

Z-Index at CheckThat! 2023: Unimodal and Multimodal Check-Worthiness Classification

Prerona Tarannum¹, Md. Arid Hasan^{1,*}, Firoj Alam² and Sheak Rashed Haider Noori¹

¹Daffodil International University

²Qatar Computing Research Institute

Abstract

In this study, we report our participation in CheckThat! lab's Task 1. The aim is to determine whether a claim made in either unimodal or multimodal content is worth fact-checking. We implemented standard preprocessing and fine-tuned the XLM-RoBERTa-large model. Additionally, we applied zero-shot learning and utilized a feed-forward network with embeddings for unimodal content. For subtask 1A submission, we used combined BERT-based models (BERT and BERT multilingual), ResNet50, and Feed Forward network and we ranked as 3rd (Arabic) and 5th (English). We used feed forward network with embeddings for subtask 1B submission and ranked as 3rd in Arabic and 6th in both English and Spanish. In further experiments, our evaluation shows that XLM-RoBERTa-large model outperforms the other models.

Keywords

Multimodal Fact-checking, Multigenre Fact-checking, Checkworthiness identification, Misinformation, XLM-RoBERTa-large, ResNet50, Feed Forward Network

1. Introduction

Social media is now considered one of the mainstream communication platforms for exchanging information among people with a few taps on the screen. While it is recognized as a primary source of information that can create a positive impact [1], it also has been exploited by malicious actors. These actors use the platforms to spread disinformation and misinformation that can be harmful to individuals, society, and organizations [2]. This includes hate speech [3], hostility [4, 5], harmful memes [6], propagandistic news and memes [7], abusive language [8], cyberbullying and cyber-aggression [9], and rumours [10].

The negative impact of such harmful and misleading information has heightened interest among researchers and organizations in identifying and curbing its spread among the public. Numerous efforts and studies have been conducted to automate this identification process [11, 12, 13, 14].

Among other research efforts, over the past few years, the CheckThat! Lab initiative has been pivotal in advancing the development of systems for detecting check-worthiness in political

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


*Corresponding author.

✉ prerona15-14134@diu.edu.bd (P. Tarannum); arid.cse0325.c@diu.edu.bd (Md. A. Hasan); fialam@hbku.edu.qa (F. Alam); drnoori@daffodilvarsity.edu.bd (S. R. H. Noori)

🆔 0000-0002-3292-1870 (P. Tarannum); 0000-0001-7916-614X (Md. A. Hasan); 0000-0001-7172-1997 (F. Alam); 0000-0001-6937-6039 (S. R. H. Noori)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

debates, tweets, and transcripts [15, 16, 17, 18]. This year, the CheckThat! Lab initiative has introduced five tasks that include seven languages – Arabic, Dutch, English, German, Italian, Spanish, and Turkish. Among the five tasks, the task 1 is check-worthiness detection in multigenre and multimodal content [19, 20].

We participated task 1, which consists of two subtasks that are estimating check-worthiness in multimodal (Subtask-1A) and multigenre (Subtask-1B) content. Subtask-1A is offered in two languages (Arabic and English) and Subtask-1B is offered in three languages (Arabic, English, and Spanish) where we participated in all languages. For our experimental setup, we choose widely used pre-trained transformer-based language models. However, challenges arise when multilingual versions of these pre-trained models are applied to tasks where the facts and claims differ by country [21]. In such cases, the potential for misinformation to spread across languages through knowledge transfer becomes evident.

For our study, we utilized the power of transformer models such as BERT (monolingual), multilingual-BERT (multilingual), and XLM-RoBERTa (multilingual) to extract text features. For visual features, we utilized the CNN-based pre-trained model, ResNet50. GPT-4 was applied for zero-shot learning, and we fine-tuned a feed-forward network to learn from the GPT embeddings (text-embedding-ada-002).

