Fraunhofer SIT at CheckThat! 2023: Can LLMs Be Used for Data Augmentation & Few-Shot Classification? Detecting Subjectivity in Text Using ChatGPT

Notebook for the CheckThat! Lab at CLEF 2023

Raphael Antonius Frick[1,*,†]

[1]*Fraunhofer Institute for Secure Information Technology SIT | ATHENE - National Research Center for Applied Cybersecurity, Rheinstrasse 75, Darmstadt, 64295, Germany,*
url=https://www.sit.fraunhofer.de/

Abstract

The fight against the spread of misinformation and rumors on the Internet has become a difficult issue lately. In some cases, it is difficult to tell whether a news article published on the Internet contains opinions or was written objectively. This year's CheckThat! 2023 Task 2 dealt with the recognition of such texts. Due to the recent rise of large language models, this work analyzed the extent to which large language models such as ChatGPT can be used to augment unbalanced data sets and whether they can serve as a reliable few-shot classifier. The proposed approaches were trained and evaluated on the English and German subtasks of the challenge. While the models trained with the augmented data were unable to outperform the BERT models trained without the additional data, the few-shot classification scheme was able to outperform across different data set splits, most notably with the English test set. On the private test sets, the proposed ChatGPT-based few-shot classifiers achieved an F_1 value of 0.73 on the English data and an F_1 value of 0.68 on the German data. However, they have not been shown to achieve stable performance over multiple data set splits.

Keywords

Subjectivity Detection, Large Language Models, Few-Shot Classification,

1. Introduction

Social media has introduced new ways how information can be disseminated, allowing individuals to share news easily and opinions with a global audience. However, this unprecedented accessibility has also paved the way for the rapid spread of fake news. The viral nature of social media platforms amplifies the reach and impact of false information, often leading to widespread misinformation and confusion.

A particular challenge associated to this is the ability to distinguish between news articles shared on the internet that are written subjectively or objectively. Subjectively written texts

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
raphael.frick@sit.fraunhofer.de (R. A. Frick)

require special attention because they do not express facts in a value-free way, but instead contain feelings and opinions of the author.

As part of this year's CheckThat! 2023 competition [1, 2], Task 2 [3] was about assessing whether a sentence in a news article was written in an objective or subjective tone. The task was offered to several languages, including Arabic, Dutch, English, German, Italian, and Turkish. In this paper, we describe the approach we used in the competition to classify news articles written in English and German. The approach takes advantage of ChatGPT, a large language model (LLM) based on GPT3.5. LLMs are deep learning models trained on large amounts of textual data so that they can produce coherent and contextually relevant responses to natural language input. They have demonstrated their remarkable capabilities on a variety of NLP tasks, including speech translation, sentiment analysis, text generation, and question answering. Despite their success, LLMs still face several challenges that warrant further investigation. One such challenge is the generation of biased content, which stems from the models' training data reflecting the biases present in the real world. In this paper, we investigate whether LLMs can be used to enrich imbalanced datasets and whether they are useful for distinguishing between objectively and subjectively written text by using them as few-shot classifiers. By using ChatGPT as a few-shot classifier, an F_1 score of 0.73 was achieved on the English private test data set, whereas an F_1 score of 0.68 was achieved on the German test set.

The remainder of the paper is structured as follows. In Section 2, an introduction to large language models and solutions to detecting objectivity in text is given. Section 3 gives a description over the data set provided by and used throughout the competition. The analyzed methods and their results on each data set are showcased in Section 4. The paper then concludes with a discussion on the achieved results.

2. Related Work

2.1. Large Language Models

Recently, large language models such as ChatGPT[1], LLama [4], and Bard[2] have gained a lot of popularity. These models were trained on large datasets comprising billions of websites and documents, and can thus recognize patterns that enable them to produce conditioned text.

Even though they are capable of synthesizing text even for complex topics, there are still some major challenges that require solving. Since the models are trained on data collected within a certain time period, the generated texts cannot refer to events happening thereafter. Based on the collected data, they try to estimate which token is most likely to follow next for a given sequence of tokens. However, this has the consequence that the texts produced are tainted with biases, e.g., in relation to gender and politics. Because the models do not include a control loop that determines whether the statements made in the synthesized texts are true or not, some texts contain fictitious statements. Most models are unimodal and consider only textual data. Recently, however, the focus has shifted from purely text-based models to models that support multiple modalities, such as visual data combined with textual data [5].

[1]https://openai.com/blog/chatgpt
[2]https://bard.google.com

2.2. Zero- & Few-Shot Learning

Machine learning algorithms, and deep learning classifiers in particular, require a large amount of labeled data when training. However, in some cases it is not possible to provide a large set of examples that could be used to train such a model.

