72
Research paper
Doi 10.7251/stp202401059L
ISSN 2566-4484

Ana Lojić, International Burch University, ana.lojic@stu.ibu.edu.ba
Zerina Mašetić, International Burch University, zerina.masetic@ibu.edu.ba
Samed Jukić, International Burch University, samed.jukic@ibu.edu.ba

USING NATURAL LANGUAGE PROCESSING (NLP) FOR CATEGORIZING PAPER TITLES FROM GOOGLE FORMS

Abstract

Modern data collection, storage, and processing rely on diverse techniques to handle various types of information, ranging from structured tables to free-form text. This paper explores the captivating application of Natural Language Processing (NLP) for categorizing titles from Google Forms or any other textual data. The process of training an NLP model will be demonstrated through a specific example. Just as we learn from our past experiences, NLP models need to be fed with relevant data and labels. This ensures accurate and efficient processing even when new titles are introduced. We will conclude with a fascinating demonstration of how NLP algorithms analyze the structure and meaning of titles. By identifying keywords and understanding the context, they can automatically classify titles into relevant categories. This dramatically simplifies data organization and analysis, empowering us to extract valuable insights faster.

Keywords: Data Mining, Classification, Natural Language Processing, Multi-Layer Perceptron

КОРИШЋЕЊЕ ОБРАДЕ ПРИРОДНОГ ЈЕЗИКА (NLP) ЗА КАТЕГОРИЗАЦИЈУ НАСЛОВА РАДОВА СА ГУГЛ ФОРМСА

Сажетак

Модерне методе прикупљања, чувања и обраде података ослањају се на разнолике технике како би се носиле са различитим типовима информација, од структурираних табела до слободног текста. Овај рад се бави примјеном обраде природног језика (NLP) за категоризацију наслова из Google Формс или било којих других текстуалних података. Кроз конкретан примјер, илустроват ћемо процес обучавања NLP модела. Баш као што и ми учимо из претходних искустава, тако и NLP моделе треба хранити релевантним подацима и ознакама. Тако осигуравамо прецизну и ефикасну обраду чак и када се унесу нови наслови.Завршит ћемо приказом како NLP алгоритми анализирају структуру и значење наслова. Идентификујући кључне ријечи и разумијевајући контекст, они могу аутоматски сврстати наслове у релевантне категорије. Ово драстично поједностављује организацију и анализу података, омогућавајући нам брже добијање драгоцијених увида.

Кључне ријечи: рударање података, класификација, обрада природног језика, вишеслојни перцептрон

1. INTRODUCTION

When it comes to the categorization and sorting of textual data collected through various forms, the main criterion is often cited as selecting one of the specified category options. Analyzing and categorizing textual data is a rather time-consuming task, susceptible to errors.

It has been observed that, during the textual registration of participants in numerous events in Bosnia and Herzegovina, the benefits of algorithms from the field of artificial intelligence (AI) [1] and machine learning (ML) [2] are minimally utilized.

AI, ML, and NLP work together to tailor systems for text processing to specific tasks. This involves adjusting models, optimizing performance, and continuous learning to adapt models to changes in language and data.

Progress in NLP often stems from innovations within machine learning, including new model architectures, pattern recognition algorithms, and improved data learning techniques.

NLP algorithms can understand the meaning and context of words and phrases, enabling subtle categorization. This is crucial when titles may have subtle differences in meaning that need to be accurately captured.

In this paper, we will analyze methods of using natural language to achieve a high degree of accuracy in the textual processing of data required for classifying areas of innovation based on the textual title submitted through Google Forms at the International Innovation Fair for Youth "Inost mladih" [3].

2. LITERATURE REVIEW AND RESULTS OBTAINED FROM SIMILAR RESEARCH STUDIES

Text analysis is a crucial component in numerous fields, including document classification and information mining. Automatic text classification offers an efficient way to organize and understand large amounts of data. Several authors have tackled similar topics and demonstrated the significance of applying NLP in numerous studies.

In a study conducted by Pang et al. (2002), the authors investigated the use of various feature extraction methods, such as tokenization, stemming, and TF-IDF (Term Frequency-Inverse Document Frequency), for sentiment classification of movie reviews [4].

In their study, Liu et al. (2010) provided an overview of feature extraction methods for text analysis, including n-grams, which are sequences of n consecutive words. These methods have proven effective in capturing local dependencies in [5].