The structure of this paper is as follows: we summarize the relevant related works for this study in Section 2. In Section 3, we report the methodology. A detailed discussion of the results of our study is provided in Section 4. Finally, we draw conclusions in Section 5.

2. Literature Review

The current information ecosystem, including online and social media, is abundant with incorrect claims. These are not only present in textual form but are also found in misleading photos and videos, casting a veil over reality. Many fact-checking organizations have been established to address this issue, including FactCheck.org¹, Snopes², PolitiFact³, and FullFact⁴. Moreover, international initiatives such as the Credibility Coalition⁵ and Eufactcheck⁶ [22] have been launched to extend these efforts.

The ClaimBuster system [23] was one of the early studies in this area, comprising 28,029 sentences transcribed from 30 historical US election debates. The transcriptions were done by students, professors, and journalists. Hassan et al. [24] described how the fact-checking platform ClaimBuster combines supervised learning to identify significant factual assertions in political debates. In their study, Thorne and Vlachos [25] reviewed automated fact-checking research and related areas, including task formulations and approaches from various papers and authors. They indicated that supervised learning methods are used for automated fact-checking and discussed the emerging field of fact-checking involving images and footage. Shi and Weninger [26] evaluated thousands of extracted claims using knowledge graph datasets,

¹<http://www.factcheck.org/>

²<http://www.snopes.com/fact-check/>

³<http://www.politifact.com/>

⁴<http://fullfact.org/>

⁵<https://credibilitycoalition.org/>

⁶<https://eufactcheck.eu/>

specifically DBpedia and SemMedDB. These datasets were compiled from politics, biology, geography, and history through a public knowledge graph.

The study by Alam et al. [2] used pretrained language models, focusing on COVID-19 topics in four different languages, and achieved robust performances. In [27], the authors benchmarked a multimodal dataset titled ‘Factify’. This dataset, collected from Twitter and news sources and was manually annotated. Nakov et al. [28] surveyed what is needed for human fact-checkers in order to support them in their work. This study shows a partial difference between what fact-checkers want and what technology has to offer. Li et al. [29] introduced the MM-COVID dataset, a multilingual and multidimensional COVID-19 fake news data repository designed to combat disinformation. This repository includes data in six different languages: English, Spanish, Portuguese, Hindi, French, and Italian. Their study further explores into cross-language fake news detection, research facilitated by multimodal data, and rapid fake news detection. Suryavardan et al. [30] studied multi-modal fact verification with the use of Vision Transformer for visual features and BERT for text features on Factify multimodal dataset consisting of 50,000 data instances.

The study of Zhuang and Zhang [31] focuses on unimodal and multimodal fact-checking using the transformer-based model on Factify dataset and achieved an F1 score of 75.59. Augenstein et al. [32] used *MultiFC* dataset, which is collected from 26 fact-checking websites in English that are accompanied by text sources and extensive metadata and are rated for accuracy by professional journalists. The study of end-to-end multimodal fact-checking and explanation generation done by Yao et al. [33] used *MOCHEG* dataset that includes the input as a claim and a substantial amount of online content, such as articles, images, videos, and tweets.

Guo et al. [34] focused on automated fact-checking focusing on three stages of the fact-checking framework - claim detection, evidence retrieval, claim verification, and justification production using neural network-based approach. Alam et al. [12] offers a snapshot of recent studies, advocating for the integration of harm and factuality into a multimodal disinformation detection system.

In recent years, remarkable research outcomes have emerged from shared tasks, including those from the CLEF CheckThat! labs’ shared tasks [35, 36, 15, 37, 38, 39]. These tasks have presented challenges concerning the automatic identification [40, 41] and verification [42, 43, 44] of claims in COVID-19 news, political debates, and tweets [17].

3. Methodology

3.1. Data

We utilized the dataset provided by the organizers of the CLEF CheckThat! 2023 lab for Task 1: Check-Worthiness in Multimodal and Unimodal Contents [19, 45, 20]. Subtask 1A comprises both text and images as input data, collected from Twitter. Subtask 1B includes only text data, which is a mix of political debates, transcriptions, and tweets on topics such as COVID-19 and politics. We present the distribution of the official datasets used in this shared task for our experiments in Table 1 (for Subtask 1A) and Table 2 (for Subtask 1B).

