SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2022
ORGANIZATION AND MANAGEMENT SERIES NO. 163

THE IMPACT OF THE COVID-19 PANDEMIC ON THE LEVEL OF SENTIMENT IN IT PROJECTS IMPLEMENTED IN THE OPEN SOURCE FORMULA

Krzysztof S. TARGIEL

University of Economics in Katowice, Faculty of Informatics and Communication;
krzysztof.targiel@ue.katowice.pl, ORCID: 0000-0001-7815-1210

Purpose: The aim of this paper is to analyze possibilities of using sentiment analysis in project management.
Design/methodology/approach: The objectives were achieved by analyzing the sentiment on the mailing lists of the open-source project run by the Apache Software Foundation. The main method used for the research was calculation of sentiment and a comparison of its course with the events taking place in the surroundings of the project.
Findings: We have found that the attitude of users, as the main stakeholders, changes with the change of external factors (caused by the COVID-19 pandemic). At the same time, using lexical methods, it is not possible to determine the cause, but only the fact that such a change has occurred.
Research limitations/implications: The main limitation in research was using only one project. It must be checked on other similar projects. Other future research direction are using other than lexical methods.
Practical implications: The obtained outcomes allow for positive thinking about the possibility of using sentiment analysis in project management. This will require the definition of appropriate indicators and the definition of methods for their use.
Originality/value The novelty of the paper is comparison stakeholders (users) sentiment with external factor influencing on project.

Keywords: project management; sentiment analysis; natural language processing.

Category of the paper: Research paper.

1. Introduction

Modern project management is based on effective communication between project stakeholders. In today projects, electronic communication is playing an increasingly important role. This became more evident during the COVID-19 pandemic. Various types of electronic

communication based on instant messengers (IM) and e-mail systems have become the basis for the projects proceeding in a situation of social isolation caused by the pandemic.

Project success depends on the opinions of stakeholders. Sentiment analysis, which involves assessing the attitude presented in the stakeholders communication, in projects can be used. Changing the attitude of stakeholders from positive to negative is an important signal that the project is going wrong. The ability to catch such a moment will allow you to take appropriate corrective actions. In our work, we tried to investigate the impact of the COVID-19 pandemic on the projects. In particular, we tried to measure this impact through changes in the level of sentiment in IT projects carried out at that time. Due to the public availability of communication, open-source projects implemented by the Apache Software Foundation were used.

Communication based on emails are especially susceptible to automatic analysis based on Natural Language Processing. Many computer methods of natural language processing (NLP) are currently being developed. They are methods of text and speech processing (Speech recognition, Word segmentation), Morphological analysis (Lemmatization, Stemming), Syntactic analysis (Parsing), Lexical semantics (Sentiment analysis, Terminology extraction) and many others like Automatic summarization and Machine translation. Sentiment analysis seems to be a particularly useful tool for analyzing communication in a project. Having such tools at our disposal, we can analyze the impact of external phenomena on IT projects implemented in the open-source format.

The aim of the study is to study the impact of the phenomenon of the COVID-19 pandemic on the sentiment observed in communication in open-source IT projects. We will try to find an answer to the research question *Can we observe change in sentiment during COVID-19 pandemic*. In order to find an answer to this question, communication in a selected project implemented by the Apache Foundation consortium will be analyzed

The work is divided into the following parts. The first section presents analysis of sentiment in open-source projects. The phenomenon of open-source projects is explained here and the concept of sentiment analysis is also explained. The next section presents methodology used in analysis of sentiment in this work. Proposed approach of sentiment analysis, used data and methods of preprocessing are explained. Next section presents visualization of main findings on level of sentiment in time of COVID-19 pandemic. The paper ends, with section which present the conclusions and directions of further research on the use of NLP in project management.

2. Sentiment analysis of Open-source projects

2.1. Open source projects

Nowadays, software development is a very complicated undertaking in which many specialist are involved. One of the most effective ways to develop software is the open-source formula. Open-source software (OSS) is computer software that is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose (Laurent, Andrew, 2008).

Some organizations which are follows open-source formula are the Linux Foundation, the Eclipse Foundation, home of the Eclipse software development platform, the Debian Project, creators of Debian GNU/Linux distribution; the Mozilla Foundation, home of the Firefox web browser and finally the Apache Software Foundation.

Mailing lists are the core means of project communication in open source projects, where they are used during software development and maintenance to discuss technical issues, propose changes, report bugs, or ask how-to questions about configuration or any other parts of the product (Obaidi, Klünder, 2021).

