8th International Scientific Conference
Technics and Informatics in Education
Faculty of Technical Sciences, Čačak, Serbia, 18-20th September 2020

Session 2: *IT Education and Practice*

Professional paper
UDC: 004.65

Analysis of Term "Web" Using Opinion Mining Techniques

Jelena Plašić[1*], Andrijana Jovičić[1], Marija Blagojević[1]
[1] University of Kragujevac, Faculty of Technical Sciences Čačak, Serbia
[*] jelena.plasic@ftn.kg.ac.rs

Abstract: *Due to the growing amount of data available on the web and the valuable information that can be obtained from their analysis, this paper aims to present one of the ways of data processing, using opinion mining techniques. In this paper, the term "web" is analyzed, and the data on which the analysis was performed are taken from various social networks on which this word is mentioned. From the achieved results, a certain successful application of the mining tool is intended in order to determine the positive or negative connotation of the text, while further work in the field of "neutral" attitudes is forthcoming.*

Keywords: *data mining; opinion; web mining; text analysis; web mining tools*

1. INTRODUCTION

Web content mining involves extracting useful information from the content of the Web page, i.e. researching text, images and graphics on the Web page to determine the relevance of the content for a given query. Due to the heterogeneity and shortcomings of structured Web data, disclosing such information is a great challenge. The process of discovering content on the Web can be divided into two points of view: an agent-based procedure and a database-based procedure. The first aims to improve the retrieval and filtering of information, and the second to model data on the Internet in a more convenient form, in order to apply standard query mechanisms for databases and data mining for analysis [1].

Opinion mining, also known as sentiment analysis, involves the research of a text, i.e. the processing of natural language to monitor the public's mood about certain products and events. In this way, companies can analyze the opinions of users about their products, services or ideas. Opinion analysis involves the use of data mining, machine learning and artificial intelligence to process changing opinions [2]. Social networks like Facebook and Instagram have increased the ability to analyze opinions.

There is a large amount of related research in this area. Authors in [3] used opinion mining techniques to analyze the term "Information technology". The results indicate the possibility of successful application of the mentioned techniques while opening up a future research question in the domain of "neutral" attitudes.

The authors in [4] analyze students attitudes with text processing in Rapid Miner.

Sentiment analysis is used with a variety of data sources, from text to social media. In general, both structured and unstructured data are analyzed. The authors in [5] present an analysis of unstructured Twitter data.

Having in mind the previous research, the goal of the paper is defined: to determine the possibility of applying two different tools in order to analyze sentiment.

2. RESEARCH TOOLS AND METHODOLOGY

Voyant Tools [6] is an online tool for analyzing various text formats (text, HTML, XML, PDF, RTF, MS Word). The "Voyant Tools" tool was used to analyze the text for the experimental part of this paper. The tool offers the ability to copy the text we want to analyze or upload the entire file. For the purposes of this paper, the text "Introduction to Web Mining" (M. Yadav, P, Mittal, "Web Mining: An Introduction," IJARCSSE, March 2013, [1]) taken from the Internet was analyzed. Some of the sentiment analysis tools are: Brand24, Clarabridge, Repustate, Lexalytics Social Mention, Social Searcher, Sentiment Analyzer.

RapidMiner [7] is a tool that provides an integrated environment for data preparation, machine learning, text analysis and predictive analytics. It is used for business and commercial applications, for research, education, training, rapid prototyping and application development and supports graphical data analysis.

The RapidMiner tool was developed on an open core model. The free version is limited to one logical processor and 10,000 rows of data and is available under the AGPL (Affero General Public License -

license for free software) license [8]. Commercial versions start at $ 2,500.

As in the free version of RapidMiner the tool has much greater possibilities of graphical data analysis compared to other programs, this tool was used for the experimental part of the work. For the purposes of this paper, using the RapidMiner tool, the CSV file downloaded from the social-searcher site was analyzed. Data analysis was performed on a set of data collected from the English-speaking area, for the term "web".

