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

**Abstract.** This study introduces the creation of a synthetic evaluation dataset for Serbian SentiWordNet through the application of Large Language Models (LLMs), with a focus on the Mistral model. Confronting the significant scarcity of sentiment analysis resources for the Serbian language, this research endeavours to bridge this gap by generating a dataset that supports the evaluation and improvement of sentiment analysis tools tailored to Serbian. Within Serbian WordNet, sentiment polarity values were automatically mapped from the English SentiWordNet using the Inter-Lingual Index (ILI). To refine these values for better alignment with the Serbian language context, an evaluation set was created. Employing a rigorous methodology, 500 synsets were initially selected from Serbian WordNet based on their alignment with the *senti-pol-sr* lexicon and values mapped from SentiWordNet. These synsets were subjected to sentiment polarity classification via the Mistral model. A balanced subset of 75 synsets was then randomly extracted and subjected to finer sentiment gradation, followed by a thorough manual review. The study's findings reveal a high degree of model reliability, with approximately 93.3% of the responses fulfilling the established acceptability criteria. This outcome highlights the efficacy of LLMs like Mistral in automating sentiment analysis processes for languages with limited resources, underscoring the significant potential for broader application in under-represented linguistic contexts.

# **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 (Amatriain, 2024; Brown et al., 2020).

**Zero-shot Learning** involves the model making predictions or annotations without having seen any explicit examples of the task during training. This capability is particularly useful for sentiment analysis in languages or contexts where annotated data are scarce, as it allows the LLM to apply its pre-existing knowledge to new, unseen tasks (Amatriain, 2024; Brown et al., 2020).

**Few-shot Learning**, on the other hand, provides the model with a small number of examples from which it can learn to perform a task. This approach is especially advantageous for refining the model's understanding and increasing its accuracy in specific applications, such as distinguishing nuanced sentiment expressions (Brown et al., 2020).

The utilization of **Prompts and Responses** with LLMs enables these learning paradigms to be applied effectively. By crafting prompts that guide the model towards the desired output, researchers can leverage the LLM's capabilities for sentiment analysis. This involves providing a prompt that clearly states the task, such as identifying the sentiment of a given text, and then allowing the model to generate a response based on its training and the context provided by the prompt. Such an approach has been instrumental in harnessing the power of LLMs for detailed sentiment analysis, offering a flexible and efficient method for analysing sentiment across diverse datasets (Amatriain, 2024).

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 proposed solution entails the creation of a synthetic dataset comprising synsets annotated by a LLM. The primary motivation behind this approach is to enhance the existing sentiment values in Serbian WordNet, particularly targeting synsets whose corresponding words are identified with specific polarity in a sentiment lexicon derived from Serbian corpora but are mapped as purely objective in sentiment.

To achieve a balanced sample conducive to effective evaluation, especially in the later application of machine learning models for sentiment classification, a set of 500 synsets was randomly selected and processed using the Mistral model. This initial processing aimed to categorize the synsets into positive, negative, and objective sentiment groups. Subsequently, an equal number of synsets from each sentiment category were chosen for finer gradation.

The results underwent a manual annotation process, where each synset, along with the values returned by the LLM, was assessed and assigned a simple 'pass' or 'fail' grade based on their alignment with the expected sentiment annotations.

Originally, the methodology was designed to incorporate a few-shot learning approach for fine sentiment gradation, utilizing examples from synsets not selected for the primary dataset—specifically, those synsets that exhibited consistent sentiment values across both Serbian and English SWN. However, the preliminary results obtained through the zero-shot approach were found to be sufficiently satisfactory, rendering the few-shot component unnecessary. Consequently, the study proceeded exclusively with the zero-shot learning paradigm, where the Mistral model was applied without prior examples specific to the task of sentiment analysis.

Manual annotation of the outputs confirmed the validity of this streamlined approach. The zero-shot methodology demonstrated a remarkable success rate of over ninety percent, affirmatively showing that even without the inclusion of few-shot learning and the additional context it provides, the LLM could effectively discern and classify sentiment within the selected Serbian synsets.

