Troubling Times for PhD Research on Text Categorization? ChatGPT for Automatic Genre Identification

Taja Kuzman[1,2][0000−0001−7436−9896]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
taja.kuzman@ijs.si

Abstract. Instruction-tuned GPT models have shown strong capabilities in natural language generation tasks, which naturally leads researchers to explore where their abilities end. In this paper, we examine whether they can be used for a text classification task, more specifically, automatic genre identification. We perform preliminary experiments by prompting a GPT-3.5 model through the ChatGPT web platform and compare its performance with a multilingual XLM-RoBERTa language model that was fine-tuned on datasets, manually annotated with genres. The evaluation encompasses test sets in two languages: English and Slovenian. The results show that the ChatGPT model outperforms the fine-tuned model when applied to the dataset which was not seen before by either of the models. Notably, even when applied on Slovenian, an under-resourced language, its performance remains on par with that on English. Laborious manual annotation campaigns and fine-tuning models occupy a central place of many PhD theses on text categorization tasks. However, these efforts are not needed when using an instruction-based GPT model. Consequently, this paper encourages researchers in the text categorization field to evaluate their methodologies in light of the emerging GPT language models, as the rapid development of these technologies might require a shift in research direction in some future PhD studies.

Keywords: Automatic Genre Identification · Text Classification · ChatGPT · Fine-Tuned Language Models

1 Introduction

The conventional approach to text classification problems in the field of natural language processing (NLP) typically involves the development of supervised machine learning systems. This approach has consistently demonstrated superior performance in classification tasks, especially when deep neural models are used [13]. It entails a development of an appropriate label schema and annotation guidelines, followed by extensive annotation campaigns which involve the manual annotation of thousands of texts, preferably by multiple annotators. For a

manual annotation to be successful and produce reliable results, it requires a significant amount of time, human effort, and financial resources. As a result, the design and results of a manual annotation campaign often represent one of the primary contributions of a PhD research focused on a text categorization task. Another significant aspect of the work of many PhD students in this field involves the development of machine learning systems for the task, achieved through training and evaluating them on manually annotated datasets. This approach is also pursued in my own PhD research, which investigates automatic genre identification, a text categorization task that focuses on genres as text categories. As part of my research, we developed a novel genre schema to facilitate accurate manual annotation and enhance the performance of trained genre classifiers. The annotation process involved manually labeling over a thousand texts with genre labels. Subsequently, we conducted numerous experiments, employing various machine learning technologies and genre datasets, to construct an optimal genre classifier (see [11,9,10]). However, recent advancements in the field have introduced highly potent large language models that do not necessitate annotated training data for specific classification tasks. In the latter part of 2022, a new generation of instruction-tuned GPT (Generative Pre-Trained Transformer) language models, specifically the GPT-3.5 model family [16], became accessible through the ChatGPT web platform. Although these models are optimized for dialogue, preliminary research demonstrated their remarkable performance also in various natural language processing tasks, including stance detection [29], implicit hate speech detection [7] and machine translation [6]. In contrast, they were surpassed by fine-tuned BERT-like large language models for some other NLP tasks [20,30,26].

Following these findings, this paper examines the performance of ChatGPT in automatic genre identification, in comparison to the pre-trained XLM-RoBERTa model [3] that was fine-tuned for genre classification. The evaluation is conducted on two test sets: the English EN-GINCO dataset and the Slovenian GINCO dataset [11]. While the texts from the datasets are sourced from the web, their genre annotations have not been made available online. This ensures that the evaluation of the performance of the ChatGPT model is not influenced by the inclusion of genre annotations in the pre-training data, and that there is no data leakage from pre-training to the test datasets.

The experiments demonstrate an impressive performance of ChatGPT on this task, as it outperforms the fine-tuned large language model (LLM) on the English test set. This study not only provides initial insights into the performance of instruction-tuned GPT models on this task but also represents one of the first investigations of their text classification capabilities in a language other than English. Despite Slovenian being an under-resourced language, the findings indicate that ChatGPT's performance in this language is comparable to that in English. Considering the continuous emergence of newer and more advanced GPT models, it is reasonable to expect further improvements in their performance for this task. Consequently, we raise the question of whether annotation campaigns, annotated datasets, and trained models will become less relevant for

PhD studies focusing on text categorization tasks. We present these findings to encourage researchers working on similar tasks to conduct similar evaluations, which would enable them to gain insights into the performance of such models and potentially prompt them to slightly change their research direction.

