Baltic J. Modern Computing, Vol. 0 (2012), No. 0, pp. 00–00

Using Large Language Models to Improve Sentiment Analysis in Latvian Language

Pauls Purvins[1], Evalds Urtans[2], Vairis Caune[2]

[1] University of Latvia, Riga, Latvia
[2] Ventspils University of Applied Sciences, Department of Computer Science, Ventspils, Latvia
me@puupuls.lv, evalds.urtans@venta.lv, vairis.caune@venta.lv

Abstract. This empirical study explores the use of large language models (LLMs) in sentiment analysis and presents a new approach to creating a dataset in Latvian language using Reddit data. Using prompt engineering for the GPT-3.5-turbo model (latest at the time of writing), we achieved 82% accuracy that exceeds previous research on Latvian Tweet Sentiment Corpus by 50% in three class sentiment classification. We also demonstrate that LLMs can partially replace human labelers, making data set creation more cost-effective, especially for larger datasets. This work contributes to sentiment analysis in non-English languages, leveraging the power of LLMs. The paper introduces a new LVReddit dataset that contains more than 90000 samples, making it the largest available sentiment dataset for the Latvian language. Our findings confirm the LLM's underlying "understanding" of language. However, LLMs occasionally deviate from response templates, making parsing challenging. Future research should investigate fine-tuned models based on novel datasets and analyze language patterns.

Keywords: Large Language Models, Sentiment Analysis, Dataset creation, Latvian Language, Deep Learning, ChatGPT, Prompt Engineering

1 Introduction

Introducing a new approach to sentiment analysis in the Latvian language, this scientific article explores the use of large language models (LLM) and prompt engineering to create a cost-effective method for generating data sets. The study shows that LLM and prompt engineering can partially replace human annotators, making the creation of large data sets more economically viable. The validation data set achieved an accuracy 82% with the zero shot method by developing prompts for the GPT-3.5-turbo model (latest model at the time of writing), which improved the previous accuracy more than two times in the sentiment analysis of three classes in this data set. This work contributes to the development of sentiment analysis using LLM capabilities and examines prompt engineering tasks in Latvian language data processing. The created LVReddit

data set contains more than 90,000 samples and is published on the GitHub repository [3], becoming the largest open data set available for sentiment analysis in the Latvian language.

2 Related work

There has been previous work on creating labeled sentiment datasets for Latvian language, mostly in the bachelor and master thesis of the students, and it is listed in Table 1. Most popular datasets in English, Lithuanian, and Estonian are listed in Table 2.

Table 1: Latvian sentiment datasets by size

Dataset	Size	Positive samples	Neutral samples	Negative samples	Data source
latvian-tweet-sentiment-corpus Peisenieks and Skadins (2014)	1177	383	627	167	Twitter
LV-twitter-sentiment-corpus Nicmanis and Paikens (2017)	2272	797	1223	252	Twitter
LV-twitter-eater human labeled Sproģis and Rikters (2020)	5420	1631	2507	1282	Twitter
LV-twitter-eater automatically labeled Sproģis and Rikters (2020)	18130	2976	14926	228	Twitter
sikzinu-analize dataset Vīksna (2018)	3682	935	2208	539	Twitter
OM dataset Spats and Birzniece (2016)	6227	3104	2617	506	Twitter
LVReddit (This work)	91821	9200	40076	42545	Reddit

Prompt engineering is a new field of study that has emerged together with recent advances in large language models. Given that this field is so new and at the same time so popular, multiple studies have explored the task of prompt engineering, including (White et al. (2023); Kojima et al. (2022); Zhou et al. (2022)).

On the topic of sentiment analysis and prompt engineering, Wang et al. (2023) explored a similar task to this study but focused on the English language, achieving similar accuracy to the one shown in this study.

3 Methodology

This section describes the prompt engineering process, metrics used, data mining, and sample labeling process.