Wang et al. (2023) compared the performance of various deep learning algorithms, such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), for classifying short texts on social media. They found that LSTM networks achieved the best results due to their ability to capture sequential information in [6].

The study by Kim (2014) demonstrates promising results in the field of sentiment analysis of Twitter data. The combination of TF-IDF weighting and SVM classifier achieved an accuracy of 89%[7]. Accuracy [8] is defined as the ratio of correctly classified items to the total number of classified items. In this case, out of 100 tweets classified by Kim's model, 89 were classified correctly (positive, negative, or neutral).

Although an accuracy of 89% is impressive, it does not provide a complete picture of the model's performance. It is important to consider other performance metrics, such as recall and F1-score. The F1-score [9] combines precision and recall into a single metric. Recall [10] measures the proportion of relevant items that were correctly classified. It is possible that Kim's model missed some relevant tweets (e.g., classifying negative tweets as neutral).

For a complete picture, additional information is needed such as:

1. What was the ratio of positive, negative, and neutral tweets in Kim's dataset?
2. What values were achieved for recall and F1-score?
3. How does Kim's model compare to other models for sentiment analysis of Twitter data?

While the 89% accuracy suggests that Kim's model was successful in accurately classifying the majority of Twitter data, for a comprehensive evaluation of its performance, additional information on other metrics and comparison with alternative models is necessary.

Xu et al. (2018) achieved promising results using an LSTM (Long Short-Term Memory) network for automatic classification of news categories. Their model reached an F1-score of 92%, indicating high precision and recall during classification.

The F1-score is the harmonic mean between precision and recall. A high F1-score means that the model is effective in both accurately identifying relevant categories (precision) and finding the majority of relevant articles (recall) [11].
LSTM networks are a type of recurrent neural networks that excel in learning sequential dependencies in data. In text, the order of words and their context can be crucial for determining the category of news. LSTM networks can learn these long-term dependencies and utilize them for improved classification.
For a better understanding of the study, it is necessary to further investigate:

- How many different news categories did the model classify?
- What was the format of news articles in the dataset (text, title, abstract, etc.)?
- How does the performance of the LSTM network compare to other models for news category classification?

The study by Xu et al. (2018) demonstrates that LSTM networks can be an effective tool for automatic news category classification. A high F1-score of 92% indicates the model's robustness. However, for a comprehensive understanding of the research, it would be useful to know more about the specifics of the study and comparisons with other models.

3. MATERIAL METHODS AND RESEARCH ORGANIZATION

During the data collection process, the Data Meaning process [12] was utilized. As a foundational basis for analysis, we used participant data from the International Exhibition of Ideas, Innovations, and Creativity "Inost mladi" spanning from 2015 to 2022. The dataset includes the following information:

1. Category Name
   - Automation
   - Informatics
   - Free Topic (business models, architecture, medicine, design...)
2. Paper Title

When using the MLP (Multiple Layer Perceptron) classifier for Natural Language Processing (NLP), it is typically employed as part of a broader system that involves multiple data processing steps.
When implementing the MLP classifier for NLP, libraries such as TensorFlow, PyTorch, or scikit-learn can be helpful as they provide ready-made implementations of these algorithms and functions for working with data. It is important to emphasize that the efficiency of the model is often linked to the quality of training data, the selection of relevant features, and the proper tuning of hyperparameters.
The MLP (Multiple Layer Perceptron) [13] is a classifier whose operation is based on the functioning of a multilayer perceptron. The multilayer perceptron is a model of artificial neural networks that maps input datasets to a set of corresponding outputs. Artificial Neural Networks (ANN) [14], in particular, attempt to mimic the functioning of the human brain by representing a set of interconnected neurons. In this study, we analyzed a total of 7,000 paper titles. This large sample size allows us to draw relevant conclusions with a high degree of confidence.
The independent variables include paper categories (such as Automation, Informatics, and Free Topic), while the dependent variables focus on the analysis of textual content within the papers, such as themes, writing styles, and structure.
For data processing and analysis, we utilized the Python programming language, which provides a rich ecosystem of libraries and tools for text processing, machine learning, and data analytics. Specifically, we employed popular libraries like TensorFlow, PyTorch, and scikit-learn for implementing machine learning models.
The data is organized into three main categories: training, validation, and testing. Training data was used for model training, validation data for parameter tuning and performance evaluation during training, while testing data was used for final model performance evaluation after training. This organization allows us to objectively assess the model's performance on new, unseen data.

