DSHacker at CheckThat! 2023: Check-Worthiness in Multigenre and Multilingual Content With GPT-3.5 Data Augmentation

Arkadiusz Modzelewski, Witold Sosnowski and Adam Wierzbicki

Polish-Japanese Academy of Information Technology, 86 Koszykowa St., 02-008 Warsaw, Poland

Abstract

This article showcases our approach to check-worthiness detection, a task within the CheckThat! Lab of the 2023 CLEF Conference. This task aimed to design a system capable of determining if a claim, provided in diverse data formats such as tweets, debate snippets, and speech transcriptions, necessitates fact-checking. Our method combined a unified framework for processing content in three languages - English, Spanish and Arabic. At the heart of our system is the XLM-RoBERTa, a pre-trained multilingual model. To enhance its performance, we applied data augmentation strategies using GPT-3.5 provided by OpenAI, which included generating paraphrases and translating text fragments to create a rich dataset. The system's effectiveness is evidenced by its performance against baseline results in all languages, notably winning first place with an F1 score of 0.641 in the Spanish category. Additionally, our exploration sheds light on the attributes of the model's performance in processing different languages, highlighting its exceptional performance in Spanish and indicating room for improvement in handling complex Arabic language structures.

Keywords

Check-Worthiness, Fact-Checking, XLM-RoBERTa, GPT, Data Augmentation, Multilingual, Multigenre

1. Introduction

At a time when disinformation spreads rapidly through various channels, natural language processing (NLP) is becoming a powerful ally in the automatic identification and verification of claims. By extracting relevant information from claims, NLP plays a crucial role in capturing intent and context - essential attributes in assessing the veracity of a claim.

Language models, especially those rooted in deep learning, are an essential subset of NLP technologies that have shown great potential in claim identification and verification. Essentially adept at generating human-like text, language models are trained to assess the probability of word sequences, making them highly flexible and adept at understanding the complexities of natural language.

One of the remarkable advancements in language models is the adoption of Transformer architectures [1], which has led to significant improvements in various NLP tasks, including

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

arkadiusz.modzelewski@pja.edu.pl (A. Modzelewski); witold.sosnowski@pja.edu.pl (W. Sosnowski); adam.wierzbicki@pja.edu.pl (A. Wierzbicki)

0009-0003-1169-831X (A. Modzelewski); 0000-0002-2241-9588 (W. Sosnowski); 0000-0003-0075-7030 (A. Wierzbicki)

claim verification [2]. The Transformer's attention mechanism allows it to focus on different parts of the input text, which is crucial in understanding the context and relationships within the text [1].

Additionally, Transformer-based language models, such as BERT [3] and GPT [4], have been fine-tuned for fact-checking tasks [5]. Fine-tuning involves training the model on a specific dataset related to the task at hand, allowing it to specialize and improve its performance in that particular task [6].

Furthermore, GPT-3, one of the latest and largest language models by OpenAI, has demonstrated the potential for few-shot learning, where the model is presented with a few examples and can generalize to perform tasks even with minimal data [7].

Now that we know how important it is to use NLP and language models to determine whether claims are true or false, let us discuss our approach to check-worthiness detection. Nowadays, information spreads quickly and sometimes can be harmful or disinformative. Our research is about making a system that can automatically check the reliability of claims. The following sections will discuss the challenge we took at CLEF 2023 Labs and how we made a system that can do this in English, Spanish, and Arabic.

1.1. Problem Overview

Claims, particularly those that include misinformation and disinformation, can spread rapidly through various social media platforms and find their way into debates and public speeches. The harmful impact of such claims cannot be underestimated, as they have the potential to mislead and manipulate public opinion. Therefore, determining the check-worthiness of claims that represent different genres and characteristics is often crucial. Traditionally determining the check-worthiness of claims relies on the expertise of professional fact-checkers, debunkers, or human annotators. However, this manual assessment process can be resource-intensive and costly. In this regard, it is essential to develop automated systems for claim identification and verification, which can act as supportive technology for fact-checking organizations and journalism. By utilizing automated systems, we can improve the speed and accuracy of claim evaluation, reducing the reliance on manual assessments. Therefore, CheckThat! Lab (held in the framework of CLEF 2023[1]) organizers introduced a task aimed at developing a solution that could assist specialists in check-worthiness identification.

