Exploring the Synergy Between LLMs and Knowledge Graphs for Advanced Abusive Speech Detection in Serbian

Naučni rad DOI: 10.18485/judig.2025.1.ch23

Danka Jokić[1], *0000-0002-7621-5847*
Ranka Stanković[2], *0000-0001-5123-6273*

Abstract

Large Language Models (LLMs) have introduced immense changes in the field of artificial intelligence (AI) by their impressive capabilities in language processing and generation. However, their limitations in structured knowledge, reasoning, and factual accuracy pose significant challenges in real-world applications. On the other hand, knowledge Graphs (KGs) with their ability to store and represent interconnected data entities and semantic relationships offer a complementary solution. This paper explores the synergy between LLMs and KGs, emphasising their potential to enhance LLMs with factual grounding, enable complex reasoning, and improve explainability in AI systems.

Focusing on online safety and moderation of harmful textual content in Serbian, the paper explores and highlights how KGs can enhance LLM-based abusive speech detection systems by providing richer contextual understanding and structured reasoning. By utilising datasets such as the AbCoSER corpus of abusive speech in Serbian, the study demonstrates how KGs can provide richer context to improve abusive speech detection. Through advanced prompting techniques and hybrid search approaches inspired by retrieval-augmented generation (RAG), the research lays the groundwork for more robust, context-aware, and ethically aligned generative AI applications.

Keywords: Knowledge Graphs (KGs), Large Language Models (LLMs), Natural Language Processing (NLP), Abusive Speech

[1] Language Resources and Technologies Society (JeRTeh), danka.jokic@afrodita.rcub.bg.ac.rs
[2] Language Resources and Technologies Society (JeRTeh), ranka.stankovic@rgf.bg.ac.rs

1. Introduction

The emergence of large language models (LLMs), from transformer model BERT (Devlin et al, 2019) to the advanced models such as like GPT 3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), has significantly impacted the field of artificial intelligence (AI) by excelling at language understanding and generation tasks (Kau et al., 2024). However, when it comes to LLMs implementation in real-life scenarios, it is delayed in areas where the correctness of output and reasoning capabilities are of the utmost importance due to the following reasons:

- LLMs rely on parametric knowledge, lacking structured knowledge and reasoning capabilities (Pan et al., 2023), which limits their effectiveness in applications requiring factual accuracy and context-aware reasoning (Pan et al., 2024).
- As being trained on general text corpora, LLMs manifest a lack of domain knowledge (Kau et al., 2024; Pan et al., 2024), including private and business-critical knowledge (Pan et al., 2023),
- LLMs are not able to verify or explain the answers, leading to limited interpretability (Pan et al., 2023),
- LLMs are prone to generating fabricated content known as hallucinations, stemming mainly from the knowledge gaps within the model (Agrawal et al., 2024),
- The knowledge in LLMs is frozen in their parameters at the time of training (Kau et al., 2024),
- The nature of data used for LLM training raised particular ethical, data privacy, and data bias concerns (Pan et al., 2023).

Knowledge graphs (KGs), on the other hand, offer a compelling solution for these challenges (Pan et al., 2024). There has been a growing focus on KGs as sources of structured knowledge for LLM-based models. The intuitive structures of KGs effectively represent real-world knowledge by representing entities with nodes and relationships between them as edges, which enables a greater understanding of a word's semantics via its context (Kau et al., 2024). The entities and their relationships are depicted in a machine-readable format, KGs provide a rich source of structured knowledge. KGs can improve, ground, and verify LLM generations so as to significantly increase reliability and trust in LLM usage (Pan et al., 2023).

