CLaC-2 at CheckThat! 2024: A Zero-Shot Model for Check-Worthiness and Subjectivity Classification

Notebook for the CheckThat! Lab at CLEF 2024

Shayne Gruman*, Leila Kosseim

Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada

Abstract

In this paper, we describe our approach to CLEF 2024 Lab 2 CheckThat! Task 1 (Check-worthiness) and Task 2 (Subjectivity), which aims to evaluate how consistent Large Language Models (LLMs) can distinguish between objective truths worth fact checking and subjective opinions. Our approach was based on prompt engineering utilizing Google's pre-trained LLM Gemini. To optimize performance, we including a random subset of the training data in the input prompt while also augmenting the test data through paraphrasing. We achieved and F1 score of 0.564 for Task 1 and 0.445 for Task 2 which ranked us 24th and 14th respectively. This work contributes to understanding the limitation of LLMs and highlights one of their major pitfalls, subjective reasoning.

Keywords

Subjectivity Detection, Large Language Models (LLM), Prompt Engineering

1. Introduction

Since the dawn of the internet, news sources have become abundant and hence personalized to their target demographic [1]. Automated methods to accurately interpret whether a claim is subjective have been an area of research far many years [2]. This paper describes a more modern approach leveraging LLMs natural language processing capabilities.

CLEF-2024 Task 1 [3] challenged participants to develop a model that could binary classify if a claim is worth fact checking or not [1]. This task is inherently objective as the sentence in question will exclusively contain a verifiable factual claim, or not be worthy of fact checking. Task 1 proposed three languages, out of which we participated in only English.

The CLEF-2024 Task 2 is also a binary classification task aimed at distinguishing whether a sentence in a news article reflects the author's subjective opinion or presents objective information[2]. Subjectivity in natural language refers to the aspects of language used to express personal opinions. However, this task faces an inevitable issue as the perception of subjectivity can be influenced by personal biases [4]. Task 2 proposed six languages, out of which we participated in only English.

With the recent surge in LLMs, we were motivated to test the subjectivity classification capabilities of such models. To do so, we used Google's LLM Gemini because of its access to the information on Google's search engine [5].

Section 2 summarizes the data and presents an overview of the English corpus distribution. Section 3 presents an overview of our model's methodology, while Section 4 describes the outcome of our approach. Finally, Section 5 analyses the results of our model. The code of the models presented in this paper is available on GitHub. [3]

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
s_frijag@live.concordia.ca (S. Gruman); leila.kosseim@concordia.ca (L. Kosseim)
https://github.com/shaynefg (S. Gruman); https://github.com/CLaC-Lab/ (L. Kosseim)

[1] https://checkthat.gitlab.io/clef2024/task1/
[2] https://checkthat.gitlab.io/clef2024/task2/
[3] https://github.com/CLaC-Lab/CLEF-2024-Task-2

2. Background

CLEF 2024 CheckThat! lab proposed 6 shared tasks: Check-worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness [6]. This paper describes the system we developed for Task 1 and Task 2 [7].

2.1. Datasets

Organisers of the CLEF CheckThat! Task 1 and Task 2 provided two different datasets. The data for Task 1 was based off on debates and political speeches [8], whereas in Task 2, it was extracted from news articles [9].

2.1.1. Datasets: Task 1 - Check-worthiness

CLEF 2024 Task 1 - Check-worthiness offered datasets in Arabic, Dutch, and English, out of which we only participated in English. As shown in Table 1, the data set consisted of three features:

1. **Sentence ID:** A unique identifier for each sentence.
2. **Sentence:** The sentence itself, which is either worthy of fact checking or not.
3. **Label:** A binary label: "No" if a fact check is not necessary, or "Yes" if it is.

