so as to uncover information hidden in the data and how to use it.

Keywords: Data mining, text mining, News Database, Mixed-Method

1. INTRODUCTION

The news in the world today is often used as a database for archiving news, while among them there are very useful knowledge and hidden relationships. The discovery of this knowledge can be very practical and interesting for decision-makers and analysts in various fields and in particular the analysis of news.Given the huge volume of these data, scientific data must be analyzed, the hidden relationships and knowledge extracted between them and accessible to managers and decision-makers.In this way, the efficiency of the process of analyzing and identifying news is increased intentionally, because the analysis of the news is out of user analysis and can be accomplished through much higher yielding data mining methods.Using these methods can give all the organizations in the country the ability to analyze the news coming to the organization, according to the model of the organization, and to provide it to the decision makers of that organization.

Therefore, in this paper, by explaining the exact application of data mining and technology methods in the analysis of published news, how to extract relationships and work outcomes in the news system is explained.Eventually, with hidden knowledge available in the database, which will be analyzed and interpreted, a framework for analyzing the news and how the hidden relationships between the news will be presented based on the predictive methods of data mining.In general, the goals that follow from this method can be summarized as follows:

- Determine the keywords and nature of the news based on the data mining model, and provide analytical reports.
- Discovery of hidden knowledge of news produced based on the time series available among them.
- Detecting and analyzing hidden relationships between keywords and the nature of news in the past.
- The discovery of important words in published news, which has a great impact on the minds of people and readers of news and determines the orientations of the nature of news based on data mining techniques.
- Correct categorization and clustering, information and data.

2. DEFINITION OF CONCEPTS

2.1. Data mining definition

It can be said that text mining uses information retrieval techniques, information extraction, and the processing of natural language. And link them to KDD algorithms and methods, data mining, machine learning, and statistical data [1].Due to different research areas, different definitions of the text can be considered for each of them.

- Text mining = Extraction of information: In this definition, the corresponding text mining is considered to be extraction of information (extracting facts from the text).
- Text mining = Discovery of text data: Text mining can be considered as methods and algorithms of machine learning and statistical fields for texts with the aim of finding useful patterns. For this purpose, preprocessing texts is essential. In many ways, methods for extracting information, processing natural language, or some simple preprocessor for extracting data from texts are used. Then data mining algorithms can be applied to extracted data [2][1].
- Text mining = KDD process.

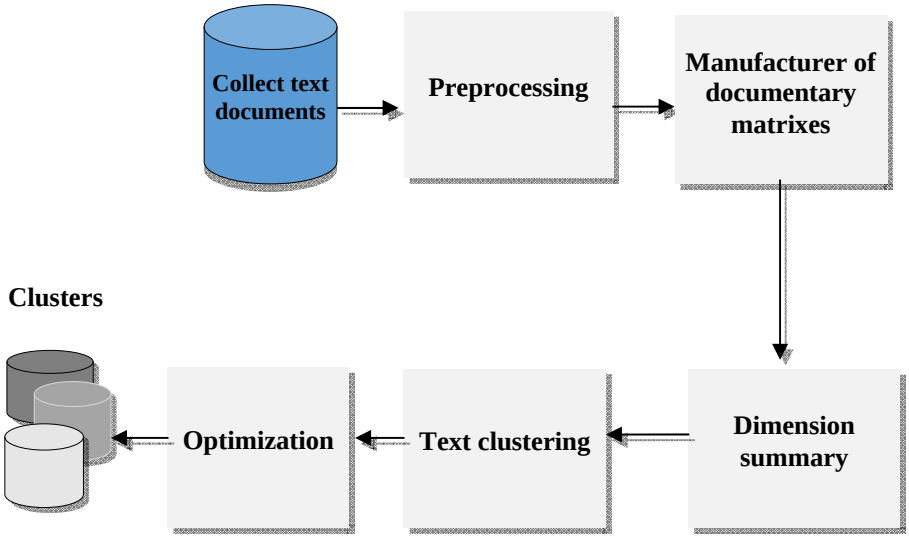

Figure 1: Common steps for text mining

2.2. Discovering Knowledge and Its Relationship with Text-mining

Knowledge Discovery in the Database (KDD), the term refers to all the ways in which it seeks out the relevance and order of the visible information. The KDD word is used to describe all stages of extracting information from the database, as well as the purpose of the primary tasks of the decision rules.In general, KDD is the process of finding information and useful patterns of data, and data mining is the use of algorithms to find useful information in the KDD process [4][3].