The LLM primary used in this research is **Mistral 7B – Instruct.** That fine tuned variant of **Mistral 7B**, a 7-billion-parameter language model designed for superior performance and efficiency. It outperforms the **Llama 2 13B** model across various benchmarks. Notably, it surpasses **Llama 1 34B** in reasoning, mathematics, and code generation (Jiang et al., 2023).

The variant used here, **Mistral 7B – Instruct,** also outperforms the **Llama 2 13B – Chat** model on both human and automated benchmarks (Jiang et al., 2023).

Released under the Apache 2.0 license, the model offers a flexible tool for researchers, with the capability of being run locally without incurring costs (Jiang et al., 2023).

# 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.

For the sake of creating balanced set of samples, definitions from random sample of 500 synsets were processed using LangChain Python library, a powerful tool for creating, experimenting with, and analysing language models and agents (Chase, 2022).

The LangChain Python library serves as wrapper for several LLM modules. In this work this Hugging Fade Hub and Transformer.

The chain used in this research was simple containing of just one prompt template with one input variable – the definition of synset, and Mistral 7B – Instruct model downloaded from Hugging Face Hub. The model was executed on local machine.

The LangChain language chain allow for prompt template with variables, which are marked by curly brackets, to be invoked with values of those variables, sent to LMM module, returning the response.

Using appropriate prompt as shown, the sample was divided into those marked positive, negative, objective and those not properly marked. There was 290 objective, 102 negative, 33 positive and 75 errors.

The prompt used for classification was refined through experimental testing on a smaller subset of synset definitions. Comparative analyses of prompt instruction texts in English and Serbian revealed that Serbian-language prompt instruction yielded more accurate results. To ensure comprehensive coverage of potential responses, the number of output tokens was set to five.

Further examination had shown that duplicates exist within the set, due some differing words from the lexicon refer to same synsets. After removing duplicates, 279 objective, 97 negative and 27 positive. At this juncture that was no reason to count improperly marked.

To facilitate more detailed sentiment analysis, a random subset comprising 25 synsets from each sentiment category was selected, forming what is referred to as the "balanced sample."

*Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite njegov sentiment.*

*Sentiment treba da bude striktno klasifikovan kao "pozitivan", "negativan", ili "objektivan". Nijedan drugi odgovor neće biti prihvaćen.*

*Tekst: {text}*

*Sentiment:*

Prompt template for determining polarity

The balanced sample was processed through two prompt templates for greater sensitivity, one for positive and one for negative. The prompt templates were experimentally designed, by trying several possibilities, and incremental changes on few small set of synset definition from Serbian Wordnet.

The singular prompt for determining both was not chosen to keep with structure of SWN, where a synset can have both POS (positive value) and NEG (negative value) above zero, as long their sum is lesser or equal to one.

During this phase other LMM models were tested. Idea was to compare differing results, simulating annotation by multiple annotators, but results lead to exclusion from further work. Mitral had proved that much better.

To accommodate the anticipated length of the output strings, the number of output tokens was increased to nine.

*Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite da li ima pozitivan sentiment.*

*Sentiment treba da bude striktno klasifikovan kao "nije pozitivan", "slabo pozitivan", "umereno pozitivan", "veoma pozitivan", ili "ekstremno pozitivan". Nijedan drugi odgovor neće biti prihvaćen.*

*Nijedan drugi odgovor neće biti prihvaćen.*

*Tekst: {text}*

*Pozitivan sentiment:*

Prompt template for fine marking of positive sentiment

*Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite da li ima negativan sentiment.*

*Sentiment treba da bude striktno klasifikovan kao "nije negativan", "slabo negativan", "umereno negativan", "veoma negativan", ili "ekstremno negativan". Nijedan drugi odgovor neće biti prihvaćen.*

*Nijedan drugi odgovor neće biti prihvaćen.*

*Tekst: {text}*

*Negativan sentiment:*

Prompt template for fine marking of negative sentiment

For each definition in balances set, two values were assigned in that way. In intention of prompt templates design limits answers to set of responses.