1.2. Task Description

CheckThat! Lab organizers introduced five different tasks at CLEF 2023. Our research focused on claims check-worthiness identification, and therefore we participated in Task 1: *Check-Worthiness in Multimodal and Multigenre Content.* This task aimed to ascertain the need for fact-checking a claim presented in a text snippet. In this task, we had two kinds of data, which were translated into two subtasks:

- **Subtask 1A (Multimodal)**: check-worthiness assessment in a multimodal approach on tweets that included both a text snippet and an image.

[1] https://clef2023.clef-initiative.eu/

- **Subtask 1B (Multigenre)**: check-worthiness assessment in a multigenre approach on a text representing a tweet or a debate/speech transcription snippet.

Both subtasks were offered in multiple languages, namely subtask 1A in Arabic and English, whereas subtask 1B in Arabic, English, and Spanish. We focused entirely on detecting check-worthy text snippets in a multilingual and multigenre approach. Accordingly, we participated in subtask 1B. For a more comprehensive understanding of the task, we recommend referring to the paper that provides a detailed description of Task 1 [8].

1.3. Our Contribution

Our goal was to develop a single predictive system to assess check-worthiness in three languages. In this regard, we focused on employing a multilingual pre-trained XLM-RoBERTa-large model [9] as the core of the proposed system. The multilingual XLM-RoBERTa-large model was fine-tuned utilizing the available data and additional datasets obtained in data augmentation. We augmented the dataset by employing GPT-3.5[2] to translate and paraphrase the existing data. Our model improved upon baseline models for all languages. Notably, our system achieved the highest performance in the Spanish language, surpassing all other proposed approaches.

2. Related Work

Identifying and detecting disinformation and misinformation have emerged as highly significant research areas. Now, researchers focus on tackling specific challenges related to identifying disinformation, misinformation, and fake news. One of these challenges is, for instance, the recognition of check-worthiness in claims [10]. Hassan et al. [11] prepared a U.S. presidential debate dataset and developed classification models to distinguish between three different categories: check-worthy factual claims, non-factual claims, and unimportant factual claims. In their research, they experimented with three different classical machine learning models, namely Multinomial Naive Bayes Classifier, Support Vector Classifier, and Random Forest Classifier [11]. Jaradat et al. [12] created ClaimRank, an online system for detecting check-worthy claims that supported English and Arabic. ClaimRank, in its system architecture, reused the neural network model proposed by [13]. In addition to using classical machine learning to detect check-worthy claims, some studies proposed the latest pre-trained models. One such solution was the proposal by Kartal and Kutlu [14] to use the BERT model and various features to prioritize claims based on their check-worthiness. Features used by the authors included domain-specific controversial topics, word embeddings, part-of-speech tags, and others [14].

Check-worthiness detection is also the subject of research within Checkthat! Labs from previous years [15, 16, 17, 18, 19]. NUS-IDS team was one of the top performing teams in subtask related to detecting check-worthiness of tweets in 2022 [19]. The NUS-IDS team utilized the multilingual system that effectively took advantage of labeled data in all available languages in provided datasets [20]. AI Rational team in the same subtask employed a pre-trained RoBERTa model with data augmentation [21]. One other team is also worth highlighting. PoliMi-FlatEarthers team fine-tuned a generative pre-trained GPT-3 model with all data in

[2]OpenAI. Available on OpenAI Platform: https://platform.openai.com/ (accessed April 25, 2023)

English [22]. They obtained the third-best performance when applying it to the English *test* dataset. The zero-shot application of the model to other languages was less successful [19].

3. Dataset

The dataset utilized in the Subtask 1B enclosed various genres and languages. The genres that these data represented were tweets and text snippets from transcriptions of debates and public speeches. Three languages were available: English, Spanish and Arabic. In addition, we enriched the data using an augmentation technique that involved the use of GPT-3.5.