Among the features that can be used to measure the quality of the patterns found in the data are: human comprehensibility, validation with statistical criteria, novelty and usefulness. The discovery of knowledge in the database can be considered as a process defined by several steps of processing.These steps should be applied to the dataset in order to extract useful patterns. These steps are performed repeatedly, and some steps require feedback from the user.A KDD user In order to select the correct subset of the data, there is a need for a high understanding of the scope of the data, a proper category of patterns, and a good yardstick for interesting patterns. Therefore, the KDD system should have interactive tools, not automated analysis systems.According to [3][4], the steps can be expressed as follows: 1) understanding the business 2) understanding the data 3) preparing data 4) modeling 5) evaluation 6) deployment. The preprocessing stage is often one of the steps that takes a lot of time, and is still very important in achieving the desired result. Especially in text mining, which requires special preprocessing methods for converting text data into a format suitable for data mining algorithms.

2.3. Related search areas

There are three basic ways to deal with this vast amount of unstructured information: information retrieval, information extraction, and natural language processing.

Information retrieval: Essentially related to the retrieval of documents. The usual work involved in data retrieval is the consideration of the user's need for the most relevant texts and documentation. This is not the search for knowledge, but only the set of words that he considers more relevant to the searcher's information needs. This method does not really bring us any knowledge or even any information [5].

Natural Language Processing: The overall purpose of the NLP is to achieve a better understanding of the natural language by computers. Robust and simple techniques for fast text processing are used. Also, linguistic analysis techniques are used to process the text.

Extraction of information: The purpose of information extraction methods is to extract specific information from text documents. Extracting information can be used as a preprocessor in the text.Extracting information

includes mapping natural language texts (e.g., reports, articles, journals, newspapers, emails, web pages and any text database) into a predefined and predefined display or templates that are filled out, Selects the key information from the original text. Once the information is extracted, then the information can be stored in the database for future use [6].

3. RELATED WORKS

There are many methods in the knowledge extraction phase. However, all of these methods may be divided into two main categories. The main two are methods based on efficiency and knowledge-based methods.In the first approach, designers are worried about the system's performance, so they will design the system that has the best performance and speed. The most common methods in this approach are statistical methods and neural networks. Statistical methods are based on any kind of statistical information that can be extracted from texts. Things like repeating words alone, repeating words together, and similar things.On the other hand, there are knowledge-based approaches that look at this from another angle. They try first to extract as much as possible existing concepts from the body of texts and, secondly, to establish relationships between these concepts.The use of this method is highly dependent on the NLP. In fact, this is the goal that the NLP also pursues and that is the understanding of the text. Devices that use these methods are currently not numerous, but DR-LINK from Syracus University is one of them [7].

3.1. Discotex Method

This method was provided by Kanya in 2007 [8]. This method provides a new framework for information mining based on the integration of the Information Extraction System (IE) and the Standard Inference (KDD) module.IE converts text documents into more structured data. In fact, it searches for specific pieces of data in natural language documents and converts a set of semi-structured text documents to a more structured database.In this method, RAPIER and BWI are used to build IE. Then the database built by the IE module is used by the KDD module to explore more knowledge. In an improved version of this method, the rules derived from the KDD module are used to predict the missing information and improve the accuracy of the IE module. Apriori and Ripper have been used to build the KDD module.

