DSHacker at CheckThat! 2024: LLMs and BERT for Check-Worthy Claims Detection with Propaganda Co-occurrence Analysis

Paweł Golik[†], Arkadiusz Modzelewski[1,2,*,†] and Aleksander Jochym

[1]*Polish-Japanese Academy of Information Technology, Poland*
[2]*University of Padua, Italy*

Abstract

This paper presents our approach to check-worthiness detection, one of the main tasks in the *CheckThat! Lab 2024* at the *Conference and Labs of the Evaluation Forum*. The challenge was to create a system to determine whether a claim found in Dutch and Arabic tweets or English debate snippets needs fact-checking. We explored fine-tuning pre-trained BERT-based models and employing a few-shot prompting technique with OpenAI GPT models. Our study compared monolingual models based on the BERT architecture with a multilingual XLM-RoBERTa-large model capable of processing data in multiple languages. Additionally, we investigated the link between propaganda detection and the check-worthiness of content. We also incorporated the recently released OpenAI GPT-4o model. Our systems' impressive performance, surpassing baseline results across all languages, is highlighted by our high-ranking positions: 3rd in Arabic, 2nd in Dutch, and 8th in English, with even better outcomes in post-deadline experiments.

Keywords

Check-Worthiness, Fact-Checking, XLM-RoBERTa, GPT-3.5, GPT-4o, Propaganda

1. Introduction

1.1. Problem Overview

Nowadays, information spreads from many online sources. As a result, it is crucial to ensure the information is accurate, as it affects public discourse and people's decisions. The spread of disinformation and misinformation threatens the integrity of public discussions, impacting areas such as news reporting, political debates, and social media interactions. It is therefore essential to have solid fact-checking systems.

Fact-checking is not just about verifying that something is true. It is also essential for helping people make good decisions based on accurate information and ensuring that the shared information is trustworthy [1]. Such efforts help create an environment where people can have thoughtful discussions and make deliberate choices. Due to the recent rapid advancements in Artificial Intelligence, automated systems can now offer assistance and support the fact-checking process.

1.2. Task Description

CheckThat! Lab at CLEF 2024 addresses issues that aid research and decision-making throughout the fact-checking process [2]. In the initial editions, *CheckThat! Lab* focused on developing an automated system to assist journalist fact-checkers during the key stages of the text verification process, which follows a structured pipeline [3, 4, 5, 6, 7]:

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
golik.pawel@gmail.com (P. Golik); arkadiusz.modzelewski@pja.edu.pl (A. Modzelewski); aleksanderjochym@gmail.com (A. Jochym)
https://amodzelewski.com/ (A. Modzelewski)
0009-0003-1254-6879 (P. Golik); 0009-0003-1169-831X (A. Modzelewski)

1. Assessing whether a document or claim is check-worthy, i.e., determining if its veracity should be checked by a journalist.
2. Retrieving previously verified claims that could aid in fact-checking the current claim.
3. Gathering further evidence from the Web, if necessary, to support the verification.
4. Making a final decision on the factual accuracy of the claim based on the collected evidence.

CheckThat! Lab Task 1 at CLEF 2024 focuses on the first step of the pipeline. Its goal is to provide an automated system for deciding whether a tweet or transcription claim needs fact-checking [2, 8]. Traditionally, this decision involves experts or human reviewers considering whether the claim can be proven true and if it could cause harm before labeling it as worth checking [2, 8].

In this scenario, we are dealing with a binary classification task. For each instance, which is a short text like a tweet or a caption from a political debate transcription, we aim to predict one of two labels: *"Yes"* or *"No,"* indicating whether the text is worth fact-checking.