2.2. Sentiment analysis

Recent years have seen a strong development of computer natural language processing methods. After the first periods of Symbolic NLP (1950s - early 1990s), and Statistical NLP (1990s - 2010s), present NLP methods have huge potential for implementation. Natural Language Processing (NLP) refer to automated machine-driven algorithms for understanding of human language and extracting information (Dinov, 2018). Common tasks for these methods include text and speech processing, morphological analysis, syntactic analysis, lexical semantics, relational semantics, and discourse (Natural language processing, 2021). Some new applications includes: automatic summarization, machine translation, natural language generation. One of the very interesting directions of NLP development in the context of project management is the analysis of sentiment.

Sentiment Analysis (SA) is defined as "the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information" (Sentiment analysis, 2021). SA is used to classification the polarity of text at the document, sentence, or word level. Text can be classified as positive, negative, or neutral. Some more sophisticated models, can classify also emotional states such as enjoyment, anger, disgust, sadness, fear, and surprise according to Plutchik wheel of emotions (Plutchik, 1980).

There are two main approaches in Sentiment Analysis:

- machine learning – the analyzed text is introduced to the input of pre-learning neural networks, on the output of which the sentiment value is obtained,
- lexical approach – has used lexicons of known sentiment-related words, their polarities, then uses those to score the sentiment of the text.

The first approach has some disadvantages in context of project communication. There are not enough data to learn neural network. This problem can be solved by using pretrained neural networks with embedded layers like GloVe for Tweeter texts (Pennington et al., 2014). The use of lexical methods and neural networks for sentiment analysis is widely discussed in the paper (Aydogan, Akcayol, 2016).

The second approach is based on lexicons. They are created in specific language, based on specific corpus. This approach was used in Tourani et al. paper (Tourani et al. 2017), where authors have used general lexicons to analyze communication in open source projects. One of the most known systems using lexicons is SentiStrength (Thelwall et al., 2010).

The use of sentiment analysis to study the overtone of short textual information has a long history. Wiebe, Wilson and Cardie (2005) worked on manual annotation of expressions of opinions and emotions in language for the large corpus. Arora et al. (2010) have contribution on the automatic feature construction. Gupta, Gilbert and Di Fabbrizio (2012) worked on emotion detection in email customer care. Saif and Turney (2010) worked on the creation of an emotion lexicon using Amazon's Mechanical Turk. Bandhakavi et al. (2017) worked on general-purpose emotion lexicons (GPELs) that associate words with emotion categories. Kumar, Kawahara and Kurohashi (2018) proposed a two-layered attention network based on Bidirectional Long Short-Term Memory for sentiment analysis. It is also worth mentioning that there are papers trying to capture the impact of the Covid-19 pandemic on public mood, to mention only the work by Ghosh (2021).

Sentiment analysis is widely used in the financial market, Bollen et al. (2011) was the first to prove that Tweeter analysis allows for predictions of financial markets. Creamer et al. (2013) showed the effect of financial information on volatility. Bukovina (2016) used also in these considerations the analysis of sentiment. In the work (Oliveira et al., 2017), sentiment indicators on Tweeter were used to forecast volatility. In the work (Seng, Yang, 2017) sentiment dictionaries were used to forecast volatility.

Beginning with Tourani et al. paper (Tourani et al., 2017), the analysis of sentiment was also used in the analysis of communication in software projects. Some more interesting works in this regard include: Islam and Zibran (2018) which compare different tools to detect sentiment among developers. A critical approach to using sentiment can be found in (Lin et al., 2018). Alesinloye et al. (2019) had identify and interpret patterns of sentiment during a release-cycle. Biswas et al. (2019) had used word embedding for sentiment analysis in software projects. The most recent literature review in Sentiment analysis tools used in software project we can find in (Obaidi, Klünder, 2021).

3. Methodology

3.1. Proposed approach of sentiment analysis

To achieve the intended goals, emails were selected for the chosen project that was implemented before and during the COVID-2019 pandemic. The messages were selected for the period which covers the period of the pandemic. These emails were processed, removing their irrelevant elements, and then using the lexical methods, the analysis of the expressions used was made.

The expressions used affect the attitude of the stakeholders during the course of the project. Their change after the outbreak of the pandemic may indicate the impact of this phenomenon on IT projects.