3.2. Textminer method

In this method, firstly, the semi- structured data is changed, for example, documents are converted to structured data stored in a database. The second component applies data mining techniques to the output of the first component. Most methods for text mining apply exploration algorithms to labels attributed to each document.These tags may be keywords extracted from the document or just a list of words in the document. In textminer, exploration algorithms on terms(meaningful sequence of words such as department of computation) combined with events (meaningful set of terms, for example, in a financial domain, purchase between company A and B) extracted from Documents are applied.The authors believe that the most important feature factors that describe a document are the terms and events expressed in the document. This information is stored in a table called EvantType.Extracting information is an important technology that has one step ahead of it. In this way, once the information is extracted, then the information can be stored in the database and searched for the query and summarized in the natural language.The first necessary step is the linguistic advance (linguistic). This step involves a number of linguistic techniques such as tokenization, a part of speech labeling, and so on. The general objectives of this paper can be divided into: 1) Managing the information stored in the text database (document collections) 2) Extracting useful knowledge.This method consists of two components of text analysis and data mining. The first component converts semi-structured data into more structured data stored in the database, and the second component applies data mining techniques to the output of the first component. The purpose of this method is to manage information (categorizing documents in appropriate categories) and exploring data to explore knowledge.Therefore, in this method, semantics and events are first extracted and then stored in the database. Then, the proper clustering algorithm (using the Rock algorithm and the lnc concept) is applied to the generated database and the documents are grouped, so that the same documents fall into one group. Then, an appropriate classification algorithm (decision tree) is used to validate the results of clustering and better exploitation of the discovered knowledge. For further details, this approach can be found in [9].

4. RESEARCH METHODOLOGY AND PROPOSED FRAMEWORK

In today's research, a variety of methods are used. These methods, or quantitatively, use the statistics and numbers to achieve the result, or using qualitative methods. Research methods If we consider the two-sided vector, we can say that one side of the vector is quantitative methods and the other is qualitative methods. Quantitative methods are carried out using statistics and figures, and qualitative methods, with the help of different types of observation and interviewing fans, collect information. The common feature of these two methods is to equip the researchers at all stages of the research with a variety of information gathering and analysis engineers. In fact, quantitative methods deal with counting and measuring aspects of social life, while qualitative methods deal with the production of reasoning descriptions and the discovery of the meanings and changes of social activists [10].

However, all research methods do not summarize the use of quantity or quality, but one can adopt an approach that uses both methods depending on the type of research. These types of methods are called hybrid [11][12][13]. In fact, the emergence and use of hybrid methods to strengthen research is carried out.The combined method has been used in this study. The reason for this is the use of news text databases. These databases are quantitatively and qualitatively gathered from a variety of news stories. We also use the CRISP standard and methodology to illustrate the research framework by using data mining techniques. Figure 2 shows the proposed research framework [14].