1.3. Our Experiments

Our approach to check-worthiness detection involved experimenting with both monolingual and multilingual models, as well as leveraging Large Language Models (LLMs) through few-shot prompting. For monolingual models, we fine-tuned BERT-based architectures tailored to specific languages like English, Dutch, and Arabic. In the multilingual approach, we utilized the XLM-RoBERTa-large model, fine-tuning it on combined datasets from multiple languages. Additionally, we employed OpenAI's GPT-3.5 and the recently released GPT-4o models for few-shot prompting, which demonstrated results comparable to fine-tuned BERT-based models. Furthermore, we explored the relationship between propaganda detection and check-worthiness by fine-tuning models on a propaganda detection dataset before applying them to the check-worthiness task. Our comprehensive experiments enabled us to surpass baseline results and achieve high-ranking positions across different languages: 3rd in Arabic, 2nd in Dutch, and 8th in English, with even better outcomes in post-deadline experiments.

2. Related Work

Detecting disinformation has become a crucial area of research. Researchers work not only on disinformation, but also address particular challenges associated with identifying disinformation, misinformation, and fake news. One such challenge is determining which claims are worthy of checking [7]. Hassan et al. [9] introduced a dataset from U.S. presidential debates and created classification models to differentiate among three distinct categories: check-worthy factual claims, non-factual claims, and insignificant factual claims. Jaradat et al. [10] developed ClaimRank, an online tool designed to identify check-worthy claims, with support for two languages: English and Arabic. Kartal and Kutlu [11] proposed a hybrid model which combines BERT with various features to prioritize claims based on their check-worthiness.

The detection of check-worthy claims has also been a research focus in past years within CheckThat! Labs [12, 13, 1, 7, 14]. Alam et al. [1] introduced a task in 2023 for check-worthiness detection in multimodal and multigenre content with a multilingual dataset with three languages: Arabic, Spanish, and English. Team *ES-VRAI* proposed different methods based on pre-trained transformer models and sampling techniques for detecting check-worthiness in a multigenre content [15]. Their approach resulted in the first position in the Arabic language [15]. Team *OpenFact* was the best-performing team on English [1]. They fine-tuned *GPT-3 curie* model using more than 7K instances of sentences from debates and speeches annotated for check-worthiness [16]. In Spanish language the best performance achieved work done by Team *DSHacker* [17]. Modzelewski et al. [17] presented a system based on fine-tuning XLM-RoBERTa on all languages with additional data augmentation. For data augmentation Team *DSHacker* utilized *GPT-3.5* model for translating and paraphrasing available training data [17].

3. Dataset

The dataset consists of texts and their corresponding gold labels annotated by human experts, forming a multilingual dataset. It includes data in four languages: English, Spanish, Dutch, and Arabic. Dataset was divided into training D_{train}, validation (dev) D_{dev}, and dev-test $D_{dev-test}$ datasets with gold labels. We made final predictions for the test datasets D_{test} published in English, Dutch, and Arabic. The English texts depict captions from political debates, while in the other languages, they represent tweets. In most datasets, the classes are imbalanced, with a predominance of texts that are not check-worthy. Table 1 provides detailed information about the datasets. For more information, refer to Hasanain et al. [8].

4. Our Approach

We experimented with two approaches for text classification: fine-tuning BERT-based models and utilizing few-shot prompting with GPT models, including recently released GPT-4o model. These methodologies are both widely used today and represent the state-of-the-art in the industry [18]. However, deciding whether fine-tuning or in-context learning yields better results is not trivial. Many scientists have addressed the challenge of comparing the two approaches fairly [18].

Table 1
Data characteristics. For average values, we report the arithmetic mean along with the standard deviation. (The '#' symbol stands for 'count.')

