SINAI at CheckThat! 2024: Transformer-based approaches for Check-Worthiness Classification

Notebook for the CheckThat! Lab at CLEF 2024

Sergiu Stoia[1,*], Jaime Collado-Montañez[1], Cristian Ibáñez-Bautista[1], Arturo Montejo-Ráez[1], María Teresa Martín-Valdivia[1] and Manuel Carlos Díaz-Galiano[1]

[1]*Department of Computer Science (University of Jaén), Campus Las Lagunillas, s/n, Jaén, 23071, Spain*

Abstract

This paper discusses the participation of the SINAI team in the CLEF-2024 CheckThat! lab Task 1 for English. The task involves assessing whether claims extracted from transcribed texts should be fact-checked. We explored two approaches to address this challenge: adjusting a Transformers-based model and using prompting-based techniques. In order to address imbalances within the data provided by the organizers, the method of class weighting is employed. Our best-performing system achieved an F1 score of 0.761 for the positive class and was ranked seventh among all twenty-six submissions in the competition.

Keywords

Fact-checking, Transformers, LLM, Fine-tuning, Prompting, Data augmentation

1. Introduction

In the era of information overload, the ability to efficiently identify and verify potentially misleading or false claims is paramount. Fact-checking has become a critical task to ensure the integrity of information disseminated across various platforms. The CheckThat! Task 1 [1] aims to advance research in this domain by focusing on the initial step of the fact-checking pipeline: determining which claims in a text are worth fact-checking.

Our participation in the lab was focused on the English dataset, where the objective was to develop methods to automatically assess the check-worthiness of claims within transcriptions of debates and speeches. We explored two distinct approaches to tackle this challenge: fine-tuning a transformer-based model [2] and leveraging the capabilities of large language models (LLMs) through prompting.

This paper presents the contribution of the SINAI team to the CheckThat! Lab. We analyze the strengths and limitations of each method, providing insights into their potential for enhancing automated fact-checking systems. Our findings contribute to the ongoing efforts [3] to refine the process of identifying claims that merit verification, ultimately supporting the broader goal of combating misinformation.

The remainder of the paper is organized as follows. Section 2 analyses the data provided by the organizers. Section 3 presents a description of the two different approaches developed for this lab. Section 4 shows the results obtained with such approaches and Section 5 summarizes our conclusion and future research directions.

2. The CheckThat! task

The CheckThat! task proposed in the CLEF forum [4, 5] is to determine whether a claim in a tweet or transcription is worth fact-checking. Traditionally, this decision involves judgments from professional

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

sstoia@ujaen.es (S. Stoia); jcollado@ujaen.es (J. Collado-Montañez); cib00005@red.ujaen.es (C. Ibáñez-Bautista); amontejo@ujaen.es (A. Montejo-Ráez); maite@ujaen.es (M. T. Martín-Valdivia); mcdiaz@ujaen.es (M. C. Díaz-Galiano)

0009-0009-2599-2820 (S. Stoia); 0000-0002-9672-6740 (J. Collado-Montañez); 0000-0002-8643-2714 (A. Montejo-Ráez); 0000-0002-2874-0401 (M. T. Martín-Valdivia); 0000-0001-9298-1376 (M. C. Díaz-Galiano)

Table 1
Class imbalance for each split

Split	No	Yes
english_train	75.94 %	24.06 %
english_dev	76.94 %	23,06 %
english_dev-test	66.04 %	33.96 %
english_test	74.19 %	25.81 %

fact-checkers or human annotators answering auxiliary questions like "does it contain a verifiable factual claim?" and "is it harmful?" before assigning a check-worthiness label. This year, the task uses multi-genre data and requires judgments based solely on the text. It is available in Arabic, English, and Spanish.

3. Data

For the task, the organisers provide multi-genre textual data in Arabic, Dutch, Spanish and English taken from tweets and transcriptions [6, 1]. In this section we present a brief analysis regarding the selected datasets used for the development of Task 1.