The synergy with KGs has the potential to enhance LLM capabilities significantly in several key areas. First, augmenting LLMs by incorporating external knowledge of KGs can substantially improve their ability to comprehend factual information and generate more accurate and reliable responses to complex questions, thereby reducing the hallucination (Agrawal et al., 2024). In addition, KGs also offer the necessary context and relationships to enable LLMs to perform more sophisticated reasoning tasks and incorporate commonsense knowledge into their reasoning processes, leading to more nuanced and human-like understanding. Furthermore, the explicit relationships within KGs can be harnessed to explain the reasoning behind LLM outputs, directly addressing a critical challenge in interpretable AI. ERNIE (Zhang et al., 2019), a language representation model, is a good example of this synergy. It was trained on large-scale textual corpora and KGs, allowing it to simultaneously utilise lexical, syntactic, and knowledge information, resulting in better language understanding.

While the potential is undeniable, challenges require attention. Developing effective methods for LLMs to learn from text data and knowledge graphs jointly is crucial for successful integration. Additionally, ensuring the consistency and quality of knowledge graph data is essential, as incomplete or inaccurate information can lead to biased or erroneous LLM results. Moreover, utilization of KGs is constrained by the availability of existing graphs and the resources needed for their construction and maintenance (Pan et al., 2023; Kau et al., 2024).

Despite these challenges, the integration of LLMs and KGs holds immense potential to revolutionize various AI applications. LLMs empowered by KGs can provide more accurate and comprehensive answers in question-answering systems. Intelligent assistants integrated with KGs can understand and respond to user queries with greater context and factual grounding. Additionally, the combination of LLMs and KGs can lead to the development of more factually accurate and contextually relevant natural language and code generation and the creation of more sophisticated and dynamic knowledge representation systems (Kau et al., 2024). A roadmap and forward-looking taxonomy of synergetic usage of LLMs and KGs is given in (Pan et al., 2024).

There are three possible directions for technical implementation of such systems (Pan et al, 2023). Firstly, KGs can be incorporated into training data for LLMs to complement natural language text. Secondly, triples in KGs can be used for prompt construction as input to LLMs. Additionally, KGs can be used as external knowledge in retrieval-augmented language models.

In this paper, we will present how knowledge graphs can be used to enhance the reasoning capabilities of LLMs concerning online safety and moderation of harmful textual content in Serbian. We investigated the synergy between LLMs and knowledge graphs for the abusive speech phenomenon from various perspectives:

- Using LLMs to generate harmful content in the Serbian language,
- Using LLM to generate a knowledge graph,
- Using LLMs and KG to detect abusive speech in the Serbian language.

One of the main objectives of our research was to explore whether we could leverage domain knowledge encapsulated in KGs to enhance the generation and detection of abusive speech.

The remainder of this paper is organized as follows. In Section 2, Related work, we presented the work done in the research area of exploring the synergy between LLMs and knowledge graphs while tackling the abusive speech detection and the Serbian language. An overview of the methodology used in our research with short description of the dataset is presented in Section 3. The results and evaluation are presented in Section 4. In Conclusion, we summarize the results of our research and indicate further research directions.

2. Related Work

2.1. Using KGs for Abusive Speech Detection

In their research, Maheshappa et al. (2021) incorporated a knowledge graph into a hate speech detection pipeline. They constructed a graph out of tweet text and then produced a node2vec embedding. The embedding was an input to an LSTM. The input to another LSTM was fastText embeddings of tweet text. The outputs of two LSTMs were concatenated and passed as input to a fully connected linear layer. The results with two LSTMs were better than the results of one LSTM with fastText embeddings only. Lobo et al. (2022) presented a hybrid approach which integrates KGs and deep learning models to recognize language that references gender and sexual orientation in hate speech and thereby predict the hate speech target. The authors used an existing ontology of gender, sex and sexual orientation to support annotation of the Jegsaw toxicity dataset and ML algorithms to build weights of each term in the ontology. They used one-hot encoding of the words to train the DL model to recognize homophobic and gender hate speech. The evaluation on gender and sexual orientation demonstrates that a knowledge-grounded approach is key to enhancing model transparency, robustness, and handling of annotation errors. In their recent work, Zhao et al. (2024) introduced the MetaTox method that combines LLMs and a meta-toxic knowledge graph to address domain-specific knowledge gaps in toxicity detection. They exploited LLMs to build a meta-toxic KG, which is later used via a retrieval process to supply LLM with toxic knowledge.