Dataset	Language	#samples	%check-worthy	Avg. #chars	Avg. #words
TRAIN	EN	$22,500$	24.06%	97.24 ± 70.50	20.59 ± 13.96
	ES	$19,948$	15.65%	166.02 ± 120.10	30.57 ± 20.66
	NL	995	40.70%	188.22 ± 77.48	33.56 ± 14.73
	AR	$7,333$	30.59%	180.05 ± 74.23	32.03 ± 13.62
DEV	EN	$1,032$	23.06%	89.31 ± 67.14	19.09 ± 13.27
	ES	$5,000$	14.08%	208.89 ± 77.84	37.14 ± 14.52
	NL	252	40.48%	194.30 ± 75.74	35.40 ± 15.23
	AR	$1,093$	37.60%	164.14 ± 71.84	28.90 ± 13.27
DEV-TEST	EN	318	33.96%	67.78 ± 49.41	15.55 ± 10.54
	ES	$5,000$	10.18%	151.49 ± 100.27	27.93 ± 18.07
	NL	666	47.45%	193.74 ± 77.99	40.20 ± 16.85
	AR	500	75.40%	200.05 ± 77.04	35.77 ± 13.82
TEST	EN	341	25.81%	78.25 ± 58.03	17.41 ± 11.69
	NL	$1,000$	39.70%	221.98 ± 74.03	39.29 ± 13.65
	AR	610	35.74%	272.78 ± 30.39	46.32 ± 6.74

4.1. Fine-tuning BERT-based Models

To provide a comprehensive overview of the effectiveness of BERT-based models in identifying check-worthy content across multiple languages, we conducted experiments using two types of models. First, we employed monolingual models that we fine-tuned exclusively on the training data from a single language. Secondly, we used multilingual models that we fine-tuned on different combinations of the training datasets from multiple languages. Our experiments allowed us to compare the performance of both multilingual and monolingual models in the context of check-worthiness detection.

We started with comprehensive hyperparameter tuning for all chosen models (mono- and multilingual). This involved fine-tuning each model on the training dataset D_{train} using every combination of hyperparameter values we specified. We used the proposed D_{dev} dataset for validation. Then, we as-

sessed each model's performance by measuring F_1 score to determine the most effective hyperparameter values.

To obtain the final models for submission, we merged the training D_{train} and validation D_{dev} datasets to form the ultimate training dataset. We then retrained the model with the best hyperparameter values using the $D_{dev-test}$ dataset as the final validation set. Please refer to our Appendix A, which presents optimal hyperparameters of each final submission.

4.1.1. Monolingual Models

For each language, we have chosen a single pretrained monolingual model available at *HuggingFace* that we later fine-tuned on the corresponding language:

- **ENGLISH (MONO-EN)** - *FacebookAI/roberta-large* - the language model (355M parameters) trained on English data in a self-supervised fashion [19].
- **DUTCH (MONO-NL)** - *DTAI-KULeuven/robbert-2023-dutch-large* - the first Dutch large (355M parameters) model trained on the OSCAR2023 dataset [20].
- **ARABIC (MONO-AR)** - *UBC-NLP/MARBERT* - trained on randomly sampled 1B Arabic tweets (with at least 3 Arabic words) from a large in-house dataset of about 6B tweets [21].

4.1.2. Multilingual Models

We also fine-tuned a multilingual *FacebookAI/xlm-roberta-large* [22] model on a combined dataset from all available languages (**MULTI-ALL**). Since Spanish was not included in the final submission, we performed another fine-tuning using only English, Dutch, and Arabic data (**MULTI-NO-ES**).

Additionally, we experimented with fine-tuning a multilingual model previously fine-tuned on a different propaganda-related dataset. We first fine-tuned the *FacebookAI/xlm-roberta-large* model on the propaganda presence binary classification task and then fine-tuned the model again on the *CheckThat! Task 1* data (**MULTI-PROP2**). Refer to Section 5 for more information.

4.2. Few-shot Prompting with GPT Models

We also employed Large Language Models (LLMs) to generate check-worthiness predictions. Our experiments included OpenAI's *gpt-4o* **(GPT-4o)** and *gpt-3.5-turbo-1106* **(GPT-3.5)** generative models. We implemented the few-shot prompting technique using the OpenAI Chat Completions API. Few-shot prompting with GPT models leverages pre-trained language models to perform specific tasks without retraining. Instead, the model is guided by providing a few examples and their expected responses within the input prompt.

Each prediction request sent to the GPT model consisted of a list of messages presented to the model. Each message contains the role and content attribute. There are three roles available:

1. **system** message helps set the behavior of the model (assistant) by providing it context and guidelines.
2. **user** messages can provide exemplary requests for the assistant. In our case - example requests for a provided text's check-worthiness evaluation.
3. **assistant** messages indicate the expected output of the assistant.