4. RESEARCH RESULTS

The following sections will present the steps through which we arrived at the results, using the MLP classifier.

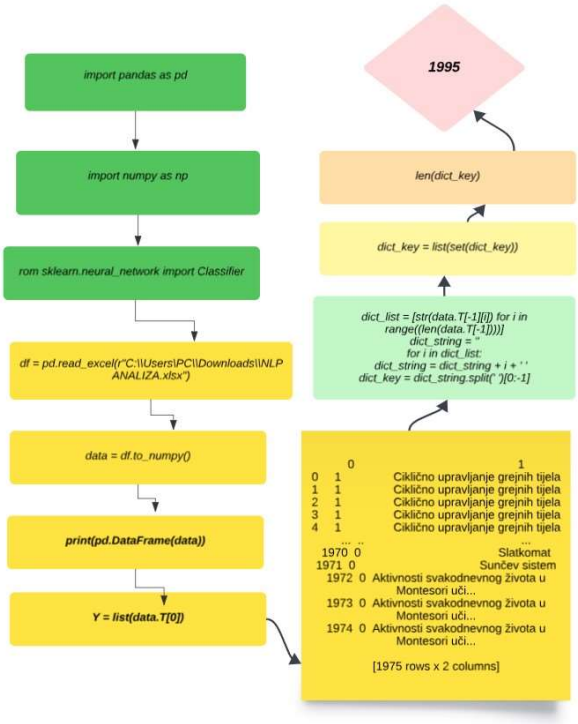
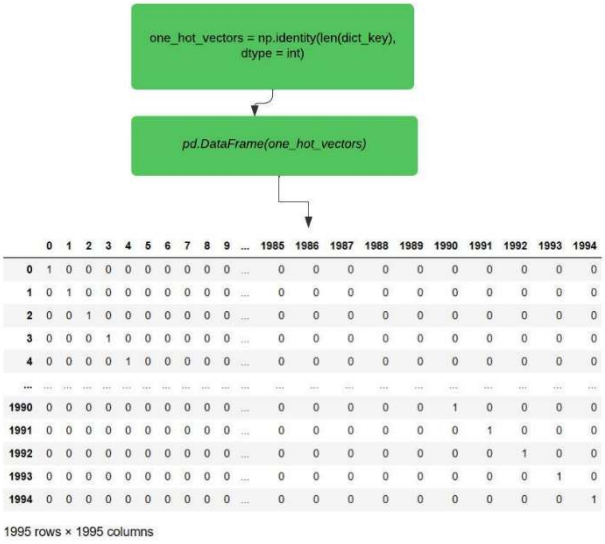


Figure 1. Flowchart 1: Procedure for Creating Input Parameters Using an Excel Database and Text Preparation

In the continuation of the paper, we will illustrate the process of text vectorization.
We generate a unit matrix using the NumPy package, whose dimensions are equal to the size of the set of unique words in the database.
The algorithm we apply for vectorization is 'one-hot encoding', meaning that each unique word is associated with a unit vector (unit, in this context, means that it contains only one unit, and everything else is zeros, analogous to vectors i, j, k in a 3D Gaussian coordinate system).
Afterwards, we create a dictionary containing all unique words as keys and the corresponding vectors as their values. All of this will be presented in Step 2.



	0	1	2	3	4	5	6	7	8	9	...	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1990	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
1991	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
1992	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
1993	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
1994	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

1995 rows × 1995 columns

Figure 2. Flowchart 2: Text Vectorization Process

In Step 3, we will use the 'dictionary' as a reference to create a list X, which is a list of aggregated unit vectors corresponding to each title in the Excel database.

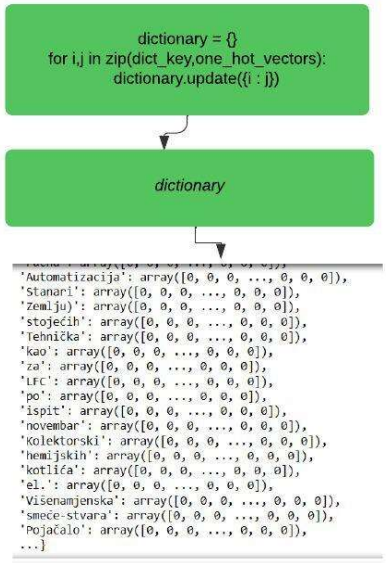


Figure 3. Flowchart 3: Process of Creating a Vector List

In Step 4, efforts will be directed towards checking the model's performance on the validation set during training to prevent overfitting. It can be observed at the end that the number of elements in the X list is equal to the number of titles in the Excel database, which is 1975 entries.