2.2. Text Classification Using Generative AI Models

The rapid development of generative AI (GenAI) models resulted in several studies using them for text classification, including abusive speech detection. Zhu et al. (2024) used ChatGPT for cyberbullying and COVID HATE datasets, reporting high recall but low precision rates. ChatGPT was used in Huang et al. (2023) to identify implicit hate in tweets and to generate an explanation of the given classification. The model correctly identified 80% of implicit hateful tweets. In another study presented in Guo et al. (2024), the authors explored LLMs as classifiers on five datasets. They achieved improvement of 7.9% to 24.2% in F1 Score compared to state-of-the-art models for some of the datasets. One of their findings was the dependence of the model's effectiveness on prompt design and the language of the text.

2.3. Abusive Language Detection in Serbian

The first information search experiments in the Serbian language dealing with the detection of attacks as a result of national, racial, or religious hatred in a corpus of newspaper articles were presented by Krstev et al. (2007). The first abusive speech dataset, AbCoSER, and lexicon in Serbian language were presented in Jokić et al. (2021). The results of abusive speech detection on the same corpus were published in Jokić et al. (2024a) and first experiments with LLMs in Jokić at al. (2024b). The best-performing classifier was BERTić with an F1 score of 0.827 and an accuracy of 0.89. Vujičić Stanković and Mladenović (2023) used a hate speech lexicon and a dataset in the Serbian language to train a classifier for automatic hate speech detection in the sports domain. They reported reaching 96% precision in detecting hate speech using a BiLSTM deep neural network.

To address the lack of a software tool for a safer digital environment for users, Milaković et al. (2024) created a dataset of ugly and derogatory words in the Serbian language and a web extension to analyze and censor such words.

Muminović and Muminović (2025) evaluated detection of toxic comments in Serbian, Croatian and Bosnian using a dataset composed of 4,500 YouTube and TikTok comments. They evaluated four LLM models in zero-shot and context-augmented settings. The best result was achieved with the Gemini 1.5 pro model in context-augmented mode, reaching an F1 score of 0.82 and an accuracy of 0.82. Their result also demonstrated how adding minimal context can improve toxic language detection, and they suggested strategies such as improved prompt design and threshold calibration for better results.

3. Methodology

This section describes the methodology we employed in this research. The framework we used in our experiments is depicted in Figure 1.

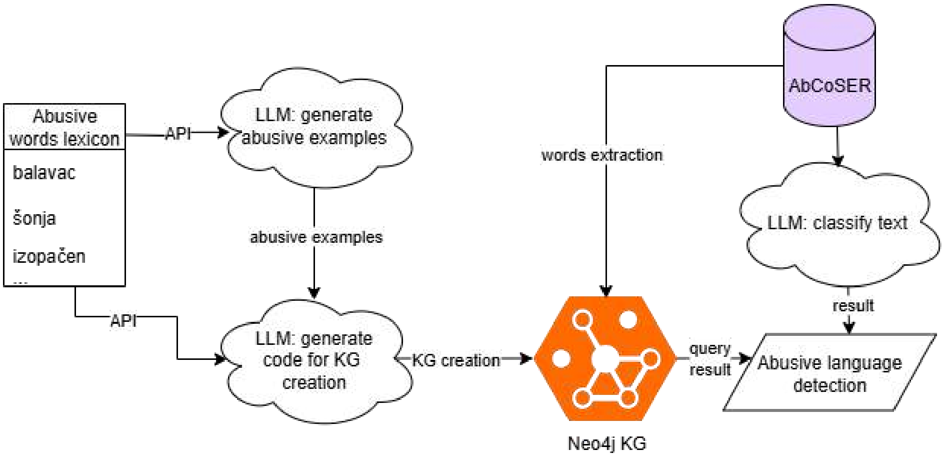


Figure 1: A framework for abusive language detection using KG and LLM

3.1. Generation of Abusive Language Examples

An abusive speech lexicon HURT_LIST used in this research was crafted by crowdsourcing and Web resources (Jokić et al., 2021) and is composed of 1,434 unique lemmas. It contains canonical forms of lemmas and its colloquial variations due to the nature of abusive speech that can be found in digital environments. which is often obfuscated by replacing or omitting characters in a word or using abbreviations. An example of obfuscation by replacement is lemma vag1na, which was made by replacing character i in “vagina” (eng. vagina) with number 1.