[3] https://github.com/Puupuls/LVRedditCorpus

Table 2: Other popular sentiment datasets

Dataset	Size	Positive samples	Neutral samples	Negative samples	Data source
Stanford Sentiment Treebank Socher et al. (2013)	10662	5331	0	5331	Movie reviews
IMDb Movie Reviews Maas et al. (2011)	50000	25000	0	25000	Movie reviews
MPQA Opinion Corpus Wiebe et al. (2005)	10657	859	7353	2412	News
Lithuanian Internet comment dataset Kapoiūtė-Dzikienė et al. (2019)	10570	2176	1873	6521	LT Internet comments
Estonian Valence Corpus Pajupuu et al. (2016)	4086	-	-	-	Internet comments and news

3.1 Classification methods

Sentiment analysis can be done in multiple ways:

1. One sentiment per sample (Document or sentence)
   Detect one sentiment per sample. "Today is a nice and sunny day but I don't like how windy it is." could be classified both as positive or negative, but it is difficult to capture nuances in cases when multiple opinions are expressed. Taboada (2016)
   Some of the most popular implementations include
   (a) Lexicon-based
   Uses a lexicon that contains words that express positive or negative sentiment and classifies based on word frequencies in classifiable text. It is easy to implement but does not account for words that have different meanings based on context.Taboada (2016)
   (b) Machine Learning-based
   Using classical machine learning techniques like SVMs. Build pretty robust solutions comparatively easily, as most libraries nowadays include easy-to-use implementations. Requires a sufficient amount of data to learn from.Taboada (2016)
   (c) Deep learning-based
   Uses a deep learning model, classically RNNs, but since the rise of transformer architecture, similar to most natural language tasks, it uses them. This method is a lot harder to implement but it can build a deeper "understanding" of language given a sufficient amount of data
   (d) LLM prompt-based
   The novel and not much-explored method uses large language model prompts to classify samples by leveraging the deep language "understanding" of large language models. Requires access to a large language model and systematic prompt engineering to achieve best results.
   Can be either zero-shot or few-shot depending on the used prompt meaning that prompts can include some samples and expected outputs (few-shot) to guide it.

2. Aspect level analysis.
   Separate sentiment instances are analyzed. Can more accurately classify texts like "Today is a nice and sunny day but I don't like how windy it is." by classifying positive sentiment about a sunny day but negative about the wind. Schouten and Frasincar (2016)

3.2 Prompt engineering

Prompt engineering is the process of designing and analyzing various prompts for large language models, with the aim of finding prompts that yield desired results. This process is analogous to the training of neural networks, where input and output are known, and the weights of the network are optimized.

In prompt engineering, instead of optimizing weights, the focus is on optimizing the prompt itself, which enables the language model to transform given input data into the desired output.

The process of prompt engineering was carried out in an iterative manner, evaluating the prompts using the latvian-tweet-sentiment-corpus dataset Peisenieks and Skadins (2014). The evaluation was performed on a validation subset of the dataset, which was created by selecting an equal number of examples from each class, reducing the dataset from 1177 to 501 examples.

Drawing on recent research in the field White et al. (2023) Kojima et al. (2022) Zhou et al. (2022) Wang et al. (2023), seven prompts in the English language were developed, and their performance was evaluated in the validation set.

These prompts, together with their results, were given to the ChatGPT tool with the prompt "Suggest better prompts that might improve 3 class classification accuracy." which gave 6 prompts in English and 3 in Latvian.

These prompts were also evaluated, and this automatic enhancement process was performed one additional time, but no improvement in resulting metrics was noticed.

This process resulted in 24 evaluated prompts.