The paper is structured as follows. In Section 2, an overview of the task of automatic genre identification is provided. Section 3 presents the ChatGPT model and the fine-tuned LLM classifier, referred to as the X-GENRE classifier. The genre-annotated test sets are described in Section 4, and the results of the model comparison are discussed in Section 5. Finally, Section 6 concludes the paper by discussing the main findings and providing suggestions for future research.

2 Automatic Genre Identification

Automatic genre identification is a text classification task which focuses on categorizing texts into genre categories, such as *News*, *Legal*, and *Promotion*. The genre categories are traditionally defined based on the author's purpose, as well as the function and conventional form of the text [18]. The development of genre classifiers has been motivated by three primary objectives. Firstly, it aims to provide search engine users with the ability to filter search results based on genre. Secondly, genre classification enables researchers to gain insights into the content and quality of large text collections that are automatically collected, by analyzing the distribution of genres within these collections. Lastly, genre information has been used to enhance other NLP tasks, as demonstrated in previous studies on language processing [14], machine translation [27], and automatic summarization [24].

This text classification task was found to be challenging for both humans [4,25] and machine learning models. Traditional non-neural methods did not achieve satisfactory results in this task. They were reported to heavily rely on a specific dataset used for training and to struggle to generalize to unseen datasets [23]. Recently, a breakthrough in this field occurred with the introduction of deep neural Transformer-based language models. These models demonstrated remarkable performance in automatic genre identification, even when applied to out-of-domain or cross-lingual scenarios [10,21]. However, the highest reported results in this task range from 0.68 to 0.76 in micro F1 score [12,11,21], while these machine learning models can achieve up to 0.99 accuracy for topic detection [13]. This discrepancy highlights a greater complexity of automatic genre identification compared to text categorization tasks that primarily rely on lexical features. Genre identification necessitates the identification of higher-level patterns, including syntactic characteristics, in addition to lexical features [9]. Moreover, despite the advancements in fine-tuned pre-trained Transformer-based language models, they still rely on manually-annotated datasets, which is a laborious and expensive process.

3 Models

3.1 X-GENRE Model – Fine-Tuned XLM-RoBERTa Model

We compare the ChatGPT model with a massively multilingual base-sized XLM-RoBERTa Transformer-based model [3], fine-tuned on genre-annotated datasets, hereinafter referred to as the X-GENRE classifier[3]. The classifier was fine-tuned on approximately 1,700 instances from three datasets, manually annotated with genre labels: English CORE [4], English FTD [22] and Slovenian GINCO [11] dataset. Each dataset has its own set of categories, which were then mapped into a unified schema – the X-GENRE schema. We chose to use multiple datasets instead of just one to ensure that the model can generalize well to unseen datasets and languages. This approach also helps to prevent that the model would be too dependent on a specific dataset, which was observed in previous studies [23].

3.2 ChatGPT Model

ChatGPT is an instruction-tuned GPT (Generative Pre-Trained Transformer) large language model, provided by the OpenAI through a web interface. It is based on the GPT-3.5 model [16]. The GPT-3.5 model was optimized for dialogue using a technique called the Reinforcement Learning with Human Feedback [2]. This approach involves leveraging human feedback to optimize the model's answers. Firstly, the GPT language model was pre-trained on a massive multilingual web text collection. The model was then fine-tuned to follow instructions using a dataset consisting of prompts and corresponding human-generated answers. Based on that, it was instructed to produce multiple answers to the prompts from the dataset, and its answers were evaluated by human annotators. This information was then used as a reward function to improve the model's performance – a reward function was trained to predict which of the outputs was rated the highest by the annotators. The reward function was further optimized with reinforcement learning using the proximal policy optimization algorithm [19]. At the time of conducting the experiments, the GPT-3.5 model was not accessible through an API. Instead, it could only be used via the ChatGPT web platform[4] which does not provide any insight into the hyperparameters used by the model. That is why we refer to the model as "ChatGPT" model, since we cannot access the GPT-3.5 model directly and we use the ChatGPT platform instead.

We use the ChatGPT Feb 13 version and conduct the experiments in the period from February 24th to March 2nd, 2023. To categorize the texts into genre categories with ChatGPT, we manually input prompts to the web platform and manually extract the corresponding categories and explanations from its answers. In the prompt, we specify the main criteria for defining genres and list the available categories for the model to choose from. In order to facilitate

[3] The X-GENRE classifier is freely available in the Hugging Face repository: https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier.