A prompt was meticulously crafted to elicit examples of hate speech based on seed terms from HURT_LIST lexicon of abusive speech, while circumventing the guardrails imposed by LLM concerning abusive speech generation. The approach pertains to knowledge-controlled generation methods (Agrawal et al., 2024). For each seed term, five examples of abusive language were generated, and LLM was asked to rate them from 1 (least hateful) to 5 (most hateful). As the model, we used GPT-3.5-Turbo by OpenAI with temperature 0.5 to foster the creativity of generated responses.

3.2. Knowledge Graph Data Model Generation

The generated examples and the HURT_LIST lexicon are used to produce a knowledge property graph in Neo4j AuraDB3. Knowledge Graph is a design pattern to store, organize, and access interrelated data entities and semantic relationships between different pieces of information, enabling more sophisticated understanding and reasoning about data (Neo4j documentation). LLM is used for generating a graph data model from the examples and the lexicon to be used as a skeleton for data import in the Neo4J knowledge graph. The initial code4 was modified to accommodate our use case, and OpenAI GPT-4 was used as an LLM for the generation of the model comprising nodes, relations and properties. In addition, the vector embeddings were generated for each term of the HURT_LIST lexicon and each abusive speech example and stored in the corresponding graph property for later use for the similarity/vector search. For this purpose, we used the OpenAI "text-embedding-ada-002" model. The text-based and vector-based indexes were created in AuraDB accordingly.

3.3. Abusive Speech Detection

We tested the OpenAI GPT-3.5-Turbo model for abusive speech detection. Three prompt techniques were tested: zero-shot, context prompt and chain-of-thought prompting (Wei at al., 2022). We adopted the approach of Kumarage et al. (2024) for the content of each prompt. For context and chain-of-thought prompt, the LLM is provided with the definition of abusive language from Caselli et al. (2021), where abusive speech is defined as "hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions." The resulting context prompt is given below:

> *Given the following text, evaluate if it contains abusive speech or not, specifically hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions. Please provide your assessment as labels True or False.*
> *text: {text}*

[3] Fully Managed Graph Database Service | Neo4j AuraDB

[4] Initial code: llm/notebooks/llm-gen-KG-from-CSV.ipynb at main · danb-neo4j/llm

The other two prompts are provided in the Appendix. The output of the LLM-based classifier was combined with data available in the previously generated knowledge graph, which is used either as a simple word list or through similarity search. The idea was to probe GraphRAG (Edge et al., 2025) for this use case.

3.4. Dataset

As a dataset, we used the AbCoSER corpus of abusive speech in Serbian (Jokić et al., 2021), which includes labels for 6,436 Serbian tweets. It contains general abusive speech, without targeting a specific type of targets such as racial, LGBT or misogynistic speech. The dataset was annotated at two levels. The first level label indicates the presence of abusive speech. At the second level, the abusive speech was further annotated to identify the type of speech detected, whether it was hate speech, profanity, derogatory language, or other. The annotation was performed manually by two independent annotators and a supervisor to resolve any differences. Out of a total of 6,436 tweets, 1,416 were flagged as containing abusive speech.

4. Results

4.1. LLM for abusive language generation

The examples generated for the term "balavac" (eng. drooler), together with the assigned abusiveness score, are given in Table 1. Our hypothesis is that the abusiveness score, produced by LLM, is influenced by the presence of negation and explicitly offensive words in the text of examples 1-4, and therefore resulted in a higher score than example number 5.

