Automated Fact-Checking for Assisting Human Fact-Checkers

Preslav Nakov[1]*, David Corney[2], Maram Hasanain[3], Firoj Alam[1], Tamer Elsayed[3], Alberto Barrón-Cedeño[4], Paolo Papotti[5], Shaden Shaar[1], Giovanni Da San Martino[6]

[1]Qatar Computing Research Institute, HBKU, Qatar
[2]Full Fact, UK
[3]Computer Science and Engineering Department, Qatar University, Qatar
[4]DIT, Alma Mater Studiorum–Università di Bologna, Forlì, Italy
[5]EURECOM, France
[6]Dipartimento di Matematica, University of Padova, Italy
{pnakov, fialam, sshaar}@hbku.edu.qa, david.corney@fullfact.org, a.barron@unibo.it,
{maram.hasanain, telsayed}@qu.edu.qa, papotti@eurecom.fr, dasan@math.unipd.it

Abstract

The reporting and the analysis of current events around the globe has expanded from professional, editor-lead journalism all the way to citizen journalism. Nowadays, politicians and other key players enjoy direct access to their audiences through social media, bypassing the filters of official cables or traditional media. However, the multiple advantages of free speech and direct communication are dimmed by the misuse of media to spread inaccurate or misleading claims. These phenomena have led to the modern incarnation of the *fact-checker* — a professional whose main aim is to examine claims using available evidence and to assess their veracity. Here, we survey the available intelligent technologies that can support the human expert in the different steps of her fact-checking endeavor. These include identifying claims worth fact-checking, detecting relevant previously fact-checked claims, retrieving relevant evidence to fact-check a claim, and actually verifying a claim. In each case, we pay attention to the challenges and the potential impact on real-world fact-checking.

1 Introduction

The spread of fake news, misinformation and disinformation on the web and in social media has become an urgent social and political issue. Social media have been widely used not only for social good, but also to mislead entire communities. To fight against such false or misleading information, several initiatives for manual fact-checking have been launched. Some notable fact-checking organizations include *FactCheck.org*,[1] *Snopes*,[2] *PolitiFact*,[3] and *FullFact*.[4]

Such fact-checking organizations are also potential beneficiaries of and/or leaders in automated fact-checking research. As misinformation and disinformation have become major concerns globally, tech companies, as well as national and international agencies began work in this area. Recently, several international initiatives have also emerged such as the *Credibility Coalition*[5] and *EUfactcheck*,[6] and some tools have been made available such as *Google Factcheck*[7] and *Hoaxy*.[8] Moreover, fact-checking is common in settings beyond online misinformation, as the verification of content's accuracy is a priority for many organizations [Karagiannis *et al.*, 2020].

A large body of research focused on developing automatic systems for fact-checking [Li *et al.*, 2016; Shu *et al.*, 2017; Lazer *et al.*, 2018; Vosoughi *et al.*, 2018; Vo and Lee, 2018]. This includes datasets [Hassan *et al.*, 2015; Augenstein *et al.*, 2019], and evaluation campaigns [Thorne and Vlachos, 2018; Nakov *et al.*, 2021a]. However, there are credibility issues with automated systems [Arnold, 2020], and thus a reasonable solution is to build tools to facilitate human fact-checkers. Yet, there has been limited work in this direction.

Thus, in this survey, we explore what fact-checkers want and what research has been done that can actually support them in their work. This is important because manual fact-checking is time-consuming. The study by Vlachos and Riedel [2014] describes the following typical sequence of fact-checking steps: (*i*) extracting statements that are to be fact-checked, (*ii*) constructing appropriate questions, (*iii*) obtaining the pieces of evidence from relevant sources, and (*iv*) reaching a verdict using that evidence.

In the current information ecosystem (including web and social media), there is a large volume of false claims not only in textual form, but also misleading or manipulated images and videos, including "deepfakes." However, here we limit our focus to automated fact-checking on text, as it remains the focus of most professional fact-checkers.

[*] Contact Author

[1] http://www.factcheck.org

[2] http://www.snopes.com/fact-check/

[3] http://www.politifact.com

[4] http://fullfact.org

[5] http://credibilitycoalition.org

[6] http://eufactcheck.eu

[7] http://toolbox.google.com/factcheck/explorer

[8] http://hoaxy.osome.iu.edu

There have been a number of surveys on "fake news" [Shu *et al.*, 2017; Lazer *et al.*, 2018; Vosoughi *et al.*, 2018; Alam *et al.*, 2021], rumors [Zubiaga *et al.*, 2018], fact-checking [Thorne and Vlachos, 2018; Kotonya and Toni, 2020], factuality [Li *et al.*, 2016; Zannettou *et al.*, 2019; Nakov *et al.*, 2021b], and propaganda [Martino *et al.*, 2020]. Unlike that work, here we study the desiderata of fact-checkers vs. the research attempts that aim to meet them.