[4] https://chat.openai.com/chat

comparison between the two models, we adopt the same genre classes as those used by the X-GENRE classifier. Additionally, we request the model to provide an explanation for its choice. After these instructions, the second part of the prompt consists of the text to be classified. We use the same English prompt for each text, regardless of whether the text is in Slovenian or English language.

Example of the prompt: *Please classify the following text according to genre (defined by function of the text, author's purpose and form of the text) and explain your decision. You can choose from the following classes: News, Legal, Promotion, Opinion/Argumentation, Instruction, Information/Explanation, Prose/Lyrical, Forum, Other. The text to classify: Shower pods install in no time... 1. Prepare the floor with the waste and the water supply pipes. 2. Attach shower equipment to the shower pod shell running flexible tails (H&C or just C) down back. 3. Move unit into position connecting water supplies on the way and the waste outlet trap. 4. Having secured the shower pod shell to the building structure doors may now be fitted.*

4 Genre-annotated test sets

To evaluate the models' performance on English and Slovenian texts, we use random samples from two genre datasets that were manually annotated by us: the English EN-GINCO dataset and the Slovenian GINCO dataset. As the prompts have to be entered manually to the ChatGPT web platform, we restrict our experiments to using 100 randomly selected instances from each dataset. It is important to note that the annotations, used in these experiments, have not been published online, which ensures that the ChatGPT model did not see them during pre-training.

The GINCO dataset[5] [11] consists of 1002 Slovenian web texts originating from two Slovenian web corpora, the slWaC 2.0 corpus [5] from 2014 and the MaCoCu-sl 1.0 corpus [1] from 2021. These web corpora were created by crawling the Slovenian top-level web domain (`.si`) and connected websites from common domains (e.g., `.com`.) The dataset was manually annotated with 24 genre categories from a GINCO genre schema by two annotators with linguistic background who followed comprehensive guidelines for genre annotation[6] (see [11] for more details on the annotation procedure). The GINCO dataset is divided into training, evaluation and test splits. The training split was included in the training data for the X-GENRE classifier. Thus, to prevent any data leakage in our experiments, we use the test split to extract a random subset for evaluation.

The EN-GINCO dataset comprises 272 English texts from the English web corpus enTenTen20[7] [8]. The dataset was prepared to evaluate the performance of genre classifiers in the out-of-domain scenario and was therefore not incorporated into the training data for the X-GENRE classifier. The annotation process

[5] The GINCO dataset is available at http://hdl.handle.net/11356/1467.

[6] The annotation guidelines are available at https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

[7] https://www.sketchengine.eu/ententen-english-corpus

for the EN-GINCO dataset involved manual annotation by the same annotators and adhered to the same schema as the GINCO dataset.

Both datasets were initially annotated with the GINCO schema. However, previous work [11] showed that the performance of classifiers trained on the GINCO dataset can be further improved by reducing the size of the label set through category merging. That is why, in our experiments, we map the original GINCO categories to a new label set with a lower granularity – the X-GENRE genre schema. Another reason for using the X-GENRE schema is that this label set is also used by the X-GENRE classifier. The new schema is a generalization of various schemata applied to different datasets, that is, the CORE [4], FTD [22] and GINCO [11] dataset. The motivation behind this schema is two-fold – it is more user-friendly than any of specific schema in the available training datasets, and it allows for merging training data from different datasets, resulting in a more robust model. The final schema consists of 9 labels: *Information/Explanation, Instruction, Legal, News, Opinion/Argumentation, Promotion, Prose/Lyrical, Forum* and *Other*[8].

Table 1. Label distribution in the English test set (EN-GINCO), the Slovenian test set (GINCO) and the dataset, used for training the X-GENRE classifier.

Labels	EN-GINCO	GINCO	X-GENRE training
Information/Explanation	25%	24%	17%
Promotion	22%	17%	16%
Opinion/Argumentation	18%	11%	14%
News	18%	29%	19%
Other	6%	7%	4%
Forum	6%	5%	8%
Instruction	5%	5%	12%
Legal	0%	1%	4%
Prose/Lyrical	0%	1%	6%

The distribution of labels in each test set is presented in Table 1. It is evident that test sets are imbalanced, with four most frequent categories representing more than 80% of instances in each test set, and five less common categories each representing 7% or less of the set. Two categories from the genre schema, namely *Legal* and *Prose/Lyrical*, are present only in the Slovenian test data. As label distribution can impact the performance of the classifier, we also add into the comparison the distribution of genre classes in the dataset used to train the X-GENRE classifier. As shown in Table 1, the training data contains a higher representation of classes that are less frequent in the test sets and some of the most frequent categories in the test sets, such as *Information/Explanation* and *News*, constitute a smaller percentage of the training dataset compared to the test sets.