3.3 Metrics

In this research accuracy (shown in Equation 1) was used as the main metric for the evaluation of the training results that are consistent with existing works in the domain Garkaje et al. (2014), Gediņš and Paikens (2013), Gulbinskis and Šmite (2010), Nicmanis and Paikens (2017), Peisenieks and Skadins (2014), Pinnis (2018), Spats (2015), Spats and Birzniece (2016), Sproģis and Rikters (2020). Also while developing prompts balanced classes with F1 score (Equation 4), precision (Equation 2), and recall (Equation 3) were used.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precission = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times precision \times recall}{precission + recall} \tag{4}$$

3.4 Data scraping

For the dataset source, the Latvian subforum of Reddit [4] was chosen since all previous data sets use Twitter and news portals. As Reddit's official API imposes limitations on max retrieved data and does not return more than the last 1000 posts, the decision was made to use pushshift.io provided Reddit API [5] that does not have this limitation.

To filter out only Latvian posts, langdetect package for Python3 [6] was used. No other filtering or modification was done to the dataset.

Code that has been used for data-gathering is published and available together with the dataset as an open-source repository on GitHub [7].

3.5 Labeling

To label the data, every sample was combined with a prompt and sent to OpenAI API. The received response was parsed according to the expected results format from the prompt and sample labeled. If the result could not be parsed, the sample was labeled as "Neutral" class as analyzing these responses and respective samples revealed that this mostly happened in cases where not enough information for sentiment classification was provided thus making the sample neither positive nor negative. This process is shown in Figure 1.

For validation, a human-labeled dataset was created by labeling a subset of the LVReddit dataset. Each prompt was labeled by 2 labelers. This subset was not used in prompt engineering to estimate the accuracy of the LVReddit dataset.

[4] `https://reddit.com/r/latvia`
[5] `https://github.com/pushshift/api`
[6] `https://pypi.org/project/langdetect/`
[7] `https://github.com/Puupuls/LVRedditCorpus`

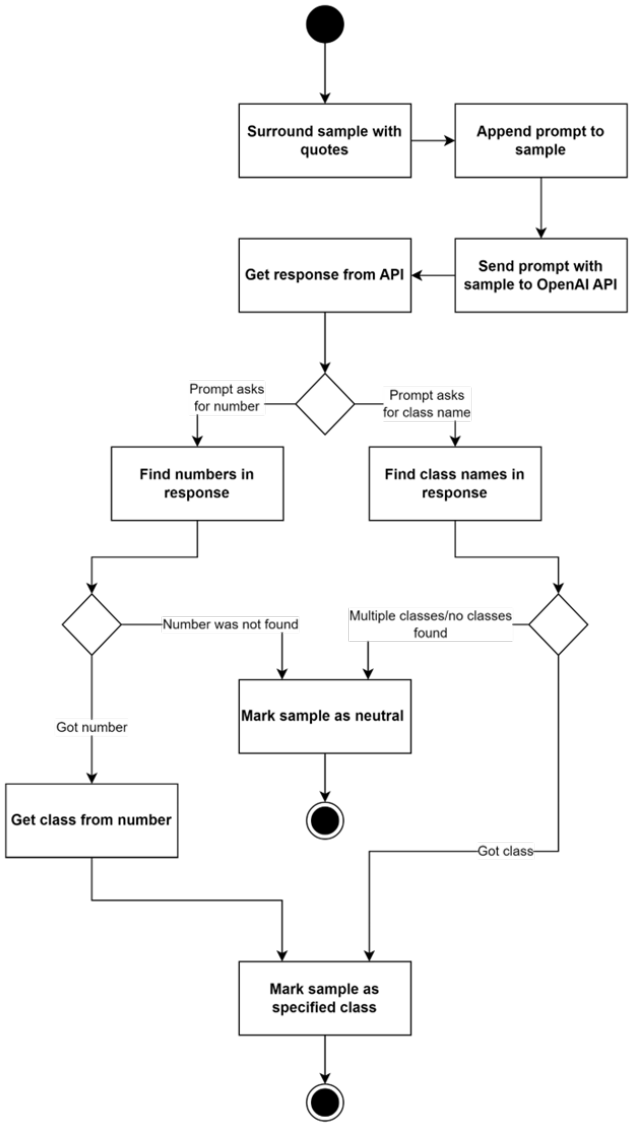

Fig. 1: Process of labeling sample form dataset

3.6 Dataset

The developed dataset comprises 3028 articles in the Latvian language, accompanied by 88793 comments also in Latvian. From now on, this data set will be known as the LVReddit data set. A comparison of its size with previous datasets is depicted in Figure 2. The LVReddit dataset is more than five times larger than the closest automatically

labeled dataset that was labeled using a lexicon-based approach and more than 14 times larger than the closest human-labeled dataset for the Latvian language.