Data analysis begins by inserting the downloaded data into the "RapidMiner" tool, when it is possible to make changes and adjust the data entry, or format the data in the table. Downloaded data includes images, statuses and videos downloaded from Facebook, Instagram, Twitter, YouTube, Vimeo, Flicker, Dailymotion, Reddit, etc. Collected user comments are positive, negative or neutral.

3. RESULTS

RapidMiner operators were used for data analysis, the data analysis procedure is shown in Figure 1.

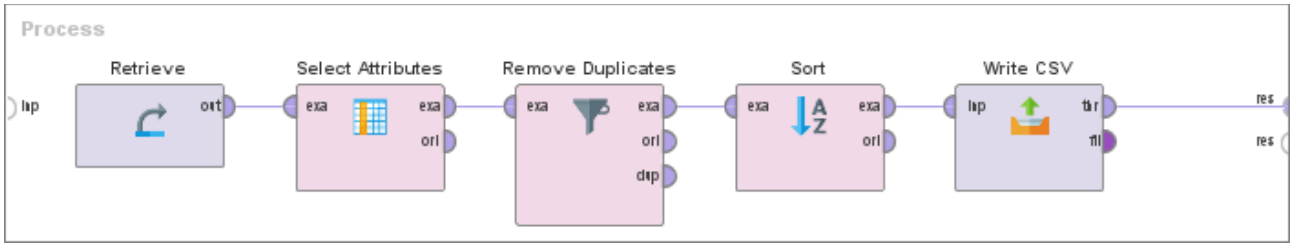

Figure 1. *Processes for data analysis*

The data is first entered into the Retrieve operator, whose task is to retrieve the entered data, which in this example is stored locally, in the data folder.

Once the data has been downloaded, the Select Attributes operator selects the attributes to be used in the analysis. The Select Attributes operator offers the ability to select the type of attribute selection. Only one attribute can be selected, which attributes will be displayed, and regular expressions can be used to customize the display of attributes. In this case, the attributes to be used in the analysis are selected.

After selecting the attribute, it is necessary to remove duplicate entries using the Remove Duplicates operator and sort the data using the Sort operator. In the given example, sorting is performed according to the network attribute, in descending order.

When sorting is complete, the results are saved to a CSV file using the Write CSV operator. The Write CSV operator offers the ability to adjust parameters such as the type of separator, whether to display attribute names in the first row, and whether to add it to an existing file.

Figure 2 shows the result of starting this process.

Figure 2. *The result of starting the data analysis process*

The RapidMiner tool offers the possibility of analyzing data based on the network from which they were downloaded, as well as a ranking list of all social networks sorted according to the amount of downloaded data.

In the given example, the data were collected from ten different social networks, and in Figure 3 it can be seen that most of the data was downloaded from Twitter, YouTube and Instagram, and the least from Reddit and Flicker.

