CheckIT!: A Corpus of Expert Fact-checked Claims for Italian

Jacopo Gili[1], Lucia Passaro[2] and Tommaso Caselli[3]

[1]*Department of Computer Science, University of Turin, Italy*
[2]*Department of Computer Science, University of Pisa, Italy*
[3]*CLCG, University of Groningen, Netherlands*

Abstract

This paper introduces *CheckIT!*, a resource of expert fact-checked claims, filling a gap for the development of fact-checking pipelines in Italian. We further investigate the use of three state-of-the-art generative text models to create variations of claims in zero-shot settings as a data-augmentation strategy for the identification of previously fact-checked claims. Our results indicate that models struggles in varying the surface forms of the claims.

Keywords

Fact-checking, Corpora, Data augmentation, Generative AI Model Evaluation

1. Introduction

The pollution of the information ecosystem by means of misleading or false information has reached unprecedented levels at a global scale. This has been possible thanks to a combination of multiple factors, among which the collapse of (local and national) journalism; an increasing sense of distrust in science and evidence-based facts; and the presence of computational amplification tools such as bots [1].

Manually fact-checking claims is an expensive operation (in terms of time and effort) and in many cases, it comes too late. Authors in [2] have shown how false and inaccurate information propagates online eight times faster than true and reliable information. Letting this kind of information free to circulate may have harmful impacts on groups and individuals as well as threaten the texture of democratic societies. It is thus urgent and critical to implement automatic solutions that can assist content moderators and information professionals to promptly react in presence of false or misleading information.

In Figure 1, we present the full fact-checking verification pipeline [3]. As it appears, multiple steps are involved: (i) assessing whether a claim is worth of being fact-checked; (ii) checking whether the claim has been previously fact-checked; (iii) if this is not the case, then evidence to evaluate the veracity of the claim must be gather (usually using reliable sources online); and (iv)

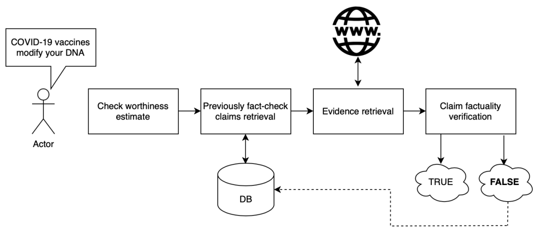

Figure 1: Fact-checking pipeline: (i) check-worthiness; (ii) previously verified claims retrieval; (iii) claim evidence retrieval; (iv) claim veracity assessment. The figure is an adaptation from [5] and [7].

finally, assessing its veracity status. Having access to a database of previously fact-checked claims is a valuable resource for fact-checkers because claims tend to be repeated (even if with small variations) over time, and this is particularly true for politicians [4, 5, 6]. The availability of such a resource can save time and contribute to mitigate the effects of misinformation.

This paper presents *CheckIT!* the first corpus of previously fact-checked claims for Italian. In its current version, *CheckIT!* is based on a collection of 3,577 claims of 317 Italian politicians and public figures, provided with evidences and veracity labels.

Contributions Our main contributions can be summarized as follows: (i) **we introduce *CheckIT!***, a fact-checking resource filling a gap in the language resource panorama for Italian for claim verification and, more generally, for misinformation detection and countering; (ii) we conducted a **feasibility study on automatic paraphrasing** in Italian, exploring the potential of leveraging advanced Natural Language Processing (NLP) techniques

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
jacopo.gili584@edu.unito.it (J. Gili); lucia.passaro@unipi.it (L. Passaro); t.caselli@rug.nl (T. Caselli)
0009-0007-1343-3760 (J. Gili); 0000-0003-4934-5344 (L. Passaro); 0000-0003-2936-0256 (T. Caselli)

for generating high-quality texts that preserve the original meaning of the claims while introducing linguistic variations; (iii) we propose an **initial framework for the automatic extension of fact-checking resources**, which enables the continuous growth and enrichment of *CheckIT!* with additional fact-checked claims and related evidence.

