Benjamin Szilágyi – Rozália Lakner

NLP-enhanced classification of remote employment opportunities

Abstract

In the rapidly evolving digital era, the labour market is undergoing significant transformations, particularly in online job searching. This progress, however, presents the challenge of efficiently filtering a large amount of information, making artificial intelligence (AI) tools, especially those using Natural Language Processing (NLP), increasingly vital.

In this study, which is aimed at identifying remote work opportunities, an automated text classification system that merges linguistics with AI has been developed. This system goes beyond merely sorting job offers; it focuses on understanding and interpreting the language of job advertisements for effective computer comprehension, while an attempt is made to preserve the semantics of the text.

The research commences with data preparation, where texts from online job portals are subjected to technical filtering and normalization, along with linguistic analysis for key feature extraction. Subsequently, the exploration of applying rule-based classification and supervised machine learning algorithms to this textual data is undertaken, which serves to demonstrate AI's proficiency in linguistic categorization.

The accuracy of these AI-driven methods in job advertisement classification is underscored by a comparative analysis of model performance, highlighting the synergy between linguistic principles and AI. The study concludes that the efficiency of online job searching for remote positions is significantly improved by this AI and NLP-enhanced system, illustrating the growing intersection of linguistics and AI in data analysis.

Keywords: NLP-based text classification, machine learning models, rule-based learning methods, text processing

Introduction

In the contemporary digital era, the job market has experienced a paradigm shift, particularly in the realm of remote work. This shift is mainly due to digitalization and the COVID-19 pandemic, especially in fields requiring high online connectivity, like IT. Developers and other IT professionals increasingly prefer remote work for its broader opportunities. This trend is mirrored on various online job advertisement platforms, where remote positions are prominently featured.

Given the vast number of listings, artificial intelligence, specifically Natural Language Processing, becomes crucial for efficiently classifying and finding suitable opportunities. This shift has necessitated the development of sophisticated tools to efficiently navigate and categorize the influx of online job opportunities. NLP presents a robust framework for addressing these needs by analysing the language of job advertisements to discern remote positions from traditional ones.

This paper investigates the application of NLP in enhancing the classification of remote employment opportunities. By integrating linguistic analysis with artificial intelligence, the research develops an automated classification system designed to shift through online job postings, selecting those that offer remote work.

The significance of NLP in this context is manifold, drawing upon its proven efficacy in text analysis and machine learning to adapt to the nuances of job advertisements. The methodologies employed, including data extraction, preprocessing, and vectorization, are fundamental to transforming raw textual information into a structured form suitable for classification.

A case study will provide the empirical foundation for the research, applying the discussed NLP techniques to actual job advertisements and evaluating their effectiveness. The study aims to demonstrate the practical usefulness of NLP-driven systems in enhancing the searchability and classification of remote jobs, reflecting the evolving requirements of the job market.

Theoretical framework and Methodological approach

Essence of NLP

In the field of artificial intelligence, NLP plays a crucial role as an advanced channel between human communication and machine interpretation. This technology is indispensable across a wide range of applications, significantly improves the ability of machines to understand human language. From basic translation services to complex sentiment and semantic analysis, NLP capabilities are essential for analysing large datasets, automating interactions, and aiding decision-making processes in a wide range of industries.

A compelling example of NLP's application is highlighted in related work focusing on customer satisfaction (Tusar & Islam 2021) – a critical factor for business success in today's competitive landscape. Many organizations, recognizing the importance of understanding and meeting customer needs, invest heavily in various strategies. However, traditional manual analysis often

falls short in accurately addressing the complex needs of customers, leading to decreased satisfaction levels, loss of loyalty, and increased marketing costs to counter these effects. An innovative solution to these challenges is found in the implementation of Sentiment Analysis, which utilizes the combined strengths of NLP and Machine Learning (ML) techniques. Sentiment Analysis is extensively applied to extract insights from the public opinion on topics, products, and services, utilizing publicly available online data. In this vein, this research explores the effectiveness of NLP techniques, such as Bag-of-Words and TF-IDF, alongside a variety of ML classification algorithms (Support Vector Machine, Logistic Regression, Multinomial Naive Bayes, Random Forest) to identify the most efficient approach for Sentiment Analysis on a large, imbalanced, and multi-class dataset. Remarkably, this study achieves an accuracy of 77% using Support Vector Machine and Logistic Regression with the Bag-of-Words technique. This related work serves as an illustrative example of how NLP can be leveraged to enhance business strategies by accurately analysing and responding to customer sentiments. The insights from such related work underscore the transformative potential of NLP in business intelligence and customer relationship management. By integrating the lessons learned from these applications of NLP, it becomes possible to not only improve customer satisfaction but also to innovate and streamline decision-making processes across various sectors.

