Hybrid Method for Emotion and Sarcasm Classification in Polish Based on English Dedicated Methods

Urszula Gumińska
Institute of Information Technology/Lodz University of Technology
Lodz, Poland *urszula.krzeszewska@dokt.p.lodz.pl*

Aneta Poniszewska-Marańda
Institute of Information Technology/Lodz University of Technology
Lodz, Poland *aneta.poniszewska-maranda@p.lodz.pl*

Remy Dupas
Universite de Bordeaux
Bordeaux, France *remy.dupas@u-bordeaux.fr*

Abstract

Artificial intelligence and natural language processing are rapidly developing fields. Natural Language Processing (NLP), like other Machine Learning (ML), Deep Learning (DL), and data processing tasks, requires a large amount of data to be effective. Despite the emergence of newer and better NLP models, text processing in languages other than English, such as Polish, is still problematic. Applications that recognize emotion or sarcasm in texts, among others, as those that could help people with intellectual disabilities face many challenges, from a lack of data in a specific language to a shortage of solutions dedicated to such problems. Responding to this need could be the hybrid model created as part of this research.
The paper presents the proposal of hybrid method used for emotion and sarcasm classification in Polish that was based on English dedicated methods together with its implementation and evaluation based on performed experiments using the proposed datasets.

Keywords: Machine Learning, Deep Learning, Natural Language Processing, Polish emotion classification, Polish sarcasm classification.

1. Introduction

Natural Language Processing (NLP) is a rapidly growing field of Artificial Intelligence (AI) [1]. As its important sub-field is used to analyse, understand and generate language. NLP studies the interaction between human and computer through natural language – used by humans in everyday communication. Natural language processing, understood as a field of computer science and linguistics, can encompass areas outside of both machine learning and deep learning, as well as take advantage of the strengths of both. Recent progress in the field of Generative Artificial Intelligence (GAN), in particular the ChatGPT tool, has further drawn the attention of researchers to issues concerning text processing.

Despite the fact that there are many solutions that are working properly, one can still encounter the problem of accessing these tools in languages other than English. Even if there are people willing to prepare such a tool for a specific natural language, there is often a lack of good quality data to do so [2]. In addition, the proposed multilingual solutions are often too large, requiring a huge investment of money and computing power to make them work.

Researchers are trying to address the lack of dedicated tools in low resourced languages by creating ones based on machine translation. Datasets for specific task are translated from English into another language to serve as learning data. While this solution has its advocates, it

also has its drawbacks, such as the fact that the model does not learn what the language actually looks like, what its dynamics are, but only, formal or semi-formal translations.

Within the scope of this work, it was decided to reverse this process and use a machine learning model as part of a classifier system, where models are dedicated to the classification of English texts. The goal is to create a hybrid tool for Polish language classification, using English language classification models. Doing so would allow the use of English language resources for classification in Polish, without the need of having a datasst for a specific classification task in low-resource language.

The paper presents the proposal of hybrid method used for emotion and sarcasm classification in Polish that was based on English dedicated methods together with its implementation and evaluation based on performed experiments using the proposed datasets. This paper is structured as follows: section 2 outlines the related works on natural language processing and data in low resourced languages used for processing, section 3 presents the architecture of proposed hybrid classifier model, section 4 describes the datasets used for the study, while section 5 deals with the experiments contacted in context of the study.

2. Related works on NLP and data in low resourced languages for processing

As part of the evolution of various fields of artificial intelligence, including NLP, as well as the development of technology, new programming languages or increased computer resources, approaches to solving problems in this field have been changing. Previously, solutions were based exclusively on a set of rules, statistical methods [3], expert systems or directly in NLP on dictionary methods [4] were popular. Nowadays, on the other hand, we are much more likely to see dedicated machine learning [5] or deep machine learning solutions [6]. However, with the development of deep machine learning methods, the resulting models are becoming larger and more expensive. They are based on unimaginably large data sets, which is often not available in languages other than English.

