The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness

Alberto Barrón-Cedeño[1], Firoj Alam[2], Tanmoy Chakraborty[3], Tamer Elsayed[4], Preslav Nakov[5], Piotr Przybyła[6,7], Julia Maria Struß[8], Fatima Haouari[4], Maram Hasanain[2], Federico Ruggeri[9], Xingyi Song[10], and Reem Suwaileh[11]

[1] DIT, Università di Bologna, Forlì, Italy
[2] Qatar Computing Research Institute, HBKU, Doha, Qatar
[3] Indian Institute of Technology Delhi, New Delhi, India
[4] Qatar University, Doha, Qatar
[5] Mohamed bin Zayed University of Artificial Intelligence
[6] Univesitat Pompeu Fabra, Barcelona, Spain
[7] Institute of Computer Science, Polish Academy of Sciences, Poland
[8] University of Applied Sciences Potsdam, Potsdam, Germany
[9] DISI, Università di Bologna, Bologna, Italy
[10] University of Sheffield, Sheffield, UK
[11] HBKU, Doha, Qatar
`https://checkthat.gitlab.io`

Abstract. The first five editions of the `CheckThat!` lab focused on the main tasks of the information verification pipeline: check-worthiness, evidence retrieval and pairing, and verification. Since the 2023 edition, the lab has been focusing on new problems that can support research an decision-making during the verification process. In this 2024 edition, we focus on new problems and —for the first time— we propose six tasks in fifteen languages (Arabic, Bulgarian, English, Dutch, French, Georgian, German, Greek, Italian, Polish, Portuguese, Russian, Slovene, Spanish, and code-mixed Hindi-English): Task 1 is on estimation of check-worthiness (the only task that has been present in all `CheckThat!` editions), Task 2 is on identification of subjectivity (a follow up of the `CheckThat!` 2023 edition), Task 3 is on identification of teh use of persuasion techniques (a follow up of SemEval 2023), Task 4 detection of hero, villain, and victim from memes (a follow up of CONSTRAINT 2022), Task 5 Rumor Verification using Evidence from Authorities (a first), and Task 6 robustness of credibility assessment with adversarial examples (a first). These tasks represent challenging classification and retrieval problems at the document and at the span level, including multilingual and multimodal settings.

Keywords: disinformation · fact-checking · check-worthiness · subjectivity · political bias · factuality · authority finding · model robustness

---

General and task coordinators appear first, in alphabetical order.

1 Introduction

During its previous five editions, the `CheckThat!` lab has focused on developing technology to assist the *journalist fact-checker* during the main steps of the verification process [31, 11, 10, 7, 6, 32, 33, 29, 30, 5, **?**]. Given a document, or a claim, it first has to be assessed for check-worthiness, i.e. whether a journalist should check its veracity. If this is so, the system needs to retrieve claims verified in the past that could be useful to fact-check the current one. Further evidence to verify the claim could be retrieved from the Web, if necessary. Finally, with the evidence gathered from diverse sources, a decision can be made: whether the claim is factually true or not. This year, we propose six tasks:

Task 1 Check-worthiness estimation: to identify claims that could be important to verify on social- and mainstream media (the only task that has been organized during all editions of the lab; cf. Section 2).

Task 2 Subjectivity in news articles: to spot text that should be processed with specific strategies [41]; benefiting the fact-checking pipeline [21, 23, 50] (cf. Section 3).

Task 3 Persuasion techniques: to identify text spans in which a persuasion technique is being issued to influence the reader (cf. Section 4).

Task 4 Detecting hero, villain, and victim from memes: to predict the role of each entity: *hero*, *villain*, *victim*, or *other* in a given meme and a list of entities (cf. Section 5).

Task 5 Rumor Verification using Evidence from Authorities: to retrieve evidence from trusted sources (authorities that have "real knowledge" on the matter) and determine if the rumor is supported, refuted, or unverifiable according to the evidence (cf. Section 6).

Task 6 Robustness of credibility assessment with adversarial examples: to discover examples indicating low robustness of misinformation detection models (cf. Section 7).

2 Task 1: Check-Worthiness Estimation

Motivation Fact-checking is a complex process. Before assessing the truthfulness of a claim, determining if it can be fact-checked at all is essential. Given the time-consuming nature of this process, it is important to prioritize claims that are important to be fact-checked.

Task definition The aim of this task is to assess whether a statement, sourced from either a tweet or a political debate, requires fact-checking [1]. To make this decision, one must consider questions such as "Does it contain a verifiable factual claim?" and "Could it be harmful?" before assigning a final label for its check-worthiness.