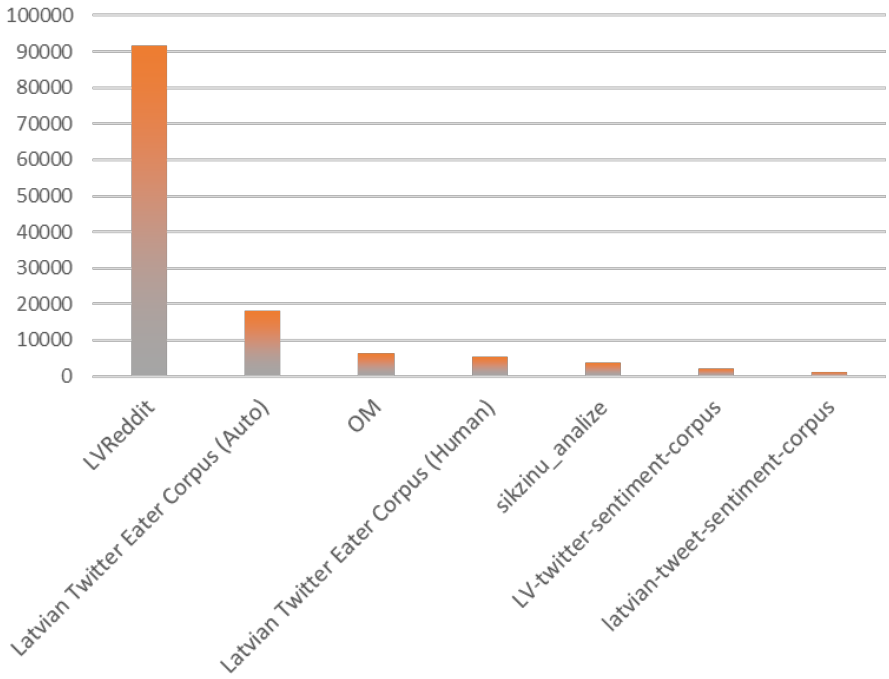

Fig. 2: Dataset size comparison with previous datasets

In total, the data set contains 91821 samples of which more than 40000 have neutral and negative sentiment as shown in Figure 3

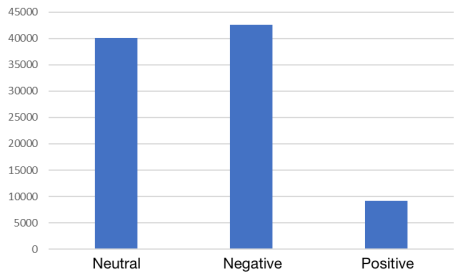

Fig. 3: Number of samples in LVReddit dataset for each class

The dataset consists of unfiltered and unmodified samples together with metadata. A subset of this data set was labeled by 2 human annotators as a validation data set for LVReddit itself. It contains 150 samples from each class, where the label matches between the two human annotators. Prompt engineering was performed on the complete data set excluding samples from the evaluation dataset.

4 Results

The top prompt achieved 82.0% accuracy in the evaluation dataset, and the top 5 prompts are listed in Table 3. The best prompt in the Latvian language achieved 78.2% accuracy. The confusion matrix for the best prompt is shown in Figure 4. F1 score for it is 0.8131.