In order to feed the data to the systems described in Section 3, an analysis of the data obtained for each language was necessary, showing notable distinctions among the different languages. Whilst the Arabic and Dutch sets only contain texts sourced from tweets, those in English are based on transcriptions, whereas the Spanish collection comprises texts from both data sources. In order to use data that accurately reflects the desired outcome, the sets in Arabic and Dutch have been discarded, and only the texts from the Spanish set based on transcriptions have been chosen to increase the amount of available data.

In supervised learning tasks, analyzing the class proportion within the dataset is crucial, since the distribution of classes can significantly impact the learning algorithm. An analysis of this ratio has been conducted for each data split, revealing a clear imbalance among classes 1.

Sequence length is an important feature to take into account since most transformer-based models have a limited amount of tokens per sequence to be trained on. Thus, we used the Tiktoken[1] library from OpenAI to calculate the average length of the provided texts. Figure 1 shows an histogram that reveals that most sequences contain less than a hundred tokens, which is short enough for most state-of-the-art models. The average tokens in the dataset is 21.06 and the standard deviation 14.38.

4. System description

In this section, we describe the different approaches presented for the official evaluation of the first task.

4.1. Transformer fine-tuning

Leveraging the power of models based on Transformers like RoBERTa-base [7] has become a prevalent approach for achieving robust and accurate results for classification tasks. When applied to binary classification tasks, RoBERTa-base demonstrates remarkable performance by capturing intricate patterns and semantic nuances within the text data.

As previously mentioned, the imbalance among classes in the data might negatively impact the performance of this fine-tuning. To address the class proportion disparities, the method of class weighting is employed. This technique rectifies imbalances within datasets by assigning greater importance to underrepresented classes, leading to more equitable and accurate predictions. This value is calculated as the ratio of negative samples to positive samples.