[8] For more details, see the definitions of the labels at https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier.

5 Results

In this section, we present a comparison between the performance of the ChatGPT model, which is guided by prompting, and the X-GENRE classifier, an XLM-RoBERTa model that was fine-tuned on the X-GENRE dataset. The models are evaluated on two test sets: one consisting of 100 instances from the English EN-GINCO dataset, and the other consisting of 100 instances from the Slovenian GINCO dataset. This also allows comparison of the ChatGPT's performance on high-resource language (English) versus low-resource language (Slovenian).

Table 2. Comparison of the ChatGPT model and the fine-tuned X-GENRE model on the two test sets.

Test set	Model	Micro F1	Macro F1	Accuracy
EN-GINCO	ChatGPT	**0.74**	**0.66**	**0.72**
	X-GENRE	0.67	0.61	0.67
GINCO	ChatGPT	0.75	0.64	0.75
	X-GENRE	**0.91**	**0.91**	**0.91**

Table 2 shows the results for the performance of the two models on the English test set (EN-GINCO) and on the Slovenian test set (GINCO). The results on the EN-GINCO dataset show the out-of-dataset performance of both models, as they were not trained on this dataset. Surprisingly, the ChatGPT model outperforms the X-GENRE classifier by 5–7 points in micro F1, macro F1 and accuracy. It achieves micro F1 of 0.74, macro F1 of 0.66 and accuracy of 0.72. The X-GENRE classifier was fine-tuned on more than 1,700 manually-annotated texts, which required laborious annotation campaigns. In contrast, the ChatGPT model was not explicitly trained for this task. Given these circumstances, the obtained results for ChatGPT are quite impressive.

In contrast, the fine-tuned X-GENRE classifier significantly outperforms ChatGPT when the models are evaluated on the Slovenian test set. It achieves micro F1, macro F1 and accuracy scores of 0.91. However, it is important to note that the X-GENRE model was trained on the training portion of the GINCO dataset, while we use the test split of the same dataset for its evaluation. While the results obtained by the X-GENRE model on the EN-GINCO test set reflect its performance across different datasets, the results on the GINCO test set indicate its performance within the same dataset, which explains its significantly higher results on the latter test set. A more interesting finding is the comparison of ChatGPT's performance on English texts and Slovenian texts. Table 2 illustrates that ChatGPT's performance on Slovenian is comparable to its performance on English, despite the fact that Slovenian has a considerably smaller presence in the training data used for pre-training and fine-tuning the ChatGPT model, since it is an under-resourced language[9].

[9] The exact distribution of these two languages in the pre-training dataset for ChatGPT is not available. However, insights into the likely language distribution in the

To further analyze the differences between the two models, we examine how their predictions differ on the label level. The analysis is conducted on instances from both test sets combined. The findings reveal a substantial level of concurrence between the models, with their predictions aligning in 68% of cases. Among the matching predictions, 62.5% were accurate, while in 5.5% of instances the label that was predicted by both models was incorrect.

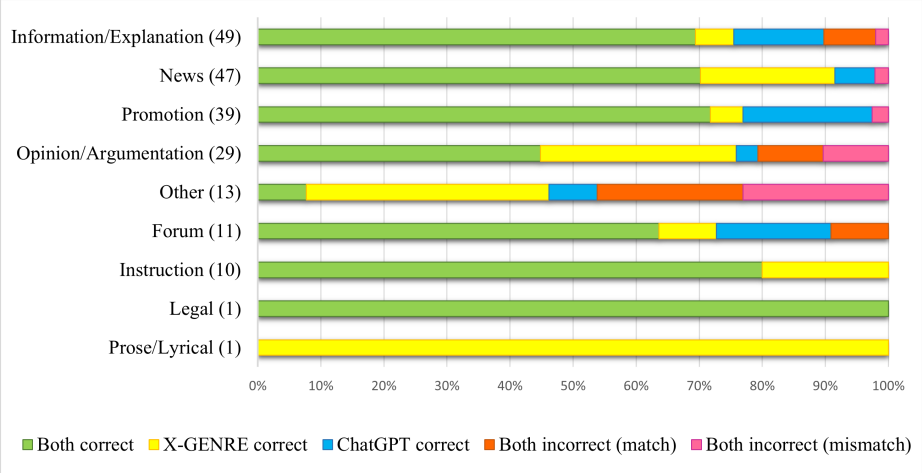

Fig. 1. The performance of both models on each genre category. The count of instances from both test sets belonging to the corresponding label is added next to the label name.