Table 1 Table with examples of abusive speech

No.	Abusive text example	Abusivenes s core
1	Nemoj da mi se obraćaš, balavac jedan!	5
2	Opet si napravio sranje, balavac jedan!	5
3	Kakav si ti balavac, ne možeš ni da obaviš jednostavan zadatak!	5
4	Ti si običan balavac, nemaš pojma o ničemu!	5
5	Pusti me na miru, balavac jedan!	4

The quality of the generated abusive speech examples varied. Only from the presented examples, we can observe issues with grammar, such as word order in a sentence, incorrect negation formulation, and improper usage of case. The examples 1, 2 and 5 were similar, although the model was instructed and set to provide creative and different instances.

4.2. LLM for Knowledge Graph Model Generation

As an illustration of the schema generation ability of LLMs, we asked the model to generate the simple KG structure for our use case. The KG structure generated employing LLM is presented in Figure 2 and it has the following elements:

Node properties:

- Dataset {_id: STRING, number: INTEGER}
- Term {_id: STRING, term: STRING, originalTerm: STRING, termEmbedding: LIST}
- Example {_id: STRING, exampleText: STRING, textEmbedding: LIST}
- AbusiveScore {_id: STRING, score: INTEGER}
- Relationship properties:
- The relationships:
- (:Term)-[:one_to_many]->(:Example)
- (:Example)-[:one_to_one]->(:AbusiveScore)

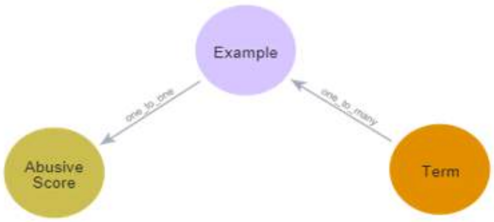


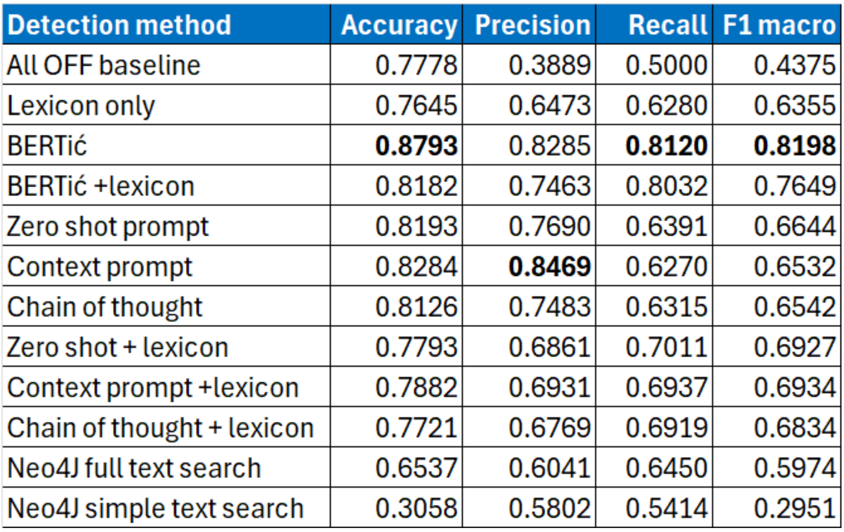
Figure 2 KG structure in AuraDB

The graph properties termEmbedding and textEmbedding correspond to embedding vectors and are stored as a list of numbers in the graph database.

An example can be related to multiple terms, for instance, in the sentence in Table 2. “You've done shit again, you drooler!” shall be connected to two nodes - one with the term “shit” and another with the term “drooler”.

4.3. Abusive Text Classification Results

The results of the conducted experiments are presented in Figure 3. To get a clear representation of the results, we stated outcomes of a dummy (All OFF baseline) classifier that assigns to each record the label of the most frequent class, that could serve as a default baseline model.