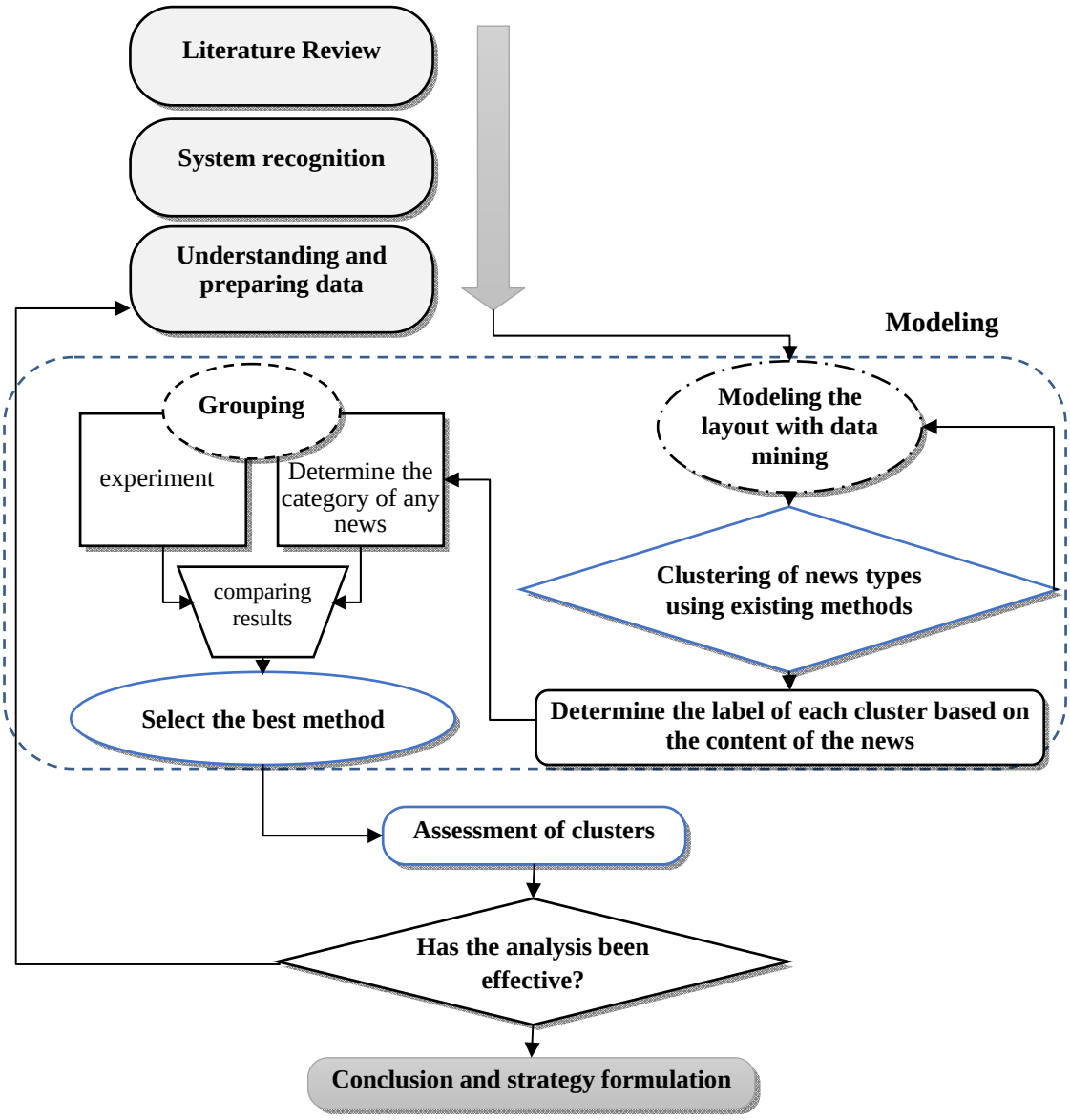

Figure 2: Proposed Research Framework

In order to analyze the research in this context, the literature review of topics such as data mining, news analysis, k-means algorithm and text mining and its various methods are discussed in detail. This survey was conducted to give a more complete coverage of the topics and gain ideas. The next phase of the research involves examining the existing system for the implementation of the analysis. For this purpose, various text and news analyzes were considered as a case study. The reason for this is the existence of high-volume data in this area and close connection with analytical topics.

During this research, in order to understand the space of the case, various types of web pages have been used in the field of news and literature reviews of this field. The next step is to identify the data and prepare it for use in the implementation phase. At this stage, it is examined which data is needed and what data in this template is appropriate for the implementation phase.

In this research, two different databases have been used. The first is the database of news type information. The second is the text newsletter database, which ultimately provides a more detailed analysis with the help of data mining algorithms for this section.In order to know what data is needed, we must first fully review the literature and fully understand the existing system. After the data is collected it is necessary to clear the data.One of the common problems with data is their low quality, the operation that fixes the quality problem is called data cleansing [15]. Removing additional records is one of the steps to clear data, duplicate or duplicate data must be identified and deleted, and confusing data should also be identified and repaired.

The next stage is modeling, in this phase two operational phases are taken. The first phase involves the use of the k-means algorithm to cluster the types of news texts in order to determine which news text is relevant to which news area.The second phase analyzes the categorization of existing news using methods such as the decision tree that provides an algorithm to specify the type of news and their categorization. The first phase is using the k-means clustering algorithm with the help of Rapid Miner software and the second phase using the same software. At this stage, we are looking to develop an analytical algorithm and improve the results.