Thus, current research builds upon this foundation, emphasizing the power of NLP to offer solutions that are not just reactive but also proactive in understanding and catering to the nuanced landscape of human language and interaction.

NLP in text classification

Text classification is a foundational aspect of NLP, utilizing both knowledge-based systems and machine learning techniques to effectively decode and categorize textual data. Knowledge-based systems are built on explicit, expert-defined rules, facilitating applications like spam detection through their clarity and simplicity. On the other hand, machine learning approaches – including Logistic Regression, Support Vector Machines (SVM), and Random Forests - are adept at pattern recognition, rendering them ideal for tasks requiring adaptability, such as sentiment analysis and opinion mining.

One example study in the field of NLP focuses on Opinion Mining, specifically targeting Bangla text and employing data from various social media platforms (Taher–Azharul Hasan–Afsana Akhter 2018). This research

leverages both linear and nonlinear SVM configurations, alongside the N-gram method, to classify documents more effectively. Unlike traditional approaches that consider single words as vectors, this work utilizes N-grams, sequences of 'n' words, as a unified vector, leading to better classification results for different 'n' values. This study underscores the utility of combining machine learning techniques with NLP methodologies, such as N-grams, to enhance the processing and understanding of large-scale textual data.

A closely related work (Hansen et al. 2023) examines the shift towards remote work catalysed by the pandemic, analysing over 250 million job vacancy postings across five English-speaking countries. This research employs a sophisticated language-processing framework, refined through 30,000 human classifications, achieving 99% accuracy in identifying postings that offer hybrid or fully remote positions. This notable precision in classifying job postings based on remote work opportunities showcases the practical application of NLP in monitoring significant societal trends.

All these related works illustrate the broad applicability and impact of NLP and machine learning in understanding and interpreting the complexities of human language and social changes. They highlight the progression of NLP systems from basic text categorization to sophisticated analyses capable of uncovering nuanced insights within vast datasets.

NLP methodologies for data analysis

A methodical approach to data analysis is the basis for successful NLP applications. It starts with data extraction and cleansing, removing extraneous elements that could cloud analysis. Subsequent preprocessing stages, including tokenization and lemmatization, further refine the text to its analytical essence. Annotation assigns critical markers to data, paving the way for machine learning algorithms to learn efficiently. Finally, vectorization translates the curated text into a numerical form, laying the foundation for complex computations and classifications.

Synthesizing theory and practice

The theoretical concepts of NLP play an importance role in the development of systems that allow effective classification of texts. These principles form the backbone of practical applications, such as organizing unstructured job advertisements and facilitating the rapid identification of remote work opportunities. The following case study demonstrates how these NLP

methodologies are employed to dissect, understand, and categorize job advertisements. This application of NLP not only illustrates the practical utility of theoretical knowledge, but also demonstrates the transformative potential of NLP in reshaping the job market in the digital age.

Case study: NLP-driven classification of remote jobs

NLP technique implementation

This section outlines the implementation of NLP techniques to identify remote employment opportunities within online job advertisements. The process began with data extraction, where a total of 324 job advertisements for 'full-stack web developer' positions were collected from three major online portals using Octoparse, a user-friendly web scraping tool. This facilitated the rapid acquisition of high-quality, relevant data, which was then prepared and cleaned for analysis.

Python (Van Rossum–Drake 2009) and its related data science modules and libraries served as the main platform of the research. Pandas framework (McKinney 2010) was used for many of the data manipulation tasks, which provided the bases to securely and efficiently transform and send data down the pipeline. The GPT-3.5 Turbo model (OpenAI developer platform, 2023) was employed to correct text inconsistencies, such as merged words, within the job descriptions. This step involved using the OpenAI API for text correction, significantly enhancing the quality of the dataset for subsequent processing stages. Preprocessing included normalization activities like lowercase conversion, punctuation removal, and tokenization, followed by more advanced procedures such as stop word removal and lemmatization (Jurafsky–Martin 2023). These were carried out utilizing the NLTK (Bird, Loper–Klein 2009) Python library which is a standard tool for data scientists to process textual data.