In our experiments, the conversation is formatted starting with a system message that clarifies the task and the concept of check-worthiness. This is followed by alternating pairs of user and assistant messages. One pair for each few-shot example, where a user message poses a question about the example's content check-worthiness, and the corresponding assistant message provides the gold label for the example, either 'Yes' or 'No'. The final message following the pairs is one user message with

the actual text to be classified by the model (See Appendix B). For each instance to be classified, we included four examples of few-shot prompting from the training dataset, two of which are check-worthy. The chosen few-shot examples were consistent in a given language. The prompt templates remained consistent for both the **GPT-4o** and **GPT-3.5** experiments.

5. Propaganda Co-occurrence Analysis

Since propaganda often involves misleading, biased, or manipulative information, such content is naturally more likely to warrant fact-checking [23, 24]. Therefore, we decided to indirectly analyze whether propaganda co-occurs with check-worthy claim. For that purpose we predicted the presence of propaganda using a model fine-tuned on a propaganda dataset. The underlying assumption was that check-worthy statements are more likely to contain propaganda techniques, given their potential to persuade or manipulate public opinion. We leveraged a multilingual *FacebookAI/xlm-roberta-large* [22] model fine-tuned by Modzelewski et al. [25] on the *IberLEF DIPROMATS 2024 Task 1a* [26] and then we employed it on the check-worthiness $D_{dev-test}$ dataset to evaluate this hypothesis **(MUTLI-PROP1)**. *IberLEF DIPROMATS 2024 Task 1a* is a binary classification task for propaganda detection in English and Spanish tweets [26].

Table 2 shows the performance metrics of the **MULTI-PROP1** model calculated for the $D_{dev-test}$ dataset. Relatively high precision shows that many texts with detected propaganda are indeed worth fact-checking. However, low recall illustrates that many check-worthy texts do not contain propaganda detectable by our model. Such results lead to an intuitive conclusion that propaganda often signals the need for fact-checking, but not all check-worthy statements necessarily rely on propaganda methods.

We then further fine-tuned this model on *CheckThat! Task 1* data and utilized it to predict check-worthiness (**MULTI-PROP2**). Due to time constraints, we did not perform hyperparameter tuning for this model. Instead, we used the hyperparameter values obtained from the search conducted during the **MULTI-ALL** experiment.

Table 2
Performance Metrics of MULTI-PROP1 experiment obtained on DEV-TEST datset.

Language	F_1 Score	Accuracy	Precision	Recall
Arabic	0.2208	0.294	0.6579	0.1326
Dutch	0.3597	0.551	0.5563	0.2658
English	0.1940	0.660	0.5000	0.1204
Spanish	0.1385	0.736	0.1037	0.2083

6. Results

Since the number of allowed submissions was limited to one submission per language, we selected the models for the final predictions based on the F_1 scores obtained on the $D_{dev-test}$ dataset. The models selected for the final submission are: **MONO-EN**, which ranked 8th on the final leaderboard for English, and **MULTI-NO-ES**, which ranked 2nd on the final leaderboard for Dutch and 3rd on the final leaderboard for Arabic. Table 3 provides the details of the models we submitted. Additionally, the table shows the baseline provided by the organisers of *ChecktThat Lab 2024 Task 1* and the score of the best team for each language.

After the submission deadline, we experimented with Large Language Models, namely **GPT-3.5** and **GPT-4o**. Moreover, our experiments that combined knowledge from propaganda classification with check-worthiness detection through models **MULTI-PROP1** and **MULTI-PROP2** have also taken place after the deadline. Unfortunately, due to time constraints and strict deadlines, we could not complete these experiments before the submission deadline. Therefore, we did not consider their results on the $D_{dev-test}$ dataset when selecting models for submission. However, some of our post-deadline

Table 3
Our final results from the *CheckThat! Lab* Task 1 official leaderboards.