Figure 1 illustrates the performance of both models on each genre label. As shown in the Figure, the genre categories *Information/Explanation*, *News*, *Promotion*, *Instruction* and *Legal* were correctly predicted by both models in at least 70% of instances. The label *Other* proved to be the most challenging for the classifiers, with almost half of the instances being predicted incorrectly. This outcome is not surprising, since this label is the least well defined and encompasses texts that do not fit into any other well-defined category. Examining instances where only one of the models is correct, we observe that ChatGPT demonstrates higher accuracy in predicting *Information/Explanation*, *Promotion*, and *Forum*. On the other hand, the X-GENRE classifier exhibits greater accuracy in predicting *News*, *Opinion/Argumentation* and *Instruction*. Based on this analysis, we can conclude that the outputs of the two models are quite complementary,

ChatGPT pre-training data can be inferred from information on the pre-training dataset for the GPT-3 model [15], a previous OpenAI model. In that dataset, English accounted for 92% of the dataset (180 billion words), while Slovenian represented only 1% (26 million words) of the data.

as expected given the different nature of the two models. Consequently, merging the outputs could be a viable approach in specific use cases.

6 Conclusions

In this paper, we analyze the performance of the instruction-tuned ChatGPT model in the task of automatic genre identification. We compare it with the X-GENRE classifier, a Transformer-based language model that was fine-tuned on more than 1,700 texts, manually-annotated with genres in three annotation campaigns and two languages. Surprisingly, the findings indicate that when evaluated on a dataset that neither model was trained on, ChatGPT outperforms the X-GENRE classifier. Despite initial expectations that the ChatGPT model would struggle with texts in languages other than English, particularly Slovenian as an under-resourced language, the results demonstrate consistent performance across both English and Slovenian genre-annotated datasets.

Until now, the prevailing approach in text classification was that the model needs to be provided with a significant amount of manually-annotated data to achieve satisfactory results for the given task. Consequently, in numerous PhD studies focused on text categorization, the primary emphasis and contribution of the research work lie in the preparation of manually-annotated data and the development of models based on this data. However, our preliminary study demonstrates that recent instruction-tuned GPT models yield comparable or even superior results to the fine-tuned models, without necessitating any dataset, manually annotated for this specific task. These results might hint at a new era for text categorization, wherein manual annotation would only be required for the creation of test sets to evaluate the models. Thus, these findings can be relevant for PhD students, engaged in text classification tasks that rely on manual data annotation and training models using the annotated data. Similar experiments could be conducted on any text classification task to ascertain whether the traditional approach is still reasonable. However, it is important to consider the capabilities of various machine learning architectures for the specific task at hand. The selection of the most suitable model ultimately depends on the nature of the research being conducted. Instruction-tuned GPT models offer the advantage of being readily applicable without the need for additional training or large annotation campaigns. In contrast, supervised classification methods, which include fine-tuned BERT-like language models, are more reliable. The X-GENRE classifier is limited to selecting a label from a predetermined set, whereas GPT models are not restricted to providing specific classes and may assign unexpected labels to texts. Thus, if the objective of the research is to automatically annotate extensive text collections, GPT models may be less suitable as they necessitate an examination of the results and the implementation of post-processing techniques to ensure that texts are annotated with a closed label set. Furthermore, if the text classification task requires classifiers that are explainable and trustworthy, non-neural machine learning algorithms may be a

preferable choice over large language models due to the inherent lack of explainability associated with deep neural technologies.

Motivated by the aforementioned findings, we plan to expand the scope of our comparison in several ways. Firstly, it is important to note that the current experiments were conducted on only 200 instances, as each prompt had to be manually provided to the ChatGPT web platform. However, recently, an official ChatGPT API has been released, which will allow for much more efficient classification experiments. This will enable us to extend the comparison to larger test sets encompassing a wider range of languages and genre-annotated datasets. Secondly, like some of previous studies [20,6], we plan to extend the comparison to include other GPT models, such as the GPT-4 model [17]. Thirdly, as extensive prompt engineering was out of the scope of this paper, we intend to explore more advanced prompting techniques in order to maximize the potential of the ChatGPT model. Specifically, we plan to investigate the efficacy of manual few-shot chain-of-thought prompting [28], which was shown to significantly enhance ChatGPT's performance on classification tasks [30]. Furthermore, we plan to further improve the performance of the fine-tuned BERT-like models, by training them on diverse datasets, larger datasets and by experimenting with different distributions of languages inside a multilingual dataset.