The other approach to implementing the hybrid model was proposed in [7]. This approach combines three different deep learning architectures: recurrent neural networks (RNN), long short-term memory (LSTM) networks and gated recurrent unit (GRU) in various combinations containing only two selected architectures or all three. This means that the hybrid approach here is primarily the use of several architectures, rather than a combination of classical statistical or data processing methods, with ML and DL-based methods. The results showed that the most important factor determining the effectiveness of the model is the appropriate selection of architectures within the prepared hybrid solution.

The second topic addressed in this paper is the problem of lack of data in low resourced languages, such as Polish. Among solutions dedicated to single languages other than English, two types of solutions are encountered in terms of data type.

The first are based on data originally collected in that language. These are solutions like BERT multilingual [8], which uses Wikipedia data in all 104 languages or single language models like BERT [9] and model trained on texts originally in Polish [10], where a dataset from the organized PolEval competition was used. However, when it comes to more specific tasks that language modelling the data collection process itself is much more time consuming and expensive. In addition, manual labelling of data often involves numerous errors due to fatigue.

The second approach focuses on using data originally in English translated into a specific low resourced language. This requires the prior creation of a machine translation model between the two languages, plus a dictionary of the translation model as well as the dataset to be used in a specific task with a language other than English. This solution is very convenient for the researcher, but, as with a dataset collected manually, it is subject to the risk of error. Translation models have a certain efficiency, which is not 100%. Nevertheless, such a solution is much faster and less costly than manual data collection and labelling. Such a solution was used,

among others, in [11] [12].

When creating hybrid model, the process of machine translation has been taken into account not at the stage of preparing the dataset, but as part of the solution, which allows the use English models for the task of classification in Polish.

3. Architecture of proposed hybrid classifier model

In the context of the prepared hybrid classifier model, it was decided to use the architecture shown in figure 1. This structure can be further divided into three main components: (1) machine translation module, (2) classification module, (3) summarization module.

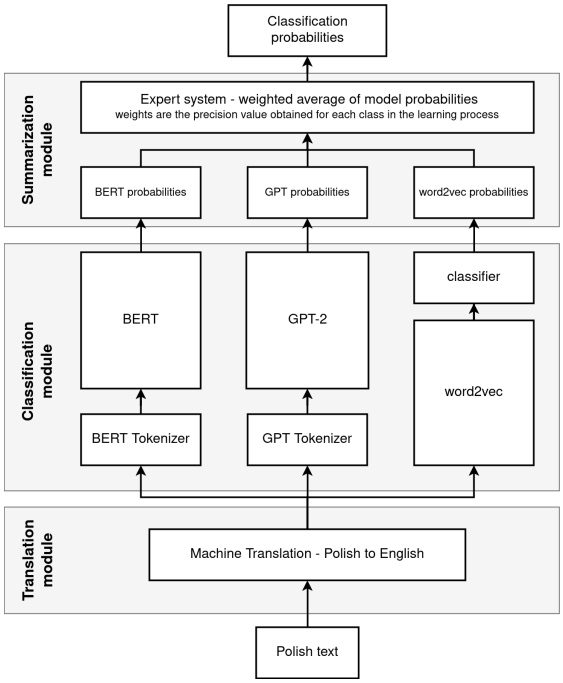

Fig. 1. Architecture of proposed hybrid classifier model.

3.1. Machine translation module

Within the first layer, which is machine translation, the transformation of original texts written in Polish into corresponding texts in English is performed. This is necessary in order to use models for English at a later stage. For this purpose, the mT5 model [13] that was trained to specifically perform translation between Polish and English was used. To build this, pre-trained model from HuggingFace was chosen.

T5 models and derivatives are classic models in the encoder-decoder architecture using transformers. This is important mainly because such a model unfortunately faces all the same problems as other models using transformers. Among which, in this case, the biggest one is the repetition of the same word or phrase. This can significantly affect the quality of the translation, especially when using tokens in the original sentences that this particular model does not know.

3.2. Classification module

The purpose of the second layer, the *classifier module*, is to determine the prediction of classes of given problems (emotion classification and sarcasm detection) by selected ML models. It was decided to use three language models, two of which are more advanced and, as in the machine

translation module, use the transformer architecture, while the third is a large simpler one. These models are as follows; (1) BERT [8], (2) GPT-2 [14], (3) word2vec [15]. Used BERT and GPT-2 models were pre-trained and available over the HuggingFace API, the word2vec model was trained from scratch.