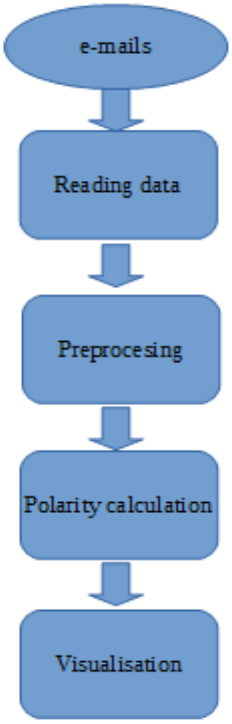

Figure 1. Approach of sentiment analysis process.

Source: own elaboration.

The flowchart of the sentiment analysis is shown in Figure 1. First we read e-mails into the computer system. Then we pre-process the data. For such prepared data polarity is calculated and finally visualization is shown.

3.2. Data selection for pandemic situation

To verify the hypotheses, the mailing list of the Apache OpenOffice project, implemented in the open-line formula, was selected. For this project communication is publicly available at "https://openoffice.apache.org/mailing-lists.html". Mailing list is maintained since 2011 till today.

OpenOffice is an open-source office suite. It was an open-sourced version of the earlier StarOffice, which Sun Microsystems acquired in 1999 for internal use. Sun open-sourced the OpenOffice suite in July 2000 as a competitor to Microsoft Office. In 2011, Oracle Corporation, then the owner of Sun, announced that it would no longer offer a commercial version of the suite and donated the project to the Apache Foundation. Apache renamed the software to Apache OpenOffice. Today the most actively developed successor projects is LibreOffice (OpenOffice.org, 2022).

Apache OpenOffice is an office productivity suite providing six productivity applications (Writer, Calc, Impress, Draw, Math, Base) based around the OpenDocument Format (ODF). OpenOffice is released on Windows, OS X, Linux. It is available in 41 languages.

3.3. Dataset

From the sixteen lists kept for the Apache OpenOffice.org project, a list of system users was selected, available at users@openoffice.apache.org. From the period including the times before and during the pandemic, files in .mbox format were downloaded from the archive, containing communication in the project. This period covered the months from January 2017 to December 2020. Each file .mbox contained communication in a given month. These files were unpacked into separate folders, which contained, as separate files, e-mails sent to the list in a given month. Then in the R system (R Development Core Team 2022) they were read in for further processing.

The final form of the data downloaded to the R system is presented in Table 1. The 'author' field contains the email address of the email author. The 'datetime' field contains the exact date of sending the email in POSIX format. The 'subject' field contains subject of the mail, and the field 'content' contains text of the mail.

Table 1.
Data structure read into R system

	Dataframe emails			
	author	**datetime**	**subject**	**contend**
Type of data	character	POSIXct	character	character

Source: own elaboration.

3.4. Data preprocessing

After reading the data into the R system in the format presented in Table 1, the data was preprocessed using the **tm.plugin.mail** package available in the R system. Preprocessing include:

- removing citation using function removeCitation() - which removes citations, i.e., lines beginning with >, from an email message,
- removing multiparts using function removeMultipart() - which removes non-text parts from multipart email messages,
- removing signatures using function removeSignature() - which removes signature lines from an email message.

Finally, a marker in the *yyyy-mm* format was added to each email message, indicating the month when the e-mail was created. It will be used to group and calculate the sentiment in the following months.

3.6. Sentiment analysis techniques

The polarity() function available in the **qdap** package was used for the sentiment calculation. This function approximate the sentiment (polarity) of text by using grouping variables (months). Mentioned function, have implemented method proposed by Hu and Liu (2004). In this method are counted words with positive and negative polarity. Polarized Terms, that are words associated with positive or negative context are stored in lexicon (Hu Liu Polarity Lookup Table: hash_sentiment_huliu). Other words are Neutral Terms - words with no emotional context. Positively polarized terms have value +1, as negatively polarized terms have value -1. The sum of the values of the polarized words in a sentence is divided by the root square of the number of all words. This gives the value of the polarity of the sentence, as it is presented in equation (1).

$$polarity = \frac{\sum_{i=1}^{n}(x_i^P + x_i^N)}{\sqrt{n}} \tag{1}$$

where:

x_i^P – value of positively polarized term, if term is with positive polarity it is is equal to 1, else 0,

x_i^N – value of negatively polarized term, if term is with negative polarity it is is equal to -1, else 0,

n – number of terms in sentence.

There are also available Negators (words that invert polarized meaning) and Valence Shifters (words that effect the emotional context). We have two kinds of Valence Shifters: Amplifiers (words that increase emotional intent) and De-Amplifiers (words that decrease emotional intent).