		F_1 Score				Official
Language	**Model**	**Winner**	**Baseline**	**DEV-TEST**	**TEST**	**Rank**
English	MONO-EN	0.8020	0.3070	0.9118	**0.7600**	8
Dutch	MULTI-NO-ES	0.732	0.4380	0.6907	**0.7300**	2
Arabic	MULTI-NO-ES	0.5690	0.4180	0.8599	**0.5380**	3

results are superior to our final submission results and, in the case of the Dutch language, even surpass the leaderboard winner.

Table 4 shows the results of all experiments we conducted. We report the F_1 scores calculated on the $D_{dev-test}$ and D_{test} datasets. We did not have access to the ground truth for the test data while developing this system. Nevertheless, we calculated the test D_{test} dataset scores after the organizers released the labels for this dataset following the submission deadline. Each record of *DEV-TEST* columns represents the F_1 scores yielded by a model fine-tuned once on the combined training dataset (D_{train} + D_{dev}) with $D_{dev-test}$ used as a validation dataset.

Table 4
Check-worthiness classification results on the $D_{dev-test}$ and D_{test} sets for all languages.

		F_1 Score			F_1 Score	
Language	**Model**	**DEV-TEST**	**TEST**	**Model**	**DEV-TEST**	**TEST**
English	MONO-EN	0.9118	0.7600	GPT-3.5	0.7343	0.6529
	MULTI-ALL	0.8932	0.7429	GPT-4o	0.8376	0.7207
	MULTI-NO-ES	0.8867	**0.7647**	MULTI-PROP2	0.8571	0.7368
Dutch	MONO-NL	0.6571	0.6182	GPT-3.5	0.4235	0.6937
	MULTI-ALL	0.6657	0.7401	GPT-4o	0.5844	**0.7915**
	MULTI-NO-ES	0.6907	0.7300	MULTI-PROP2	0.6667	0.7336
Arabic	MONO-AR	0.7740	0.5254	GPT-3.5	0.8640	0.5539
	MULTI-ALL	0.8040	0.5568	GPT-4o	0.8916	0.5523
	MULTI-NO-ES	0.8599	0.5380	MULTI-PROP2	0.8347	**0.5680**

6.1. Results

6.1.1. English Language Results

The submission of the **MONO-EN** model earned us the 8$^{\text{th}}$ position on the leaderboard. Notably, the difference between the winning model's F_1 score and ours was small—just 0.042. Interestingly, the **MULTI-NO-ES** model, an *xlm-roberta-large* fine-tuned on English, Dutch, and Arabic data, achieved a better F_1 score than our chosen submission, with a score of 0.7647. However, the difference may not be statistically significant. Most of the remaining models nearly matched our best result.

6.1.2. Dutch Language Results

In Dutch, we came very close to winning with the **MULTI-NO-ES** model, achieving an F_1 score of 0.73, just 0.02 points behind the actual winner. Two of our post-deadline models surpassed our submission result - **MULTI-PROP2** with a slightly better F_1 score of 0.7336 and **GPT-4o** with an impressive 0.7915. The results from **GPT-4o** even outperformed those of the leaderboard winner. Interestingly, the monolingual approach **MONO-NL** performed relatively poor, with an F_1 score of 0.6182.

6.1.3. Arabic Language Results

Similarly, we submitted the predictions of the **MULTI-NO-ES** model for the Arabic language. This submission secured us third place on the leaderboard. The difference in F_1 scores between the **MULTI-**

NO-ES model and the 1st ranked model on the leaderboard is 0.031. The monolingual model performed worse than the multilingual systems, but the difference is minor. Interestingly, **GPT-3.5** produced one of the best results - 0.5539. Our top post-deadline model, **MULTI-PROP2**, nearly matched the F_1 score of the leaderboard winner, with a difference of just 0.002.