The remainder of the paper is structured as follows: Section 2 describes the data collection, the veracity label harmonization process, and presents an analysis of the dataset. In Section 3, we discuss the results of our paraphrases experiments with text generation tools for Italian (`mT5`, `Camoscio`, `ChatGPT`) as a strategy to extend the variability of expressions of previously fact-checked claims. Our efforts have been mainly focused on assessing the quality of these generative tools. Related work is discussed in Section 4. Finally, Section 5 concludes the paper and draws directions of further development.[1]

2. *CheckIT!*: Data Collection and Analysis

CheckIT! has been obtained by collecting all available fact-checked claims from *Pagella Politica*[2] and structuring them into a unified representation format. *Pagella Politica* is a web-based news outlet fully dedicated to fact-checking and analysis of political news in Italy since October 2012. *Pagella Politica* aims to provide accurate information and it aims to empower readers with knowledge, fostering a deeper and more informed engagement with the political landscape. To gather all claims we have obtained access to *Pagella Politica*'s public APIs, and scraped claims covering a period from October 3rd 2012 to April 26th 2023. In our harmonization process, we retained 3,577 claims out of 4,547, with 17 common attributes (see Table A in Appendix A for details). For each claim, the evidence text has been split into sentences and all hyperlinks have been extracted and stored separately.

As for the veracity verdicts, *Pagella Politica* has changed its labelling scheme since its firsts appearance: they have moved from a five-label scheme ("Vero" [True], "Ni" [Mostly true], "C'eri quasi" [Half True], "Pinocchio andante" [False], "Panzana pazzesca" [Pants on fire]) to verbose verdicts with explanations (e.g., "the politician is right"). However, mapping the verbose verdicts to the original five labels was impossible, especially for non-experts and for the very nuanced difference between the labels "Ni" [Mostly true], "C'eri quasi" [Half True]. In addition to this, verbose verdicts are not optimal for training machine learning classifiers. To avoid losing 270 of the most recent claims, we decided to simplify the labeling scheme. We thus reduced the label granularity from five to three, by collapsing "Ni" [Mostly true] and "C'eri quasi" [Half True] into "Impreciso" [Imprecise], and "Pinocchio andante" [False] and "Panzana pazzesca" [Pants on fire] into "Falso" [False] to separate certainly true and false information from the imprecise one. Subsequently, we manually analyzed the verbose verdicts and assigned the corresponding label.

At the end of this operation, we have the following label distributions: **1,255** claims labelled as "**Vero**" [True], **1,512** labelled as "**Impreciso**" [Imprecise], and **810** labelled as "**Falso**" ["False"]. The label distribution is not perfectly balanced, with the majority class being "Impreciso". While on the one hand it is comforting to see that, in absolute terms, politicians do not overtly lie, on the other hand it is not surprising to observe that politicians may manipulate data and news as a propaganda strategy to convince the audience of their arguments. The tendency of the last 16 months is worrying as 61.14% (192 out of 314) of the claims have been fact-checked as false.

The distribution of the claims over time is not well balanced as illustrated in Table 1. Early years see a rich activities, while this diminishes in more recent times (see also Figure 1 in Appendix B. Election years (2013, 2018, and 2022 for national Parliament elections; 2014 and 2019 for European Parliament elections) contain the majority of the fact-checked claims.

		Verdict		Year		Verdict	
Year	True	Imprecise	False	Year	True	Imprecise	False
2012∗	143	121	54	2018	25	88	29
2013	315	294	97	2019	79	180	76
2014	263	255	83	2020	63	153	92
2015	144	191	60	2021	59	84	95
2016	40	73	24	2022	83	16	136
2017	22	42	28	2023∗	19	4	46

Table 1
CheckIT!: Distribution of verdict labels per year. Years marked with ∗ cover less than 12 months.

When focusing on the most debated topics, the large majority of the claims (79.54%) concern four main areas: Social Issues (983), Economics (919), Institutions (599), Foreign Affairs (350), clearly corresponding to topics of public interest as they directly affect the lives of citizens and the working of the democratic institutions. Some of these topics face peaks of fact-checking in correspondence of relevant events. For example, 21.71% of the claims concerning Foreign Affairs are registered in 2015 during the European migrant crisis[3]; 12.71% of claims related to Social Issues are in 2020 during the first phases of the COVID-19 pandemics; 10.63% of claims for Economics are in 2019, when the citizens' income ("Reddito di Cittadinanza") was introduced. An overview of the distribution of the topics and the corresponding verdict

[1] Code and data: https://github.com/Jj-source/Check-It.
[2] https://pagellapolitica.it/fact-checking
[3] https://en.wikipedia.org/wiki/2015_European_migrant_crisis

labels is presented in Table 2.