2 What Fact-Checkers Want

Recently, Full Fact carried out extensive interviews with professional fact-checkers from 24 organizations in 50 countries [Arnold, 2020]. The report discussed key challenges they face where they believe technology can help. These include monitoring potentially harmful content, selecting claims to check, creating and distributing articles, and managing suggestions from readers (such as tip lines serving WhatsApp or Signal).

The same report revealed that most fact-checkers do *not* believe that tools to automate the verification of claims, i.e., the last step of a typical fact-checking pipeline [Vlachos and Riedel, 2014], will be used in the foreseeable future. Some believe that the required intuition and creativity can never be automated, even if some parts of their work can be supported.

This sets up a twin challenge for Artificial Intelligence (AI) practitioners: *first*, to develop practical tools that solve the problems fact-checkers face, and *second*, to demonstrate their value to fact-checkers in their day-to-day work. In the meantime, there is a recognised need for tools to help with finding claims, including previously fact-checked claims, and in providing relevant evidence to help write fact-checking articles.

2.1 Finding Claims Worth Fact-Checking

Choosing which claims to check is a complex process. Fact-checking is time-consuming and it often takes effort to determine whether a claim can even be checked, let alone whether it is misleading. Fact-checkers have to balance the potential harm that a misleading claim may cause (including risk to health, risk to democratic processes, and risk of exacerbating emergency situations) against the effort required to check a claim. Fact-checkers are also committed to being non-partisan, and thus it is important that such tools do not introduce any unfair bias. In many countries, governments choose not to publish reliable official statistics, thus making certain statistics-related claims virtually impossible to verify.

While simple algorithms can often decide whether content is viral, it is much harder to estimate the "checkworthiness" of a claim. For example, breaking news stories are often both popular and accurate. Given the limited resources of fact-checking organizations, many claims that are check-worthy nonetheless remain unchecked; thus, using historic lists of claims that were or were not checked is *not* a reliable indication of whether similar claims are worth fact-checking.

Claims may be found in many sources, including news websites, social media (text, audio, or video), and broadcast media. To monitor such a range of sources, fact-checkers often use a variety of technologies, such as news alerts, automatic speech recognition and translation tools, all of which typically depend on underlying AI technologies.

2.2 Detecting Previously Fact-Checked Claims

Misleading claims are often repeated in multiple channels, independently of any fact-checks or rebuttals.[9] Once a claim has been established as misleading, the ongoing spread of repeats or copies of the claim can be minimised by its rapid detection. In the simplest cases, these could be simple "copy and paste" repeats that are relatively easy to detect, but more often they will be paraphrases of the original or endlessly evolving variations. Given the resources required to write fact-checking articles, it is preferable to respond to multiple repeats of a claim with a single fact-checking article.

The number of fact-checking initiatives continues to grow. The *Duke Reporters' Lab* lists 305 active fact-checking organizations.[10] While some of them have debunked just a couple of hundred claims, others such as *PolitiFact*, *FactCheck.org*, *Snopes*, and *Full Fact* have each fact-checked thousands or even tens of thousands of claims.

Moreover, manual fact-checking often comes too late. It has been shown that "fake news" can spread six times faster than real ones [Vosoughi *et al.*, 2018], and that over half of the spread of some viral claims happens within the first ten minutes of their posting on social media [Zaman *et al.*, 2014]. To counter this, quickly detecting that a new viral claim has already been fact-checked allows for a timely action that can limit the spread and the potential harmful impact. The problem is made harder by the transient nature of many claims. For example, a claim about infection rates may be wrong today but correct next week, and thus re-using previous checks should be done carefully.

For journalists, the ability to discover quickly whether a claim has been previously fact-checked could be revolutionizing as it would allow them to put politicians on the spot during live events. In such a scenario, automatic fact-checking would be of limited utility as, given the current state of technology, it is not credible enough in the eyes of a journalist.

Finally, false claims often originate in one language and then get translated to other languages. Tools that can spot repeated claims across languages would be useful to address this. More generally, multi-lingual tools can help fact-checkers around the world, even those with limited resources.

2.3 Evidence Retrieval

Fact-checking is often limited by the time available: there are typically far more claims to verify than what is practically possible. Even if full automation remains out of reach (see the next section), tools that support fact-checkers in their manual verification process are to be welcomed.