6.1.4. GPT Few-shot Prompting Results

The few-shot prompting approach using GPT models showed similar results to the fine-tuning approach with BERT-based models. Across all languages, the **GPT-4o** model consistently outperformed or performed similarly to the **GPT-3.5** model. Specifically for English, **GPT-4o** performed noticeably better than **GPT-3.5** and achieved results only slightly lower than fine-tuned BERT-based models. In the case of Dutch, **GPT-4o** was also superior to **GPT-3.5**, and it emerged as the best model among those we tested for this language. An interesting observation from our experiments is that the differences between Dutch on $D_{dev-test}$ and D_{test} results were more pronounced for GPT models than other models. For Arabic, **GPT-3.5** and **GPT-4o** showed nearly identical performance, comparable to experiments using BERT-based models.

6.1.5. Propaganda Detection Transfer Learning Results

The results of a multilingual model fine-tuned on *DIPROMATS 2024 Task 1a* data further fine-tuned on English, Dutch, Arabic, and Spanish using *CheckThat! 2024 Task 1* data (**MULTI-PROP2**) in one case outperformed all other approaches, but most probably the difference was statistically insignificant. We observed negligibly better results for Arabic and slightly worse results for the remaining languages compared to the **MUTLI-ALL** model.

6.2. Results Discussion

Our results show that the outcomes on the D_{test} and $D_{dev-test}$ datasets are significantly different. The explanation for this phenomenon is not straightforward. One possible reason is overfitting to the $D_{dev-test}$ data when selecting the best model during the final fine-tuning. However, this doesn't explain the discrepancy observed with the GPT models with few-shot prompting, which were not fine-tuned yet still showed significant mismatches between the $D_{dev-test}$ and D_{test} results. A potential reason for GPT large models differences is that the $D_{dev-test}$ and D_{test} data may have differed significantly.

The GPT models, particularly **GPT-4o**, showed results comparable to the fine-tuned BERT-based models. The few-shot prompting approach with OpenAI API offers a more straightforward solution for such tasks, delivering similar performance at a lower time cost.

The monolingual models generally performed worse than the multilingual models in our experiments. Thus, using a language-specific model is not always the best choice. Adding Spanish to the training data did not always improve performance for the multilingual models.

Model obtained from fine-tuning on the *IberLEF DIPROMATS 2024* propaganda detection task and further fine-tuning on check-worthiness data did not yield any significant performance improvement compared to multilingual models fine-tuned exclusively on the check-worthiness data. Limitation of this approach was that we did not perform hyperparameter tunning for **MULTI- PROP2**. Hyperparameters optimization could significantly change our results.

7. Conclusions and Future Work

In our work, we experimented with monolingual and multilingual approaches for various languages for check-worthy claims detection. For the monolingual approach, we utilized BERT models tailored to specific languages, optimizing hyperparameters and fine-tuning each model separately for each language. For the multilingual approach, we used *XLM-RoBERTa-large*. First, we optimized and fine-tuned it on the entire dataset. Then, we excluded Spanish from the training data in a second experiment. Additionally,

we employed two LLMs, namely *GPT-3.5-turbo* and recently released *GPT-4o* for each language, using few-shot prompting to classify texts. We also fine-tuned a model on the *IberLEF DIPROMATS 2024 Task 1* dataset and used this model to predict whether the data from *CheckThat! Lab 2024 Task 1* contained propaganda. With this analysis, we aimed to indirectly determine whether check-worthy data also includes propaganda. We later utilized the mentioned model (*XLM-RoBERTa-large* fine-tuned on the *IberLEF DIPROMATS 2024 Task 1a* for binary propaganda classification) and further fine-tuned it for check-worthiness classification. Our final submissions for all languages were: English - *RoBERTa-large* fine-tuned, and optimized exclusively on English data; Dutch - *XLM-RoBERTa-large*, fine-tuned and optimized on all data except Spanish; Arabic - *XLM-RoBERTa-large*, fine-tuned and optimized on all data except Spanish.

From our experiments we can conclude that GPT models, particularly GPT-4o, showed results comparable to the fine-tuned BERT-based models. Moreover, language-specific BERT-based model performed better only on English dataset. In other languages we got better results by utilizing multilingual model.