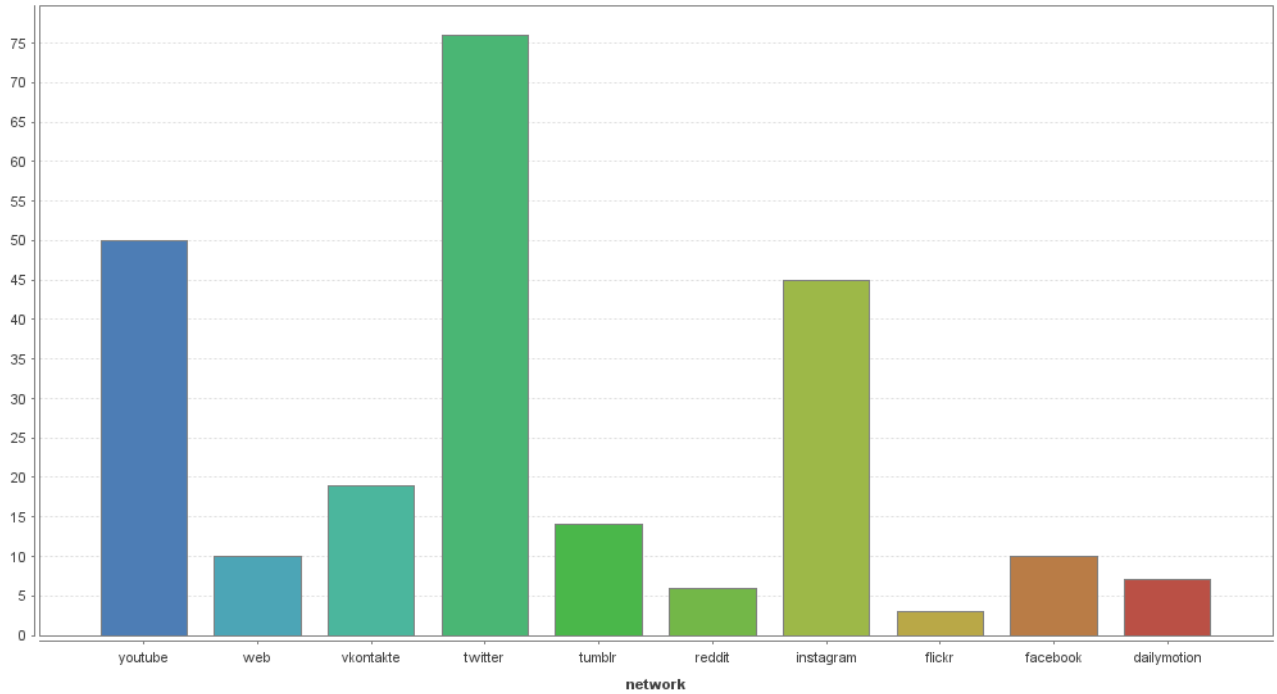

Figure 3. *Data analysis based on the network from which they were downloaded*

Data types are determined based on the Web sites from which they were downloaded. The collected data can be divided into four groups:

- links,
- videos,
- status,
- pictures

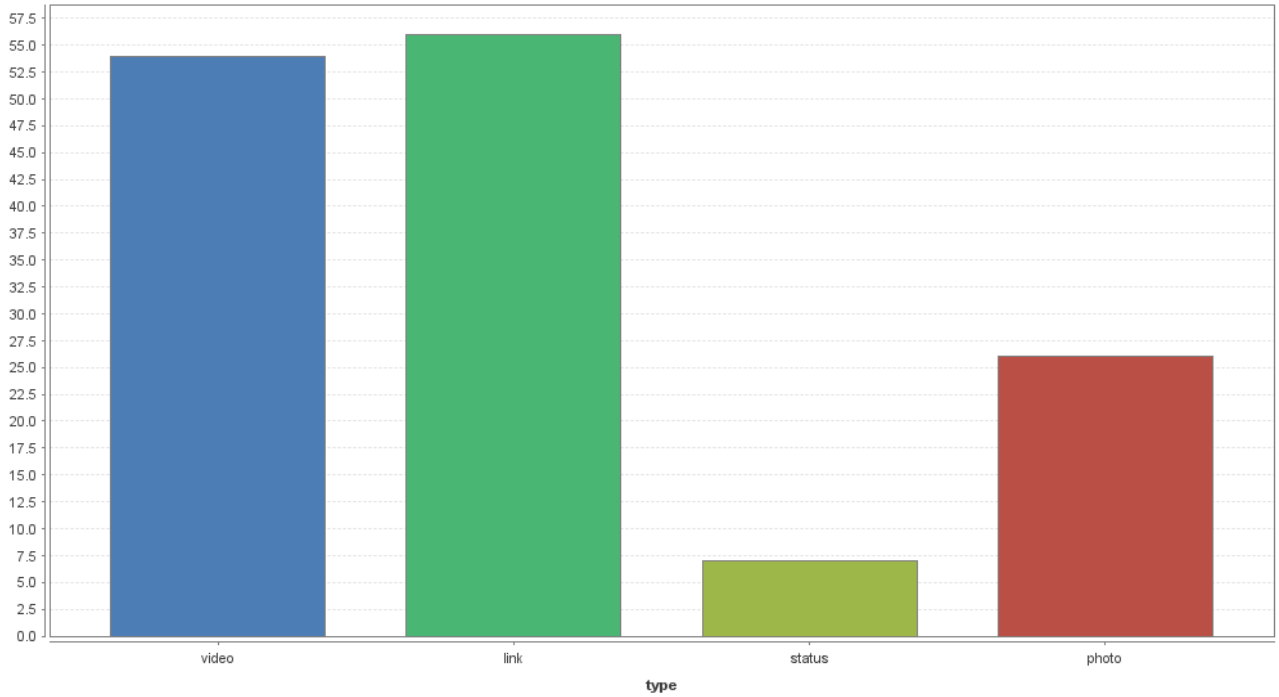

Figure 4. *Data analysis based on data type*

Figure 4 shows that the most common data types are link and video, while the least common type is status. The reason for this statistic is that most of the data comes from sites that serve for the distribution of multimedia data (video, images).

The sites that are used for discussion, publication of opinions and publication of tutorials are selected.

Sentimentality is the basic feeling, attitude, assessment or emotion associated with an opinion. It can be divided into three basic groups:

- positive,
- negative,
- neutral.