[1] https://github.com/openai/tiktoken

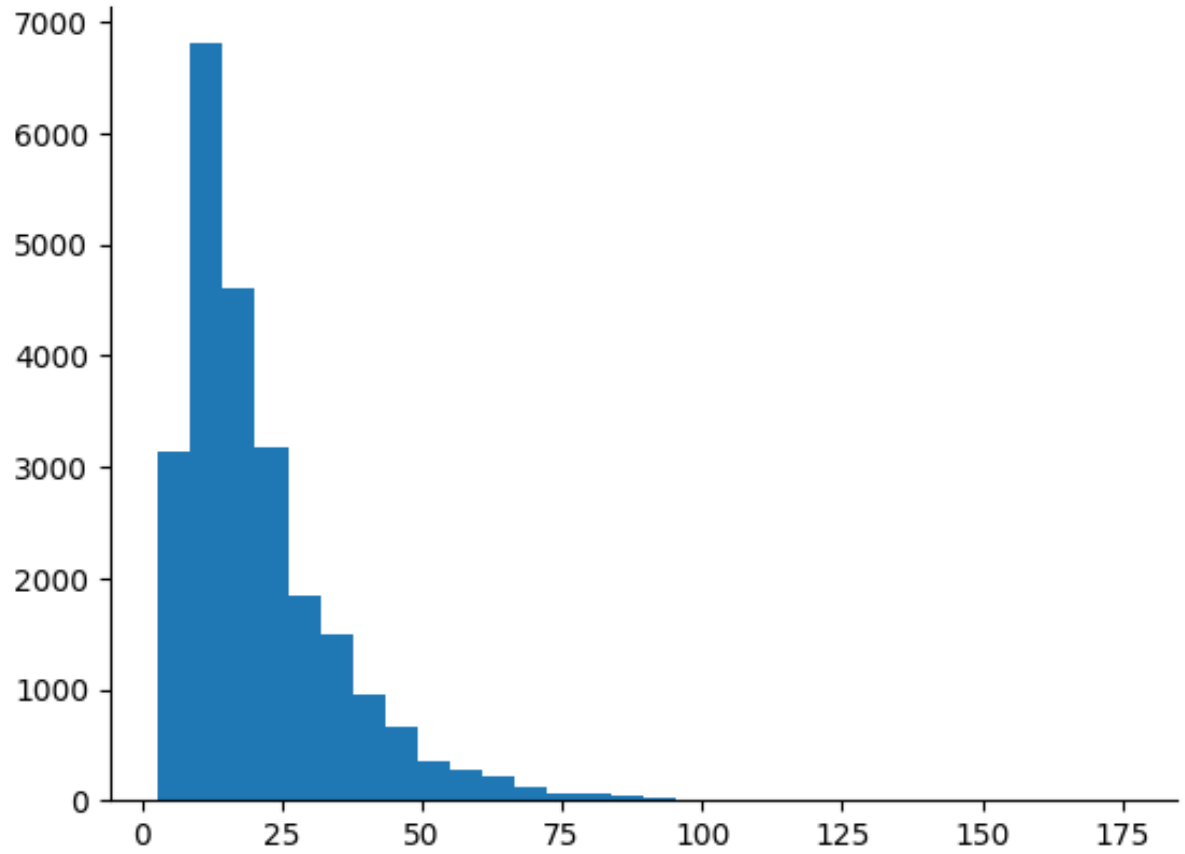

Figure 1: Token count histogram.

Table 2
Finetuning evaluation with dev-test set

Experiment	F1 positive class	macro precision	macro recall	macro F1
RoBERTa fine-tuning with original set	0.8965	0.9408	0.9117	0.9240
RoBERTa fine-tuning with augmented set	0.8899	0.9416	0.9048	0.9197

Two finetunings were conducted using the same base model. The first finetuning was performed exclusively with the original English dataset, whereas the second expanded the original training data by incorporating translated texts from the Spanish set. These experiments were performed with a learning rate of 1e-06 and a batch size of 8, using both train and dev sets combined as training data and dev-test for evaluating the checkpoints obtained during the execution. To safeguard against overfitting and unnecessary training cycles, the strategy of early-stopping was implemented. This technique stopped the training process if the model shows no improvement over three consecutive epochs. The results from the best checkpoints of both experiments showed minimal differences 2, obtaining F1 score of 0.896 for the positive class by using the original dataset.

4.2. LLM prompting

In addition to text and class labels, the organizers provided a sentence_id field that can be sorted to get consecutive sentences in a longer paragraph transcription. An example of this is shown in Table 3 where all the sentences belong to the same intervention from a presidential debate between candidates George Bush and Michael Dukakis in September 1988[2]:

[2]https://www.presidency.ucsb.edu/documents/presidential-debate-winston-salem-north-carolina

Table 3
Consecutive sentences

Sentence_id	Text	class_label
16	I think we've seen a deterioration of values.	No
17	I think for a while as a nation we condoned those things we should have condemned.	No
18	For a while, as I recall, it even seems to me that there was talk of legalizing or decriminalizing marijuana and other drugs, and I think that's all wrong.	No
19	So we've seen a deterioration in values, and one of the things that I think we should do about it in terms of cause is to instill values into the young people in our schools.	No
...	...	...

Table 4
Duplicated texts with different class_label

Sentence_id	Text	class_label
40	We've been dealing with him; he's been dealing drugs to our kids.	Yes
64	We've been dealing with him; he's been dealing drugs to our kids.	No
19096	We are better off than we were four years ago.	Yes
20136	We are better off than we were four years ago.	No
27908	That will not help us compete with China.	No
27910	That will not help us compete with China.	Yes
29254	I do not say that.	Yes
29256	I do not say that.	No
34258	We don't know who the rebels are.	Yes
34260	We don't know who the rebels are.	No

[BUSH:] *"I think we've seen a deterioration of values. I think for a while as a nation we condoned those things we should have condemned. For a while, as I recall, it even seems to me that there was talk of legalizing or decriminalizing marijuana and other drugs, and I think that's all wrong. So we've seen a deterioration in values, and one of the things that I think we should do about it in terms of cause is to instill values into the young people in our schools. We got away, we got into this feeling that value- free education was the thing."*

Further analysis highlights the need of utilizing context to determine the model prediction as there are similar sentences labeled differently as shown in Table 4.