Tools that automatically retrieve relevant data from trusted sources may save fact-checkers a lot of time. This is especially true if the evidence is hidden in large text documents, audio-visual recordings and streams, or is in a language that the fact-checker is not familiar with. Thus, combining automatic transcription, summarization, translation, and search can make sources of evidence available to fact-checkers that would be impossible or impractical to access otherwise.

[9] President Donald Trump repeated one false claim over 80 times: http://tinyurl.com/yblcb5q5.

[10] http://reporterslab.org/fact-checking/

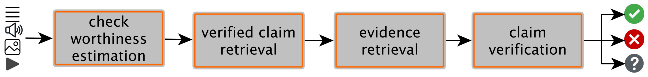

Figure 1: A fact-checking pipeline.

2.4 Automated Verification

On first consideration, the automated verification of claims seems like the ultimate application of AI to fact-checking. If such technologies can be developed and deployed, they would allow fact-checking organizations to be faster and to provide a more comprehensive coverage than manual fact-checking could ever achieve. However, many claims are not simply *correct* or *incorrect*, but may be *partially correct*, or *correct but misleading* without extra context, etc. One key role of professional fact-checkers is to help their audience gain full understanding of a claim, with all its nuances and complexity, rather than simply applying a binary classification.

Fact-checkers can only have an impact if they are trusted by their readers. They therefore take great care to only publish fact-checks after meticulous research, and adhere to strict editorial standards, as outlined, e.g., in the fact-checkers' code of principles[11] by the *International Fact-Checking Network*. This leads to a major hurdle before adopting fully automated verification methods: such methods will inevitably be imperfect, and publishing incorrect fact-checks could seriously damage the reputation of the responsible fact-checking organization. They may be more valuable as internal tools by presenting the evidence, reasoning and conclusion regarding a claim, before the (human) fact-checker writes and publishes their fact-checking article.

3 What Technology Currently Offers

Fact-checking is not a straightforward or routine process. It requires a chain of steps that go from sensing media and spotting check-worthy claims all the way through to concluding whether the claim is true, partially-true, false, misleading, or perhaps impossible to judge. Figure 1 shows a typical fact-checking pipeline, partially derived from [Barrón-Cedeño *et al.*, 2020]. Below, we discuss each step in this pipeline.

3.1 Finding Claims Worth Fact-Checking

As fact-checkers are flooded with claims, they need to decide what is worth fact-checking. This has encouraged the development of AI solutions, e.g., as part of shared tasks such as the *CLEF CheckThat! lab* 2018-2021 [Nakov *et al.*, 2018; Elsayed *et al.*, 2019; Barrón-Cedeño *et al.*, 2020; Nakov *et al.*, 2021a], and inside dedicated fact-checking organizations such as Full Fact.[12] The problem is widely tackled as a ranking one, where the system has to produce a check-worthiness scores. The score increases the system's transparency and provides fact-checkers with the ability to prioritize or to filter claims. Fact-checkers can also provide feedback on how reflective this score is of the actual check-worthiness of a claim, which can be later used to tune the system.

ClaimBuster [Hassan *et al.*, 2017] is the first system for check-worthiness detection, and it was used by fact-checkers in the *Duke Reporters' Lab* project.[13] It was trained on a manually annotated dataset to distinguish between *non-factual sentences*, *unimportant factual claims*, and *check-worthy factual claims*, using features based on sentiment, named entities, part-of-speech tags, words, and claim length. Konstantinovskiy *et al.* [2021] developed a more detailed schema and dataset for check-worthiness annotation of TV shows. Gencheva *et al.* [2017] created a dataset of political debates, derived by observing which sentences were fact-checked by fact-checkers, and modeled the sentence structure and the context of the claim. The dataset was used in the *ClaimRank* system [Jaradat *et al.*, 2018], and was extended to multitask learning from nine fact-checking organizations [Vasileva *et al.*, 2019]. Further extensions were used for the *CLEF CheckThat! lab*, where the participants developed models based on pre-trained transformers such as BERT and RoBERTa [Hasanain and Elsayed, 2020; Nikolov *et al.*, 2020; Williams *et al.*, 2020]. The task was also modeled using positive unlabeled learning [Wright and Augenstein, 2020].

During a recent general election, *Full Fact* used a fine-tuned BERT model to classify claims made by each political party, according to whether they were *numerical claims*, *predictions*, *personal beliefs*, etc. This allowed fact-checkers to rapidly identify the check-worthy claims, and thus to focus their efforts in the limited time available while voters are making their final decisions.