Table 1
Sample instance from the Dev-Test data set of Task 1 - Check-worthiness

ID	37436
Sentence	He's been a professor for a long time at a great school.
Label	No

Each instance was extracted from debates and political speeches. The English corpus consisted of a total of 23,851 instances with 22,851 for training, 1,032 for development and only 318 for testing. As shown in Table 2, the data sets were not balanced with 76% of the Train set being not worthy of a fact check versus 24% being worthy of a fact check. The Dev-Test set had a slightly different distribution with 66% being not worthy of a fact check versus 34% being worthy of a fact check.

Table 2
English corpus data distribution across Train, Dev, and Dev-Test sets

	# Sent.	% No	% Yes
Train	22,501	76	24
Dev	1,032	77	23
Dev-Test	318	66	34
Total	**23,851**	**76**	**24**

2.1.2. Datasets: Task 2 - Subjectivity

CLEF 2024 Task 2 - Subjectivity provided datasets in Arabic, Bulgarian, English, German, and Italian, out of which we participated in only English. As shown in Table 3, the data set consisted of 3 features:

1. **Sentence ID:** A unique identifier for each sentence.
2. **Sentence:** The sentence itself, which is either subjective or objective.
3. **Label:** A binary label, "OBJ" for objective, or "SUBJ" for subjective.

Each sentence was an objective or subjective extract from a news article. The English corpus consisted of a total of 1,049 instances.

Table 3
Sample instance from the Dev-Test data set of Task 2 - Subjectivity

ID	8745d4da-91c9-4538-acee-b0e7b1c413fd
Sentence	Who will redistribute the hoarded wealth that has exploded beyond all reason in the author's lifetime?
Label	SUBJ

Table 4
English corpus data distribution across Train, Dev, and Dev-Test sets

	# Sent.	% OBJ	% SUBJ
Train	830	64	36
Dev	219	48	52
Dev-Test	243	48	52
Total	**1,049**	**61**	**39**

As shown in Table 4, the Train set was not balanced with 64% instances from the training set being objective versus 36% being subjective, whereas the Dev-Test set had a distribution of 61% as objective 39% as subjective.

2.2. Related Work

LLMs have demonstrated a general ability to understand natural language [10]. Vijayan [11] proposed a solution that leverages LLMs for binary classification. In this experiment, the Sarcasm dataset from OpenAI evals [4] (created from News Headlines Dataset For Sarcasm Detection) was tested on a multitude of LLMs [11]. Their result show that gpt-3.5 preformed the best with and F1 score of 0.9, while Cohere and j2-ulltra preforming worse with F1 scores 0.79 and 0.68 respectively [11].

Each of the LLMs involved in Vijayan experiment was fine tuned to the specific task using EasyLLM [5]. The aim of their experiment was to evaluate the performance of the LLMs without changing any default hyperparameters [11]. Therefore, no hyperparameters were modified.

In terms of subjectivity detection, previous research has followed two main approaches: lexicon-based [12], and machine-learning methods (ML)[13]. In the former, researchers use a list of subjective words, and the frequency of these words determines a document's subjectivity score. The list of words is manually compiled and is static. Alternatively, machine-learning methods also rely on the idea that certain words are inherently subjective, however, these models utilize ML algorithms to dynamically determine the list of subjective words [14].

3. Methodology

To participate in both Task 1 and Task 2, our approach was based on three key components:

Leveraging LLMs: LLMs provide an extensive understanding of general language patterns as they are trained on massive data sets. Because of modern processing power and transformers, LLMs are now able to grasp a larger context window [15]. We used Gemini because of its access to information on Google's search engine [5] . Given Gemini's free API available on Google AI Studio [6], we were able to perform binary classification on the corpus.

[4] https://github.com/openai/evals/tree/main/evals/registry/data/sarcasm
[5] https://www.easyllm.tech/
[6] https://ai.google.dev/aistudio

Paraphrasing: In an attempt to ensure the LLM has confidence in it's classification, we asked Gemini to generate two semantically identical paraphrases of each instance in the test set. The prompt we used to paraphrase the test set can be seen in Table 5.