Topic	True	Verdict Imprecise	False
Environment	44	59	18
Social Issues	291	402	290
Economics	291	459	169
Justice (Civil and Criminal)	62	63	30
Foreign Affairs	124	163	63
Institutions	266	233	91
Other	50	65	34
Not Specified	127	64	214

Table 2
CheckIT!: Distribution of verdict labels per topic.

CheckIT! contains 317 unique Italian politicians/public figures. The corpus has a very long tail, with the large majority of politicians being attributed only one claim. An aspect to consider in this dataset concerns the popularity and the roles that politicians have. The top 10 politicians are all prominent figures in the Italian political sphere. They are (former) secretary of major political parties, Prime Ministers, ministers, or popular party leaders. This top 10 covers 52.80% of all the fact-checked claims. On the other hand, only 18.92% (60) of the politicians appear in at least 10 claims. A statistic we are not able to provide in full given the current version of *CheckIT!* is the distribution of the claims per political party. Although we know that there are 16 political parties, more than 1,000 claims lack this information, i.e., it was not available through the APIs. Table 3 shows the distributions of the verdict labels for the top 10 political figures.

Politician	True	Verdict Imprecise	False
Matteo Renzi	169	186	73
Matteo Salvini	68	137	109
Beppe Grillo	62	125	55
Giorgia Meloni	39	69	64
Silvio Berlusconi	32	60	52
Luigi Di Maio	40	64	37
Renato Brunetta	39	46	27
Enrico Letta	52	38	12
Alessandro Di Battista	31	40	13
Laura Boldrini	52	26	1

Table 3
CheckIT!: Distribution of verdict labels for the top 10 politicians.

Are the claims biased? Documentation of potential biases in datasets has gained increasing awareness in the NLP community. From what we have seen so far, the dataset does not seem to present major biases in terms of political orientations, i.e., sovra-representation of a political party or side. The top 10 politicians (Table 3) are quite evenly distributed among the three major political areas that characterizes Italy in the past 10 years: three for the center-left/left, three for the M5S area, and four for the center-right/right. As a way to estimate the presence of potential biases, we have run a simple machine learning experiment to estimate the prediction of the veracity labels from the claims themselves. Previous work has shown that this is not an easy task (if even possible) [8, 9]. We have thus split *CheckIT!* into a Train (80%) and Test (20%) and trained two linear Support Vector Machine (SVM) models. We have used a simple TF-IDF vectorization[4] in both cases. In the second experiment, we have concatenated the names of the politicians to the text of their claims. Both SVMs are further compared with a Dummy classifier implementing majority voting. Results are summarized in Table 4.

Model	Label	P	R	Macro-F1
Dummy	True	0.0	0.0	0.195
	Imprecise	0.416	1.0	
	False	0.0	0.0	
SVM - claims only	True	0.458	0.392	0.422
	Imprecise	0.457	0.573	
	False	0.387	0.294	
SVM - claims & politicians	True	0.456	0.411	<u>0.425</u>
	Imprecise	0.449	0.553	
	False	0.411	0.300	

Table 4
Claim veracity prediction. Underscore figures indicate the best result.

As expected, the results are way far from being satisfying. Although the SVMs seem to learn something, when compared to the Dummy classifier, their overall macro-F1 is well below 0.5. A slight improvement in the False class can be observed when the names of the politicians are concatenated with the claims. However, this appears to be an effect of the data split (out of 317 unique entities, 121 appear both in our train and test splits). While on one hand, these results further confirm a limited presence of bias in the data, they further support previous results on the difficulty of assessing the veracity of a claim from the claim itself, especially when it is uttered using formally correct language [10].

3. Automatic Paraphrases of Fact-checked Claims

This battery of experiments is devoted to evaluate the use of generative language models to enrich fact-checking datasets by varying the expression of the claims. This

[4]We have used word uni- and bigrams, character n-grams, with a range of 2-5, and stop-word removals.

data augmentation approach plays a pivotal role for the development of robust systems for the identification of previously fact-checked claims (step (ii) in Figure 1), and thus reducing the manual workload of professional fact-checkers. In particular, we generate five alternative versions of the *CheckIT!* claims using three generative models, namely `mt5`, `Camoscio`, and `ChatGPT`.

mt5 The only available Italian model for paraphrase generation is `aiknowyou/mt5-base-it-paraphraser`. This model is based on `mt5` and fine-tuned on Tapaco and STS Benchmark datasets for Paraphrasing. `mt5` [11] is a multilingual variant of `T5` [12] that was pre-trained on a new Common Crawl-based dataset covering 101 languages. The TaPaCo Corpus, used for fine-tuning, is a freely available paraphrase corpus for 73 languages extracted from the Tatoeba database.