Future work could incorporate a detailed linguistic analysis of texts to understand the different linguistic features among check-worthy texts and those that do not require checking. By identifying specific linguistic features and patterns, we could develop more nuanced systems that better differentiate between these types of texts. This analysis could involve examining rhetorical devices and stylistic elements that are prevalent in check-worthy claims.

8. Limitations and Ethics

We acknowledge that our research may raise a number of ethical issues. The first shortcoming is a lack of clear explainability of our models' results. Each model generates check-worthiness ratings without explanations as to why a statement was rated check-worthy or not. Users may need explanations to understand the basis for the model's decisions. One of the most crucial tools for our research was fine-tuning BERT-based models and utilizing GPT-based Large Language Models. As a result, if BERT or GPT-based models were trained on data containing any bias, disinformation, or misinformation, these problems may affect the results of our experiments. The next potential shortcoming is a possible specialization of our systems to detect specific types of check-worthy information. Our systems may possibly not handle more subtle content or other off-topic statements well. We did not check whether the dataset included one topic check-worthy information, and we relied on the work of the workshop organizers to cover more than a specific type.

References

[1] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, et al., Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content, Working Notes of CLEF (2023).

[2] 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.

[3] P. Nakov, A. Barrón-Cedeno, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9, Springer, 2018, pp. 372–387.

[4] T. Elsayed, P. Nakov, A. Barrón-Cedeno, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the clef-2019 checkthat! lab: automatic identification and verification of claims, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International

Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10, Springer, 2019, pp. 301–321.
[5] A. Barrón-Cedeno, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media, in: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42, Springer, 2020, pp. 499–507.
[6] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, et al., Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 264–291.
[7] P. Nakov, A. Barrón-Cedeño, G. da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, et al., Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022, pp. 495–520.
[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: 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.
[9] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: Proceedings of the 24th acm international on conference on information and knowledge management, 2015, pp. 1835–1838.
[10] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, Claimrank: Detecting check-worthy claims in arabic and english, arXiv preprint arXiv:1804.07587 (2018).
[11] Y. S. Kartal, M. Kutlu, Re-think before you share: a comprehensive study on prioritizing check-worthy claims, IEEE transactions on computational social systems 10 (2022) 362–375.
[12] P. Atanasova, L. Marquez, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. Da San Martino, P. Nakov, et al., Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness, in: CEUR WORKSHOP PROCEEDINGS, volume 2125, CEUR-WS, 2018, pp. 1–13.
[13] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness (2019).
[14] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, Y. S. Kartal, F. Alam, G. Da San Martino, et al., Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates., 2021.
[15] H. T. Sadouk, F. Sebbak, H. E. Zekiri, Es-vrai at checkthat! 2023: Analyzing checkworthiness in multimodal and multigenre (2023).
[16] M. Sawiński, K. Węcel, E. P. Księżniak, M. Stróżyna, W. Lewoniewski, P. Stolarski, W. Abramowicz, Openfact at checkthat! 2023: head-to-head gpt vs. bert-a comparative study of transformers language models for the detection of check-worthy claims, Working Notes of CLEF (2023).
[17] A. Modzelewski, W. Sosnowski, A. Wierzbicki, Dshacker at checkthat! 2023: Check-worthiness in multigenre and multilingual content with gpt-3.5 data augmentation, Working Notes of CLEF (2023).
[18] M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, Y. Elazar, Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12284–12314. URL: https://aclanthology.org/2023.findings-acl.779. doi:10.18653/v1/2023.findings-acl.779.
[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL:

http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[20] P. Delobelle, F. Remy, Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion, 2023.
[21] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, ARBERT & MARBERT: Deep bidirectional transformers for Arabic, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7088–7105. URL: https://aclanthology.org/2021.acl-long.551. doi:10.18653/v1/2021.acl-long.551.
[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[23] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.
[24] S. Shaar, F. Alam, G. Da San Martino, A. Nikolov, W. Zaghouani, P. Nakov, A. Feldman, Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection, in: A. Feldman, G. Da San Martino, C. Leberknight, P. Nakov (Eds.), Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Online, 2021, pp. 82–92. URL: https://aclanthology.org/2021.nlp4if-1.12. doi:10.18653/v1/2021.nlp4if-1.12.
[25] A. Modzelewski, P. Golik, A. Wierzbicki, Bilingual propaganda detection in diplomats' tweets using language models and linguistic features., in: IberLEF@ SEPLN, 2024.
[26] P. Moral, J. Fraile, G. Marco, A. Peñas, J. Gonzalo, Overview of dipromats 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers, Procesamiento del Lenguaje Natural 73 (2024).