Social media companies are also working on combating misinformation and disinformation on their platforms. *Facebook* described a proprietary tool to identify claims that should be fact-checked.[14] They leverage flags by the users for a post indicating that it is potentially false, as well as features from the content of the replies, to predict whether the post contains false information. The model is updated using feedback from fact-checkers.

3.2 Detecting Previously Fact-Checked Claims

Interestingly, despite the importance of detecting whether a claim has been fact-checked before, it has been explored only recently. Shaar *et al.* [2020] formulated the task, and released two specialized datasets: (a) on tweets, which are to be compared to claims in *Snopes*, and (b) on political debates, to be matched to claims in *PolitiFact*. They further proposed a learning-to-rank approach based on a combination of BERT and traditional BM25, matching the input to the entire fact-checking article. Follow-up work explored the role of context for (b), including using neighboring sentences, co-reference resolution, and reasoning over the target text with Transformer-XH [Shaar *et al.*, 2021]. The task was also featured in the *CLEF CheckThat! Lab* [Barrón-Cedeño *et al.*, 2020; Nakov *et al.*, 2021a]. Vo and Lee [2020] explored a multi-modal setup, where tweets with claims about images were matched against the *Fauxtography* section of *Snopes*. *Full Fact* is currently trialling a similar tool internally.

[11] http://www.poynter.org/ifcn-fact-checkers-code-of-principles/

[12] http://fullfact.org/blog/2019/dec/how-we-use-ai-help-fact-check-party-manifestos/

[13] http://reporterslab.org/tech-and-check

[14] http://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works

Recently, *Google* has released the *Fact Check Explorer*,[7] which is an exploration tool that allows users to search a number of fact-checking websites, such that use *ClaimReview* from `schema.org`,[15] for the mentions of a topic, a person, etc. However, the tool cannot handle complex claims, as it uses *Google Search*, which is not optimised for long queries.

3.3 Evidence Retrieval

Evidence retrieval aims to find external evidence to help fact-checkers decide on the factuality of an input claim. When the input consists of a check-worthy claim and a (potentially closed) data collection, the process could finish in the production of a ranking of the relevant data —as in a *standard* retrieval scenario— or in the extraction of specific pieces of evidence, e.g., a text snippet or a recording.

When dealing with a closed reference collection, the task can be addressed as a ranking problem, e.g., based on BM25 or on some kind of similarity over vectorial representations between the input claim and the documents in the collection. Recent work has also combined document-level and sentence-level similarity to improve relevant document retrieval [Akkalyoncu Yilmaz *et al.*, 2019].

Once a relevant document has been found, it is possible to further extract relevant snippets representing arguments in favour or against the target claim, to be presented to the human fact-checker [Alshomary *et al.*, 2020].

It is also possible to further generate snippets to brief the fact-checkers with some relevant background knowledge about the target claim. Fan *et al.* [2020] achieved this by first generating and retrieving relevant *passage briefs*, then identifying and retrieving documents based on *entity briefs*, and finally generating and answering *question answering briefs* decomposed from the claim.

The *CLEF-2013 INEX lab* [Bellot *et al.*, 2013] included a shared task that asked to retrieve evidence snippets from a pool of 50k books to *confirm* or to *refute* a claim. They found that entity matching was one of the most important features.

The *CLEF CheckThat! lab* also featured tasks on claim evidence retrieval, at the document and also at the passage level, which was offered in Arabic [Elsayed *et al.*, 2019; Barrón-Cedeño *et al.*, 2020].

The *Fact Extraction and Verification shared task* (*FEVER*) focused on extracting an evidence sentence related to a claim from Wikipedia articles and determining whether it *supports*, *refutes*, or *provides no enough information* about the claim [Thorne *et al.*, 2018]. As in *INEX*, named entities were among the key pieces of information, and they were often used to compose the queries to retrieve the most relevant articles [Malon, 2018; Hanselowski *et al.*, 2018]. A typical system to solve the task starts with document retrieval, e.g., using BM25, followed by sentence retrieval based on the similarity between the input claim and each sentence in the top-n retrieved documents, which can be measured using TF.IDF, Word Mover's Distance, or BERT. Finally, it would use natural language inference to decide on the verdict. More recent work has used specialized neural semantic matching networks for each of these steps [Nie *et al.*, 2019].

Evidence retrieval might need to go beyond text, e.g., when verifying a claim about an image or a video. In such cases, reverse image search can help find other contexts where the multimedia content was used [Zlatkova *et al.*, 2019]. This allows to check whether these contexts agree with the claim, and to detect out-of-context content, e.g., an image or a video from one event portrayed as being from a different event, as well as potentially manipulated images/videos. Popular tools for this include *TinEye*,[16] *Google Image Search*, and *Yandex Image Search*. Relevant research tools are also being developed in two EU projects: *WeVerify*[17] and *InVID*.[18]

3.4 Automated Verification

Automatic claim verification approaches can be divided into explainable and non-explainable.