Data The dataset is comprised of multigenre content in Arabic, English, Dutch and Spanish. The Arabic and Dutch datasets consist of tweets that were collected using keywords related to COVID-19 and vaccines, following the annotation schema described in [2]. The dataset for English consists of transcribed sentences from candidates during the US presidential election debates and annotated by human annotators [4]. We use essentially the same dataset reported in [4], with some updates that reflect improved annotation accuracy. The Spanish dataset consists of tweets collected from Twitter accounts and transcriptions from Spanish politicians, which are manually annotated by professional journalists who are experts in fact-checking. These datasets include 8.9k, 1.9k, 23.9k and 30k instances in Arabic, Dutch, English, and Spanish, respectively [1, 28]. We split them into training (∼74%), development (∼12%), and development-test (∼15%) sets (an average estimate from all languages) to facilitate training, parameter tuning, and to obtain initial results on the development-test set. For the evaluation of systems in this lab edition, new test sets containing ∼500 instances per language will be released.

Evaluation This is a binary classification task and we evaluate it on the basis of the F_1-measure on the check-worthiness class.

3 Task 2: Subjectivity Detection

Motivation Verifiable facts are not only communicated in objective and neutral statements, but can also be found in subjectively colored ones. While objective sentences can be directly considered for verification, subjective ones require additional processing steps, e.g., extracting an objective version of the contained claims.

Task definition Given a sentence from a news article, determine whether it is subjective or objective. This is a binary classification task and is offered in Arabic, English, German, Italian and in a cross-lingual setting.

Data For training and validation we provide 1.9k sentences in Arabic, 1.3k in English, 1.3k in German, and 2.2k in Italian from last year's iteration [12]. About 300 new sentences are being collected and labelled for each language to be used as novel test sets. The dataset for the cross-lingual setting will be compiled from the individual datasets of the aforementioned languages.

Evaluation We use macro-averaged F_1-measure as the evaluation metric.

4 Task 3: Detection of Persuasion Techniques in News Articles

Motivation A major characteristic of disinformation is that it is not just about lying, but also about convincing people to think or to act in a specific way.

Thus, it is conveyed using specific rhetorical devices: persuasion techniques (e.g., emotional appeals, logical fallacies, personal attacks). Here, we aim to detect the use of such techniques in news articles in various languages.

Task definition Given a set of news articles and a list of 23 persuasion techniques organized into a 2-tier taxonomy, including logical fallacies and emotional manipulation techniques that might be used to support flawed argumentation [35], the task consists of identifying the spans of texts in which each technique occurs. This is a multi-label multi-class sequence tagging task.

Data We will use an existing corpus, consisting of $2k$ news articles in 9 languages annotated with $48K$ instances of persuasion techniques [36], as our training dataset. A new test dataset of ~ 500 news articles in Arabic, Bulgarian, English, Portuguese, and Slovene will be provided.

Evaluation The task is evaluated using an extension of the F_1-measure taking into account partial overlaps between predicted and golden spans [9], and an evaluation at both coarse- and fine-grained level with respect to the type of persuasion technique is envisaged.

5 Task 4: Detecting the Hero, the Villain, and the Victim from Memes

Motivation Memes, characterized by their diverse multimodal nature, are frequently employed to communicate intricate concepts effortlessly on social media. However, this simplicity can sometimes oversimplify intricate concepts, leading to the potential delivery of harmful content, often wrapped in humor. While previous studies identified various types of harm caused by memes [24, 38, 43, 47], they largely overlook nuanced analyses like "narrative framing", especially in resource-constrained settings. Current approaches have limitations in addressing multimodality and reasoning about visual and semantic elements in memes, as noted in prior findings [45]. Identifying narrative roles in memes is crucial for in-depth semantic analysis, especially when examining their potential connection to harmful content like hate speech, offensive material, and cyberbullying [44].

Task definition The task aims to determine the roles of entities within memes, categorizing them as "hero", "villain", "victim", or "other" in a multi-class classification setting that considers systematic modeling of multimodal semiotics [45].

Data We already have the `HVVMemes` dataset [46], including $6.9k$ labeled instances. Additionally, we will introduce a new test dataset of 500 instances for the following languages: Arabic, English, and code-mixing of Hindi and English.

Evaluation The macro-averaged F_1-measure will primarily assess model performance. Two role-label experts will annotate each official test set, overseen by a consolidator following guidelines from previous work [46].