For positive:

* "nije pozitivan" (not positive)
* "slabo pozitivan" (slightly positive)
* "umereno pozitivan" (moderately positive)
* "veoma pozitivan" (very positive)
* "ekstremno pozitivan" (extremely positive)

For negative:

* "nije negativan" (not negative)
* "slabo negativan" (slightly negative)
* "umereno negativan" (moderately negative)
* "veoma negativan" (very negative)
* "ekstremno negativan" (extremely negative)

Top of Form

Resulting dataset is stored in form coma separated values file[[2]](#footnote-2).

Columns in that file are as follows:

* **ILI**: Inter-Lingual Index, which connects the synset to its equivalents in other languages' WordNets.
* **Definition**: The gloss of the synset, providing its meaning or explanation.
* **Lemma\_names**: The lemmas (base forms) of the words contained within the synset.
* **Sentiment\_SWN**: The sentiment value mapped from SentiWordNet (SWN), indicating the original sentiment score in the English version.
* **Sentiment\_lexicon**: The sentiment value derived from a sentiment lexicon created for the Serbian language, reflecting local sentiment nuances.
* **Sentiment\_sa**: The initial classification by the Large Language Model, categorizing the sentiment as positive, negative, or neutral.
* **Sentiment\_sa\_positive**: Fine-grained sentiment classification for positive sentiment, indicating the degree of positivity ranging from "not positive" to "extremely positive."
* **Sentiment\_sa\_negative**: Fine-grained sentiment classification for negative sentiment, indicating the degree of negativity ranging from "not negative" to "extremely negative."

Due the limited number of samples, 75 in total, it was possible to preform manual evaluation on the whole set. This assessment was facilitated using Visual Studio Code's Data Wrangler extension, a tool that enabled a detailed comparison between the synset definitions and their corresponding fine-grained sentiment responses.

# Results

The summary of outputs generated by LMM using prompt templates for fine grading of sentient are shown on Table 1 and 2.

|  |  |
| --- | --- |
| **Row Labels** | **Count of sentiment\_sa\_negative** |
| ekstremno negativan | 1 |
| Nije negativan | 60 |
| Umereno negativan | 1 |
| umjereno negativan | 2 |
| veoma negativan | 11 |
| **Grand Total** | **75** |

Table 1. Negative sentiment

|  |  |
| --- | --- |
| **Row Labels** | **Count of sentiment\_sa\_positive** |
| Nije Ova iz | 1 |
| Nije pozitivan | 39 |
| Nijedan O | 6 |
| Nijedan Tekst | 1 |
| Učimo se držati | 1 |
| Umereno pozitivan | 3 |
| Umereno pozitivan The | 2 |
| Umjereno pozitivan | 6 |
| veoma pozitivan | 12 |
| Veoma pozitivan O | 1 |
| Veoma pozitivan R | 2 |
| Veoma pozitivan Ov | 1 |
| **Grand Total** | **75** |

Table 2. Positive sentiment

While output of LMM were not limited as suggested in prompt template that within easy correctable limits. Output generated from negative template prompt is more alike to suggested, with just addition of dialects. While outputs from positive template prompt contain one nonsensical answer “Učimo se držati“, and some that could be loosely inferred as not positive. For manual examination all of those were considered non positive. Also, minor typos, like extra letter, article or casing were ignored for manual examinations.

Furthermore, there no instances of ether slightly positive or slightly negative among the outputs.

For content there were two synsets clearly incorrectly marked, and three that are questionable.

Cleary incorrect are:

* **BILI-00000941**, synset meaning specific kind of mourning, characteristic for Serbia, involving loud wailing and sombre singing. It is marked as purely objective (not positive not negative), white it is obviously negative sentiment, even strongly negative.
* **ENG30-04525038-n**, synset meaning corduroy, which is marked a mildly positive. As type of textile, it should be objective.

For that are questionable are **ENG30-01215137-v** (“arrested”) as mildly negative, **ENG30-00309647-n** (“expedition”) as very positive and **ENG30-03135152-n** (“Christian cross”) as very positive.