Camoscio The second method we used to generate paraphrases is based on instruction-based models. Specifically, we used `Camoscio` [13], an Italian version of Alpaca [14] obtained by instruction-tuning LlAMA on Italian data automatically translated with `ChatGPT`. To obtain the paraphrases, we used the following prompt: "Scrivi 5 parafrasi di questa frase: *claim*" ('Write 5 paraphrases of this sentence: *claim*') where *"claim"* is one of the original claims belonging to *CheckIT!*.

ChatGPT The third method consists in directly prompting `ChatGPT` APIs[5] with the following text: "Parafrasa le seguenti frasi: *claims*" ('Paraphrase the following: *claims*') where "claims" are the original claims belonging to *CheckIT!*.

For all models, we have used the default parameters. For `ChatGPT`, the temperature was left to 1 and `max_token` to 2,000.

3.1. Evaluation Metrics

To assess the goodness of the generated texts, we conducted a comprehensive evaluation encompassing comparisons between the model-generated paraphrases, the original sentences, and paraphrases by three human annotators.

In all evaluation settings, we use four automatic metrics, Cosine Similarity (*Cos*), BLEU [15], ROUGE [16], and BERTScore [17], to gain multiple perspectives on the models' performance and gauge both the fidelity and the variations with respect to the original claims exhibited by the models. In particular, *Cos* will return the the semantic similarity between the two texts based on word frequency distributions. BLEU, although commonly used for Machine Translation, will assess the overlap of n-grams (word sequences) between the claim and the paraphrases as a proxy for text variation. Similarly, ROUGE,[6] which returns the overlap of n-grams and the longest common subsequence, will also assess the variations of the generated text with respect to the original claims. Finally, BERTScore, which calculates the similarity between two sentences or texts by utilizing contextualized embeddings from pre-trained language models and comparing the embeddings of overlapping words between the candidate and reference sentences, will help us to better assess the semantic similarity.

3.2. Evaluation Settings and Results

Overall, we have four evaluation blocks. The first block is based on 10% (i.e., 357) of the claims in *CheckIT!*. In this case, we compared the automatically-generated paraphrases against the original claims.

The latter three are based on a subset of 50 claims that have been independently paraphrased by the human annotators.[7] Annotators were given basic instructions which closely resembled the prompts of `Camoscio` and `ChatGPT`: "Provide a paraphrase for each of the following sentences." In the second evaluation block, we compare human-generated paraphrases (a total of 150 instances corresponding to 3 different variants per claim) with the original claims. In the third evaluation block, we evaluate the human-generated paraphrases with respect to each other: for each data point, we compared the four metrics between all the combinations of annotators (e.g., A1 *vs.* A2; A2 *vs.* A3; A1 *vs.* A3, and so on). Note that some of the metrics (i.e., ROUGE and BERTScore) are not symmetric, thus results may vary. In the fourth evaluation block, we compared the automatically-generated paraphrases against human-generated ones.

Block I: Machines *vs.* Claims The summary of the results is in Table 5. `Camoscio` produced a considerable number of empty paraphrases. To ensure fair comparisons, we excluded these empty paraphrases from the metrics calculation. Overall, we notice a trend of higher variation in generation for `ChatGPT`. Despite the high average cosine similarity with the original texts, `ChatGPT` displayed better performances for creative rephrasing. Surprisingly, `mT5` does not perform very well, as indicated by the high scores across all metrics. Differences between the training materials and the *CheckIT!* data may have had an impact. Finally, `Camoscio` is the worst performing models. Out of 1,785 possible paraphrases for the 357 claims considered, it fails to generate an output 1,320 times. The few successful cases are almost exact

[5] We used `GPT 3.5-turbo`.

[6] ROUGE is a set of metrics: ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSum.

[7] All annotators are also the author of this paper.

Metric	ChatGPT	mt5	Camoscio∗
BERT-P	0.80	0.88	0.91
BERT-R	0.79	0.82	0.85
BERT-F1	0.79	0.85	0.88
BLEU	0.13	0.27	0.59
Cos	0.93	0.92	0.95
ROUGE-1	0.56	0.71	0.87
ROUGE-2	0.32	0.58	0.82
ROUGE-L	0.48	0.68	0.86
ROUGE-LS	0.48	0,68	0.86

Table 5
Generated paraphrases *vs.* claims.