It was decided to choose these particular architectures for several reasons. First, it was important to include models that use different (albeit small) types of text processing – transformers and word2vec. In addition, selected transformers models utilize different blocks: BERT uses encoders, while GPT-2 uses decoders. In addition, the task for which these models were created are different. Here, it is particularly noteworthy that the GPT-2 genitive model is used for classification instead of predicting the next words in a sentence.

Another important point was that the final hybrid model could be run on a local environment. At the time when large language models with milions of parameters are available it is worth considering more manageable models. Among other things, this justifies the use of the GPT-2 model, rather than successive versions of this model.

Finally, it was still necessary to balance the number of classification models. It can be assumed that the more models are used the better the result would be. However, as already mentioned, a reasonable balance must be preserved between the potential increase in efficiency and the size of the hybrid model, hence only three classifier models were used. Here it should also be mentioned, in the case of emotion classification, the word2vec model was trained to differentiate between only two emotions, which the BERT and GPT-2 models had the biggest problem with – joy and love.

3.3. Summarization module

The last layer is the *summarization model*. Its task is to decode, based on the obtained predictions from the three previously mentioned models (BERT, GPT-2 and word2vec), the final probabilities and the final assignment of classes within the given problem to texts.

This module consists of a simple expert system created based on individual models results. Within this system, a weighted average of the predictions from each model is calculated, where the weights are the precision values obtained on the test sets by a given model when fine-tuning. Importantly, for sarcasm classification, the scores of all three models are always taken into account, while for emotion, the score of the word2vec model is only taken into account when at least one of the BERT or GPT-2 models indicated one of the two classes that the word2vec model was taught to distinguish. This can be stated in the form of a sentence: if the BERT model indicates one of the emotions {A, B} or if the GPT-2 model indicates one of the emotions {A, B} count the weighted average from all models. It means that after receiving 3 probability vectors, one from each model, the values of specific classes are calculated based on the precision of these models. For example, for the precision values for a single class of 0.2, 0.7 and 0.5, respectively, and the corresponding received probabilities of 0.8, 0.6 and 0.1, the final probability value calculated would be $\frac{0.2 \cdot 0.8 + 0.7 \cdot 0.6 + 0.5 \cdot 0.1}{0.2 + 0.7 + 0.1} = 0.63$. The result of such a summary for every class is the final result of the classification of the whole hybrid model.

4. Datasets

To conduct the study it was necessary to select four datasets:

1. The first dataset serves as a training dataset for the emotion classification task – the selected models need to be further trained for a specific emotion classification task, since their base versions cannot handle such a task, it contains texts in English.
2. The second dataset serves as a training set for the task of sarcasm classification – similar to the first dataset, it is needed because the base models can not perform well on the task of detecting sarcasm in texts, it is a set of texts in English.

3. The third dataset was used as test set for emotion classification – the dataset on which the whole hybrid model was tested, contains texts in Polish.
4. The last dataset is a test set for the classification of sarcasm – the final set on which the hybrid model was tested, contains texts in Polish.

4.1. Dataset for emotion classification fine-tuning

As a dataset for fine-tuning in order to classify emotions in English, the emotion dataset created within the [16] was used. The dataset contains more than 400,000 texts collected using the Tweeter API and the corresponding emotion from among six basic ones: joy, love, sadness, fear, surprise and anger. A few sample sentences along with their assigned classes are provided in the table 1.

Table 1. Example texts form emotion dataset with assign classes.

id	text	label
1	"ive been feeling a little burdened lately wasnt sure why that was"	sadness
2	"i have the feeling she was amused and delighted"	joy
3	"i keep feeling pleasantly surprised at his supportiveness and also his ease in new situations"	surprise

For the process of fine-tuning the model, such a large dataset is not needed. Therefore, it was decided to limit it to 40,000 texts. Importantly, the original dataset is not balanced (it contains different numbers of texts assigned to specific emotions), which could bias the results of the experimenters. Therefore, as part of the research conducted, it was decided that the dataset used should contain equalized numbers of texts for all classes. The texts were selected randomly, and also different models were trained on separately randomly selected subsets of the original dataset. The final distribution of classes is presented in the figure 2.