Given the substantial number of synsets accurately labelled throughout our sentiment analysis process, we present a selection of illustrative examples below. These examples are intended to demonstrate the criteria for what is considered correct labelling in the sentiment classifications of positive, negative, and neutral sentiments.

Examples:

* **ENG30-01220336-n**: Synset associated with actions like 'kleveta' (defamation), 'klevetanje' (slander), and 'omalovažavanje' (disparagement), described as "oštar napad na čiju ličnost ili dobro ime" (a harsh attack on someone's personality or good name). This synset's sentiment was finely graded as "Nije pozitivan" (Not positive), reflecting the absence of positive sentiment, and more precisely, "Ekstremno negativan" (Extremely negative), accurately capturing the negative connotations of the described actions.
* **ENG30-06828389-n**: This synset refers to "karakter nalik na zvezdu (\*) koji se koristi u štampanim tekstovima" (a character resembling a star (\*) used in printed texts), with lemma names including 'asterisk', 'zvezdica' (little star), and 'zvezda' (star). The sentiment for this synset was precisely classified as "Nije pozitivan" (Not positive) and "Nije negativan" (Not negative), indicating its objective nature without inherent positive or negative sentiment.
* **ENG30-10407310-n**: Pertains to "onaj koji voli i brini svoju zemlju" (one who loves and defends their country), with lemma names including 'domoljub' (patriot), 'patriota' (patriot), and 'rodoljub' (patriot). This synset's sentiment was finely graded as "Veoma pozitivan" (Very positive) and "Nije negativan" (Not negative), effectively capturing the positive connotations associated with patriotism.

# Conclusion

The analysis of the output generated by the LMM, specifically the Mistral model, indicates that 70 out of 75 responses met the acceptability criteria defined for this study. This outcome, representing a substantial majority of the dataset, underscores the Mistral model's reliability and efficacy in performing sentiment classification for the Serbian language. With an approximate success rate of 93.3%, it can be confidently concluded that the dataset is both workable and of sufficient quality for the intended task of evaluating proposed corrections to sentiment polarity within the Serbian WordNet.

A significant area for future investigation involves the augmentation of prompt templates with specific examples, transitioning the current zero-shot learning approach to a few-shot learning paradigm. This modification is anticipated to refine the model's capacity for nuanced sentiment gradation, offering a more precise and contextually aware analysis. It is particularly recommended that this enhancement be applied selectively to the fine sentiment graduation templates, where the potential for increased accuracy and sensitivity to linguistic subtleties is most pronounced (Brown et al., 2020).

Further, the exploration of advanced prompting methodologies, such as the "chain-of-thought" approach, warrants attention. This technique, by facilitating a more structured and logical progression of thought within the model's processing, may significantly improve the model's ability to extract relevant data from responses. The potential of such advanced prompting strategies to overcome the inherent challenges in accurately identifying and classifying sentiment expressions within complex linguistic contexts presents a promising avenue for research (Amatriain, 2024).

This study serves as a "proof-of-concept" for utilizing the Mistral model to generate additional synthetic datasets for Serbian, particularly in areas where resources are scarce or challenging to acquire from natural corpora. The morphological richness of the Serbian language, characterized by its extensive inflectional system, poses a significant challenge in accurately capturing named entities across all possible flexions. The successful application of the Mistral model for sentiment analysis underscores its potential as a valuable tool in addressing these challenges.

# References

Amatriain, X. (2024). *Prompt Design and Engineering: Introduction and Advanced Methods* (arXiv:2401.14423). arXiv. https://doi.org/10.48550/arXiv.2401.14423

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

Chase, H. (2022). *LangChain* [Computer software]. https://github.com/langchain-ai/langchain

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

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B*.

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)
2. [https://github.com/sasa5linkar/SWN-synth-eval-set/blob/main/balanced\_[sample2](https://github.com/sasa5linkar/SWN-synth-eval-set/blob/main/balanced_sample2.csv).csv](https://github.com/sasa5linkar/SWN-synth-eval-set/blob/main/balanced_sample2.csv) [↑](#footnote-ref-2)