6 Task 5: Rumor Verification using Evidence from Authorities

Motivation Several existing studies addressed rumor verification in social media by exploiting evidence extracted from propagation networks or the Web [34, 20, 18]. Finding and incorporating authorities for rumor verification in Twitter was proposed recently [17, 16, 15]. In the previous edition of the lab, we offered the task of *Authority Finding in Twitter* [19]; this year, we offer a follow-up task with the objective of retrieving evidence from timelines of authorities, and, accordingly, deciding whether the rumors are supported, refuted, or unverifiable.

Task definition Given a rumor expressed in a tweet and a set of authorities (one or more authority Twitter accounts) for that rumor, represented by a list of tweets from their timelines during the period surrounding the rumor, the system should retrieve up to 5 evidence tweets from those timelines, and determine if the rumor is supported (true), refuted (false), or unverifiable (in case not enough evidence to verify it exists in the given tweets) according to the evidence. This task is offered in both Arabic and English.

Data The dataset comprises 160 Arabic rumors expressed in tweets selected from the AuFIN [17, 19] and AuSTR [16, 15] datasets, and 693 timelines of authority Twitter accounts comprising about $34k$ annotated tweets in total. The same data will be automatically translated to English and validated manually. The data will be split into 60%, 20%, and 20% of the rumors for training, development, and testing respectively.

Evaluation The official evaluation measure for evidence retrieval is Mean Average Precision (MAP). The systems get no credit if they retrieve any tweets for unverifiable rumors. Other evaluation measures to be considered are Recall@5 and Precision@5. For rumor classification, we use the F_1-measure. Additionally, we also consider a strict evaluation where the rumor label is considered correct **only if** at least one retrieved authority evidence is correct.

7 Task 6: Robustness of Credibility Assessment with Adversarial Examples

Motivation The aim of the task is to assess the *robustness* of text classifiers in the misinformation detection domain, i.e. their resilience to input data that were purposefully prepared to elicit a misguided response, known as *adversarial examples* (AEs). The vulnerability of deep learning models to AEs has been initially shown for image classification [13, 48], but such weaknesses exist for text as well, even though finding them is more challenging [51]. However, exploring this area is of paramount importance, especially in the case of misinformation detection challenges, where motivated adversaries are active [39].

Task definition The task is realized in five domains: style-based news bias assessment (HN), propaganda detection (PR), fact checking (FC), rumour detection (RD) and COVID-19 misinformation detection (C19). For each domain, the participants are provided with three victim models, trained for the corresponding binary classification task, as well as a collection of 400 text fragments. Their aim is to prepare adversarial examples, which preserve the meaning of the original examples, but are labelled differently by the classifiers.

Data The task is based on the publicly available corpora with expert-annotated credibility used in the BODEGA framework [40]. HN uses news articles [37] gathered for SemEval-2019 Task 4 [25]; PR is based on the corpus accompanying SemEval-2020 Task 11 [8], with 14 propaganda techniques annotated in 371 newspapers articles by professional annotators; FC uses the claims-evidence pairs gathered for FEVER [49]; RD is based on the augmented dataset of rumors and non-rumors for rumor detection [14], created from Twitter threads. Additionally, C19 will use a previously unreleased dataset [22, 27].

Evaluation The quality of the adversarial examples will be assessed using the BODEGA score [40], which combines the change in the classifier's decision with the similarity between the original and modified example: character-level through Levenshtein distance [26] and semantic using *BLEURT* [42].

8 Conclusions

The seventh edition of the CheckThat! lab at CLEF provides a diverse collection of challenges to the research community interested in developing technology to support and understand the journalistic verification process. The tasks go from core verification tasks such as assessing the check-worthiness of a text to understanding the strategies used to influence the audience and identifying the stance of relevant characters on questionable affairs. For the first time, the lab looks at the impact of data purposefully shaped to disguise classifiers for different relevant tasks. As in every year, the evaluation framework for all tasks is freely released to the community in order to foster the development of technology against disinformation and misinformation.