Table 1

Data splits and distributions of Subtask 1A: Check-Worthiness of multimodal content

Class labels	Train	Dev	Dev-Test	Test	Total
Arabic					
No	1,421	207	402	792	2,822
Yes	776	113	220	203	1,312
Total	2,197	320	622	995	4,134
English					
No	1,536	184	374	459	2,553
Yes	820	87	174	277	1,358
Total	2,356	271	548	736	3,911

Table 2

Data splits and distributions of Subtask 1B: Check-Worthiness of multigenre unimodal content

Class label	Train	Dev	Dev-Test	Test	Total
Arabic					
No	4,301	789	682	377	6,149
Yes	1,758	485	411	123	2,777
Total	6,059	1,274	1,093	500	8,926
English					
No	12,818	4,270	794	210	18,092
Yes	4,058	1,355	238	108	5,759
Total	16,876	5,625	1,032	318	23,851
Spanish					
No	5,280	2,161	4,296	4,491	16,228
Yes	2,208	299	704	509	3,720
Total	7,488	2,460	5,000	5,000	19,948

3.2. Preprocessing

The datasets for CheckThat! lab Task 1 were collected from multiple sources including Twitter. These datasets contain numerous symbols, URLs, and invisible characters. To cleanse this noisy data, we underwent several preprocessing steps. We began by removing extraneous characters and URLs, followed by the elimination of stopwords, hashtags, and usernames from the data.

3.3. Model

We used only deep learning models to run both multimodal and multigenre classification experiments. For multimodal classification, we used transformer-based BERT [46] and XLM-RoBERTa [47] models for text input, ResNet [48] for image input, and a Feed Forward (FF) network for the fusion of both representations. The parameter sizes of BERT and XLM-RoBERTa

Table 3

Official results on the test set and overall ranking of Task 1: Check-Worthiness in Multimodal and Multigenre Content. Feed Forward Network (FF). **Bold** indicates our systems.

Language	Model	F1 (postive class)	Rank
Subtask-1A: Check-Worthiness in Multimodal Content			
Arabic	BERT-m + ResNet50 + FF	0.301	3 rd
	Best system	0.399	1 st
	Baseline	0.299	-
English	BERT + ResNet50 + FF	0.495	5 th
	Best system	0.712	1 st
	Baseline	0.474	-
Subtask-1B: Check-Worthiness in Multigenre Content			
Arabic	FF + embeddings	0.710	3 rd
	Best system	0.809	1 st
	Baseline	0.625	-
English	FF + embeddings	0.838	6 th
	Best system	0.898	1 st
	Baseline	0.462	-
Spanish	FF + embeddings	0.496	6 th
	Best system	0.641	1 st
	Baseline	0.172	-

Table 4

Detailed results on the test set of **Task 1A: Check-Worthiness in Multimodal Content**. **Bold** indicates positive class F1 score. *Underline* indicates best F1 score for each language. * indicates the model trained on both the training and development sets. XLM-R: XLM-RoBERTa-large.

Class label	Model	Accuracy	Precision	Recall	F1 Score
Arabic					
No	BERT-m + ResNet50 + FF	45.03	79.52	41.67	54.68
Yes			20.34	58.13	30.14
No	XLM-R + ResNet50 + FF	31.76	86.45	16.92	28.30
Yes			21.67	89.66	34.90
English					
No	BERT + ResNet50 + FF	51.77	66.77	45.10	53.84
Yes			40.85	62.82	49.50
No	XLM-R + ResNet50 + FF	48.10	70.59	28.76	40.87
Yes			40.44	80.14	53.75

are the largest of the transformer-based models.⁷ The size of the network and the number of parameters determine computation time and learning performance. For these two models, we used the base and multilingual versions of the BERT model and the large version of the XLM-RoBERTa⁸ model. Although ResNet50 has only more than 23 million trainable parameters, it provides comparatively better performances. Our rationale for choosing different models was to understand and report their performance in different languages.