Annotation of the dataset was carefully carried out to tag listings as 'Remote' or 'Not Remote,' using criteria developed in collaboration with project leadership. This binary classification was crucial for the study's focus on remote job identification. Additionally, vectorization of the dataset was performed using both TF-IDF and OpenAI's Ada-002 model embeddings to prepare the data for machine learning analysis, leveraging these methods' ability to capture semantic nuances and importance within the text.

The text processing steps are shown in Figure 1–4.

description
Power Platform Developer Manchester, Ruddington or London – Hybrid working model (2 days office based, 3 days remote) Very competitive day rate Fantastic opportunity to secure and initial 6 months day rate contract inside IR35 as a Power Platform Developer with Smart DCC. The Power Platform Developer role will be working with the Enterprise and End to End Architecture teams, this is a hands-on role working within the DCC's EIT function. The Power Platform Developer is the DCC subject matter expert for automating many of the DCC's manual processes. Power Platform Developer is fundamental to the success of reducing aoperation complexity and introducing a standards-based approach to using Microsoft 365 developement platforms. What will you be doing? • Process

Figure 1. Raw extracted text from a job advertisement.

cleaned_descriptipon
Power Platform Developer Manchester, Ruddington or London – Hybrid working model (2 days office based, 3 days remote) Very competitive day rate Fantastic opportunity to secure and initial 6 months day rate contract inside IR35 as a Power Platform Developer with Smart DCC. The Power Platform Developer role will be working with the Enterprise and End to End Architecture teams, this is a hands-on role working within the DCC's EIT function. The Power Platform Developer in the DCC subjec matter expert for automating many of the DCC's manual processes. Power Platform Developer is fundamental to the success of reducing operational complexity and introducing a standards-based approach to using Microsoft 365 development platforms.

Figure 2. Text after cleaning using GPT model.

power platform developer manchester ruddington london – hybrid working model 2 day office based 3 day remote competitive day rate fantastic opportunity secure initial 6 month day rate contract inside ir35 power platform developer smart dcc power platform developer role working enterprise end end architecture team handson role working within dcc's eit function power platform developer dcc subject matter expert automating many dcc's manual process power platform developer fundamental success reducing operational complexity introducing standardsbases approach using microsoft 365 development platform • process automation reporting ensure business function able replace historic process auditable automation utilising microsoft power platform

Figure 3. Text after tokenization.

[-0.021912286058068275, -0.0009061856544576585, 0.007242659106850624,
-0.03044510819017887, -0.028424536809325218, 0.024902187287807465,
-0.01160463783890009, -0.02001458778977394, -0.017024686560034752,
-0.03877314180135727, 0.018622029572725296, 0.005689685698598623,
-0.003751028561964631, 0.00536885190901613, -0.02300448715686798,
0.006136805284768343, 0.00948167126657763, -0.015440993942320347,
-0.0078024123795330524, -0.009563586674630642, -0.005484897643327713,
-0.005413222126662731, -0.0020052131731063128, -0.012649054639041424,
-0.029489431530237198, 0.01417131070047617, 0.006850149482488632,
-0.01697007566690445, 0.008539647795259953, -0.003559893299762964,
0.013044977560639381, -0.0199190191924572, 0.009078922681510448,
-0.013946044258773327, -0.0159870944917202, -9.519429295323789e-05,
-0.011229193769395351, 0.015604824759066105, 0.01899064891040325,
-0.01154320128262043, 0.024724705144762993, 0.0017799466149881482,
-0.01915447786450386, -0.009092574939131737, -0.0037203102838248014,
0.019482139497995377, 0.010212081484496593, -0.016260145232081413,
0.0099868150279375, 0.039182718843221664, -0.00016607005090918392,
0.020792780444025993, -0.016533195972442627, -0.02350963093340397,

Figure 4. Section of the text after embedding vectorization.

AI method evaluation

The study employed both rule-based and machine-learning classification methods to categorize job advertisements. The rule-based approach utilized tokenized document scoring to differentiate between ’Remote’ and ’Not Remote’ classifications based on predefined expressions and their accumulated scores. This method, while straightforward, provided a foundational understanding of the dataset’s characteristics and the effectiveness of simple linguistic rules in preliminary classification tasks. Table 1 showcases a rule-based classification example where a job advertisement for a Hybris Developer is correctly classified as ’Remote’ based on a comparison of tokenized text scores - 15 for ’Remote’ and 4 for ’Not remote’. This demonstrates the rule-based system’s ability to categorize job advertisements effectively using predefined linguistic expressions.