Acknowledgments

The work of F. Haouari is supported by GSRA grant #GSRA6-1-0611-19074 from the Qatar National Research Fund. The work of T. Elsayed was made possible by NPRP grant #NPRP-11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The work of F. Alam, M. Hasanain and R. Suwaileh is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund (a member of Qatar Foundation). The work of P. Przybyła is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Alam, F., Barrón-Cedeño, A., Cheema, G.S., Hakimov, S., Hasanain, M., Li, C., Míguez, R., Mubarak, H., Shahi, G.K., Zaghouani, W., Nakov, P.: Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content. In: Aliannejadi et al. [3]
2. Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., Nakov, P.: Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In: Findings of EMNLP 2021. pp. 611–649 (2021)
3. Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, Michalis (eds.): Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum. CLEF 2023, Thessaloniki, Greece (2023)
4. Arslan, F., Hassan, N., Li, C., Tremayne, M.: A benchmark dataset of check-worthy factual claims. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 14, pp. 821–829 (2020)
5. Barrón-Cedeño, A., Alam, F., Galassi, A., Da San Martino, G., Nakov, P., , Elsayed, T., Azizov, D., Caselli, T., Cheema, G., Haouari, F., Hasanain, M., Kutlu, M., Li, C., Ruggeri, F., Struß, J.M., Zaghouani, W.: Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023) (2023)
6. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F.: CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In: Advances in Information Retrieval. pp. 499–507. ECIR '20 (2020)
7. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Sheikh Ali, Z.: Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). pp. 215–236. LNCS (12260), Springer (2020)
8. da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P.: SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020). pp. 1377–1414 (2020)
9. Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P.: Fine-grained analysis of propaganda in news article. 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). pp. 5636–5646. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
10. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: CheckThat! at CLEF 2019: Automatic identification and verification of claims. In: Advances in Information Retrieval – European Conference on IR Research. pp. 309–315. ECIR '19 (2019)
11. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 301–321. LNCS (2019)
12. Galassi, A., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Caselli, T., Kutlu, M., Struss, J., Antici, F., Hasanain, M., Köhler, J., Korre, K., Leistra, F., Muti, A., Siegel, M., Mehmet Deniz, T., Wiegand, M., Zaghouani, W.: Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles. In: Aliannejadi et al. [3]
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6572
14. Han, S., Gao, J., Ciravegna, F.: Neural language model based training data augmentation for weakly supervised early rumor detection. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019. pp. 105–112. Association for Computing Machinery, Inc (2019)
15. Haouari, F., Elsayed, T.: Are Authorities Denying or Supporting? Detecting Stance of Authorities Towards Rumors in Twitter (2023). https://doi.org/10.21203/rs.3.rs-3383493/v1
16. Haouari, F., Elsayed, T.: Detecting Stance of Authorities towards Rumors in Arabic tweets: A Preliminary Study. In: Proceedings of the 45th European Conference on Information Retrieval (ECIR'23) (2023)
17. Haouari, F., Elsayed, T., Mansour, W.: Who can verify this? finding authorities for rumor verification in Twitter. Information Processing & Management **60**(4), 103366 (2023)
18. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: ArCOV19-Rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In: Proceedings of the Arabic Natural Language Processing Workshop. pp. 72–81. WANLP '21 (2021)
19. Haouari, F., Sheikh Ali, Z., Elsayed, T.: Overview of the CLEF-2023 CheckThat! lab task 5 on authority finding in twitter. In: Aliannejadi et al. [3]
20. Hu, X., Guo, Z., Chen, J., Wen, L., Yu, P.S.: MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2901–2912. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023)
21. Jerônimo, C.L.M., Marinho, L.B., Campelo, C.E.C., Veloso, A., da Costa Melo, A.S.: Fake news classification based on subjective language. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. pp. 15–24 (2019)
22. Jiang, Y., Song, X., Scarton, C., Aker, A., Bontcheva, K.: Categorising Fine-to-Coarse Grained Misinformation: An Empirical Study of COVID-19 Infodemic. CoRR **abs/2106.1** (2021)

23. Kasnesis, P., Toumanidis, L., Patrikakis, C.Z.: Combating fake news with transformers: A comparative analysis of stance detection and subjectivity analysis. Inf. **12**(10), 409 (2021)
24. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. NeurIPS'20 (2020)
25. Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., Potthast, M.: SemEval-2019 Task 4: Hyperpartisan News Detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 829–839. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019)
26. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**, 707–710 (1966)
27. Mu, Y., Jiang, Y., Heppell, F., Singh, I., Scarton, C., Bontcheva, K., Song, X.: A Large-Scale Comparative Study of Accurate COVID-19 Information versus Misinformation. In: TrueHealth 2023: Workshop on Combating Health Misinformation for Social Wellbeing (2023)
28. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Mubarak, H., Nikolov, A., Kartal, Y.S., Beltrán, J.: Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum. CLEF '2022 (2022)
29. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Beltrán, J., Wiegand, M., Siegel, M., Köhler, J.: Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. CLEF '2022 (2022)
30. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Beltrán, J.: The CLEF-2022 CheckThat! Lab on fighting the COVID-19 infodemic and fake news detection. In: Advances in Information Retrieval – European Conference on IR Research. pp. 416–428. ECIR '22 (2022)
31. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Gencheva, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 lab on automatic identification and verification of claims in political debates. In: Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum. CLEF '18 (2018)
32. Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., Hamdan, B., Ali, Z.S., Babulkov, N., Nikolov, A., Shahi, G.K., Struß, J.M., Mandl, T., Kutlu, M., Kartal, Y.S.: Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Candan, K., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association. LNCS (12880) (2021)
33. Nakov, P., Martino, G.D.S., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., Nikolov, A., Shahi, G.K.,