Acknowledgements This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT /A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

References

1. Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Pla Sempere, L., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., Zaragoza, J.: Slovene web corpus MaCoCu-sl 1.0 (2022), Slovenian language resource repository CLARIN.SI
2. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451 (2020)
4. Egbert, J., Biber, D., Davies, M.: Developing a bottom-up, user-based method of web register classification. Journal of the Association for Information Science and Technology **66**(9), 1817–1831 (2015)

5. Erjavec, T., Ljubešić, N.: The slWaC 2.0 corpus of the Slovene web. T. Erjavec, J. Žganec Gros (ur.) Jezikovne tehnologije: zbornik **17**, 50–55 (2014)
6. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv preprint arXiv:2302.09210 (2023)
7. Huang, F., Kwak, H., An, J.: Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. arXiv preprint arXiv:2302.07736 (2023)
8. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V.: The TenTen corpus family. In: 7th international corpus linguistics conference CL. pp. 125–127. Lancaster University (2013)
9. Kuzman, T., Ljubešić, N.: Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In: Mladenić, D., Grobelnik, M. (eds.) Odkrivanje znanja in podatkovna skladišča - SiKDD: 10 October 2022, Ljubljana, Slovenia. Institut "Jožef Stefan" (2022)
10. Kuzman, T., Ljubešić, N., Pollak, S.: Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments. In: Fišer, D., Erjavec, T. (eds.) Jezikovne tehnologije in digitalna humanistika: zbornik konference. p. 100–107. Institute of Contemporary History (2022)
11. Kuzman, T., Rupnik, P., Ljubešić, N.: The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In: Proceedings of the Language Resources and Evaluation Conference. pp. 1584–1594. European Language Resources Association, Marseille, France (2022)
12. Laippala, V., Rönnqvist, S., Oinonen, M., Kyröläinen, A.J., Salmela, A., Biber, D., Egbert, J., Pyysalo, S.: Register identification from the unrestricted open Web using the Corpus of Online Registers of English. Language Resources and Evaluation pp. 1–35 (2022)
13. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning based text classification: A comprehensive review. arXiv preprint arXiv:2004.03705 (2020)
14. Müller-Eberstein, M., van der Goot, R., Plank, B.: Genre as Weak Supervision for Cross-lingual Dependency Parsing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4786–4802 (2021)
15. OpenAI: GPT-3 Dataset Statistics: Languages by Word Count. https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv (2020), accessed: June 27, 2023
16. OpenAI: ChatGPT General FAQ. https://help.openai.com/en/articles/6783457-chatgpt-general-faq (2023), accessed: March 3, 2023
17. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
18. Orlikowski, W.J., Yates, J.: Genre repertoire: The structuring of communicative practices in organizations. Administrative science quarterly pp. 541–574 (1994)
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155 (2022)
20. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv preprint arXiv:2302.06476 (2023)
21. Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Biber, D., Egbert, J., Pyysalo, S., Laippala, V.: Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In: 16th

Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, EACL 2021. pp. 183–191. Association for Computational Linguistics (ACL) (2021)
22. Sharoff, S.: Functional text dimensions for the annotation of web corpora. Corpora **13**(1), 65–95 (2018)
23. Sharoff, S., Wu, Z., Markert, K.: The Web Library of Babel: evaluating genre collections. In: LREC. Citeseer (2010)
24. Stewart, J.G., Callan, J.: Genre oriented summarization. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science (2009)
25. Suchomel, V.: Genre Annotation of Web Corpora: Scheme and Issues. In: Proceedings of the Future Technologies Conference. pp. 738–754. Springer (2020)
26. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355 (2018)
27. Van der Wees, M., Bisazza, A., Monz, C.: Evaluation of machine translation performance across multiple genres and languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
28. Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain of Thought Prompting Elicits Reasoning in Large Language Models. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)
29. Zhang, B., Ding, D., Jing, L.: How would Stance Detection Techniques Evolve after the Launch of ChatGPT? arXiv preprint arXiv:2212.14548 (2022)
30. Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D.: Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. arXiv preprint arXiv:2302.10198 (2023)