Table 1. Rule-based classification example

Original text	Tokenized text	Class label	Prediction	Remote score	Not remote score
Hybris Developer (100% remote) We offer : Interesting and challenging job in one of the biggest company in Polish automotive industry A lot of independence in action and the opportunity to implement your own ideas Private medical care Company discounts 100% remote work or if you prefer hybrid work – we have office in convenient location in Warsaw	offer interesting challenging job one biggest company polish automotive industry lot independence action opportunity implement idea private medical care company discount 100 remote work prefer hybrid work office convenient location warsaw	Remote	Remote	15	4

For a more sophisticated analysis, supervised machine learning models were trained using the pre-processed dataset, with both TF-IDF vectors and embedding vectors serving as inputs. For this purpose, Scikit learn (Pedregosa, et al., 2011) Python library was used, which contains the toolset to prepare, train, test and deploy machine learning models. Among the supervised machine learning methods, three classification algorithms were selected based on the literature guidelines. The chosen models proved to be effective for classifying textual data in several cases. The models were as follows:

- Logistic Regression (Uddin, Khan, Hossain & Moni 2019)
- Support Vector Machine (Hassan, Ahamed & Ahmad 2022)
- Random Forest (Occhipinti, Rogers & Angione 2022)

The training process involved hyperparameter tuning with GridSearchCV and ten-fold cross-validation (Gareth, Witten, Hastie & Tibshirani 2013) to optimize model performance and mitigate overfitting. Model evaluation was comprehensive, incorporating metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess classification effectiveness rigorously. These metrics are widely used by experts in the field of machine learning (Brown 2018).

Comparative performance analysis

The comparative analysis of the classification methods used in this study underlines the distinct advantages and limitations inherent to both rule-based and machine-learning approaches in the context of remote job advertisement classification. A summary of the model metrics is shown in Table 2.

Table 2. Evaluation of trained models

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)	ROC AUC (%)
Rule-based	96.92	94.29	100	97.06	N/A
TF-IDF LR	98.46	100	97.22	98.59	100
TF-IDF SVM	76.92	82.86	76.32	79.45	86.95
TF-IDF RF	78.46	74.29	83.87	78.79	87.14
Embedding LR	76.92	74.29	81.25	77.61	86.1
Embedding SVM	76.92	71.43	83.33	76.92	86.57
Embedding RF	81.54	80	84.85	82.35	91.43

Rule-based classification, as a method, offered the benefit of providing immediate insights with relatively minimal computational resources, acting as an effective preliminary filter for identifying remote opportunities. The inherent simplicity of rule-based systems, relying on manually defined rules and expressions, makes them particularly transparent and easy to understand. Legislation has a similar requirement for norm clarity (Arató 2022; Arató–Balázs 2022). However, this reliance also introduces constraints on their adaptability and scalability, as the manual adjustment of rules to accommodate new data or variations in job advertisement formats can be labour-intensive and less dynamic.

In contrast, machine learning models, especially the TF-IDF Logistic Regression method (TF-IDF LR), demonstrated superior performance

in accurately identifying remote job opportunities. The TF-IDF Logistic Regression stood out with an overall accuracy of 98.46% and a perfect recall rate of 100%, showcasing its exceptional reliability. This method's capability to flawlessly differentiate between remote and non-remote listings, as indicated by its perfect ROC AUC score of 100% (Fig. 5), underscores the significant potential of machine learning techniques in enhancing the precision of job classification systems. Such models are adept at handling the nuanced linguistic features present in job descriptions and adapting effectively to the dataset's complexity, offering a scalable solution for the evolving demands of the job market.