A. Optimal Hyperparameter Values

This appendix includes the optimal hyperparameter values for our best models.

Table 5
Optimal hyperparamter values used in our models. Legend: **lr** - learning_rate; **bs** - batch_size; **nte** - num_train_epochs; **ws** - warmup_steps; **wd** - weight_decay

Language	Model	lr	bs	nte	ws	wd
English	MONO-EN	10^{-5}	32	6	300	0.001
Dutch	MONO-NL	10^{-5}	8	6	12	0.040
Arabic	MONO-AR	10^{-5}	16	6	165	0.001
ALL	MULTI-ALL	10^{-5}	32	4	900	0.001
	MULTI-ALL-NO-ES	10^{-5}	32	4	900	0.001
	MULTI-PROP2	10^{-5}	32	4	900	0.001

B. Few-shot Prompting Templates

In this appendix, we present the prompt messages included with each text classification request. The prompts are provided only in Spanish for brevity.

1. System prompt: *You are a fact-checking expert. Your task aims to assess the check-worthiness of a presented text. As a fact-checker, you know that to decide whether a text is check-worthy, you must answer several auxiliary questions such as "Does the text contain a verifiable factual claim?" or "Is the text harmful?". Please provide only the final label: "Yes" if the text is check-worthy and "No" otherwise.*

2. Pairs of user and assistance prompts for few-shot prompting:
   a) Example 1:
      - User content: *Answer whether the following text in Spanish is worth fact-checking. Answer using only a single English word: Yes or No. TEXT: Mañana, viernes, no puedes perderte el gran acto de cierre de campaña en Madrid. A las 19.00 h en el Pabellón 1 de IFEMA (Madrid). Con Kiko Veneno y O'Funk'illo en concierto y la intervención de @Pablo_Iglesias_, @AdaColau, @Irene_Montero_, @agarzon:. ¡Te esperamos!*
      - Assistant content: *No*
   b) Example 2:
      - User content: *Answer whether the following text in Spanish is worth fact-checking. Answer using only a single English word: Yes or No. TEXT: tve_tve vuelve a quedar en evidencia. Desplaza al minuto 18 la denuncia del #CGPJ ante las críticas de Iglesias y habla de "diferencias". Exigimos al responsable de edición del telediario explicaciones y a Sánchez que deje de utilizar rtve a su antojo @Enric_Hernandez*
      - Assistant content: *Yes*
   c) Example 3:
      - User content: *Answer whether the following text in Spanish is worth fact-checking. Answer using only a single English word: Yes or No. TEXT: El PSOE demostró su apoyo a Torra durante la moción #PorLaConvivencia Lroldansu "Cs sigue siendo el referente incontestable del Constitucionalismo en Cataluña. No podemos dejar el futuro de 7,5 millones de catalanes en manos de Torra" #ActualidadCs*
      - Assistant content: *No*
   d) Example 4:
      - User content: *Answer whether the following text in Spanish is worth fact-checking. Answer using only a single English word: Yes or No. TEXT: Pedro Sánchez ha dado el visto bueno a la apertura de las 'embajadas' catalanas en Argentina, México y Túnez. Empiezan las cesiones a sus socios separatistas. Incrementan gasto en sus majaderías, despreciando la urgencia de las política sociales.*
      - Assistant content: *Yes*

3. Used final user prompt: *Answer whether the following text in Spanish is worth fact-checking. Answer using only a single English word: Yes or No. TEXT: <Here we provided text to classify by LLM>*