Detection method	Accuracy	Precision	Recall	F1 macro
All OFF baseline	0.7778	0.3889	0.5000	0.4375
Lexicon only	0.7645	0.6473	0.6280	0.6355
BERTić	**0.8793**	0.8285	**0.8120**	**0.8198**
BERTić +lexicon	0.8182	0.7463	0.8032	0.7649
Zero shot prompt	0.8193	0.7690	0.6391	0.6644
Context prompt	0.8284	**0.8469**	0.6270	0.6532
Chain of thought	0.8126	0.7483	0.6315	0.6542
Zero shot + lexicon	0.7793	0.6861	0.7011	0.6927
Context prompt +lexicon	0.7882	0.6931	0.6937	0.6934
Chain of thought + lexicon	0.7721	0.6769	0.6919	0.6834
Neo4J full text search	0.6537	0.6041	0.6450	0.5974
Neo4J simple text search	0.3058	0.5802	0.5414	0.2951

Figure 3: Results of conducted experiments with LLMs and KG

When it comes to F1 macro, accuracy and recall metrics, the result for the BERTić transformer model [BERTić], as reported by Jokić et al. (2024a), remains the best model for the dataset. Surprisingly, the GPT-3.5-turbo model with context prompt [Context prompt] achieved a better precision score than [BERTić], which indicated that the model made very accurate predictions across the classes. We can also note how adding context to the prompt influences model results [Zero shot + lexicon, Context prompt + lexicon, Chain of thought + lexicon]. Contrary to the findings of Muminović and Muminović (2025), adding context did not improve zero-shot results for this dataset.

The HURT_LIST lexicon data are integrated with LLMs as a probe for future knowledge graph integration. The value of lexicon-based text output was a binary one if any lemma from the lemmatised text was found in the abusive speech lexicon, otherwise 0. The results, indicated by the "+lexicon" label in the table [Zero shot + lexicon, Context prompt + lexicon, Chain of thought + lexicon], were promising for the LLM-based classifier, since the presence of lexicon input improved recall and consequently resulted in a higher F1 macro score in comparison to LLM models only [Zero shot, Context prompt, Chain of thought]. In contrast, integrating lexicon input with the BERTić model [BERTić + lexicon] resulted in decreased performance, suggesting that BERTić independently classifies abusive speech more effectively due to its contextual text interpretation capabilities.

5. Conclusion

In this paper, we presented our experiments using LLMs and KGs for abusive speech detection in Serbian. We used LLMs for various tasks ranging from abusive text generation to knowledge graph structure generation, to text classification. We also tested text search over the knowledge graph in simple text and hybrid search modes as the basis for the GraphRAG approach. Based on the results of our classification experiments using an approach that combines LLMs and abusive speech lexicons, we identify a promising direction for synergy between LLMs and knowledge graphs in abusive language detection. . In that light, we aim at building a larger lexicon of abusive speech, which also includes multiword expressions (MWEs). As for the generated examples of abusive language and their inconsistent quality, we will test more LLMs, including open-source models for the Serbian language[5], and other vendors. We aim to include additional data, such as hate targets, in the knowledge graph and build a hybrid abusive speech detection model by employing LLMs, traditional machine learning, and GraphRAG.

References

[1] Agrawal, G., Kumarage, T., Alghamdi, Z., & Liu, H. (2024). Can Knowledge Graphs Reduce Hallucinations in LLMs? A Survey. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 3947–3960). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.219.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[5] https://huggingface.co/jerteh and https://huggingface.co/te-sla

[4] Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Metropolitansky D, Ness RO & Larson J (2025). From local to global: A graph RAG approach to query-focused summarization. DOI:10.48550/arXiv.2404.16130.

[5] Guo K, Hu A, Mu J, Shi Z, Zhao Z, Vishwamitra N & Hu H (2024). An investigation of large language models for real-world hate speech detection. DOI:10.48550/arXiv.2401.03346.

