AraFacts: The First Large Arabic Dataset of Naturally-Occurring Professionally-Verified Claims

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali
Computer Science and Engineering Department, Qatar University
{zs1407404, wm1900793, telsayed, a.alali}@qu.edu.qa

Abstract

We introduce **AraFacts**, the first large Arabic dataset of naturally-occurring claims collected from 5 Arabic fact-checking websites, e.g., Fatabyyano and Misbar, covering claims since 2016. Our dataset consists of 6,222 claims along with their factual labels and additional metadata, such as fact-checking article content, topical category, and links to posts or Web pages spreading the claim. Since the data is obtained from various fact-checking websites, we standardize the original claim labels to provide a unified label rating for all claims. Moreover, we provide revealing dataset statistics and motivate its use by suggesting possible research applications. The dataset is made publicly available for the research community.

1 Introduction

Fake news and misinformation are considered among the greatest threats to nations. The spread of fake news can cause manipulation in public opinion, which has adverse consequences to politics and journalism. Moreover, the recent COVID-19 pandemic revealed how medical misinformation could easily harm the health of people (Islam et al., 2020). Notable development has been made in automated fact-checking systems over the past years. However, one of the main limitations of Arabic automated fact-checking systems is the lack of Arabic datasets. Recently several Arabic fact-checking websites (e.g., Fatabyyano and Misbar) have emerged to help combat the spread of rumors and fake news, especially over social media. They constitute a valuable resource of naturally-occurring fact-checked claims in the Arab world. The claims are annotated and verified by professional fact-checkers and journalists, making them a reliable source of information. While the verified information about those rumors is posted on those websites and their corresponding social media accounts, they are not gathered and unified as a dataset that can be used to develop automated systems. In this paper, we introduce **AraFacts**, the first large Arabic Fact-checking dataset of naturally-occurring and professionally-verified claims. We crawl 6,222 claims along with their factual labels, description, and additional metadata, from 5 different fact-checking websites, and make them available to the research community in a normalized form. The contributions of our work are two-fold:

- We extract and publicly share[1] about 6k claims with their original and normalized factual labels along with their metadata.
- We propose several research applications for which our dataset can be utilized.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 describes the data collection process. Section 4 provides data analysis. Section 5 outlines some research applications, and Section 6 concludes.

2 Related Work

Several datasets of naturally occurring claims have been previously proposed, but mostly in English. For example, MultiFC dataset (Augenstein et al., 2019) includes 38,918 claims and their metadata from 26 different fact-checking websites. Similarly, Liar dataset (Wang, 2017) consists of 12,836 claims collected from PolitiFact.[2] FakeNewsNet (Shu et al., 2017a,b) is a mutli-dimensional data repository that contains 23,196 fact-checked articles from PolitiFact and GossipCop.[3] It also includes additional social context related to the checked claims.

Other datasets that focus on topic-specific claims were published recently. FakeCovid (Shahi and Nandini, 2020) dataset covers 5,182 multilingual fact-checked claims from 92 different fact-

[1] https://gitlab.com/bigirqu/AraFacts/
[2] https://www.politifact.com/
[3] https://www.gossipcop.com/

checking websites, all related to COVID-19. Moreover, PUBHEALTH dataset Kotonya and Toni (2020) consists of 11,832 claims public health-related claims attained from 8 different websites.

Over the past few years, there has been an increased interest in task-specific Arabic applications leading to the release of several Arabic datasets. Most of these are built for tasks such as check-worthiness of tweets, evidence retrieval, and claim verification (Barrón-Cedeno et al., 2020; **?**).

Additionally, the Arabic News Stance dataset (ANS) (Khouja, 2020) supports claim verification and stance prediction. The dataset is generated using existing Arabic news titles from ANT corpus (Chouigui et al., 2017). The limitation with ANS is that it requires manual annotation to generate claims from news titles. Elmadany et al. (2020) address this limitation with the AraNews dataset. AraNews uses Arabic articles from online news data to automatically generate Arabic manipulated text.

Baly et al. (2018) proposed a corpus consisting of 422 fact-checked Arabic claims and documents associated with the claims. The corpus supports multiple tasks such as stance detection and fact-checking. **AraFacts** differs from their work in two ways; it covers a larger number of claims and all claims are professionally fact-checked.

While several datasets were constructed for different tasks that are related to fact-checking, to the best of our knowledge, **AraFacts** is the largest Arabic dataset that leverages emerging Arabic fact-checking websites as a source of annotated professionally-verified information.