Table 3: Top 5 prompts of 24 prompts tested and accuracies

Prompt	Accuracy
Based on the tone of the text, what is your overall impression? Choose one of the following: Positive, Negative, or Neutral.	82.0%
What is the general sentiment of this sentence? Choose one of the following: Positive, Negative, or Neutral.	79.0%
Balstoties uz teksta toni, kāda ir kopējā noskaņa - pozitīva, neitrāla vai negatīva?	78.2%
Based on the tone of the text, would you categorize the overall sentiment as positive, neutral, or negative?	78.0%
Does the author's language in this sentence indicate a positive, neutral, or negative sentiment?	77.4%

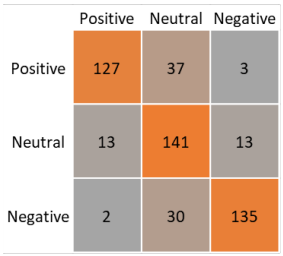

Fig. 4: Confusion matrix for the best-performing prompt on the evaluation data set.

In the human-labeled subset of the LVReddit dataset, the best prompt achieved a precision of 70.4% while the inter-human precision was 74.0%. The table 4 shows the accuracies achieved on previously existing Latvian datasets achieving new best results in 4 out of 6 datasets.

Table 4: Accuracy of sentiment analysis compared to previous work

Dataset	LLM prompt	OM lexicon	Published result
LVReddit	**70,4**%*	43,0%*	-
latvian-tweet-sentiment-corpus	**82,0**%	60,4%	35,5%
LV-twitter-sentiment-corpus	**62,2**%	53,4%	-
LV-twitter-eater val.	**65,1**%	49,7%	53,9%
sikzinu_analize	62,0%	54,0%	**72,6**%
OM	72,6%	**73**%	62%

* Using validation subset

5 Further research

In the realm of language modeling, there exist several intriguing avenues for future exploration. One such direction involves leveraging the rich LVReddit data set and the advanced LVBERT (Znotins and Barzdins (2020)) model to drive further advances in natural language understanding. In addition, an area of interest lies in the development of classification heads specifically tailored for the GPT model, enhancing its ability to perform fine-grained tasks. Delving deeper, the investigation of various language-specific features and their influence on model performance offers valuable insights for refining language models. Moreover, exploring the accuracy of multi-turn interactions within these models holds promise for more comprehensive and context-aware responses. Furthermore, the integration of other state-of-the-art language models such as BLOOM, LLaMA, or GPT-4 could improve the results. Lastly, the application of language models for aspect-level sentiment analysis represents a fascinating domain where their capabilities can be harnessed to extract sentiment nuances at a granular level. These potential research avenues offer exciting prospects to advance the field of language modeling and its practical applications.

6 Conclusions

Large language models are applicable to various tasks and offer significant improvements in natural language processing, even for languages other than English. Comparing the results with previous findings, the GPT-3.5-turbo model (latest model at the time of writing) using zero-shot prompting demonstrates a notable enhancement in sentiment analysis of three classes, achieving a 82% accuracy in the latvian-tweet-sentiment-corpus dataset. This surpasses the 76% accuracy obtained by Jānis Peisenieks (2014) for binary analysis and the accuracy 35. 5% for three-class sentiment analysis.

To further improve the accuracy of the results, a specially designed model with answer classification could be used, as discussed by Wang et al. (2023), which was not explored within the scope of this study.

By comparing the prompts in Latvian and English, the Latvian prompts produced comparable or sometimes even better results than the English prompts. These findings can be attributed to the language proficiency of the model, allowing it to operate effectively in a specific language. This similarity in results demonstrates the model's ability to accurately process and analyze language, regardless of the specific language used.

The development of language models facilitates the creation of new datasets, enabling the construction of labeled datasets with human-level accuracy at significantly lower costs. On platforms such as Upwork, where individuals can be hired for various tasks, the hourly rate for labeling can range from $10 to $25. Similarly, on Amazon Mechanical Turk[8], the label can start at 2 cents per example (resulting in a cost of $1800 for labeling the entire dataset). Labeling all LVReddit data sets using the gpt-3.5-turbo model cost approximately $35.