5. IMPLEMENTING AND REVIEWING THE RESULTS

Available data for analysis and implementation in the project is related to various news from Internet news agencies. The databases included information gathered from about 100 Internet news sites in three different languages: Persian, English, and Arabic. In total there are two different tables in this database. A table of news information that contains information such as group identifier, source identifier, Report ID, Type Id, Identity, News Address, Release Date, Creation Date, File Counter, File Path, File Address, and News Status. The second table, called the text table, also contains information such as the text ID, the news ID, the file id, the language ID, the text, and the type of text.

The total number of data in the table contains 163486 unique lines and in the text table 447407 is the unique line. The reason for the distinction between the number of rows in the two tables is the type of data in the text table. In this way, each unique newsletter in the news table has different types of text stored in the text table, such as the title of the news, the news text and the summary of the news for a specific news.

The link operation is used to identify the valid news between the two tables. As a result, the number of rows obtained from this operation is 447,392. This means that different rows are identified between the two tables, and the total number is obtained accordingly. By doing this, it becomes clear that there are some additional data in both tables. The number of news identifiers in the news table, but not found in the text table, is 3367, which means that there is an existing news item but there is no text for the news. There are 15 numbers in the news table that are not referenced in the text table. These 15 records were in fact six distinct news identities, so this news item should be removed from the text table and finally the link table (This operation is shown in Figure 3).

```
delete
from dbo.News_Text
where NewsId in (select t2.NewsId
                 from dbo.News_Text t2
                 where t2.NewsId not in (select t1.NewsId
                                         from dbo.News_News t1))
```

Figure 3: Delete news that has not been found in the newsletter.

In the news table, there were 3367 records that did not find their news identifier in the text table, so this news item should be removed from the news table and eventually the link (This operation is shown in Figure 4).

```
delete
from dbo.News_News
where NewsId in (select t1.NewsId
                 from dbo.News_News t1
                 where t1.NewsId not in (select t2.NewsId
                                         from dbo.News_Text t2))
```

Figure 4: Delete the news in the news table but not found in the text table.

Given that the main purpose of the project is the text-mining of existing information from the Persian news data in the news agencies, all non-Persian records should be cleared from the text table. In this regard, 7462 records and the English text as well as 2948 records and Arabic texts were identified and deleted(This operation is shown in Figure 5).

```
DELETE
FROM News_Text
WHERE LanguageId = '3'
```

```
DELETE
FROM DBO.News_Text
WHERE LanguageId = '2'
```

Figure 5: Delete the news table records in English and Arabic.

The language id column that represents the language of the news was deleted. This column contains distinctive values for Persian, Arabic, and English languages. Now that there is only one value for all Persian texts, deleting this column does not cause an abnormality in the database.

The "News Address" and "Source ID" columns have a peer-to-peer connection, so the "News Address" column has been deleted.

The type identifier column in the news table has six distinct values and we only know about its two values (0 and 1). Due to the small number of remaining records, there are two solutions: (a) Removing the remaining records; (b) Not considering this column in data analysis. Also, in the text table, all records where the "File ID" column has a zero opposite value were deleted because these records are merely labeled news images and do not contain meaningful content for analysis (This operation is shown in Figure 6).

```
DELETE
FROM News_Text
WHERE FileId != '0'
```

Figure 6: Deletes the news that their file id is zero.

To communicate between news and text tables, the shared field between them (news ID) is used in the link table. In the link table, all records will appear that have a subscriber identity between the news table and the text table. So the "number of two-table link records" is the same as the "number of link records of the two tables after the deletion".

```
SELECT *
FROM dbo.News_Text t1
INNER JOIN dbo.News_News t2
ON t1.NewsId = t2.NewsId
```

Figure 7: Displays all records that have a news item with the same value between news and text tables.

All changes made to identify and preprocess the data in Table 1 are displayed.