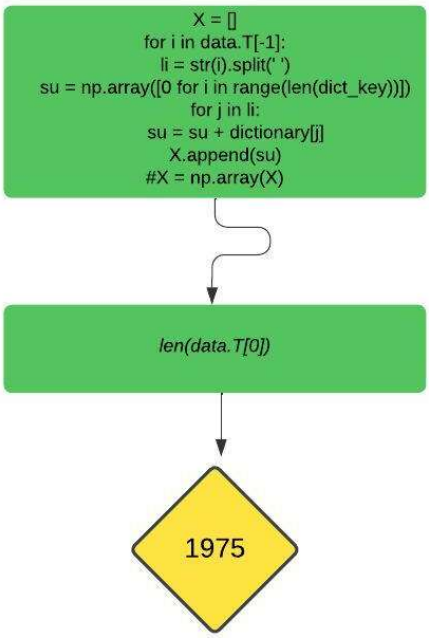


Figure 4. Flowchart 4: Procedure for Testing the Number of Entries

In Step 5, using the MLP classifier, we will calculate the percentage of classification accuracy based on the data. By using the MLPClassifier object from the SciPy package, we create a Multilayer Perceptron Neural Network with the 'Adam' optimizer and logistic (sigmoidal) activation function. The network has 15 hidden layers, trained for 1000 generations, and the algorithm is allowed to stop training before the 1000th iteration if it reaches an appropriate tolerance level.

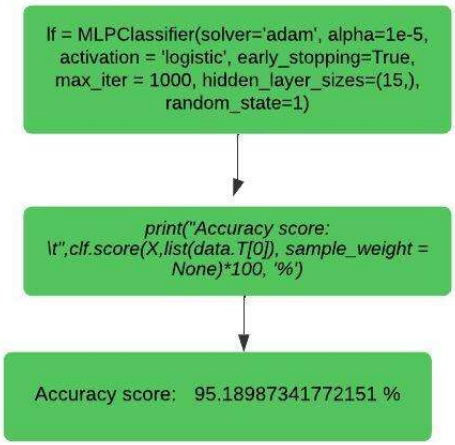


Figure 5. Flowchart 5: Accuracy Calculation

In the following sections, on Figure 1, we will visually present the Loss and Validation Scores Chart during Training Iterations.

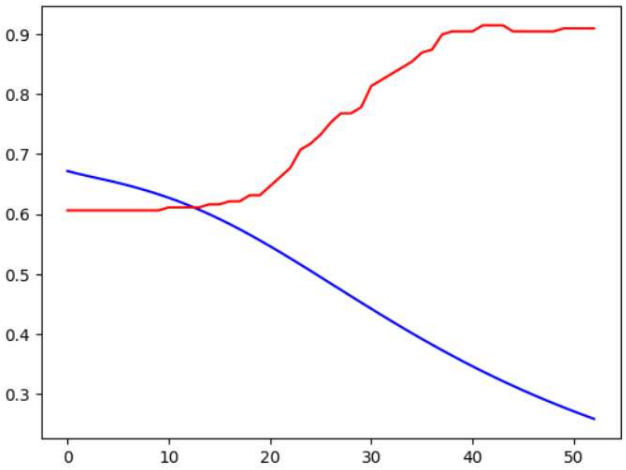


Figure 6. Loss and Validation Scores Chart during Training Iterations

This Fig. 6 illustrates the changes in model loss and validation scores throughout the neural network training process. The x-axis represents the training iterations, while the y-axis displays the model loss (blue line) and validation scores (red line).

Model loss indicates the amount of error the model has during predicting the target variable on the training dataset. A lower loss value signifies a better model prediction. Validation scores provide insight into the model's performance on an independent dataset not used for training. This line represents the accuracy or another performance metric of the model on a dataset it hasn't seen during training.

Analyzing this chart allows us to track how the model's loss decreases over iterations, indicating that the model is improving its prediction. Additionally, we monitor how validation scores increase or stabilize, suggesting that the model generalizes well to new, unseen data.