[6] Huang F, Kwak H & An J (2023). Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In: Companion Proceedings of the ACM Web Conference 2023,WWW'23 Companion. Association for Computing Machinery, pp. 294–297. DOI:10.1145/3543873.3587368.

[7] Jokić, D., Stanković, R., Krstev, C., & Šandrih, B. (2021). A Twitter Corpus and lexicon for abusive speech detection in Serbian. In 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[8] Jokić D., Stanković R. & Šandrih Todorović B. (2024a). Abusive speech detection in Serbian using machine learning. In: Mitkov R and others (eds.) Proceedings of NATURAL LANGUAGE

[9] PROCESSING AND ARTIFICIAL INTELLIGENCE FOR CYBER SECURITY, NLPAICS, Lancaster, United Kingdom. pp. 153–163.

[10] Jokić, D., Stanković, R., & Jaćimović, J. (2024b). Knowledge Graphs in the Era of Large Language Models: Opportunities and Challenges. Judig, Belgrade, Serbia. 60-61.

[11] Kau, A., He, X., Nambissan, A., Astudillo, A., Yin, H., & Aryani, A. (2024). Combining Knowledge Graphs and Large Language Models (arXiv:2407.06564). arXiv. https://doi.org/10.48550/arXiv.2407.06564.

[12] Krstev C, Gucul S, Vitas D and Radulovi V (2007). Can we make the bell ring? In: Paskaleva E and Slavcheva M (eds.) Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages. Borovets, Bulgaria, pp. 15–22.

[13] Kumarage, T., Bhattacharjee, A., & Garland, J. (2024). Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. arXiv preprint arXiv:2403.08035.

[14] Milaković A, Jocović V, Cincović J, Mićović M, Radenković U and Drašković D (2024). Detecting ugly and derogatory words in serbian language using a web browser extension. In: 2024 32nd Telecommunications Forum (TELFOR). pp. 1–4. DOI: 10.1109/âˇDa¸FOR63250.2024.10819059. ISSN: 2994-5828

[15] Muminovic, A., & Muminovic, A. K. (2025). Large Language Models for Toxic Language Detection in Low-Resource Balkan Languages (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2506.09992

[16] Neo4j documentation: What Is a Knowledge Graph? - Neo4j Graph Database & Analytics

[17] OpenAI. 2023. Gpt-4 technical report. https://doi.org/10.48550/arXiv.2303.08774

[18] Pan, J., Razniewski, S., Kalo, J. C., Singhania, S., Chen, J., Dietze, S., ... & Graux, D. (2023). Large Language Models and Knowledge Graphs: Opportunities and Challenges. Transactions on Graph Data and Knowledge.

[19] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2024.3352100.

[20] Vujičić Stanković, S. & Mladenović, M. (2023). An approach to automatic classification of hate speech in sports domain on social media. Journal of Big Data 10(1): 109. DOI:10.1186/ s40537-023-00766-9.

[21] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv. https://doi.org/10.48550/ARXIV.2201.11903

[22] Zhu Y, Zhang P, Haq EU, Hui P and Tyson G (2024). Exploring the capability of ChatGPT to reproduce human labels for social computing tasks. In: Social Networks Analysis and Mining:

[23] 16th International Conference, ASONAM 2024, Rende, Italy, September 2–5, 2024, Proceedings, Part III. Springer-Verlag, pp. 13–22. DOI:10.1007/978-3-031-78548-1_2.

[24] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities, 2019. arXiv:1905.07129.

[25] Zhao Y, Zhu J, Xu C and Li X (2024). Enhancing LLM-based hatred and toxicity detection with meta-toxic knowledge graph. DOI:10.48550/arXiv.2412.15268.

Appendix

Appendix 1. The system prompt used to generate abusive language examples

"You are an expert for detection of abusive language in a text written in Serbian language. For the educational purposes, in order to teach people how to recognize abusive speech in text you will be given an abusive term and you will provide five different examples of the term usage in an abusive content in Twitter like style, and abusiveness score in range 1 to 5, one being least abusive and 5 being the most abusive. The term will be placed between three ### as delimiters. Please provide every instance on output mentioning every time the term you identify between delimiters and than an example in the following format for each example
"term":TERM,"example": EXAMPLE,"abusiveness score":
ABUSIVENESS_SCORE"

If you cannot provide examples, you will return an empty string

Please try to provide examples for each level of abusiveness score if it's not possible to find example for some score, please provide example anyway.