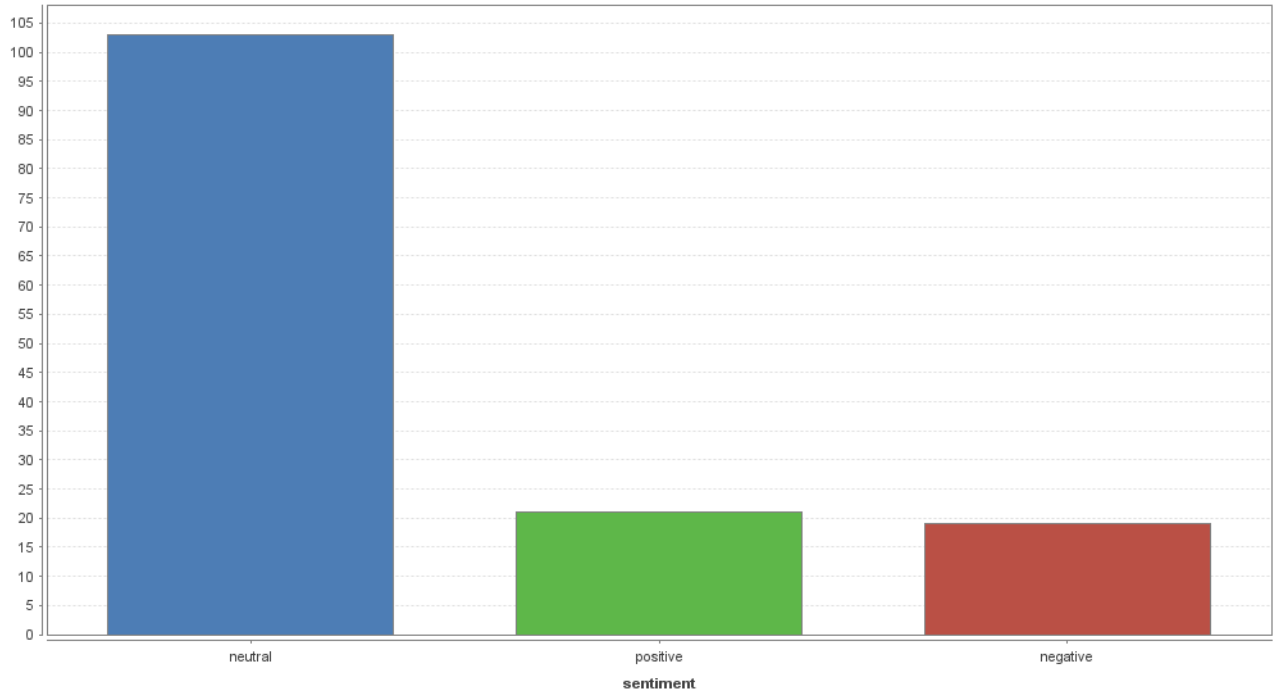

Figure 5. *Data analysis based on sentimentality*

In Figure 5 it can be seen that most neutral opinions (103), 21 positive and 19 negative. There is a possibility that neutral opinions can be classified as positive or negative opinions, because the programs that take comments are not advanced enough to recognize sarcasm or ambiguous expression of people.

Figure 6 shows the key words in the document, i.e. the words that are most often repeated with the links between them, so e.g. if any word is selected, the words associated with it will be bolded, as in Figure 6.

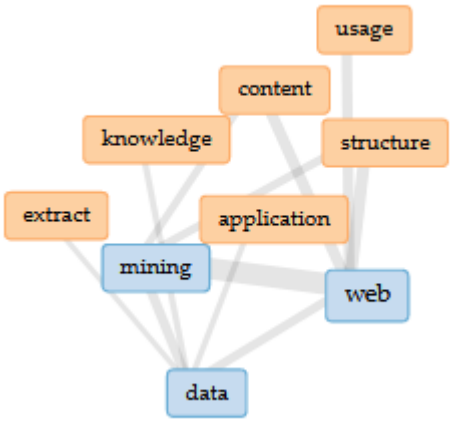

Figure 6. *Links between the most commonly used words in a document*

In order for the user to have an insight into the analyzed text, this tool also offers one of the windows in which the content to be analyzed is displayed. The user can thus review the text and can mark each word in the content and see how often it is mentioned in the entire text (Figure 7).

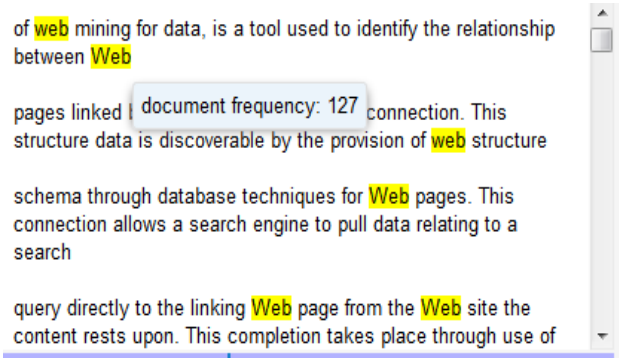

Figure 7. *The appearance of the word "web" and* the *total number of impressions*

Figure 8 shows the trend of the number of most frequently used words. The X axis presents the entire document that can be viewed by the number of pages, while the Y axis presents the number of repetitions. The disadvantage of this analysis may be the presentation of conjunctions that are inevitable in any text, so this type of analysis does not always bring useful results. The graph presents

an analysis of the most commonly used words, there are 5 of them and they are presented in various colours in order to more easily notice and see the results.