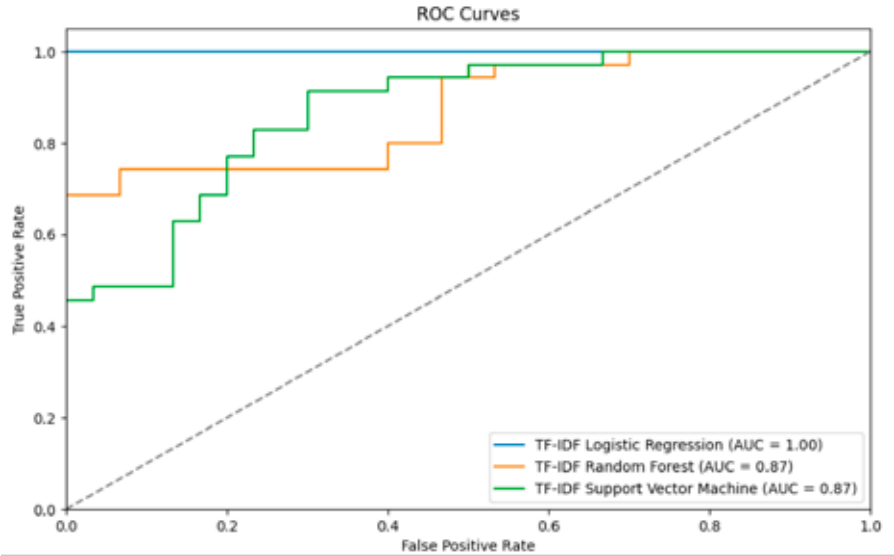

Figure 5. ROC curves of TF-IDF models.

Following closely, the Rule-based classification method demonstrated considerable efficacy with an accuracy of 96.92% and an F1 score of 97.06%. Its precision ensured that identified remote job advertisements were genuinely pertinent, providing a high degree of confidence in the classified results. Despite its slightly lower recall compared to the logistic regression model, the rule-based approach remains valuable in contexts where avoiding false positives is critical, highlighting the importance of selecting the right method based on specific project needs.

The evaluation of these models based on comprehensive performance metrics illuminated the most effective techniques for classifying remote job opportunities, revealing the essential role of AI and NLP in refining the job search process. While machine learning models like TF-IDF Logistic Regression exhibit a clear advantage in terms of accuracy and adaptability, the rule-based approach maintains its relevance through its simplicity and the immediate clarity it provides. The machine learning models utilizing embedding vectors offer a detailed linguistic analysis, as demonstrated by the Embedding Logistic Regression (Embedding LR), which provides a robust framework for understanding

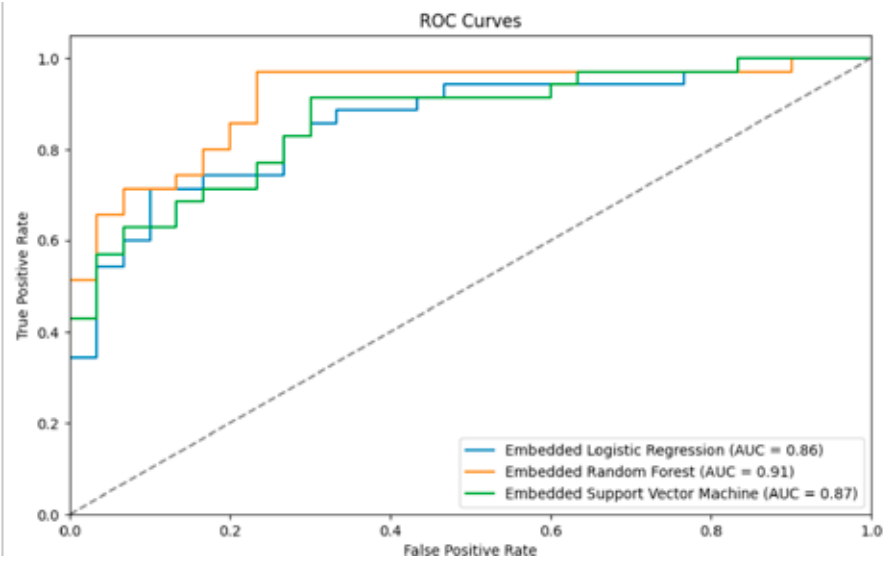

Figure 6. ROC curves of Embedded models.

contextual language in job advertisements. The Embedding Random Forest (Embedding RF) model, with an accuracy of 81.54% and a ROC AUC of 91.43%, highlights the practical application of embeddings in dealing with complex language patterns, marking a substantial advancement from traditional rule-based and TF-IDF methodologies.The analysis reveals that while rule-based and TF-IDF approaches yield higher performance metrics, embedding-based models also demonstrate considerable effectiveness in classifying job advertisements. The comparative advantage of rule-based and TF-IDF methods lies in their higher accuracy and recall rates. However, the

embedding models, with their contextual language processing capabilities, still deliver robust results, indicating their potential for application in more complex linguistic scenarios. This underscores the merit of incorporating a diverse array of methodologies in developing comprehensive classification systems for remote job advertisements.

This analysis confirms the importance of leveraging the strengths of all classification strategies to meet the varying requirements of remote job advertisement identification. As the digital job market continues to expand, the integration of these advanced NLP techniques into classification systems is key, promising a more efficient and accessible way for job seekers and employers alike in navigate the remote work landscape.