The English dataset comprised snippets of transcriptions from debates and public speeches, while the Arabic and Spanish datasets included tweets accompanied by relevant metadata. Our approach focused solely on detecting check-worthiness using text data. As such, we did not use any tweets' metadata. The dataset for each language was divided into *train*, *dev* and *dev_test* dataset. All observations included ground truth labels. During the final phase of the Checkthat! Lab, we got additional unlabeled *test* dataset. It was the final dataset for which we had to generate predictions and submit them for evaluation. For a comprehensive description of the subtask dataset, please refer to the paper by Alam et al. [8].

4. Our Approach

4.1. Data Preparation and Augmentation

As mentioned in Section 3, the datasets provided for Spanish and Arabic contained tweets and corresponding metadata fields, but in the preprocessing step, we decided to discard any additional metadata. We only utilized messages included in tweets and text snippets in English that represented transcripts of debates or public speeches. We employed textual data alongside a binary target label, where the positive class denoted a check-worthiness of the claim.

In the data preparation phase, data augmentation was performed using a variant of OpenAI's GPT-3.5 named *gpt-3.5-turbo*, which is considered highly efficient and cost-effective. Custom prompts were created for generating text translations and paraphrases, thereby enriching the dataset.

Prompts that we utilized in order to produce synthetic paraphrases are as follows:

- **English**: *"Please generate a paraphrase without any additional text or explanation for the following text: <text>"*
- **Spanish**: *"Please generate a paraphrase in Spanish without any additional text or explanation for the following text: <text>"*
- **Arabic**: *"Please generate a paraphrase in Arabic without any additional text or explanation for the following text: <text>"*

We adopted a similar approach for creating synthetic translations:

- **English**: *"Please translate the following text from English to Spanish without any additional text or explanation: <text>"*

- **Spanish**:
  - *"Please translate the following text from Spanish to English without any additional text or explanation: <text>"*
  - *"Please translate the following text from Spanish to Arabic without any additional text or explanation: <text>"*
- **Arabic**: *"Please translate the following text from Arabic to Spanish without any additional text or explanation: <text>"*

Due to resource constraints, we adopted specific rules for translation between languages to ensure quality:

1. The datasets being translated should share the same genre to maintain contextual consistency.
2. The languages involved in translation should belong to the same language family, aiding in generating natural translations due to structural similarities.

We identified English and Spanish as belonging to the Indo-European language family [23]. Table 1 outlines the translation process and justifications.

Table 1
Translation between languages and its reasoning

Original Language	Destination Language	Justification
Arabic	Spanish	Same genre
Spanish	Arabic	Same genre
English	Spanish	Indo-European language family
Spanish	English	Indo-European language family

The most significant challenges arose when generating translations from Spanish to Arabic and from Arabic to Spanish. In these instances, many observations lacked complete translation into the target language as specified. Nonetheless, we decided to include all the generated data in the training set.

After performing data augmentation, we proceeded to remove duplicates from both the original and augmented datasets. While exploring the data, we observed that certain observations were present in both the *train* and *dev* datasets. To ensure the effectiveness of fine-tuning the pre-trained models, it was imperative to eliminate these duplicate instances.

In the hyperparameter tuning phase, we augmented the training dataset exclusively. However, for building the final model, we utilized all the labeled data at our disposal, which included the training, *dev*, and *dev_test* datasets. We applied augmentation techniques to each of these datasets for the training of the final model.

4.2. Model

The primary goal of this stage was to develop a model capable of detecting check-worthy claims. For that, we adopted a unified approach creating a single model for check-worthiness detection

across three given languages. In this regard, we utilized the multilingual RoBERTa-large [9] encoder, a pre-trained model provided by *HuggingFace*, known as *XLM-RoBERTa-large*[3]

The input text was initially tokenized, resulting in an array of tokens, with special tokens such as *[CLS]* denoting the start, *[EOS]* representing the end, and *[SEP]* separating sentences. This tokenized array was then processed through the *XLM-RoBERTa-large* model, generating an array of embeddings corresponding to the input tokens. Subsequently, the embedding array was passed through a fully connected layer, which was followed by a normalization step. During the training phase, the normalized output served as input along with the relevant labels for the binary cross-entropy loss function, enabling the loss calculation.

We predicted the final class label during inference by selecting the highest value from the normalized output.