Here, zero-shot and few-shot learning can be used to solve this problem. Few-shot learning is a type of machine learning in which the model is trained using a small number of examples. The goal is to make predictions for an NLP task without having seen a single labeled example (in zero-shot learning [6]) or only a few examples (in few-shot learning [7]) that are specific to the task at hand. While large language models have proven useful in text generation, they can also be used as zero and few-shot classifiers, as shown in the work of [8, 9] and [10]. Therefore, this work investigated whether large language models are also suitable for discriminating between subjectively and objectively written texts.

3. Data Set Description

The goal of Task 2 of this year's CheckThat! Lab Challenge was to predict whether a text snippet was written in a subjective or objective tone. The task covered the languages Arabic, Dutch, English, German, Italian, and Turkish. However, we only participated in the English and German language variant of the task.

The data sets of both languages consisted of text snippets gathered from news articles. Each of these were either written subjectively or objectively, resulting in a binary classification problem. Labels were created by human annotators and provided along with the challenge data set. Examples from the data set are shown in Table 1.

The class distributions from each of the provided the data sets can be viewed in Table 2. As can be seen, the data sets are slightly imbalanced, with most samples belonging to the objectivity class. This is to be expected, as most news articles are more likely to be written in a value-free manner and do not contain opinions.

Table 1
Instances of subjectively (SUBJ) and objectively (OBJ) written sentences for Task 2

	Instance	Class
1.	Marko Mihailović, the 29-year-old figurehead of Belgrade Pride, led the city's winning bid.	OBJ
2.	This is the strongest case for stakeholder capitalism.	SUBJ
3.	Als Ergänzung zur Spritze bringt Pfizer eine Corona-Tablette auf den Markt.	OBJ
4.	> Erlange die legale Steuerfreiheit & entziehe den Satanisten die Macht!	SUBJ

4. Methods and Results

Large language models such as ChatGPT can be used to generate text based on a specific instruction. It can be used for multiple use cases, such as code generation, but also to solve specific tasks, including text analysis. In this work, ChatGPT was used to analyze how well it performs in enriching data sets and whether it can solve the task of subjectivity and objectivity

Table 2
Class distribution of the CheckThat! Lab 2023 Task 2 English (E) and German (G) data set

	Total	Objectivity	Subjectivity
Train_E	830	532	298
Dev_E	219	113	106
Test_E	243	116	127
Train_G	800	492	308
Dev_G	200	123	77
Test_G	291	194	97

Table 3
Classification results provided by each method for Task 2

		English		German	
		macro F1	SUBJ F1	macro F1	SUBJ F1
Dev	BERT	**0.76**	**0.75**	0.77	0.69
	BERT + GPT	0.70	0.67	0.76	0.67
	GPT	0.71	0.74	**0.78**	**0.74**
Test	BERT	0.69	0.65	**0.73**	**0.63**
	BERT + GPT	0.64	0.57	**0.73**	**0.63**
	GPT	**0.73**	**0.77**	0.68	0.65

itself by using it as a few-shot classifier. For this, ChatGPT (GPT-3.5-turbo) was accessed using its official API[3].

4.1. Using Large Language Models for Dataset Augmentation

Since the data sets provided for each language had a slight class imbalance, the goal was to use ChatGPT to synthesize additional samples to mitigate potential negative side effects. For this purpose, each sample was taken from the data set and ChatGPT was instructed to transform it into a sentence representing the opposing class. Here, the following instruction was used:

User: *Rewrite a text that was written in a subjective tone or in an objective tone so that it represents the other class.*
Subjectivity: "{sentence}"
Objectivity:

GPT: *{synthesis}*

The Table 4 shows the synthesis results of some examples. While the synthesized sentences correctly express the target class, the texts written in a subjective tone were often written from a personal perspective. As a result, these sentences often begin with "I wonder...", "In my opinion...", or "To me...". The subjectively written sentences in the training set provided

[3]https://openai.com/blog/openai-api

Table 4
Instances of synthesized subjectively (SUBJ) and objectively (OBJ) written sentences for Task 2

	Instance	Class
1.	"I was just drowning."	SUBJ
GPT	"I felt like I was being swallowed up by the water."	OBJ
2.	It is not entirely clear how BA.4.6 has emerged, but it's possible it could be a recombinant variant.	OBJ
GPT	In my opinion, it seems that the origin of BA.4.6 is not fully understood, but it is plausible that it could be a recombinant strain.	SUBJ
3.	Akzeptieren Sie keine Erpressung.	SUBJ
GPT	Eine Erpressung sollte nicht akzeptiert werden.	OBJ
4.	Der andere Angeklagte bekundete, er könne sich an den ganzen Vorgang nicht erinnern.	OBJ
GPT	"Ich habe den Eindruck gewonnen, dass der andere Angeklagte sich nicht an den Vorgang erinnern konnte."	SUBJ

as part of the competition did not follow this style of writing. Consequently, the synthesized sentences do not represent the data from the provided data set well. The same also applies for sentences synthesized from the German data set.