⁷110 million parameters in *BERT multilingual* and 550 million in *XLM-RoBERTa large*

⁸XLM-RoBERTa trained on the multilingual dataset

Table 5

Detail results on the test set of **Task 1B: Check-Worthiness in Multigenre Content**. **Bold** indicates positive class F1 score. Underline indicates best F1 score for each language. * indicates the model trained on both the training and development sets. XLM-R: XLM-RoBERTa-large.

Class label	Model	Accuracy	Precision	Recall	F1 Score
Arabic					
No	FF + embeddings	63.40	37.80	75.61	50.41
Yes			88.19	59.42	71.00
No	Zero-shot	31.00	22.13	42.28	29.05
Yes			73.57	27.32	39.85
No	XLM-R	65.80	41.18	91.06	56.71
Yes			95.18	57.56	<u>71.74</u>
No	XLM-R*	63.60	39.58	91.06	55.17
Yes			94.93	54.64	69.36
English					
No	FF + embeddings	90.25	88.09	98.57	93.03
Yes			96.39	74.07	83.77
No	Zero-shot	49.06	65.58	48.10	55.49
Yes			33.54	50.93	40.44
No	XLM-R	91.20	89.22	98.57	93.67
Yes			96.51	76.85	<u>85.57</u>
No	XLM-R*	90.25	88.09	98.57	93.03
Yes			96.39	74.07	83.77
Spanish					
No	FF + embeddings	89.16	94.54	93.32	93.93
Yes			47.09	52.46	49.63
No	Zero-shot	47.10	89.91	46.83	61.58
Yes			10.29	49.51	17.04
No	XLM-R	93.44	95.70	97.06	96.37
Yes			70.34	61.49	65.62
No	XLM-R*	93.58	95.80	97.11	96.45
Yes			70.98	62.48	<u>66.46</u>

3.4. Experiments

3.4.1. Subtask-1A

For Subtask-1A, we utilized a transformer-based pretrained model for text input and a CNN-based pretrained model for image input. The output from each model was concatenated and passed through a simple feed-forward network. We deployed two different architectures to train and evaluate each language. For the first experimental setup, focusing on the Arabic language, we chose the multilingual version of BERT for text input and ResNet50 for image input.

For the first experimental setup focused on the English language, we chose the BERT base version for text input and ResNet50 for image input. For the second experimental setups in both languages, we selected XLM-RoBERTa-large for text input and ResNet50 for image input.

3.4.2. Subtask-1B

Transformer Models We used the Transformer Toolkit [49] for transformer-based models. For Subtask-1B, we fine-tuned the XLM-RoBERTa model [47] with a learning rate of $1e - 5$, a maximum sequence length of 128, and a batch size of 16. We utilized a model-specific tokenizer available with the toolkit for our study. During training on the train split only, we set the epochs to 4, 14, and 5 for Arabic, English, and Spanish, respectively. However, when we merged the train and development sets for training, we ran 3, 14, and 4 epochs for Arabic, English, and Spanish, respectively.

Zero-shot Learning We used GPT-4 [50] for zero-shot learning which is also a transformer-based model. We simply used the test set for evaluating the GPT-4 model without using any kind of training data. As for the prompt, we used the similar format discussed in [51], as also shown in Listing 1. Our prompt was relatively simple, which can be explored further in future studies.

Listing 1: Example of zero-shot prompt.

```
Classify the following text into one of the two categories: Yes or No.\n\n  
text: {inputText}\n  
label:\n
```

Feed Forward Network with Embedding First, we extract the embeddings using OpenAI’s text-embedding-ada-002 model for each data split. We then fine-tune a feed-forward network on the embeddings extracted from the training set to train our model. Our feed-forward model utilizes the Rectified Linear Unit (ReLU) activation function. We have set the learning rate to 0.001 and the hidden layer size to 500. We validate our model using the validation set and finally, we evaluate the model using the test set.