4.3. Experimental Setup

We started our experiment by creating the training dataset, we combined all the provided *train* datasets from all available languages: English, Spanish, and Arabic. Moreover, we incorporated text snippets generated by GPT-3.5, as previously described in the Subsection 4.1. Next, we formed the validation dataset by combining the respective *dev* datasets for each language.

Once we established the training and validation datasets, we tokenized the text snippets and adjusted their length to 128 tokens by either truncating or padding them. Next, we focused on finding the best hyperparameters for training the model. This involved exploring different values for the batch size (ranging from 4 to 8), learning rate (ranging from 1e-7 to 1e-4), and weight decay (ranging from 1e-4 to 0.1). Additionally, we implemented a linear warmup for the initial 6% of the training steps. The hyperparameter search was conducted using the combined validation sets for all languages. Through this process, we identified the optimal hyperparameters as follows: a batch size of 8, a learning rate of 7.48e-06, and a weight decay of 2.65e-4. During the fine-tuning phase of the experiment, models with the highest F1 score for the positive class were considered the best.

With the optimal hyperparameters determined, our final approach involved training the model on the combined *train*, *dev* datasets, and all augmented data. In the end, we prepared the final predictions on the *test* dataset. The *test* dataset was utilized for the final evaluation and determination of our score, as showcased on the leaderboard.

5. Results

We present our official results and position on the final leaderboard in Table 3. As shown in the table, our model achieved remarkable performance in the Spanish language.

Table 2 shows the comparison between our *dev_test* results and our results on the official leaderboard. It is evident that the *dev_test* results are not fully representative of the final leaderboard performance. For instance, while the model exhibited exceptionally high performance in English in the *dev_test* with a score of 0.946, it did not reflect similarly in the leaderboard with a score of 0.819. This could be due to the differences in data distribution between the *dev_test*

[3]https://huggingface.co/xlm-roberta-large

set and the final *test* set used for the leaderboard. It underlines the importance of ensuring that the model is well-generalized and not overfitting to a specific dataset.

Table 2
Our dev_test results compared with our results on official Leaderboard

Language	Dev_Test	Leaderboard
Arabic	0.483	0.633
English	0.946	0.819
Spanish	0.688	0.641

5.1. Performance in Spanish

In the Spanish language, our model significantly exceeded the baseline and surpassed all the competing teams in terms of F1 score. The F1 score over the positive class was recorded at 0.641, which is almost four times higher than the baseline score of 0.172. This clearly demonstrates the effectiveness of our model in accurately assessing the check-worthiness of multigenre content in Spanish. This superior performance can be attributed to several factors including the robustness of the underlying model, quality of the training data and the fine-tuning strategies we employed. Moreover, it indicates that our model is well-suited for the Spanish language and is capable of capturing the nuances and contextual information necessary for this task.

5.2. Performance in English

Moving on to the English language, our model also performed significantly better than the baseline with an F1 score of 0.819. However, it was outperformed by other teams and secured the 9th position. Despite this, the large margin between our score and the baseline score of 0.462 reflects that our model is capable of effectively identifying check-worthy statements in English. Further optimization and tuning could potentially improve the ranking among other competitors.

5.3. Performance in Arabic

In Arabic, our model's performance was slightly above the baseline, achieving an F1 score of 0.633 compared to the baseline score of 0.625. This relatively modest improvement over the baseline suggests that there might be certain challenges that our model faces when processing Arabic content. We hypothesize that GPT-3.5, which is the backbone of our model, performs worse in generating synthetic Arabic texts compared to English and Spanish. The Arabic language has complex morphological structures and right-to-left script, which might pose challenges for the model. Further investigation is needed to identify the specific areas where the model can be optimized for better performance in Arabic.