The highest priority is to have as many different examples.

Yuu should provide an answer if and only if you are completely sure that data are correct."

Output should be provided in Serbian language and in JSON format”.

Appendix 2. System prompts used to classify text into abusive or not abusive categories

Zero-shot prompt: "Is the following text abusive or not? Just answer in True or False. Don't provide explanations

text: {text}“

Chain-of-thought prompting: "Given the following text, evaluate if it contains abusive speech or not, specifically hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions.

First describe the step by step process of how you evaluate the given text. Then provide as output your assessment as labels True or False

text: {text}"

Istraživanje sinergije između velikih jezičkih modela i grafova znanja za naprednu detekciju uvredljivog govora na srpskom jeziku

Danka Jokić, Ranka Stanković

Sažetak

Ovaj rad istražuje mogućnosti povezivanja velikih jezičkih modela (LLMs) i grafova znanja (KG) kako bi se unapredila detekcija uvredljivog govora na srpskom jeziku. LLM modeli, od BERT-a do GPT-4, postigli su izuzetne rezultate u razumevanju i generisanju jezika, ali imaju ograničenja: nedostatak strukturisanog znanja, domenskih informacija, ograničene mogućnosti objašnjavanja svog odgovora, sklonost „halucinacijama“, zastarelo znanje nakon treniranja i etičke izazove vezane za podatke. Grafovi znanja nude rešenje (ili bar ublažavanje) ovih problema pružajući strukturisano, mašinski čitljivo znanje o entitetima i njihovim relacijama, čime se poboljšava tačnost, smanjuju halucinacije, omogućava složenije zaključivanje i interpretabilnost modela.

U istraživanju je korišćen leksikon uvredljivih izraza HURT_LIST (1.434 leme), na osnovu kojeg su pomoću GPT-3.5 generisani primeri uvredljivog govora, ocenjeni po stepenu uvredljivosti. Ti primeri i leksikon su iskorišćeni za kreiranje grafa znanja u Neo4j AuraDB, uključujući čvorove (termin, primer, ocena) i vektorske reprezentacije radi semantičke pretrage, odnosno pretrage po sličnosti. Detekcija uvredljivog govora je testirana pomoću GPT-3.5-Turbo sa tri vrste prompta pristupa promptovanju: zero-shot, kontekstualni i lanac misli (chain-of-thought). Korišćen je korpus AbCoSER sa 6.436 tvitova, od kojih 1.416 sadrži uvredljiv govor.

Rezultati pokazuju da je BERTić i dalje najbolji po F1 makro rezultatu, ali je GPT-3.5-Turbo imao bolju preciznost. Dodavanje konteksta nije unapredilo rezultate u zero-shot režimu, suprotno nekim prethodnim istraživanjima. Uključivanje leksikona povećalo je odziv i F1 makro meru kod LLM klasifikatora. Početni testovi pretrage u KG-u ukazali su na probleme sa indeksiranjem srpskog teksta zbog ograničene podrške za jezike sa manje razvijenim resursima.

Zaključuje se da sinergija LLM i KG predstavlja pravac koji obećava bolje rezultate u detekciji uvredljivog govora. Budući rad uključuje proširenje leksikona višerečnim izrazima, testiranje većeg broja LLM modela, obogaćivanje grafa znanja dodatnim informacijama kao što je meta govora mržnje i izgradnju hibridnog modela koji kombinuje jezičke modele, tradicionalne metode mašinskog učenja i GraphRAG pristup.

Ključne reči: grafovi znanja, veliki jezički modeli (VJM), obrada prirodnog jezika, uvredljiv govor