For the classification, BERT-based models[11] were fine-tuned on both, the provided sentences of the train set and the synthesized sentences by ChatGPT. The English model was based on *bert-base-cased* and the German model on *bert-base-german-cased*. During training, the Adam algorithm [12] was used as an optimizer because it has an adaptive learning rate mechanism. As the initial learning rate, a value of 0.0004 was set. The model was fine-tuned over five epochs, using a batch size of 24. To ensure optimal performance on the private test split of the competition data set, only the model with the highest performance on the development split was retained.

As it can be seen in Table 3, the models fine-tuned without any additional data provided by ChatGPT performed better on all data sets. One reason for this could be the difference between the writing style of the artificially generated examples and the writing style in the provided data set. Therefore, in the case of subjectivity detection, it is not advisable to supplement the data with additional data from ChatGPT.

4.2. Using Large Language Models as Few-Shot Classifiers

In the second approach, ChatGPT was used exclusively to automatically classify the provided samples without using a separate classifier. Here, one can distinguish between zero- and few-shot classification. In contrast to the few-shot classification, zero-shot classification takes only a description of the task as input, without resorting to examples for reference. As the data set splits already contain labels, a few-shot classification scheme was chosen. It exploits the ability to self-define the output of the GPT model via its API. In this way, a fake chat history was first built that mimicked the classification responses ChatGPT would have returned based on sample text snippets. This chat history was then used to perform analysis on any sentences from the development and test sets. For the classification, the following instructions were used:

User: *Classify, whether a text was written in a subjective tone or in an objective tone.*
Text: "First by habit one thinks of those for which we have traditional images: The machine, the boss, the pork barrel, the spoils system, the politician everywhere in his popular character, acquiring merit and power by spending public money; doing things for his people with the money of other people, taking care at the same time to do enough for himself with everybody's money."

Class:

Simulated System Response: *subjectively*

User: *Classify, whether a text was written in a subjective tone or in an objective tone.*
Text: "Garina, who was there, recalls that Belgrade "looked like a war zone"."
Class:

Simulated System Response: *objectively*

User: *Classify, whether a text was written in a subjective tone or in an objective tone.*
Text: "{sentence}"
Class:

GPT: *{prediction}*

Table 3 showcases the results of the few-shot classifier on the test set. While it showed the lowest performance on the English development data set, it outperformed all models on the test data set. However, the opposite was true when evaluating the German data set. Here, it achieved the best F_1 values on the development data set, but performed significantly worse on the test data set than the fine-tuned BERT model or the model trained on additional artificially generated training data. Since the data in the private test data sets may differ from the data in the development data set in terms of certain characteristics, such as general writing style, the fine-tuned models may generalize less well on unseen data. However, since the ChatGPT-based few-shot classifier does not require any further training process, its performance is less stable across multiple data sets in contrast to fine-tuned models.

5. Conclusion

ChatGPT and other large language models such as LLama and BARD have attracted a lot of attention recently. After being trained on large corpora such as documents, web pages, and more, they have been shown to perform well on text generation and some analysis tasks. In this work, we investigated whether they can help identify news articles that report on a topic influenced by their opinions as part of the CheckThat! Lab 2023 competition. While ChatGPT was able to synthesize new data that reflected the target class very well, it also introduced several stylistic patterns that may affect the model that uses this data to fine-tune it. Therefore, the models using additional training data were unable to outperform the fine-tuned BERT models adapted to the downstream task. When ChatGPT was used as a few-shot classifier, performance varied dramatically depending on the split of the data set and the language present in the data. It performed best for the private test set of the English data set, but worse for the development set. For the German data set, it performed best on the development set but worst on the private test set. This indicates that using ChatGPT as a few-shot classifier bears the risk of achieving less stable performances across different data sets than, for example, fine-tuned

models. Overall, an F_1 value of 0.73 was achieved on the English test set and a value of 0.68 on the German test set. Thus, the few-shot classifier was still able to outperform the competition's baseline models which featured a macro-F_1 score of 0.72 on the English test set and a score of 0.64 on the German test set.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of "ATHENE – CRISIS" and "Lernlabor Cybersicherheit" (LLCS).

References

[1] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The CLEF-2023 CheckThat! Lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517.

[2] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, , T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouani, Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[3] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. Struss, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, T. Mehmet Deniz, M. Wiegand, W. Zaghouani, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles, in: M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF 2023, Thessaloniki, Greece, 2023.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[5] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. Mohammed, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, F. Wei, Language is not all you need: Aligning perception with language models, 2023. doi:10.48550/arXiv.2302.14045.

[6] M. F. Naeem, M. G. Z. A. Khan, Y. Xian, M. Z. Afzal, D. Stricker, L. Van Gool, F. Tombari, I2mvformer: Large language model generated multi-view document supervision for zero-

shot image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15169–15179.

[7] J. Li, B. Chiu, S. Feng, H. Wang, Few-shot named entity recognition via meta-learning, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 4245–4256.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[10] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, W. X. Zhao, Large language models are zero-shot rankers for recommender systems, 2023. arXiv:2305.08845.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[12] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.