With this in mind, we took two different prompting approaches, both with GPT-3.5-turbo [8]: a baseline prompt where no context is provided and another one where we concatenate all messages previous to the one being predicted:

1. The no-context prompt utilized is: *Check-worthiness definition: The process of finding whether a given TEXT contains verifiable factual claims prone to be fact-checked. You are an expert in check-worthiness. Determine if the TEXT contains a verifiable factual claim that is subject to fact-checking. Before responding, systematically consider these auxiliary questions: "Does it contain a verifiable factual claim?" and "Is it harmful?" Respond only with Yes if the TEXT contains such a claim, and No if it does not.\nTEXT:<Text>*
2. The context prompt utilized is: *Check-worthiness definition: The process of finding whether a given TEXT contains verifiable factual claims prone to be fact-checked. You are an expert in check-worthiness. You will be provided with several sentences but only the last one is relevant to the task; the rest of the text is only provided as extra context if needed. Before responding, systematically consider these auxiliary questions: "Does it contain a verifiable factual claim?" and "Is it harmful?"*

Table 5
Prompting evaluation with dev-test set for the positive class.

Experiment	Accuracy	Precision	Recall	F1
Prompting without context	0.8270	0.8354	0.6111	0.7059
Prompting with context	0.7390	0.7778	0.3241	0.4575

Table 6
Evaluation with test set

Experiment	F1 positive class
* RoBERTa fine-tuning with original set	0.7613
RoBERTa fine-tuning with augmented set	0.7305
Prompting without context	0.6233
Prompting with context	0.4647

Respond with either Yes or No based solely on your analysis of the last sentence. Yes means the last sentence should be fact-checked, No means it shouldn't.\nTEXT:<Text>

The results obtained with both approaches in the dev-test dataset for the positive class are shown in Table 5. Prompting with context underperformed significantly with respect to the no-context version, which scored an F1-score of 0.7059.

5. Results

In this section, we report the results obtained during the evaluation cycle. As only the last submission was selected, the best result from all the experiments we performed was submitted. This result corresponds to the RoBERTa-base fine-tuning approach using the original dataset 2.

Surprisingly, the metrics obtained using the gold labels provided by the organizers 6 show a considerable drop in the F1 score of the positive class compared to the results we obtained with the dev-test set.

The fine-tuned models showcased superior efficacy in determining which texts are worth fact-checking, as it is reflected in the F1 score of the positive class. This is due to the fundamental difference in training processes: while fine-tuning adjusts the model parameters to a specific task through iterative learning, the prompting approach relies solely on inference without any training process. This crucial distinction underpins the disparities in performance observed between the two methods. The fine-tuning process allows the model to adapt and specialize, leveraging task-specific information and fine-grained adjustments.

6. Conclusions and future work

In this paper we presented the systems developed by the SINAI team at the CheckThat! Lab to tackle Task 1: Check-worthiness detection. We compare two different approaches: A RoBERTa-based finetuning with and without data augmentation from the Spanish dataset provided by the organizers, and GPT-3.5-turbo prompting approaches including a baseline and a context-aware prompt. Results show transformer finetuning as a promising technique to tackle this task as we ranked 7th out of 26 participants scoring 0.761363 F1-score for the positive class.

Regarding future work, we aim to use balancing techniques such as downsampling of the majority class or data augmentation with external resources. We would also like to further explore the context-aware prompting approach as we believe extra context is important given the duplicated examples with different labels, even though it did not achieve good results the way we applied it. Other prompting

techniques such as few-shot learning [9] and chain of thought [10] could be useful to tackle this task too.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government.

References

[1] 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.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[3] E. Lazarski, M. Al-Khassaweneh, C. Howard, Using nlp for fact checking: A survey, Designs 5 (2021) 42.

[4] 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, R. Suwaileh, 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.

[5] 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.

[6] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, et al., Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 611–649.

[7] 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.

[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, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[9] A. Parnami, M. Lee, Learning from few examples: A summary of approaches to few-shot learning, 2022. arXiv:2203.04291.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed,

A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.