Conclusions and future directions

This paper contributes valuable insights into the application of NLP techniques for classifying job descriptions aimed at remote work opportunities, demonstrating that even with a relatively modest dataset, our results are both impressive and encouraging. An investigation of the effectiveness of rule-based classification systems and machine learning methods, particularly the TF-IDF logistic regression model, highlights an efficient path forward in automating the classification process. Remarkably, this model distinguishes itself by delivering high accuracy and reliability, showcasing its potential in settings where minimizing false negatives is essential.

This research reveals an important finding: effective and precise models can be developed without the exhaustive need for large datasets. This aspect of our research is particularly noteworthy as it suggests a strategic advantage in the use of NLP for job classification. By demonstrating the feasibility of achieving significant results with smaller datasets, our approach sheds light on a more accessible and sustainable methodology for model development and deployment. This efficiency not only alleviates the burden of gathering, annotating, and maintaining voluminous data sources but also reduces the computational demands typically associated with training and implementing sophisticated models.

The challenges associated with the embedding methods are due to the scale of the data set relative to the complexity of the vector spaces, which highlight the potential for future improvements and optimisation of the approach. The limitations observed with embedding methods encourage further research and the search for innovative solutions that can circumvent these limitations.

Looking ahead, expanding the dataset size emerges as a promising direction, potentially enabling more refined model tuning and improved generalization capabilities. Furthermore, the exploration of advanced deep learning techniques, such as recurrent neural networks (RNNs) and transformer-based models, could unlock new dimensions in semantic analysis, enhancing the ability of classification systems to interpret job descriptions with greater nuance and accuracy.

In summary, this paper makes a compelling case for the strategic use of NLP techniques in job classification, especially in contexts constrained by dataset size or computational resources.

References

Arató, Balázs 2022. Norm clarity in the light of Hungarian case law. *Magyar Nyelvőr,* 46 (special issue), 81–90. https://doi.org/10.38143/Nyr.2022.5.81

Arató, Balázs – Balázs, Géza 2022. The linguistic norm and norm of legal language. *Magyar Nyelvőr,* 46 (special issue), 91–103. https://doi.org/10.38143/Nyr.2022.5.91

Bird, S. – Loper, E. – Klein, E. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

Brown, J. 2018. Classifiers and their Metrics Quantified. *Molecular informatics,* 37(1–2), 1700127. https://doi.org/10.1002/minf.201700127

Gareth, J. – Witten, D. – Hastie, T. – Tibshirani, R. 2013. *An Introduction to Statistical Learning: with Applications in R.* New York: Springler.

Hansen, S. – Lambert, P. J. – Bloom, N. – Davis, S. J. – Sadun, R. – Taska, B. 2023, March. Remote Work across Jobs, Companies, and Space. Cambridge: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w31007

Hassan, S. – Ahamed, J. – Ahmad, K. 2022. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers,* 3(2666–4127), 238–48.

Jurafsky, D. – Martin, J. H. 2023, January 7. *Speech and Language Processing.* Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

McKinney, W. 2010. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference.* Austin TX. https://doi.org/10.25080/Majora-92bf1922-00a

Occhipinti, A. – Rogers, L. – Angione, C. 2022. A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications,* 201(0957–4174), 117193.

OpenAI developer platform 2023. Retrieved 10 30, 2023, from https://platform.openai.com/docs/introduction.

Pedregosa, F. – Varoquaux, G. – Gramfort, A. – Michel, V. – Thirion, B. – Grisel, O. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12(Oct), 2825–30.

Taher, A. – Azharul Hasan, K. – Afsana Akhter, K. 2018. N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1–5). IEEE.

Tusar, M. – Islam, M. 2021. A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)* (pp. 1–4). IEEE.

Uddin, S. – Khan, A. – Hossain, M. – Moni, M. 2019. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making,* 19(1), 281.

Van Rossum, G. – Drake, F. L. 2009. *Python 3 Reference Manual.* CreateSpace.

Benjamin Szilágyi
BSc Student
Óbuda University, Hungary,
Alba Regia Technical Faculty
E-mail: szilben1997@gmail.com
https://orcid.org/0009-0005-9690-4382

Rozália Lakner
Associate Professor
Óbuda University, Hungary,
Alba Regia Technical Faculty
E-mail: lakner.rozalia@amk.uni-obuda.hu
https://orcid.org/0000-0001-5665-479X