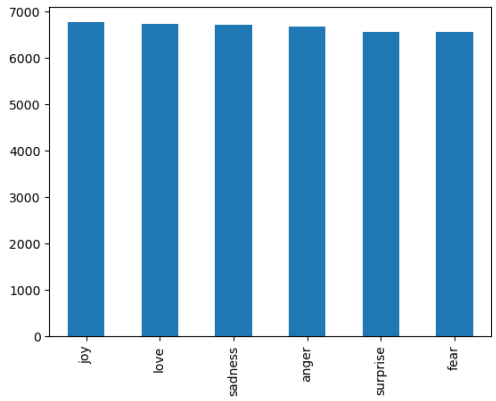

Fig. 2. Distribution of classes for fine-tuning models for emotion classification.

The dataset was divided into training, validation and test datasets, where 80% is the training set and both validation and test are 10% of the set. The data was randomly splitted between the three sets.

4.2. Dataset for sarcasm detection fine-tuning

News Headlines Dataset For Sarcasm Detection [17] [18] was used for fine-tuning task for sarcasm classification. The dataset contains 28,619 texts for two classes – containing sarcasm and without sarcasm in it. The authors claim that usage of headers make he texts self-contained.

It means that they don't need additional context to be assumed as sarcastic. The dataset is in a form of JSON file where each record consists of three attributes: (1) is_sarcastic – 1 if the record is sarcastic otherwise 0, (2) headline – the headline of the news article and (3) article_link – link to the original news article.

Table 2. Example texts from sarcasm dataset with assign classes.

id	label	headline	article_link
1	0	amanda peet told her daughter sex is 'a special hug'	https://www.huffingtonpost.com/entry /amanda-peet-told-her-daughter-sex-is-a-special-hug_us_59131898e4b0a58297e12f68
2	1	ford develops new suv that runs purely on gasoline	https://www.theonion.com/ford-develops-new-suv-that-runs-purely-on-gasoline-1819575454
3	1	cocksucker beats up motherfucker	https://local.theonion.com/cocksucker-beats-up-motherfucker-1819567714

Unlike the fine-tuning dataset for emotion classification, in this dataset the texts are evenly distributed between classes and there is no need to interfere with the number of examples used. The only change that took place in the use of this dataset was the removal of the article link. In the case of this problem, it is not needed, nor does it add any additional information from the point of view of model training. Classes distribution for this dataset are shown in figure 3.

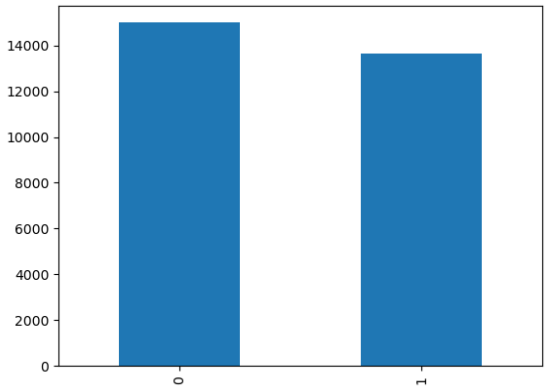

Fig. 3. Distribution of classes for fine-tuning models for sarcasm classification.

4.3. Polish dataset for emotion classification

For the final verification of the hybrid model, a dataset manually prepared and tagged by Lodz University of Technology students was used. The dataset contains 1,538 texts that are comments of the most popular videos on the YouTube platform during the 3-week period from the beginning of 2023, when the data was collected. Each text is assigned one of six categories corresponding to six basic emotions. As with the English-language dataset, these are: joy, love, sadness, fear, anger and surprise.

The quality of data in the dataset is not the best in. In consequence two main flaws of the data are: (1) quantity, (2) emotion labels inequality. The main problem is the inequality of emotions assigned to comments from the dataset. As visible in figure 4 love and joy emotions are dominating the dataset. These emotions alone form over 75% of the whole dataset. At the same time, emotions such as fear have only 15 samples in the entire dataset.