Table 3
Official results and ranks in Subtask 1B: Check-Worthiness in Multigenre Content

Arabic			**English**			**Spanish**		
Rank	Team	F1	Rank	Team	F1	Rank	Team	F1
1	ES-VRAI	0.809	1	OpenFact	0.898	**1**	**DSHacker**	**0.641**
2	Accenture	0.733	2	Fraunhofer SIT	0.878	2	ES-VRAI	0.627
3	Z-Index	0.71	3	Accenture	0.86	3	CSECU-DSG	0.599
5	**DSHacker**	**0.633**	**9**	**DSHacker**	**0.819**	4	NLPIR-UNED	0.589
–	Baseline	0.625	–	Baseline	0.462	–	Baseline	0.172

6. Limitations

While our study presents promising results, there are several limitations that need to be acknowledged. One of the critical limitations of this study was the lack of investigation into the effects of data augmentation on the model. Due to budget and time constraints, we did not conduct experiments to compare the model's performance with and without data augmentation techniques. Data augmentation, being one of the key aspects of the model, warrants further investigation to understand its actual contribution to the performance of the model in check-worthiness detection.

Another limitation was not exploring the multilinguality aspect in-depth. Our study used a single model with a dataset that combined training data from different languages. However, we did not compare its performance with dedicated models that were trained on specific languages. Such an analysis would have been insightful in understanding the pros and cons of using a single multilingual model versus multiple monolingual models.

7. Future Work

Given the aforementioned limitations, future work should involve:

- **Investigating Data Augmentation:** A systematic investigation into the effects of data augmentation on model performance. Comparing the model with and without data augmentation will be instrumental in understanding its role in improving check-worthiness detection.
- **Exploring Multilinguality:** Conducting experiments to compare the performance of a single model trained on combined datasets from different languages with models trained on language-specific datasets. This will help in identifying the best approach for multilingual check-worthiness detection.
- **Handling Biases:** Ensuring that the systems are unbiased and fair. Further research could explore techniques for identifying and mitigating biases within the models.
- **Explainability and Interpretability:** Building transparent models that can provide justifications for their predictions.

These future research directions will not only address the limitations of the current study but will also pave the way for more sophisticated and efficient check-worthiness detection systems.

8. Conclusions

In this study, as part of CheckThat! 2023 Lab's Subtask 1B, we implemented a multilingual XLM-RoBERTa-large model with GPT-3.5-based data augmentation to assess the check-worthiness of statements in multilingual and multigenre content. Our model exhibited remarkable performance, particularly in Spanish, where it surpassed all competitors with an F1 score of 0.641. Although the performance was commendable in English, it ranked 9th. In Arabic, the improvement was modest, hinting at challenges faced by the model in processing complex Arabic structures.

In summary, the study signifies a substantial step in automated fact-checking systems, particularly for the Spanish language, by employing pre-trained multilingual models with data augmentation. Future research should focus on overcoming limitations and refining performance across languages.

References

[1] 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).

[2] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[5] A. Sathe, J. Park, Automatic fact-checking with document-level annotations using bert and multiple instance learning, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), 2021, pp. 101–107.

[6] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. Smith, Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, arXiv preprint arXiv:2002.06305 (2020).

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[8] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content, in: Working Notes of CLEF 2023—Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[10] P. Nakov, A. Barrón-Cedeño, G. da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, M. Wiegand, M. Siegel, J. Köhler, Overview ofnbsp;thenbsp;clef–2022 checkthat! lab onnbsp;fighting thenbsp;covid-19 infodemic andnbsp;fake news detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, p. 495–520. URL: https://doi.org/10.1007/978-3-031-13643-6_29. doi:10.1007/978-3-031-13643-6_29.

[11] 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.

[12] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, Claimrank: Detecting check-worthy claims in arabic and english, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2018, pp. 26–30.

[13] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, 2017, pp. 267–276.

[14] Y. S. Kartal, M. Kutlu, Re-think before you share: A comprehensive study on prioritizing check-worthy claims, IEEE Transactions on Computational Social Systems (2022).

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer, 2022, pp. 495–520.

[20] M. Du, S. D. Gollapalli, S.-K. Ng, Nus-ids at checkthat! 2022: identifying check-worthiness of tweets using checkthat5 (2022).

[21] A. Savchev, Ai rational at checkthat! 2022: using transformer models for tweet classification, Working Notes of CLEF (2022).

[22] S. Agresti, S. A. Hashemian, M. J. Carman, Polimi-flatearthers at checkthat! 2022: Gpt-3 applied to claim detection (2022).

[23] T. V. Gamkrelidze, V. V. Ivanov, The early history of indo-european languages, Scientific American 262 (1990) 110–117.