Table 5
Exact prompts used to paraphrase the test set

Prompt
"Paraphrase this sentence into two semantically identical sentences <*sentence*>"

Prompt Engineering: We engineered the prompt to optimize accuracy by including a random subset of the training set, the initial test sentence and its two paraphrased test sentences. We asked Gemini to classify each of these three test sentences: for task 1, as "Yes" or No"; for task 2, as "Objective" or "Subjective". The final label for the original sentence was determined by a majority vote from the three sentences (1 original sentence + 2 paraphrased sentences). Additionally, we included a random subset of 600 training instances in the input prompt to give the model an idea on how to classify the test sentence. The exact prompts we used can be seen in Table 6 and 7.

Table 6
Prompts used to query Gemini for task 1

Method	Prompt
1. Gemini Classification	"should this sentence be labeled 'Check-worthy' or 'Not check-worthy'? <*sentence*>"
2. Context Addition	"Based on the training set given <*training subset*>, should this sentence be labeled 'Check-worthy' or 'Not check-worthy'? <*sentence*>"
3. Dataset Augmentation (Ensemble)	"Based on the training set given <*training subset*>, should these sentence be labeled 'Check-worthy' or 'Not check-worthy'? Sentence1: <*original test sentence*>, Sentence2: <*1st paraphrased sentences*>, Sentence3: <*2nd paraphrased sentences*>"

Table 7
Prompts used to query Gemini for task 2

Method	Prompt
1. Gemini Classification	"should this sentence be labeled 'Objective' or 'Subjective'? <*sentence*>"
2. Context Addition	"Based on the training set given <*training subset*>, should this sentence be labeled 'Objective' or 'Subjective'? <*sentence*>"
3. Dataset Augmentation (Ensemble)	"Based on the training set given <*training subset*>, should these sentence be labeled 'Objective' or 'Subjective'? Sentence1: <*original test sentence*>, Sentence2: <*1st paraphrased sentences*>, Sentence3: <*2nd paraphrased sentences*>"

4. Results

We experimented with three main methods in the development process of the model. Method 1 used Gemini solely for sentence classification. Method 2 included a random subset of the training set in the input prompt to feed the model context. To optimize accuracy, the input prompt token limit was maximized by including 600 random training sentences, along with the test sentence. Method 3 was

an ensemble combining method 2 along with data set augmentation through paraphrasing. Each test sentence was paraphrased into two additional sentences using Gemini. Classification for method 3 was performed by a majority vote from the original test sentence and its respective two paraphrased sentences.

Table 8
Task 1 and Task 2 F1 and accuracy scores on theDev-Test English corpus

Task	**Task 1**		**Task 2**	
Method	F1 Score	Accuracy	F1 Score	Accuracy
1. Gemini Classification	0.500	0.464	0.393	0.352
2. Context Addition	0.553	0.611	0.444	0.509
3. Dataset Augmentation (Ensemble)	**0.564**	0.482	**0.445**	0.374

Table 9
Task 1 and Task 2 F1 scores on the Test English corpus

Task	**Task 1**	**Task 2**	
Team	F1 Score	Mcro F1	F1 Score
1st Place	0.802	0.744	0.600
(baseline)	0.307	0.635	0.450
CLaC-2	0.494	0.450	0.370

Tables 8 show the results of our model on the test set. These results show statistically significant differences between Methods 1 and 2 as well as Methods 1 and 3. However, the differences between methods 2 and 3 are statistically insignificant. Comparing the model across both datasets, we see a notable difference between the model's performance on Task 2 (Subjectivity) and on Task 1 (Check-worthiness). The F1 score of Task 1 was 0.564, whereas Task 2 was 0.445. We speculate that the data set of task 1 is inherently more objective than that of task 2. Task 1's check-worthiness could be done by flagging for binary markers, such as whether the sentence contains a verifiable factual claim. This makes it easier for Gemini to perform classification on Task 1.