4. Results and Discussion

The official results and rankings evaluated by the lab organizers are presented in Table 3. The evaluation metric F1-score with respect to positive class is considered for Task 1. We also reported the best system and the baseline system along with our system in Table 3. Overall, the performance of the multimodal systems is relatively lower than that of unimodal multigenre systems, and the performance for Arabic is comparatively lower.

The detailed classification results for two subtasks and each language are reported in Tables 4 and 5. We re-ran all the experiments and reported the detailed results once the submission period had ended and the gold set had been made available. From the reported results, we can conclude that XLM-RoBERTa-large pretrained language model performs better for subtask-1B. Although the XLM-RoBERTa-large language model provides a better F1-score with respect to the positive class for subtask-1A, it did not perform satisfactorily by a large margin for the negative class.

5. Conclusion

As part of the shared task, the organizers of CLEF CheckThat! Lab 2023 provided multimodal (images and texts) and multigenre datasets. We began by cleaning the data using standard preprocessing steps, followed by conducting comparative experiments. Our investigation of transformer-based models (e.g., BERT, XLM-RoBERTa, etc.) found that the XLM-RoBERTa large version model outperforms other language models. In the submissions for subtask 1A, we ranked 3rd (Arabic) and 5th (English). For subtask 1B submissions, we ranked 3rd in Arabic and 6th in both English and Spanish. Our future study includes exploring GPT-based large language models.

References

- [1] A. Perrin, Social media usage. pew research center 2015: 52-68, 2020.
- [2] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, 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, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 611–649.
- [3] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* 51 (2018) 1–30.
- [4] S. Brooke, “Condescending, Rude, Assholes”: Framing gender and hostility on Stack Overflow, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 172–180.
- [5] S. Joksimovic, R. S. Baker, J. Ocumpaugh, J. M. L. Andres, I. Tot, E. Y. Wang, S. Dawson, Automated identification of verbally abusive behaviors in online discussions, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 36–45.
- [6] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4439–4455.
- [7] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval ’21, Association for Computational Linguistics, Online, 2021, pp. 70–98.
- [8] H. Mubarak, K. Darwish, W. Magdy, Abusive language detection on arabic social media, in: Proceedings of the first workshop on abusive language online, 2017, pp. 52–56.
- [9] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, 2018, pp. 1–11.

- [10] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, *Information Sciences* 497 (2019) 38–55.
- [11] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, *CSUR* 53 (2020) 1–40.
- [12] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, *arXiv preprint arXiv:2103.12541* (2021).
- [13] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. D. Pietro, P. Nakov, A survey on computational propaganda detection, in: C. Bessiere (Ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, ijcai.org, 2020, pp. 4826–4832.
- [14] T. H. Afridi, A. Alam, M. N. Khan, J. Khan, Y. K. Lee, A multimodal memes classification: A survey and open research issues, in: *5th International Conference on Smart City Applications, SCA 2020*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 1451–1466.
- [15] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, LNCS*, 2019, pp. 301–321.
- [16] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeño, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media, *CEUR Workshop Proceedings*, 2020.
- [17] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, *LNCS* (12880), 2021.
- [18] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The CLEF-2022 CheckThat! Lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), *Advances in Information Retrieval*, Springer International Publishing, Cham, 2022, pp. 416–428.
- [19] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The CLEF-2023 CheckThat! Lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval*, Springer Nature Switzerland, Cham, 2023, pp. 506–517.
- [20] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouni, 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.