This chart helps us understand the model learning process and evaluate its ability to generalize to new, unseen data. The analysis of the graph was performed using data visualization tools through Python, utilizing the Matplotlib library. Data on model loss and validation results were collected during the neural network training process, then organized and prepared for visualization. The graph illustrates changes in model loss and validation results throughout the training iterations, enabling tracking of model prediction improvement and its ability to generalize to new data.

5. DISCUSSION

The demonstrated accuracy rate of 95.19% in title classification underscores the proficiency of the natural language processing (NLP) methodologies employed in this study. By leveraging NLP techniques, we were able to discern patterns and extract meaningful insights from textual data, thereby facilitating effective title classification.

It's crucial to acknowledge the limitations posed by the relatively small dataset utilized in this study. While our results are promising, they should be interpreted with caution, considering the constraints imposed by the dataset size. Moreover, the applicability of our findings may vary across different datasets with distinct characteristics and complexities.

In light of these considerations, future research endeavors will focus on enhancing the robustness and generalizability of our approach. This includes evaluating the performance of the model using additional performance metrics such as precision, recall, F1-score, and others. Such comprehensive evaluation metrics are particularly pertinent in scenarios involving multiple classes or imbalanced class distributions.

Furthermore, the significance of our findings extends beyond the realm of title classification. The methodologies and insights derived from this study hold implications for various NLP applications, including sentiment analysis, document classification, and information retrieval. By elucidating effective strategies for processing textual data, our research contributes to the advancement of NLP methodologies and their practical applications across diverse domains.

6. CONCLUSION

This experiment focused on analyzing the significance of applying NLP for categorizing submitted papers for the International Innovation Fair Inost Mladih.

The mentioned natural language processing method enabled the automation of the title categorization process, contributing to time savings and resources that would otherwise be required for manual review and labeling of titles.

Furthermore, a quick and secure analysis of a large amount of data was generated, resulting in faster data retrieval at a specific moment.

Applying NLP for categorizing paper titles from Google Forms aids in automating this task, making the organization of information collected through surveys or forms more efficient.

LITERATURE:

[1] H. Sheikh, C. Prins, S. Schrijvers, "Artificial Intelligence: Definition and Background", In: *Mission AI. Research for Policy,* 15-41, 2023.

[2] J.G. Carbonell. et al. An Overview of Machine Learning, 1983.

[3] http://www.savezinovatorars.org/index.php?option=com_content&view=article&id=383%3Aodrzana-25-izlozba-ideja-inovacija-i-stvaralastva-inost-mladih-2023&catid=52%3Aaktuelnosti&Itemid=80&lang=sr_lat, posjećeno dana 01.07.2023.

[4] B. Pang, L. Lee, "A sentimental education: Sentiment analysis using supervised and unsupervised learning", In Proceedings of the ACL 2004 on Empirical Methods in Natural Language Processing (EMNLP) , pp. 271-278, 2004.

[5] D. Tang, F. Wei, N. Yang, M, Zhou, T. Liu, B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification", In Proceedings of the 52nd *Annual Meeting of the Association for Computational Linguistics* (ACL) (Volume 1: Long Papers), pp. 1139-1148, 2014.

[6] D.L. Wang, J. Gan, J. Q. Mao, F. Chen, L. Yu, „Forecasting power demand in China with a CNN-LSTM model including multimodal information," *Energy*, 263, 2023.

[7] Y. Kim, Yi-I Chiu, K. Hanaki, D. Hegde, S. Petrov, "Temporal Analysis of Language through Neural Language Models", In Proceedings of the ACL 2014 *Workshop on Language Technologies and Computational Social Science, Baltimore,* MD, USA. Association for Computational Linguistics, pp. 61–65, 2014.

[8] https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall, Visited on October 9, 2023.

[9] https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score, Visited on October 7, 2023

[10] https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall, Visited on October 9, 2023.

[11] G. Xu, Y. Meng, X. Qiu, Z., Yu, X. Wu, "Sentiment Analysis of Comment Texts Based on BiLSTM", *IEEE*, 7, 51522–51532, 2019.

[12] A. Muhamed, "How LSTM (Long Short-Term Memory) cells learn to categorize texts", *Becoming Human: Artificial Intelligence Magazine,* 2020.

[13] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, L. Trigg, "Weka-a machine learning workbench for data mining", *Data mining and knowledgediscovery handbook*, Springer, pp. 1269-1277, 2009.

[14] H. A. Bourlard et al., *Connectionist Speech Recognition*, Springer, 1994.