Metric	A1	A2	A3
BERT-P	0.76	0.78	0.83
BERT-R	0.71	0.72	0.80
BERT-F1	0.73	0.75	0.81
BLEU	0.05	0.07	0.16
Cos	0.83	0.86	0.93
ROUGE-1	0.35	0.44	0.61
ROUGE-2	0.16	0.22	0.38
ROUGE-L	0.28	0.35	0.56
ROUGE-LS	0.28	0.35	0.56

Table 6
Human paraphrases *vs.* claims.

Metric	A1-A2	A1-A3	A2-A1	A2-A3	A3-A1	A3-A2
BERT-P	0.80	0.78	0.81	0.80	0.76	0.76
BERT-R	0.81	0.76	0.80	0.76	0.78	0.80
BERT-F1	0.80	0.77	0.80	0.78	0.77	0.78
BLEU	0.10	0.06	0.10	0.07	0.05	0.07
Cos	0.89	0.85	0.89	0.87	0.85	0.87
ROUGE-1	0.45	0.36	0.45	0.42	0.36	0.42
ROUGE-2	0.22	0.17	0.22	0.19	0.17	0.19
ROUGE-L	0.37	0.29	0.37	0.35	0.29	0.35
ROUGE-LS	0.37	0.29	0.37	0.35	0.29	0.35

Table 7
Comparison across annotators.

repetitions of the original claims, as highlighted by the scores of the various measures and a manual inspection. Clear evidence of this parroting behavior is shown by the BLEU score.

Metric	ChatGPT	mt5	Camoscio∗
BERT-P	0.85	0.80	0.86
BERT-R	0.87	0.81	0.86
BERT-F1	0.86	0.80	0.85
BLEU	0.25	0.16	0.29
Cos	0.84	0.81	0.87
ROUGE-1	0.60	0.53	0.62
ROUGE-2	0.40	0.31	0.41
ROUGE-L	0.54	0.49	0.56
ROUGE-LS	0.54	0.49	0.56

Table 8
Generated paraphrases *vs.* human.

Block II: Humans *vs.* Claims Scores are reported in Table 6. In general, it seems that humans introduce more superficial variations, as highlighted by BLEU and ROUGE. However, there is an increasing adherence to the original formulation of the claim among the annotators. Notably, A1 exhibited a greater propensity for variation in their paraphrasing, while A3 tended to produce paraphrases closer to the original texts, as evidenced by the higher BLEU and ROUGE-LS. Clearly, the closer in wording to the original claim, the bigger the impact also on the more semantic oriented measures such as BERTScore and *Cos*. While A1 and A2 present close performances, A3 achieves the highest results. It appears that divergent interpretations of what a paraphrases of a claim is and how to do it have affected the results, suggesting that more precise instructions will be needed in the future to achieve more varied results.

Block III: Human *vs.* Human As we delved into the comparison among the annotators (Table 7), we found that A1 and A2 produced paraphrases that were notably more similar to each other in comparison to those produced by A3. This clearly indicates that distinct stylistic preferences have been adopted.

Block IV: Machines *vs.* Humans We evaluated the quality of the generated paraphrases by comparing them to the three human-produced paraphrases, considering the latter as references. A summary of these results is presented in Table 8. Surprisingly, the automatically generated paraphrases have a higher degree of similarity and lexical overlap with the manually generated ones. The results for `Camoscio` are quite unexpected, as it seems to qualify as the best second system after `ChatGPT`. However, this is a distortion due to the measures and the manual paraphrases. As we have seen in Table 6, A3 is very conservative, generating paraphrases close to the original claim. This is also the behavior of `Camoscio`, as observed in Table 5. On the other hand, `mT5` and `ChatGPT` appears to be more suitable candidates for this task.

4. Related Work

Automatic fact-checking is a growing field of research and previous work has already investigated multiple aspects. Early work has focused on detecting rumors in Social Media [18, 19], or on the identification of the stance of a document with respect to a claim [20, 21, 22]. Following Figure 1, the claim detection step is one of the easiest and one of the most controversial subtask. While the identification of claims is comparable to Attribution Detection [23, 24], the check-worthiness status of claims is challenging since it involves some level of subjectivity. To address this issue, previous work has collected data from authoritative sources run by professional fact-checkers (e.g., PolitiFact, Snopes) or have seen the direct involvement of human experts for the veracity labelling [3, 4, 25, 26, 27, 28, 29, 30].

