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

**Abstract. SentiWordNet is lexical resurce for English language, that can be used for**

**Summary of the methodology used**

**Key findings**

**Implications of the research**

# **Introduction**

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.). It has already been used in the creation of a hybrid framework for sentiment analysis in Serbian (Mladenović et al., 2015).

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.

# Methodology

The *senti-pol-sr* is a polarity lexicon for the Serbian language, annotated at the word level rather than by senses(Stanković et al., 2022). It includes words that exhibit clear polarity, categorized as either positive or negative, and does not contain words considered to be objective.

In this research, the lexicon was employed to select a sample suitable for annotation by LMM. This was achieved by identifying all synsets from the Serbian WordNet containing literals (words) present in the *senti-pol-sr* lexicon and simultaneously having a neutral sentiment value (0,0) as mapped from SWN.

Three primary reasons are posited for discrepancies between the sentiment values in Serbian and those derived from SWN. Firstly, while a word may convey a polarizing sentiment, the actual sense it is used in may not. Secondly, the sentiment values in SWN, generated through machine learning methods, may be inaccurate. Thirdly, and most pertinent to this study, is the possibility that while a sense is considered objective in English, it carries sentiment in Serbian.

The initial analysis identified 2,956 synsets within the Serbian WordNet that contained literals annotated with clear polarity in the senti-pol-sr lexicon, with 1,511 exhibiting positive sentiment and 1,445 negative. Given the substantial volume, processing all these synsets with LLM was deemed impractical. Consequently, a random sample of 500 synsets was selected for further investigation.

Since the further steps involve using LMM, the next we will.

* 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*.

Mladenović, M., Mitrović, J., Krstev, C., & Vitas, D. (2015). Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, *46*(3), 599–620. https://doi.org/10.1007/s10844-015-0372-5

Stanković, R., Košprdić, M., Ikonić Nešić, M., & Radović, T. (2022). Sentiment Analysis of Serbian Old Novels. *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, 31–38. https://aclanthology.org/2022.salld-1.6

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