Other thing that is really important is te fact that in this dataset texts are written in colloquial language and significant portion of the texts contain emojis. It is also the case that these texts

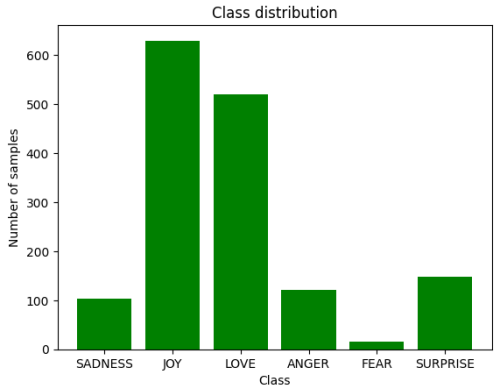

Fig. 4. Distribution of classes for final emotion classification in Polish dataset.

contain typos and often do not have Polish characters. Few example texts with its translations and assign label are show in the table 3.

Table 3. Example texts form Polish emotion dataset with translations and labels.

id	text	label
1	znam juz dawno na pamiec , spiewam pod prysznicem <3 (ang. I have known it by heart for a long time , I sing it in the shower <3)	love
2	o koniach to sie nie spodziewalem (ang. about the horses I didn't expect it)	surprise

4.4. Polish dataset for sarcasm detection

Two datasets were used to create a dedicated dataset in Polish containing sarcasm. The first 1,000 texts was from [19] and contained only sentences with sarcasm. The dataset was created to analyse the possibility of automating the recognition of hate speech in Polish. It was collected from the Polish forums and represents various types and degrees of offensive language, expressed towards minorities. This dataset contains a lot of different fields but for this study only: irony_sarcasm and text were used.

The second 1,000 texts was from PolEval 2019 Task6 cyberbullying dataset.As the dataset contains harmful and non-harmful texts for this research it was decided to select only non-harmful texts. This dataset contains only text and label filed so both of them were used for creation of final dataset for Polish sarcasm classification. The texts from both datasets were selected randomly.

5. Experiments

The following experiments were conducted in the context of the study:

1. Fine-tuning and checking the results of emotion classification in English for BERT and GPT-2 models.
2. After evaluation of the results of emotion classification in English of GPT-2 and BERT models, two emotions were selected with which the models perform the worst in differentiating, and the word2vec model was trained to distinguish between these two emotions, the results were examined.
3. Fine-tuning and checking the results of sarcasm classification in English for the BERT and GPT-2 models, trained and verified the results for the word2vec model.

4. Test the final hybrid method for emotion and sarcasm classification using English texts (without translation).
5. Test the final hybrid method for the classification of emotions and sarcasm using texts in Polish (with translation).

Using the API proposed by HuggingFace, it was decided to use models from the following checkpoints: (1) 'bert-base-uncsed' for BERT model and (2) 'gpt2' for GPT-2 model.

6. Conclusions

Artificial intelligence and natural language processing are rapidly developing fields. Despite the importance of newer and better NLP models, text processing in low-resource languages is still problematic. Applications that recognize emotion or sarcasm in texts are those that could help people with intellectual disabilities by providing proper conversation explanations. Unfortunetly, they face many challenges like a lack of data in a specific language or a shortage of solutions dedicated to such problems. Response to this need could be the hybrid model created as part of this research. Unfortunately, the results obtained for texts originally in Polish are not satisfactory. Although there is an improvement in the results of the hybrid model in relation to the single models for English-language classifications, the results do not allow to unequivocally recognize the prepared architecture as effective. Therefore, it is worth considering what could have contributed to such a result, what should be changed and what future work can be undertaken to achieve better performance. The final hybrid model have problems with Polish text classification for both emotion and sarcasm recognition tasks. The main reason for that could be the data itself. As described, the datasets in Polish for sarcasm were created as a cluster of two other datasets. These despite being labelled as containing sarcasm, do not necessarily contain it. In addition, the dataset for emotions was an unbalanced, where joy and love classes contained the most texts. Additionally, after manual reassignmenf of mis-assigned emotions to texts, it turned out that some texts contained more than one emotion, and were described by the one less represented in the dataset.