Evidence retrieval requires the identification of relevant passages from external resources that can be used to verify the claim. Two mainstream automatic verification methods are employed: Stance Detection and Natural Language Inference (NLI) [25, 31, 32]. They make use of unstructured data (i.e., textual sources) and assume that evidence is available for every claim and make a closed world world assumption, i.e., evidence is available only in one source. Complimentary methods make use of structured data, where evidence can be retrieved inside a knowledge graph [33].

Each of the subtasks involved in the fact-checking pipeline is framed as a classification task, with a varying number of labels: from a binary classification for the check-worthiness, to rich multi-class classification tasks for the veracity of the claim. For *CheckIT!* , we have opted for a three-way classification of the claim, in line with most of the previous work. The advantage of (more) fine-grained veracity classifiers is that it allows to capture also misleading or imprecise information and avoiding to reduce the world into a black or white picture.

5. Conclusions and Future Work

This work has introduced *CheckIT!*, an expert-curated fact-checked repository of claims by politicians and prominent public figures in Italy. *CheckIT!* covers 10 years of claims and it is the first publicly available dataset for fact-checking in Italian. In our analysis of *CheckIT!*, we have observed a drop in the numbers of fact-checked claims suggesting that manual fact-checking is increasingly difficult to conduct and that automated assisted tools are more and more needed.

We have conducted a preliminary investigation of three state-of-the-art automatic text generation tools for claim paraphrases. By combining multiple automatic measures, it appears that `ChatGPT` and `mT5` are the two best candidate to further explore, while `Camoscio` presents non-trivial issues with respect to failure to produce an output and variations of the generated texts.

Future work will focus on three aspects: conduct a qualitative (human-based) evaluation of the two best models; evaluate the generated paraphrases for previously fact-checked claim retrieval on the line of [5]; finally evaluate the generated paraphrasis against the topics.

Acknowledgments

The authors want to thank *Pagella Politica* and its director, Giovanni Zagni, for the access to the APIs that made the data collection for *CheckIT!* possible.

This work has been partially supported by the EU H2020 TAILOR project, GA n. 952215.

References

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[2] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151.

[3] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508. doi:10.3115/v1/W14-2508.

[4] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, et al., Claimbuster: The first-ever end-to-end fact-checking system, Proceedings of the VLDB Endowment 10 (2017) 1945–1948.

[5] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a known lie: Detecting previously fact-checked claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618. URL: https://aclanthology.org/2020.acl-main.332. doi:10.18653/v1/2020.acl-main.332.

[6] S. Shaar, F. Alam, G. Da San Martino, P. Nakov, The role of context in detecting previously fact-checked claims, in: Findings of the Association

for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1619–1631. URL: https://aclanthology.org/2022.findings-naacl.122. doi:10.18653/v1/2022.findings-naacl.122.

[7] P. Nakov, A. Barrón-Cedeño, G. da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, M. Wiegand, M. Siegel, J. Köhler, Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 495–520.

[8] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.

[9] S. Volkova, K. Shaffer, J. Y. Jang, N. Hodas, Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 647–653. URL: https://aclanthology.org/P17-2102. doi:10.18653/v1/P17-2102.

[10] T. Schuster, R. Schuster, D. J. Shah, R. Barzilay, The Limitations of Stylometry for Detecting Machine-Generated Fake News, Computational Linguistics 46 (2020) 499–510. URL: https://doi.org/10.1162/coli_a_00380. doi:10.1162/coli_a_00380.

[11] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[13] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[14] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, 2023.

[15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[16] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[18] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 69–76. URL: https://aclanthology.org/S17-2006. doi:10.18653/v1/S17-2006.

[19] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 845–854. URL: https://aclanthology.org/S19-2147. doi:10.18653/v1/S19-2147.

[20] D. Küçük, F. Can, Stance detection: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–37.

[21] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A survey on stance detection for mis- and disinformation identification, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1259–1277. URL: https://aclanthology.org/2022.findings-naacl.94. doi:10.18653/v1/2022.findings-naacl.94.

[22] J. Zheng, A. Baheti, T. Naous, W. Xu, A. Ritter, Stanceosaurus: Classifying stance towards multicultural misinformation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2132–2151. URL: https://aclanthology.org/2022.emnlp-main.138.

