Developing a Lithuanian dataset of malicious social media communications

Edgaras Baranauskas, Jevgenijus Toldinas

Computer Department, Kaunas University of Technology, Studentų 50, Kaunas, Lithuania

Aim of work

Creating a dataset of malicious social media interactions in Lithuanian necessitates considering privacy, linguistic variety, and annotation quality. When collected messages contain personal data, they must be processed in accordance with the GDPR, which includes an acceptable legal basis and protections for scientific research. The Lithuanian dataset containing more than 600 labelled messages, approximately 30% of which were neutral was created by Čepulionytė [1]. Four non-exclusive labels represented aggressive, insulting, toxic, and malicious content. The dataset description lacks a detailed account of source selection and annotator agreement. The Common Language Resources and Technology Infrastructure CLARIN-LT [2] brings together language-oriented resources, advanced tools, and their metadata at both national and European levels, focusing primarily on resources related to the Lithuanian language. Lithuanian is a language characterized by extensive inflection and morphological complexity, which complicates text classification compared to English. Non-standard orthography, mixed Cyrillic/Latin characters, or internet vernacular that evades traditional morphological tools like Morfuoklis [3] are frequently used in malicious communication on Lithuanian social networks. Skaržauskienė [4] collected 7,550 messages from 151 public Telegram channels, comprising original material in both English and Lithuanian. The process of channel selection included keyword searches, snowball sampling through shared links, and links found within Facebook groups. Three autonomous coders categorized the messages utilizing a four-dimensional framework and a specified codebook. Roughly 5% remained indeterminate owing to vagueness or insufficient context, underscoring the necessity to document ambiguous situations. To address the issues mentioned above, we suggest using OpenAI to generate a corpus of synthetic datasets in Lithuanian. We created several prompt scripts that then were used to ask OpenAI to generate dataset of ham and malicious social media Lithuanian messages in accordance with the GDPR. Finally, we developed the prompt intended to accelerate the development of a corpus of Lithuanian datasets focused on malicious communications on social media. In our experiments Claude, ChatGPT, Copilot, and Grok OpenAI tools were used. We utilize machine learning algorithms to develop and assess models on a per-dataset basis from the corpus. The final experiment summarizes obtained results. The models were evaluated and compared using machine learning model metrics.

References

[1] Čepulionytė, A.; Toldinas, J.; Lozinskis, B. A Multilayered Preprocessing Approach for Recognition and Classification of Malicious Social Network Messages. Electronics 2023, 12, 3785. https://doi.org/10.3390/electronics12183785

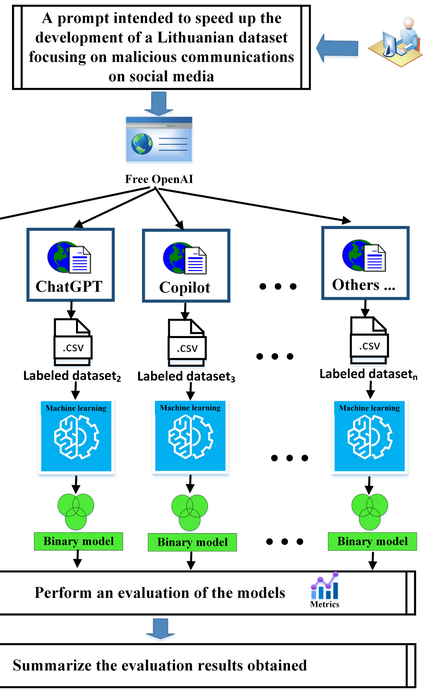
[2] CLARIN-LT repository. Hosted at CLARIN-LT, the Republic of Lithuania. Available online: https://clarin-repo.lt/info/legal (accessed on 7 10 2026).

[3] An online tool for morphological analysis and synthesis of the Lithuanian language. Available online: https://sitti.vdu.lt/morfuoklis/lt/apie (accessed on 7 10 2026).

[4] Skaržauskienė, A., Mačiulienė, M., Gulevičiūtė, G., Zelenkauskaitė, A., Diržytė, A., & D'Antonio Maceiras, S. (2026). Malicious creativity in Telegram's (software) anti-vaccination ecosystem: profiling actors and early misinformation dynamics. Creativity Studies, 19(1), 132–141. https://doi.org/10.3846/cs.2026.22247

Using OpenAI to generate a corpus of synthetic datasets in Lithuanian

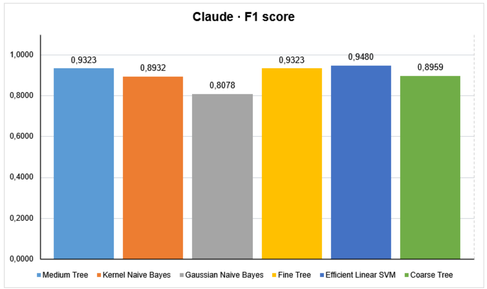
Lithuanian is a language characterized by extensive inflection and morphological complexity, which complicates text classification compared to English. Non-standard orthography, mixed Cyrillic/Latin characters, are frequently used in malicious communication on Lithuanian social networks. To address the issues mentioned above, we suggest using OpenAI to generate a corpus of synthetic datasets in Lithuanian as follows.

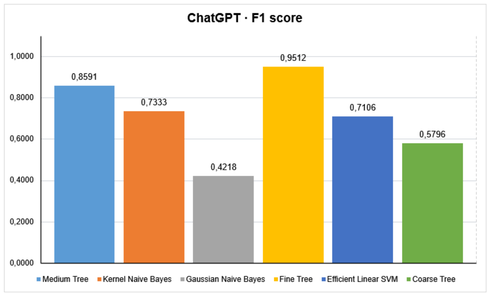


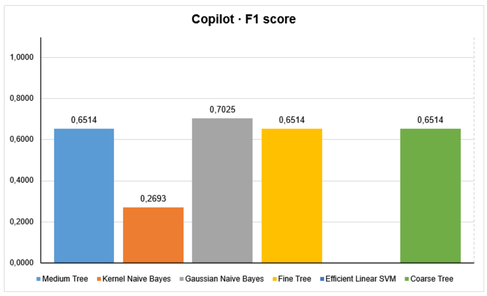
Parameters of the generated datasets

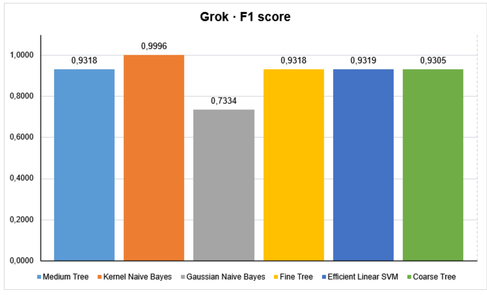
Open AI	Number of records			Percent	
	Total	Scam	Ham	Scam	Ham
Claude	50000	22500	27500	45,00%	55,00%
ChatGPT	50000	5000	45000	10,00%	90,00%
Copilot	50000	4959	45041	9,92%	90,08%
Grok	50000	19000	31000	38,00%	62,00%

Experimental results









Experimental results