3 Data Collection

In this section, we describe the process of selecting fact-checking websites, crawling the verification articles, and constructing the **AraFacts** dataset.

3.1 Fact-checking websites

We chose fact-checking websites that are either verified by the International Fact-Checking Network (IFCN) or popular in the Arab region. IFCN certification indicates that the website complies with IFCN's code of ethics. The following websites were selected as data sources: **1) Fatabyyano**[4] is an IFCN-certified fact-checking organization that launched in 2016. Fatabyyano collaborates with Facebook as a third-party fact-checker to debunk fake-news, rumors, and conspiracy theories using Facebook's claim rating system.[5] **2) FactuelAFP Arabic**[6] is the Arabic bureau of the French press news service. Their team consists of journalists and fact-checkers. It aims to debunk false statements, videos, or images that appear online. It is certified by IFCN and collaborates with Facebook. **3) Misbar**[7] is a popular independent Arabic fact-checking platform. It uses an 8-point claim rating system to label claims online. **4) Maharat-news fact-o-meter**[8] is an IFCN-certified fact-checking website that focuses on investigating rumors online and in real-life using a 3-point claim rating system. **5) Verify-Sy**[9] is a media platform that specializes in detecting and debunking false news and media by experienced journalists.

3.2 Data Extraction

From every fact-checking website, we crawled all fact-checking articles and extracted their claims, labels, and metadata. We also extracted the claim type using some indicative search keywords. The claim type identifies if the claim refers to textual information or if the claim is paired with visual information (such as a video or an image). For example; The claim *MISB_2941* is referring to a fake image shown in Figure 2.

While parsing the content of the HTML pages, some challenges were faced. One challenge was that websites like Verify-Sy do not explicitly label the veracity of the claim. To address this challenge, we reviewed the website's editorial policy and inspected a subset of the claims manually. We concluded that all claims are, in fact, false. Moreover, not all metadata fields were available on all websites. Table 1 shows the number of crawled claims per website and a summary of available metadata.

3.3 Veracity and Category Normalization

Another faced challenge was the varying claim rating system adopted by the different websites. We proposed a normalized claim rating to achieve a standard rating for all claims in **AraFacts**. We also used a normalized rating for the categories of the claims.

To normalize the veracity label of the claim, we used the label normalization method presented by

[4] https://Fatabyyano.net/

[5] http://bit.ly/2LnD1Rk

[6] https://factcheck.afp.com/ar/list

[7] https://misbar.com/

[8] https://maharat-news.com/fact-o-meter

[9] https://www.verify-sy.com/

Table 1: Summary of the amount of claims extracted from each website and some of their metadata

Website	Number of Claims	Image or Video Claims	Extracted Metadata			
			Pages with Claim URLs	Avg. no. of claim URLs	Pages with Evidence URLs	Avg. no. of Evidence URLs
Misbar	2,952	1,974	2,946	6.4	2,945	2.6
Fatabyyano	1,503	930	905	2.2	1,449	4.7
FactuelAFP	973	824	514	2.4	461	3.8
Verify-sy	707	403	179	1.1	86	0.3
Maharat-news	87	10	62	1.1	0	0.0

Khouja (2020) with some variation. We set our own claim rating scheme consisting of four labels *(False, Partly-False, Sarcasm, True)* and mapped original labels to them. To do so, we referred to the source websites' methodology, manually inspected a subset of the claims, then finally did the mapping of 27 distinct original labels to four normalized labels. We adopted a similar mapping technique to map 35 distinct topical categories to 8 normalized categories: *(Politics, News, Health, Social, Religion, General Sciences, Arts & Culture, and Other)*. More details on our claim and category normalization can be found in **AraFacts** repository.

The normalization process was performed by two authors of this work independently; then, disagreements were discussed and resolved.

3.4 Dataset Construction

After extracting claim information and normalizing the veracity labels and categories, we assigned a unique ID to each claim. We finally merged the crawled information for each website into one database with the following schema: **1**) **Claim-ID:** ID of the claim. **2**) **Claim:** Text of the claim. **3**) **Source:** Name of the fact-checking website from which the claim was crawled. **4**) **Description:** Detailed description of the claim. **5**) **Source-label:** The veracity label of the claim as it appears in the fact-checking website. **6**) **Normalized-label:** Normalized claim label. **7**) **Source-category:** Topical category of the claim as it appears in the fact-checking website. **8**) **Normalized-category:** Normalized topical category of the claim. **9**) **Date:** Article publication date. **10**) **Source URL:** URL of the article. **11**) **Claim URLs:** URLs to web pages spreading the claim. **12**) **Evidence URLs:** URLs referenced by the fact-checker to justify their annotation. **13**) **Claim type:** Indicates whether the claim refers to text, an image or a video.