[23] S. Pareti, PARC 3.0: A corpus of attribution rela-

tions, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 3914–3920. URL: https://aclanthology.org/L16-1619.

[24] C. Scheible, R. Klinger, S. Padó, Model architectures for quotation detection, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1736–1745. URL: https://aclanthology.org/P16-1164. doi:10.18653/v1/P16-1164.

[25] A. Hanselowski, C. Stab, C. Schulz, Z. Li, I. Gurevych, A richly annotated corpus for different tasks in automated fact-checking, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 493–503. URL: https://aclanthology.org/K19-1046. doi:10.18653/v1/K19-1046.

[26] I. Augenstein, C. Lioma, D. Wang, L. Chaves 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4685–4697. URL: https://aclanthology.org/D19-1475. doi:10.18653/v1/D19-1475.

[27] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, J. Glass, Automatic fact-checking using context and discourse information, Journal of Data and Information Quality (JDIQ) 11 (2019) 1–27.

[28] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7740–7754. URL: https://aclanthology.org/2020.emnlp-main.623. doi:10.18653/v1/2020.emnlp-main.623.

[29] L. C. Passaro, A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event, Inf. Sci. 615 (2022) 657–677. URL: https://doi.org/10.1016/j.ins.2022.07.128. doi:10.1016/j.ins.2022.07.128.

[30] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at EVALITA 2023: Overview of the multimodal fake news detection and verification task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper32.pdf.

[31] J. Maillard, V. Karpukhin, F. Petroni, W.-t. Yih, B. Oguz, V. Stoyanov, G. Ghosh, Multi-task retrieval for knowledge-intensive tasks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1098–1111. URL: https://aclanthology.org/2021.acl-long.89. doi:10.18653/v1/2021.acl-long.89.

[32] M. Arana-Catania, E. Kochkina, A. Zubiaga, M. Liakata, R. Procter, Y. He, Natural language inference with self-attention for veracity assessment of pandemic claims, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1496–1511. URL: https://aclanthology.org/2022.naacl-main.107. doi:10.18653/v1/2022.naacl-main.107.

[33] J. Kim, K.-s. Choi, Unsupervised fact checking by counter-weighted positive and negative evidential paths in a knowledge graph, in: Proceedings of the 28th international conference on computational linguistics, 2020, pp. 1677–1686.

Appendix A: *CheckIT!* Attribute Descriptions

Attribute	Value	Attribute	Value
`id`	unique id of the claim	`date`	timestamp of fact-checking
`link`	Pagella Politica URL	`content`	fact-checking evidence
`statement_date`	timestamp of the claim	`source`	URL of the news outlet/platform where the claim has appeared
`statement`	the claim	`verdict`	veracity label of the claim
`verdict_ext`	verbose veracity judgment of the claim	`politician`	full name of the politician or public figure owning the claim
`political_party`	Political party membership at the time of the claim	`platform`	Name of the news outlet/platform where the claim has appeared
`politicians_in`	the name(s) of any politician(s) mentioned in the claim (other than the owner of the claim)	`macro_area`	broader topic of the claim
`tags`	keywords to describe the content of the claim	`links`	list of URLs used to retrieve evidence, write the `content`, and the verdict
`versione`	versioning of the dataset		

Table A
CheckIT!: List of the attributes used to represent the data.

Appendix B: Verdict distribution overview

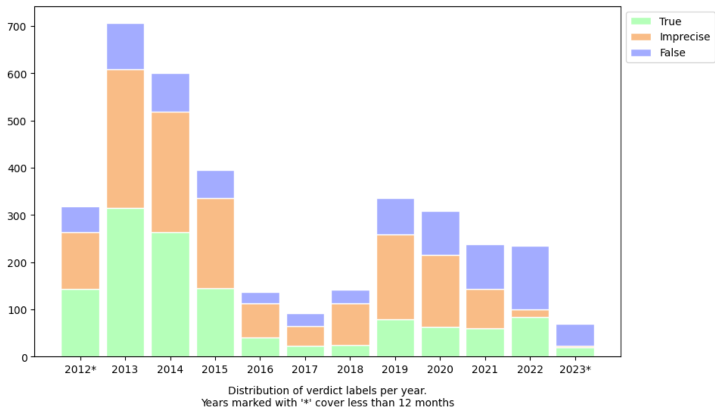

Figure 1: *CheckIT!*: Distribution of verdict labels per year (histogram)