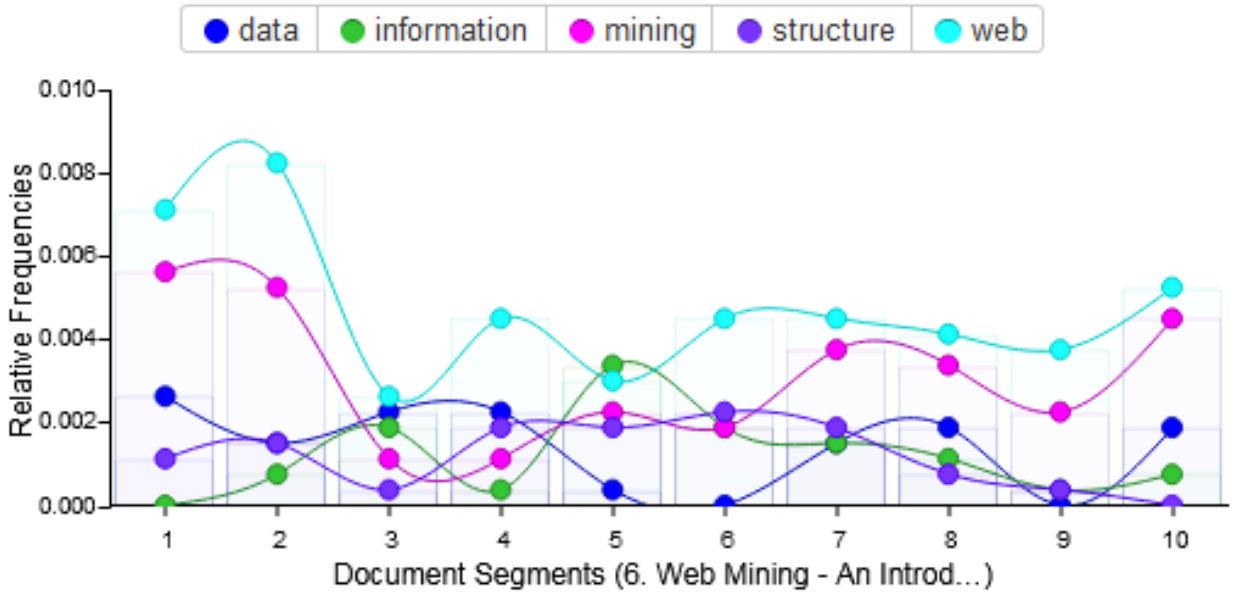

Figure 8. *The trend of the most frequently used words*

Figure 9 shows the summary results of the analysis of most frequently used words. The tools used at universities in order to detect plagiarism, i.e. the degree of identity with the content found on the Internet, or in the database, are based on a similar principle and technologies. For such systems, such technologies are used, i.e. the display of connections between words, so the identity of two or more textual contents is checked.

	Term	Count
1	web	127
2	mining	83
3	data	38
4	information	32
5	structure	32
6	content	20
7	usage	17
8	page	14
9	pages	14

Figure 9. *Summary results of the analysis of most frequently used words*

4. CONCLUSION

Bearing in mind presented results it could be concluded in several directions:

- the possibilities of opinion mining in analysis of different terms and getting valuable knowledge about using selected terms and context of using;
- the use of term "web" in context of status and trends of use;
- types of possible use in different sites (e.g. links, videos, status, pictures)

Future work is related to appliance of order technique of sentiment analysis in order to get more precise information about analysed term.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and these results are parts of the Grant No. 451-03-68/2020-14/200132 with University of Kragujevac - Faculty of Technical Sciences Čačak.

REFERENCES

[1] Yadav, M., & Mittal, P. (2013). *Web mining: An introduction*. IJARCSSE, March 2013.

[2] Liu, B., & Yhang, L. (2012). *A survey of opinion mining and sentiment analysis*. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA, 415-463.

[3] Jovičić, A., Plašić, J., Blagojević, M., & Ranković, A. (2020). *Analysis of term in information technologies using sentiment mining techniques*. 6th International conference on Knowledge management and informatics, 13.-14. 2020. January, Kopaonik

[4] Blagojević, M., & Kuzmanović, B. (2016). *Text processing in analysis of students' attitudes*.

International conference on information technology and development of education, ITRO 2016, Zrenjanin, June 10 2016. 97-100.

[5] Kharde, V., & Sonawane, S. (2016). *Sentiment analysis of Twitter data: A survey of techniques,* International Journal of Computer Applications. 139(11), retrieved from: https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf

[6] Voyant tool, retrieved from: https://voyant-tools.org/, last access, 05.07.2020.

[7] Rapid miner software, retrieved from: https://rapidminer.com/, last access, 05.07.2020.

[8] AGPL (Affero General Public License - license for free software), retrieved from: https://www.gnu.org/licenses/agpl-3.0.en.html, last access, 05.07.2020.