Struß, J.M., Mandl, T.: The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: Advances in Information Retrieval – 43rd European Conference on IR Research. ECIR '21, vol. 12657, pp. 639–649 (2021)

34. Nielsen, D.S., McConville, R.: Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3141–3153. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022)
35. Piskorski, J., Stefanovitch, N., Bausier, V.A., Faggiani, N., Linge, J., Kharazi, S., Nikolaidis, N., Teodori, G., De Longueville, B., Doherty, B., Gonin, J., Ignat, C., Kotseva, B., Mantica, E., Marcaletti, L., Rossi, E., Spadaro, A., Verile, M., Da San Martino, G., Alam, F., Nakov, P.: News categorization, framing and persuasion techniques: Annotation guidelines. Tech. Rep. JRC-132862, European Commission Joint Research Centre, Ispra (Italy) (March 2023)
36. Piskorski, J., Stefanovitch, N., Da San Martino, G., Nakov, P.: SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 2343–2361. Association for Computational Linguistics, Toronto, Canada (Jul 2023)
37. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A Stylometric Inquiry into Hyperpartisan and Fake News. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 231–240. Association for Computational Linguistics (2018)
38. Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, M.S., Nakov, P., Chakraborty, T.: Detecting harmful memes and their targets. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2783–2796. Association for Computational Linguistics, Online (Aug 2021)
39. Przybyła, P., Saggion, H.: ERINIA: Evaluating the Robustness of Non-Credible Text Identification by Anticipating Adversarial Actions. In: NLP-MisInfo 2023: Workshop on NLP applied to Misinformation, held as part of SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing. CEUR-WS.org (2023)
40. Przybyła, P., Shvets, A., Saggion, H.: Verifying the Robustness of Automatic Credibility Assessment. arXiv preprint arXiv:2303.08032 (mar 2023)
41. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. pp. 105–112. EMNLP '03 (2003)
42. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning Robust Metrics for Text Generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892. Association for Computational Linguistics, Online (jul 2020)
43. Shang, L., Zhang, Y., Zha, Y., Chen, Y., Youn, C., Wang, D.: Aomd: An analogy-aware approach to offensive meme detection on social media. Inf. Process. Manage. **58**(5) (sep 2021)
44. Sharma, S., Alam, F., Akhtar, M.S., Dimitrov, D., Da San Martino, G., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., Chakraborty, T.: Detecting and understanding harmful memes: A survey. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 5597–5606. International Joint Conferences on Artificial Intelligence Organization (7 2022), survey Track

45. Sharma, S., Kulkarni, A., Suresh, T., Mathur, H., Nakov, P., Akhtar, M.S., Chakraborty, T.: Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2149–2163. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023)
46. Sharma, S., Suresh, T., Kulkarni, A., Mathur, H., Nakov, P., Akhtar, M.S., Chakraborty, T.: Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations. pp. 1–11. Association for Computational Linguistics, Dublin, Ireland (May 2022)
47. Suryawanshi, S., Chakravarthi, B.R.: Findings of the shared task on troll meme classification in Tamil. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. pp. 126–132. Association for Computational Linguistics (Apr 2021)
48. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv: 1312.6199 (dec 2013)
49. Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A.: The Fact Extraction and VERification (FEVER) Shared Task. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER) (2018)
50. Vieira, L.L., Jerônimo, C.L.M., Campelo, C.E.C., Marinho, L.B.: Analysis of the subjectivity level in fake news fragments. In: Proceedings of the Brazillian Symposium on Multimedia and the Web. pp. 233–240. WebMedia '20, ACM (2020)
51. Zhang, W.E., Sheng, Q.Z., Alhazmi, A., Li, C.: Adversarial Attacks on Deep-learning Models in Natural Language Processing. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(3) (2020)