Additionally, we extracted the content of the fact-checking article and included it in **AraFacts**. Figure 1 shows an example claim with some attributes.

ClaimID: MIS_2550
Claim:
ترامب خلال مقطع فيديو يدعو إلى ترقب مفاجأة بداية يناير/كانون الأول الجاري، والاستمتاع بالعرض
Description:
تداولت حسابات وصفحات على موقع التواصل الاجتماعي تويتر، منذ تاريخ 29 ديسمبر/كانون الأول المنصرم، مقطع فيديو للرئيس الأميركي دونالد ترامب، ادعت فيه بأنّه قال هناك مفاجأة ستحدث بداية يناير/كانون الثاني عام 2021 تم التوقيع عليها، وعلى الجميع انتظارها والاستمتاع بالعرض، وأوضحت المنشورات أن هناك قنبلة سيفجرها ترامب مع بداية شهر يناير الجاري
Source: Misbar
Date: 2021-01-02
Source_label: مضلل
Normalized_label: Partly-false
Source_category: سياسة
Normalized_category: Politics
Source_url: https://misbar.com/factcheck/2021/01/02/فاجأة-يناير-التي-أعلن-عنها-ترامب-كانت-من-وعود-حملته-الانتخابية

Figure 1: An example claim from **AraFacts**.

Figure 2: An example claim from **AraFacts** that has a fake image. Claim MISB_2941 refers to this image claiming that UN decided to open applications for refugee resettlement in Greece.

4 Data Analysis

In this section, we perform further analysis on the collected claims. Table 1 gives an overall summary of the collected claims and some of the ex-

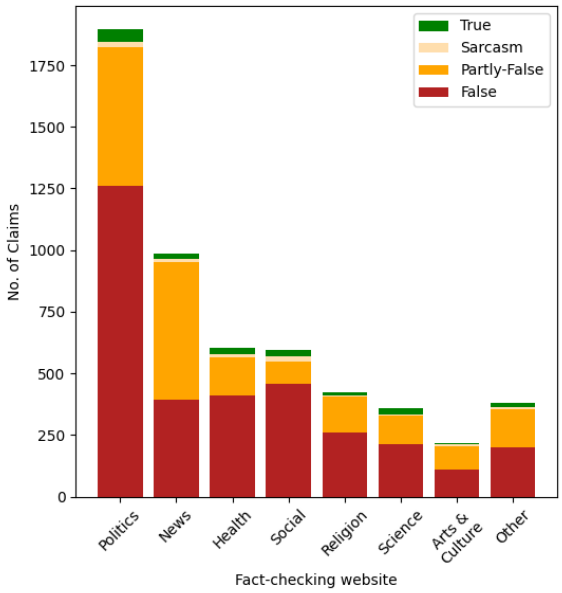

Figure 3: Distribution of normalized labels and categories.

Table 2: Top 5 source domains of claims.

Domain	% of URLs
facebook.com	54.5%
twitter.com	25.0%
perma.cc	4.1%
archive.vn	3.1%
youtube.com	3.1%

tracted metadata. In addition to the number of claims, for each website, we report the number of claims that have a refer to visual information (image or video), number of claims that have embedded claim URLs, average number of embedded URLs per claim, number of claims that have embedded evidence URLs, and average number of evidence URLs per claim.

Figure 3 illustrates the distribution of the normalized labels over the normalized categories. Most claims are political or news, and false claims are expectedly dominating in all categories.

We also examined the original source domains of the claims from the embedded claim URLs. Table 2 lists the most frequent 5 claim URLs. It is worthwhile noting that social media platforms constitute the main source of claims and rumors in the dataset.

Figure 4 shows the distribution of the claims over publication time. We notice that the majority of the claims were published in 2020; this is due to the

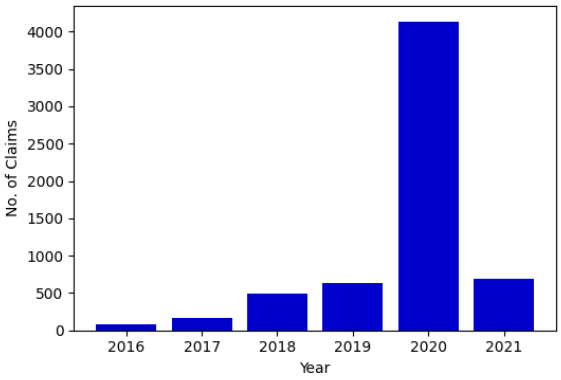

Figure 4: Claims distribution over time.