The sentiment values are constrained to be between -1 and 1 using equation (2).

$$sentiment = \left[\left[\left(\left(1 - \frac{1}{1 + exp(polarity)}\right) \cdot 2\right)\right]\right] - 1 \quad (2)$$

For each month in the selected time period, the following were counted: the number of emails (total.sentences), the number of words (total.words), the average value of sentiment (ave.polarity), standard deviation of sentiment (sd.polarity).

4. Visualization of results

First, we should glimpse at the data. On Figure 2. we can see number of emails in each month. The red line identifies the adopted month of the beginning of the COVID-19 pandemic in March 2020. We can see that the activity on the mailing list did not change drastically after the outbreak of the pandemic. Only at the end of the 2020 year is a significant decline in activity observed, which may be caused by pandemic situation.

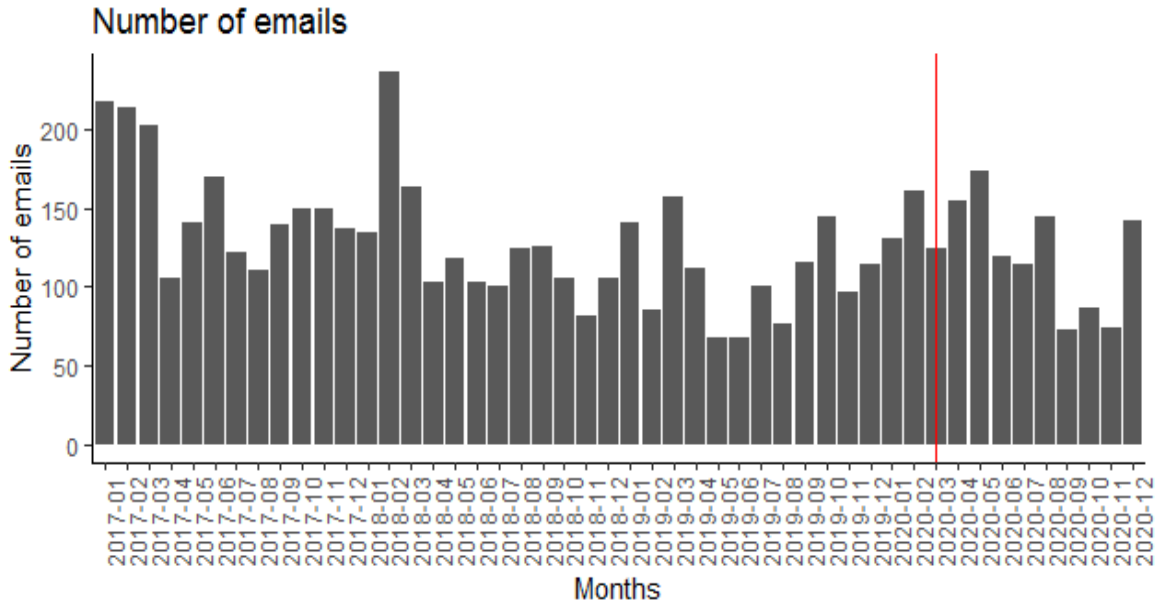

Figure 2. Number of emails in month.

Source: own elaboration in **ggplot2.**

The main topic is the level of sentiment. We can see it in Figure 3. Here the vertical (red) line marks the start of the pandemic, i.e. March 2020. In the same drawing, the horizontal (blue) line marks the neutral level of sentiment. When the sentiment is above this line, we observe a positive attitude to the project of its users. However, if it drops below, it indicates a negative attitude to the project.