Table 1: Database information after preprocessing

Explanation	Number
Number of records in the news table (unique line: News ID)	163486
Number of text table records (Unique row: Text ID)	447407
The number of link records between the two tables	447392
The number of news IDs in the news table is not found in the text table.	3367
The number of news IDs in the text table is not referenced in the news table.	15
The number of text table records that have unique news identifiers.	160125
The number of records in the news table after the removal of the News ID that is not available in the text table.	160119
The number of text table records after deleting the News ID that is not in the news table.	447392
The number of link records for two tables, after deletion.	447392
The number of text table records that contain the English text.	7462
The number of text table records that contain the tag of the images.	18109
The number of final records until the end of stage 9	418873

The table consists of linking two news and text tables, including information, news ID, text ID, type of text, text, group ID, source ID, status, and type identifier. The number of data columns dropped from 20 columns to 8 columns. The main reason for this diminution is the lack of sufficient information from the column or data redundancy.

6. CONCLUSION

In this paper, a hybrid method for analyzing the news data of Persian news agency has been presented. The proposed method follows the CRISP standard and during its implementation uses qualitative and quantitative data mining algorithms and the Kmeans algorithm.

During the process of this research, the concepts that are needed in the project are first described in detail. The next step is to examine the existing system for implementing the analysis.The next step is to recognize the data and prepare it for use in the implementation phase. Finally, at the implementation stage, we tried to fully explain the simple implementation of the research using the RapidMine software with simple and understandable code on the database and the proposed method.Using this hybrid analysis method, in addition to providing an adequate analysis of the news and finding hidden relationships between them, can have several sub-advantages, such as establishing a native method for analyzing news, a method compatible with the Persian language, establishing more semantic analyzes On common vocabulary in Persian news and eventually helping decision makers of civilian and military organizations to make the right decisions.

REFERENCES

[1] N. Aggarwal, A. Kumar, H. Khatter, and V. Aggarwal, “Analysis the effect of data mining techniques on database,« Advances in Engineering Software, vol. 47, pp. 164-169, 2012.
[2] Dutt, A., Ismail, M. and Herawan, T. (2017). A Systematic Review on Educational Data Mining. IEEE Access, 5, pp.15991-16005.
[3] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In Knowledge Discovery and Data Mining, pages 82–88, 1996.
[4] Parsaei, M. R., Rostami, S. M., & Javidan, R. (2016). A hybrid data mining approach for intrusion detection on imbalanced NSL-KDD dataset. International Journal of Advanced Computer Science and Applications, 7(6), 20-25
[5] Delen, D. Fast, A. Hill, T. Elder, J. Miner, G. Nisbet, S, Practical Text Mining and statical analysis for non- structured text data applications’ 2010.
[6] M. S. Deshpande and D. V. Thakare, “Data mining system and applications: A review,« International Journal of Distributed and Parallel systems (IJDPS), vol. 1, pp. 32-44, 2010.
[7] Alguliev, R., Aliguliyev, R., “Experimental investigating the F-Measure as similarity measure for automatic text summarization”, Applied and Computational Mathematics, Vol. 6, No. 2, pp. 278-287, 2007.
[8] Kanya, N., & Geetha, S. (2007). Information Extraction-a text mining approach.
[9] Hearst, M. A. (1999, June). Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 3-10). Association for Computational Linguistics.
[10] V.Kumar and M.Joshi .What is datamining? http://wwwusers.cs.umn.edu /~mjoshi/hpdmtut/sld004.htm, 2003.
[11] Lincoln y.s. And guba e.g, 2005, Paradigmatic Controversies, Contradictions and Emerging Confluences, Handbook of Qualitative Research, pp. 163-188.
[12] Greene. J. C, caracelli. V. J, graham. W. F, 1989, Toward a Conceptual Framework for Mixed-Method Evaluation Designs, Educational Evaluation and Policy Analysis, pp. 255-274.
[13] Bimonte, S., Sautot, L., Journaux, L., & Faivre, B. (2017). Multidimensional model design using data mining: A rapid prototyping methodology. International Journal of Data Warehousing and Mining (IJDWM), 13(1), 1-35.
[14] Niaksu, O. (2015). CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing, 3(2), 92.
[15] Windarto, A. P., & Wanto, A. (2018, September). Data mining tools| rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia. In IOP Conference Series: Materials Science and Engineering (Vol. 420, No. 1, p. 012089). IOP Publishing.