- [21] K. Singh, G. Lima, M. Cha, C. Cha, J. Kulshrestha, Y.-Y. Ahn, O. Varol, Misinformation, believability, and vaccine acceptance over 40 countries: Takeaways from the initial phase of the covid-19 infodemic, *Plos one* 17 (2022) e0263381.
- [22] M. Stencel, Number of fact-checking outlets surges to 188 in more than 60 countries, *Duke Reporters' LAB* (2019) 12–17.
- [23] 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, CIKM '15*, Association for Computing Machinery, Melbourne, Australia, 2015, pp. 1835–1838.
- [24] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster, in: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1803–1812.
- [25] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, *arXiv preprint arXiv:1806.07687* (2018).
- [26] B. Shi, T. Weninger, Discriminative predicate path mining for fact checking in knowledge graphs, *Knowledge-based systems* 104 (2016) 123–133.
- [27] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, et al., Benchmarking multi-modal entailment for fact verification, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [28] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated fact-checking for assisting human fact-checkers, in: *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI '21*, 2021, pp. 4551–4558.
- [29] Y. Li, B. Jiang, K. Shu, H. Liu, Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation, *arXiv preprint arXiv:2011.04088* (2020).
- [30] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, et al., Factify 2: A multimodal fake news and satire news dataset, *Proceedings http://ceur-ws.org ISSN 1613 (2022) 0073*.
- [31] Y. Zhuang, Y. Zhang, Yet at factify 2022: Unimodal and bimodal roberta-based models for fact checking, in: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, CEUR, 2022.
- [32] I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4685–4697.
- [33] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models, *arXiv preprint arXiv:2205.12487* (2022).
- [34] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, *Transactions of the Association for Computational Linguistics* 10 (2022) 178–206.
- [35] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, M. Kutlu, W. Zaghouani, C. Li,

- S. Shaar, H. Mubarak, A. Nikolov, Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets (2022).
- [36] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouni, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the CLEF-2018 Check-That! lab on automatic identification and verification of political claims, in: CLEF, Lecture Notes in Computer Science, Springer, Avignon, France, 2018, pp. 372–387.
- [37] T. Elsayed, P. Nakov, A. Barrón-Cedeño, M. Hasanain, R. Suwaileh, G. Da San Martino, P. Atanasova, CheckThat! at CLEF 2019: Automatic identification and verification of claims, in: Advances in Information Retrieval, ECIR '19, Springer International Publishing, Cologne, Germany, 2019, pp. 309–315.
- [38] S. Shaar, F. Alam, G. Da San Martino, A. Nikolov, W. Zaghouni, P. Nakov, A. Feldman, Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection, in: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF '21', Association for Computational Linguistics, Online, 2021, pp. 82–92.
- [39] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
- [40] P. Atanasova, L. Màrquez, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, W. Zaghouni, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 1: Check-worthiness, in: CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018.
- [41] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, G. Da San Martino, Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness, in: [52], 2019.
- [42] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghouni, S. Kyuchukov, G. Da San Martino, P. Nakov, Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality, in: CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018.
- [43] M. Hasanain, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, P. Nakov, Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality, in: [52], 2019.
- [44] S. Shaar, F. Haouari, W. Mansour, M. Hasanain, N. Babulkov, F. Alam, G. Da San Martino, T. Elsayed, P. Nakov, Overview of the CLEF-2021 CheckThat! lab task 2 on detecting previously fact-checked claims in tweets and political debates, 2021.
- [45] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, , T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouni, Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos,

- G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023.
- [46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '19*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.
- [47] 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, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20*, Association for Computational Linguistics, Online, 2020, pp. 8440–8451.
- [48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP '20*, Association for Computational Linguistics, Online, 2020, pp. 38–45.
- [50] OpenAI, Gpt-4 technical report, arXiv (2023).
- [51] A. Abdelali, H. Mubarak, S. A. Chowdhury, M. Hasanain, B. Mousi, S. Boughorbel, Y. E. Kheir, D. Izham, F. Dalvi, M. Hawasly, et al., Benchmarking Arabic AI with large language models, arXiv preprint arXiv:2305.14982 (2023).
- [52] L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), *Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings*, 2019.