Figure 5: Most frequent 100 words in claims.

fact the most fact-checking websites in our dataset were launched in the past two years. Figure 5 shows the most frequent 100 words in the claim text. We notice a mix of political figures, religious entities, and country names among others.

Finally, we analyze the types of extracted claims. Figure 6 shows the frequency of the different types of claims (Text claims, Image claims, and Video claims) for each fact-checking website. Interestingly, the two most contributing fact-checking websites (Misbar and Fatabyyano) contain more visual claims than textual.

5 Example Use Cases

This section provides example use cases and research problems that can be supported by **AraFacts**.

5.1 Claim Verification

One of the main automated fake news detection tasks is claim verification. The task is defined as follows: given a claim, predict its veracity. The metadata can be used as features to classify the claim, such as claim source URLs. While we provide four normalized veracity labels of the claim, the original veracity labels can still be utilized following

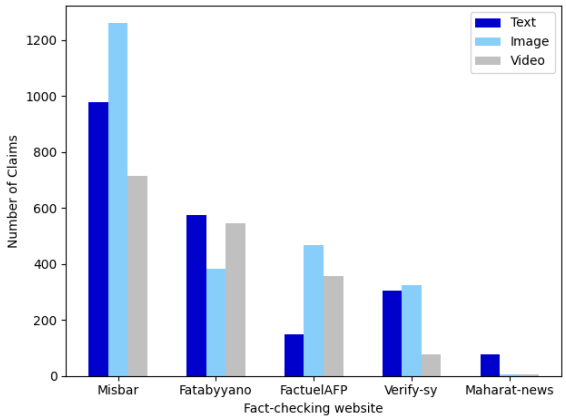

Figure 6: Type of claims published by each fact-checking website

a Multi-Task Learning (MLT) approach similar to Augenstein et al. (2019) MLT claim verification approach.

5.2 Claim Retrieval

The claim retrieval task is defined as follows: given a claim, check whether this claim has already been checked. The importance of this problem stems from the fact that many posted claims over social media platforms are just repetitions of previously fact-checked old claims. Having a system that addresses this problem can help combat the spread of those circulating claims over social media.

5.3 Evidence Retrieval

The task is defined as follows: given a claim, retrieve evidential sentences that help in verifying or debunking the claim. As we provide the content of the fact-checking articles and the URLs of evidence pages of claims, **AraFacts** can be extended by further annotating evidence sentences from the evidence pages to support such task.

5.4 Image-based Fact-checking

The task is defined as follows; given a claim that is mainly about an image, predict the factuality of the claim using the image and claim. This task has been explored for English claims (Zlatkova et al., 2019), but it has not been studied in Arabic yet. While **AraFacts** does not include the images related to the claims, we indicate the claim type and provide the URL to the fact-checking articles where the images can be extracted and paired with the claims.

6 Conclusion

In this paper, we introduce **AraFacts**, the first large Arabic dataset of naturally-occurring claims covering about 6k claims published since 2016 and already annotated by professional fact-checkers from 5 different Arabic fact-checking websites over several topical categories. The dataset supports many research tasks such as claim verification, claim retrieval, also evidence retrieval. We made our dataset publicly available to the research community, and we plan for periodic crawling to augment the newly verified claims from existing and new fact-checking websites. We believe that **AraFacts** will serve as a valuable resource for future studies on fact-checking, such as fake image detection. Our future work will focus on evaluating and utilizing **AraFacts** for several fact-checking tasks to contribute the Arabic-based work in that domain.

Acknowledgments

This work was made possible by NPRP grant No.: NPRP11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. 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)*, pages 4677–4691.

Ramy Baly, Mitra Mohtarami, and James Glass. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of NAACL-HLT*, pages 21–27.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. Ant corpus: an arabic news text collection for textual classification. In *2017 IEEE/ACS*

14th International Conference on Computer Systems and Applications (AICCSA), pages 135–142. IEEE.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, Tariq Alhindi, et al. 2020. Machine generation and detection of arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621.

Jude Khouja. 2020. Stance prediction and claim verification: An arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2017b. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. 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)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.