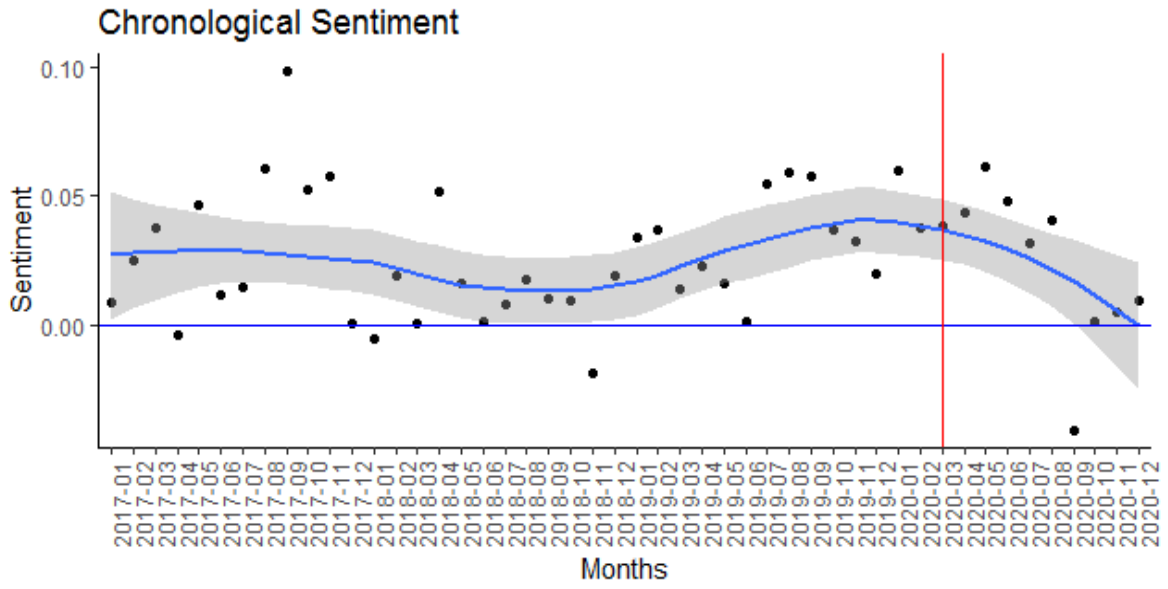

Figure 3. Level of sentiment in month.

Source: own elaboration in **ggplot2.**

The average values of sentiment in each month, are presented on Figure 3 as a points. To better present tendency we use locally estimated scatterplot smoothing method (loess), presented on this figure as a solid (blue) line. Method loess was proposed by Cleveland, Grosse and Shyu (Cleveland et al., 1992). There is also presented band of local standard deviation which is generated by package **ggplot2** (Wickham, 2016), in which the graph was made.

In the months under consideration, the general attitude of OpenOffice users is positive. The outbreak of the pandemic did not change the attitudes of users. Only at the end of 2020, we will observe a significant decrease in user attitudes. The analysis does not show what it is caused by, only we can note a significant decrease in this sentiment. Probably by referring to the content of the emails sent then, one can find the reasons for this decline, but it is no longer the subject of sentiment analysis. Her task was to note significant changes in sentiment.

Another findings which we can see on Figure 3, is increasing standard deviation band at the end of 2022 year. It is another phenomenon which may be caused by pandemic situation, but SA it does not determine it, it only notices these phenomena.

5. Conclusion and recommendations

Summarizing the presented work, it should be stated that there are observed changes in the attitude of users during the life of the project. They can be used for project management as long as they are presented as certain indicators that the project manager can track. The main findings here is the statement that Sentiment Analysis, in the form in which it was used, does not identify the cause of a change in sentiment, but only the fact of such a change itself. Further research,

for example on the content of sent e-mails, may help to find the reasons for changes in the attitude of stakeholders towards the project.

In the presented study, changes in sentiment were analyzed and compared with changes in the global situation caused by the Covid-19 pandemic. While such changes are observed in the attitude of users, there were difficulties in quantifying such a phenomenon. The first problem was to determine the exact moment when the pandemic began. As the projects under consideration were global, the designated start point was very similar due to the different start times of the pandemic in different countries. Another problem encountered during the research was the specificity of the IT industry. While large changes are observed in the level of user sentiment, such large changes were not observed in the developers' reactions (results not presented in this paper). This was due to the fact that social isolation did not adversely affect the work of developers.

References

1. Alesinloye, J.A., Groarke, E., Babu, J., Srinivasan, S., Curran, G., Dennehy, D. (2019). *Sentiment analysis of open source software community mailing list: a preliminary analysis.* Proceedings of the 15th International Symposium on Open Collaboration, pp. 1-5. OpenSym ’19. New York, NY, USA: Association for Computing Machinery.
2. Arora, S., Mayfield, E., Penstein-Rosé, C., Nyberg, E. (2010). *Sentiment Classification Using Automatically Extracted Subgraph Features.* Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, Los Angeles, pp. 131-139.
3. Aydogan, E., Akcayol, M.A. (2016) *A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques.* Proceedings of the 2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Vol. 37. C. Badica, M. Cosulschi, A.M. Florea, P. Koprinkova-Hristova, T. Yildirim. New York: IEEE.
4. Bandhakavi, A., Wiratunga, N., Massie, S., Deepak, P. (2017). Lexicon Generation for Emotion Detection from Text. *IEEE Intelligent Systems, Vol. 32*, pp. 102-108, doi: 10.1109/MIS.2017.22.
5. Biswas, E., Vijay-Shanker, K., Pollock, L. (2019) *Exploring Word Embedding Techniques to Improve Sentiment Analysis of Software Engineering Texts*. 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), IEEE, pp. 68-78.
6. Bollen, J., Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, Vol. 2*, pp. 1-8.
7. Bukovina, J. (2016). Social media big data and capital markets–An overview. *Journal of Behavioral and Experimental Finance, Vol. 11*, pp. 18-26.