Our main goal was to evaluate how consistent LLMs can conceptualize subjectivity using the default version of Gemini. The difference between our model and the related work is that our model is not fine tuned. In turn, our output is more representative of the Gemini's true opinion.

4.1. Error Analysis

With the abundance of biases on the internet, and hence in LLM training sets, it is of utmost priority that LLM parent corporations take necessary precautions to produce ethical AI. To do so, the model's output is often filtered to safeguard against misinformation. This practice is effective yet results in overly cautious models that do not engage in complex ethical opinions and will inadvertently hinder the model's capacity for subjective reasoning. This pitfall is evident in the disparity between results shown in Table 9.

5. Conclusion

Many issues come to light when deciphering subjectivity and check-worthiness in text. This paper describes our approach using Googles pre-trained LLMs, Gemini, which show a significant improvement in classification when including context in the input prompt. However, dataset augmentation did not seem to improve results.

Although our model was able to classify Task 1 with an F1 score of 0.564, the model still struggled on Task 2 with a F1 score of 0.445. We speculate that Task 1 (check-worthiness classification) is inherently more objective than Task 2 (subjectivity classification), which would explain the higher scores on Task 1.

This work is just scratching the surface of LLM intricacies. Future iterations of this project can implement more concise prompts as well experimenting with other LLMs.

Acknowledgements

The authors would like to thank the organisers of the CLEF-2024 CheckThat! shared task and the anonymous reviewers for their comments on the previous version of this paper.

References

[1] J. Kavanagh, W. Marcellino, J. S. Blake, S. Smith, S. Davenport, M. Gizaw, News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms, Technical Report, RAND Corporation, Santa Monica, CA, USA, 2019. URL: https://www.rand.org/pubs/research_reports/RR2960.html.

[2] R. Ellen, Learning subjective nouns using extraction pattern bootstrapping, in: Proceeings of the Seventh CoNLL conference, HLT-NAACL, Edmonton, Alberta, 2003. URL: https://aclanthology.org/W03-0404.pdf.

[3] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[4] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, P. Kazienko, Learning personal human biases and representations for subjective tasks in natural language processing, in: 2021 IEEE International Conference on Data Mining (ICDM), 2021, pp. 1168–1173. doi:10.1109/ICDM51629.2021.00140.

[5] M. Imran, N. Almusharraf, Google gemini as a next generation ai educational tool: A review of emerging educational technology, Smart Learning Environments 11 (2024). URL: https://doi.org/10.1186/s40561-024-00310-z. doi:10.1186/s40561-024-00310-z.

[6] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[7] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[8] M. Hasanain, R. Suwaileh, S. Weering, C. Li, T. Caselli, W. Zaghouani, A. Barrón-Cedeño, P. Nakov, F. Alam, Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content, in: [3], 2024.

[9] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles, in: [3], 2024.

[10] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Miani, A comprehensive overview of large language models, arXiv preprint (2023). URL: http://arxiv.org/pdf/2307.06435.

[11] K. Vijayan, Finetuning Large Language Modles for Binary Classification, EasyLLM (2023). URL: https://www.easyllm.tech/docs/blog/finetuning-large-language-models-for-binary-classification.html.

[12] B. Liu, Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing (2010). URL: https://www.researchgate.net/profile/Bing-Liu-120/publication/228667268_Sentiment_analysis_and_subjectivity/links/5472bbea0cf24bc8ea199f7c/Sentiment-analysis-and-subjectivity.pdf.

[13] A. S. Samaneh Karimi, A language-model-based approach for subjectivity detection , Sage Journals (2016). URL: https://journals.sagepub.com/doi/10.1177/0165551516641818.

[14] D. M. P. Kushal Dave, Steve Lawrence, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, Handbook of Natural Language Processing (2003). URL: https://dl.acm.org/doi/proceedings/10.1145/775152.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiserłukasz, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017). URL: http://arxiv.org/pdf/2307.06435.