Another big problem with the data may be the difference between the formality of the extras in English and in Polish. The datasets used to fine-tune models for English contain formal or semi-formal texts. At the same time, the used Polish texts are from the YouTube portal – they are non-formal texts. It has already been observed in studies that models learned on formal data, with a concrete structure like the Wikipedia data, have problems with texts written in colloquial language, like the tweets. This difference between the formality of the texts may have further compromised the final classification results. The last concern is the quality of the translation. Despite the fact that the mT5 model initially seemed to do very well in translating sentences into English, after detailed analysis, it turned out that there are cases in which it had big problems. When an emoticon other than a smiley face was used in the text, the model was completely unaware of how to handle such a translation. This resulted in a large part of the sentences being translated only halfway. Since the translation model is the basis of prepared solution, this had a huge impact on the final results.

In the future, there is a plan to prepare similar classification model with models learned on Polish datasets. This type of action would specifically address the problem of capturing the dynamics of the language and culture of Polish native speakers and eliminate the problem of the quality of machine translation.

References

1. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications 82, 3713–3744 (2023)

2. Hrkut, P., Toth, S., Duracik, M., Mesko, M., Krsak, E., Mikusova, M.: Data Collection for Natural Language Processing Systems. In: Intelligent Information and Database Systems. ACIIDS 2020. CCIS Vol. 1178. Springer (2020)
3. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. In: Information Retrieval, Vol. 1, pp. 69–90 (1999)
4. Tang, H., Yan, D., Tian, Y.: Semantic dictionary based method for short text classification. In: Journal of China Universities of Posts and Telecommunications, Vol. 20, Supplement 1, pp. 15–19 (2013).
5. Agarwal, J., Christa, S., Pai, A., Kumar, M. A., Prasad G.: Machine Learning Application for News Text Classification. In: Proceedings of 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 463–466 (2023)
6. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep Learning Based Text Classification. ACM Computing Surveys (CSUR) 54(3), 1–40 (2020)
7. Sunagar, P., Kanavalli, A.: A Hybrid RNN based Deep Learning Approach for Text Classification Internet. Journal of Advanced Computer Science and Applications13(6) (2022)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805 (2019)
9. Kleczek, D.: PolBERT: attacking Polish nlp tasks with transformers. In: Ogrodniczuk, M., Kobylinski, L. (eds), Proceedings of PolEval 2020 Workshop. Institute of Computer Science, Polish Academy of Sciences (2020)
10. Wawer, A., Sobiczewska, J. Predicting Sentiment of Polish Language Short Texts. In: Recent Advances in Natural Language Processing (RANLP 2019), Bulgaria. INCOMA Ltd, pp. 1321–1327 (2019)
11. Sazzed, S., Jayarathna, S.: A Sentiment Classification in Bengali and Machine Translated English Corpus. In: Proceedings of IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), CA, USA, pp. 107–114 (2019)
12. Tebbifakhr, A., Bentivogli, L., Negri, M., Turchi, M.: Machine Translation for Machines: the Sentiment Classification Use Case. ArXiv, abs/1910.00478 (2019)
13. Linting, X., Noah, C., Roberts, A., Kale, M., Rami, A.-R., Siddhant, A., Barua, A., Colin, R.: mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–498 (2021)
14. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. (2018)
15. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of 1st International Conference on Learning Representations, ICLR 2013, USA (2013)
16. Saravia, E., Liu, H.C., Huang, Y.H., Wu, J., Chen, Y.S.: CARER: Contextualized Affect Representations for Emotion Recognition. In: Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3687–3697 (2018)
17. Misra, R., Prahal, A.: Sarcasm Detection using News Headlines Dataset. In: AI Open, Vol. 4, pp. 13–18 (2023)
18. Misra, R., Grover, J.: Sculpting Data for ML: The first act of Machine Learning. Kindle Edition, ISBN 978-0-578-83125-1 (2021)
19. Troszynski, M., Wawer, A.: Will the computer recognize a hater? The use of machine learning (ML) in qualitative data analysis (in Polish). In: Przeglad Socjologii Jakosciowej, Vol. 13(2), pp. 62–80 (2017)