8. Cleveland, W.S. Grosse, E., Shyu, W.M. (1992). Local regression models. In: J.M. Chambers, T.J. Hastie (Eds.), *Statistical Models*. Wadsworth & Brooks/Cole.
9. Creamer, G.G., Ren, Y., Nickerson, J.V. (2013). *Impact of Dynamic Corporate News Networks on Asset Return and Volatility.* Proceedings of the 2013 International Conference on Social Computing, pp. 809-814.
10. Dinov, I.D. (2018). Natural Language Processing/Text Mining. In: I.D. Dinov (ed.), *Data Science and Predictive Analytics: Biomedical and Health Applications Using R* (pp. 659-95). Cham: Springer International Publishing.
11. Ghosh, D. (2021). Impact of the Covid-19 Pandemic on the Expression of Emotions in Social Media. *Multiple Criteria Decision Making , Vol.* 15, pp. 23-35
12. Gupta, N., Gilbert, M., Di Fabbrizio, G. (2012). Emotion Detection in Email Customer Care, *Computational Intelligence, Vol. 29*, pp. 10-16, doi: 10.1111/j.1467-8640.2012.00454.x.
13. Hu, M., Liu, B. (2004). *Mining and Summarizing Customer Reviews*. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle: ACM.
14. Islam, M.R., Zibran, M.F. (2018). *A comparison of software engineering domain specific sentiment analysis tools.* 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), pp. 487-91.
15. Kumar, A., Kawahara, D., Kurohashi, S. (2018). *Knowledge-enriched Two-layered Attention Network for Sentiment Analysis.* Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2, pp. 253-258.
16. Laurent, S., Andrew, M. (2008). *Understanding Open Source and Free Software Licensing*. O'Reilly Media.
17. Lin, B., Zampetti, F., Bavota, G., Di Penta, M., Lanza, M., Oliveto, R. (2018). *Sentiment analysis for software engineering: how far can we go?* Proceedings of the 40th International Conference on Software Engineering, pp. 94-104. ICSE ’18. New York, NY, USA: Association for Computing Machinery.
18. *Natural language processing*. Available online https://en.wikipedia.org/wiki/ Natural_language_processing, 15.12.2021.
19. Obaidi, M., Klünder, J. (2021) Development and Application of Sentiment Analysis Tools in Software Engineering: A Systematic Literature Review. *Evaluation and Assessment in Software Engineering*, pp. 80-89.
20. Oliveira, N., Cortez, P., Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications, Vol. 73*, pp. 125-144.
21. *OpenOffice.org*. Available online https://en.wikipedia.org/wiki/OpenOffice.org, 10.10.2022.

22. Pennington, J., Socher, R., Manning, Ch.D. (2014). *GloVe: Global Vectors for Word Representation*.
23. Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. New York: Harper and Row.
24. R Development Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
25. Saif, M., Turney, P. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26-34, Los Angeles, CA.
26. Seng, J.-L., Yang, H.-F. (2017). The association between stock price volatility and financial news? A sentiment analysis approach. *Kybernetes, Vol. 46, Iss. 8*, pp. 1341-1365.
27. *Sentiment analysis*. Available online https://en.wikipedia.org/wiki/Sentiment_analysis, 14.04.2021.
28. Thelwall, M., Buckley, K., Paltoglou, G., D. Cai, D., Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, Vol. 61, Iss. 12,* pp. 2544-2558.
29. Tourani, P., Yiang, Y., Adams, B. (2017). *Monitoring sentiment in open source mailing lists: exploratory study on the apache ecosystem.* Proceedings of 24th Annual International Conference on Computer Science and Software Engineering, CASCON'14, pp. 34-44.
30. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.
31. Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation, Vol. 39, Iss. 2-3*, pp. 165-210.