Creating a Synthetic Evaluation Dataset for Serbian SentiWordNet Using Large Language Models

**Abstract. Brief overview of the problem and research objective**

**Summary of the methodology used**

**Key findings**

**Implications of the research**

# **Introduction**

* ~~Background on sentiment analysis and its importance~~
* ~~Overview of SentiWordNet and its role in sentiment analysis~~
* ~~Challenges in creating SentiWordNet for under-represented languages like Serbian~~
* ~~Introduction to the concept of synthetic evaluation datasets~~
* ~~Objectives of the paper~~

Sentiment Analysis is the process of computationally determining the emotional tone behind words to understand the attitudes, opinions, and emotions expressed by them. One of the two main methods for Sentiment Analysis is by using sentiment lexicons. Sentiment lexicons are specialized dictionaries that associate words and phrases with sentiment values, facilitating the automated analysis of emotions in text (Liu, 2010).

A prominent example of such a lexicon is SentiWordNet (SWN), which extends the English WordNet dictionary by assigning to each synset (a set of cognitive synonyms) sentiment scores that reflect the collective emotional tone of the words (Baccianella et al., 2010).

Synsets containing the same meanings in different languages are interconnected through the Inter-Lingual Index (ILI), enabling these associations in various languages' WordNets.

It had proven that by using such a connection the sentiment values expressed in the SWN can be applied to other languages save English(Denecke, 2008). Serbian WordNet contains sentient values gained by direct mapping of synsets using ILI to SWN (Mladenovic et al., n.d.).

Such lexicon could be improved by replacing mapped values with those more representative of the Serbian language. But to evaluate such improvements the evaluation dataset – a subset of synsets from Serbian WordNet already annotated with sentiment polarity – is needed.

SNW has such an evaluation dataset, Micro-WNO a manually labelled subset of synsets from Princeton WordNet. It is publicly available online[[1]](#footnote-1).

Creating a comparable evaluation dataset for the Serbian language manually would necessitate a significant effort, involving either a small number of expert annotators or a larger group of less skilled annotators. Given the absence of such resources, an alternative approach becomes imperative.

Synthetic evaluation datasets are artificially created collections of data designed to test and validate computational models, particularly in domains where real-world data may be scarce, biased, or too sensitive to use. These datasets are generated through algorithms or simulations that aim to mimic the statistical properties of real data, allowing researchers to conduct robust evaluations under controlled conditions (Lu et al., 2024).

The advent of Large Language Models (LLMs) had allowed for creation of much better synthetic datasets. For purposes of NLP tasks, among them sentiment analysis, LLMs have proven that can perform adequate annotation with just a few examples (Brown et al., 2020).

This raises the question of whether LLMs could be employed not just to create an evaluation dataset, but to annotate the entirety of Serbian WordNet with sentiment polarity values. The decision to focus on creating a small evaluation dataset stems from the prohibitive computational expense associated with annotating the entire network.

The solution proposed here is to synthetic dataset, a set of synsets annotated by an LLM.

# Literature Review

* Studies related to SentiWordNet and its applications
* Research on synthetic dataset generation for sentiment analysis
* Few-shot learning and its relevance to sentiment analysis
* Overview of Large Language Models (LLMs) in sentiment analysis
* Gap in research for Serbian SentiWordNet and the need for evaluation datasets

# Methodology

* Overview of the Serbian WordNet and selection of synsets for evaluation
* Detailed description of the few-shot learning approach using LLMs
  + Selection criteria for training examples
  + Configuration of the LLM
* Process of labelling the selected synsets to create the synthetic evaluation dataset
  + Criteria for label assignment
  + Validation process for the assigned labels
* Development of evaluation metrics for assessing the synthetic dataset

Idea: Use the Sr-pol-set to find synstes that POS and NED <0.5. Which should show diffence of polarity between Serbain and English.

Use that set to choose some, selected pos, neg and obj

# Implementation

* Preparation of the few-shot learning environment
* Selection and preparation of the LLM
* Step-by-step process of creating the synthetic dataset
  + Data processing and cleaning
  + Few-shot learning execution
  + Labelling and validation

# Results

* Performance evaluation of the few-shot learning approach
* Discussion on the reliability and validity of the synthetic evaluation dataset

# Discussion

* Implications of the findings for the development of Serbian SentiWordNet
* Challenges encountered during the research and how they were addressed
* Potential for applying the methodology to other under-represented languages
* Suggestions for improving the dataset generation process

# Conclusion

* Summary of the research findings
* Contribution of the study to the field of sentiment analysis and language resources
* Future research directions

# References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource. *Proceedings of the 7th International Conference on Language Resources and Evaluation*. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. *Proceedings - International Conference on Data Engineering*, 507–512. https://doi.org/10.1109/ICDEW.2008.4498370

Liu, B. (2010). Sentiment Analysis and Subjectivity. In F. J. Damerau & N. Indurkhya (Eds.), *Handbook of Natural Language Processing, Second Edition* (pp. 629–661). Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business.

Lu, Y., Shen, M., Wang, H., Wang, X., Rechem, C. van, & Wei, W. (2024). *Machine Learning for Synthetic Data Generation: A Review*.

Mladenovic, M., Mitrovic, J., & Krstev, C. (n.d.). *Developing and Maintaining a WordNet: Procedures and Tools*.

1. https://github.com/aesuli/Sentiwordnet/blob/master/data/Micro-WNop-WN3.txt [↑](#footnote-ref-1)