Comparing the results obtained through the prompts with the results of the previous dataset, the language model's prompts outperformed the previous results in three out of five datasets. In two datasets, the improvement was lower, at 0.4% and 10.6%, respectively.

The differences observed in different data sets could be attributed to variations in the length and content of the messages, different data selection approaches, and variations in the methodology to determine the ground truth values.

References

Garkaje, G., Zilgalve, E., Dargis, R. (2014). Normalization and automatized sentiment analysis of contemporary online latvian language, *Baltic HLT*, Vol. 268, pp. 83–86.

Gediņš, K., Paikens, P. (2013). Automātiskā teksta emocionālās noskaņas noteikšana latviešu valodā, *LU Archive* .
`https://dspace.lu.lv/dspace/handle/7/21072`

Gulbinskis, I., Šmite, D. (2010). Digitālo tekstu sentimenta analīze, *LU Archive* .
`https://dspace.lu.lv/dspace/handle/7/18978`

Kapoiūtė-Dzikienė, J., Damaeviius, R., Woźniak, M. (2019). Sentiment analysis of lithuanian texts using traditional and deep learning approaches, *Comput.* **8**, 4.
`https://api.semanticscholar.org/CorpusID:59342838`

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., Iwasawa, Y. (2022). Large language models are zero-shot reasoners, *ArXiv* **abs/2205.11916**.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C. (2011). Learning word vectors for sentiment analysis, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 142–150.
`http://www.aclweb.org/anthology/P11-1015`

Nicmanis, D., Paikens, P. (2017). Sabiedrības attieksmes modelēšana, izmantojot sentimenta analīzi, *LU Archive* .
`https://dspace.lu.lv/dspace/handle/7/35260`

Pajupuu, H., Altrov, R., Pajupuu, J. (2016). Identifying polarity in different text types.
`https://api.semanticscholar.org/CorpusID:201818495`

Peisenieks, J., Skadins, R. (2014). Uses of machine translation in the sentiment analysis of tweets, *Baltic HLT*, Vol. 268, pp. 126–131.

Pinnis, M. (2018). Latvian tweet corpus and investigation of sentiment analysis for latvian, *Baltic HLT*.

Schouten, K., Frasincar, F. (2016). Survey on aspect-level sentiment analysis, *IEEE Transactions on Knowledge and Data Engineering* **28**, 813–830.

[8] https://www.mturk.com

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank, *Conference on Empirical Methods in Natural Language Processing*.
`https://api.semanticscholar.org/CorpusID:990233`

Spats, G. (2015). Application of opinion mining for written content classification in latvian text, *RTU Noslēgumu darbu reģistrs* .

Spats, G., Birzniece, I. (2016). Opinion mining in latvian text using semantic polarity analysis and machine learning approach, *Complex Syst. Informatics Model. Q.* **7**, 51–59.

Sproģis, U., Rikters, M. (2020). What can we learn from almost a decade of food tweets, *Baltic HLT*.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics.

Vīksna, R. (2018). Emocionālās ekspresijas noteikšana sīkziņās latviešu valodā, *RTU Noslēgumu darbu reģistrs* .
`https://github.com/RinaldsViksna/sikzinu_analize`

Wang, Z., Xie, Q., Ding, Z., Feng, Y., Xia, R. (2023). Is chatgpt a good sentiment analyzer? a preliminary study, *ArXiv* **abs/2304.04339**.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt, *ArXiv* **abs/2302.11382**.

Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation* **39**, 165–210.
`https://api.semanticscholar.org/CorpusID:382842`

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J. (2022). Large language models are human-level prompt engineers, *ArXiv* **abs/2211.01910**.

Znotins, A., Barzdins, G. (2020). Lvbert: Transformer-based model for latvian language understanding, *Baltic HLT*.