Explainable approaches, also known as *reference-based approaches*, are more relevant to assisting human fact-checkers. They verify the input claim against a trusted source such as tables [Chen *et al.*, 2020] or a database [Ahmadi *et al.*, 2019], or using inference over a knowledge graph, possibly while also using Horn rules [Gad-Elrab *et al.*, 2019]. This includes two approaches that we discussed above: finding previously fact-checked claims that can verify the input claim [Shaar *et al.*, 2020], and fact-checking it against Wikipedia [Thorne *et al.*, 2018; Nie *et al.*, 2019].

Non-explainable approaches make a prediction based on the content of documents retrieved from the Web [Popat *et al.*, 2016; Karadzhov *et al.*, 2017; Augenstein *et al.*, 2019], or on social media by modeling the message and its propagation, the users and their reactions over time, links to media sites, etc. [Castillo *et al.*, 2011; Shu *et al.*, 2017; Vosoughi *et al.*, 2018; Nguyen *et al.*, 2020]. This further includes analysis of the language used in the claims based on lexicons such as LIWC [Rashkin *et al.*, 2017], or using perplexity analysis [Lee *et al.*, 2021]. Fact-checking has also been done using masking in BERT-style transformers [Lee *et al.*, 2020].

While automatic verification is hard, there are promising results for certain kinds of claims. For example, an explicit claim about a numerical value, such as "*In 2017, global electricity demand grew by 3%.*", can be verified automatically using official statistics, even when this requires applying a complex formula [Karagiannis *et al.*, 2020]. Success here depends on the availability of reliable data, presented in a consistent format, which varies widely between countries and fields. Similarly, simple claims can be verified with promising accuracy when good evidence is available, e.g., for popular entities on the Web [Augenstein *et al.*, 2019].

While the accuracy and the scope of automated fact-checking algorithms keeps improving, two problems prevent their adoption in fact-checking organizations. First, even on the original datasets, their effectiveness is not high enough to allow automatic decisions. Second, most claims in the public realm are more complex, e.g., that COVID-19 vaccines have been developed too quickly and are still experimental.[19]

[15] http://schema.org/ClaimReview

[16] http://tineye.com/

[17] http://weverify.eu/tools/

[18] http://www.invid-project.eu

[19] http://fullfact.org/online/covid-19-survival-rate-less-998/

To verify such claims, fact-checkers might need to interview experts, to collaborate with other fact-checkers, to understand the context and the framing of the claims, to track down and to verify multiple sources and pieces of evidence — all of which require human-level intelligence. The general verification of arbitrary claims requires deep understanding of the real world that currently eludes AI. Indeed, most methods are designed to assist fact-checkers with suggestions, and assume that a human user will assess the verification output.

3.5 Some Real-World Systems

Below, we present a brief overview of some notable systems that cover multiple steps of the fact-checking pipeline, while also offering a suitable user interface.

AFCNR: The system accepts a claim as an input, searches over news articles, retrieves potential evidence and presents to the user a judgment on the stance of each piece of evidence towards the claim and an overall rating of the claim's veracity given the evidence [Miranda *et al.*, 2019]. The system was extensively tested by eleven journalists from BBC.

BRENDA: This is a browser extension, which allows users to fact-check claims directly while reading news articles [Botnevik *et al.*, 2020]. It can take either the full page opened in the browser or a highlighted snippet inside the page. In the first scenario, the system applies check-worthiness identification in order to decide which sentences to fact-check.

ClaimPortal: [20] After retrieving tweets in response to a query, the system [Majithia *et al.*, 2019] scores them for check-worthiness using *ClaimBuster* and tries to verify each tweet using previously fact-checked claims from *PolitiFact*.

Squash: The system is developed at the *Duke Reporters' lab*, this system (*i*) listens to speech, debate and other events, (*ii*) transcribes them into text, (*iii*) identifies claims to check, and then (*iv*) fact-check them by finding matching claims already fact-checked by humans.[21]

Full Fact's system (*i*) follows news sites and social media, (*ii*) identifies and categorizes claims in the stream, (*iii*) checks whether a claim has been already verified, and then (*iv*) enriches the claims with data to support the fact-checker. It is in daily use in the UK and several countries in Africa.[22]

We believe that the prototypes presented above are good examples of the steps taken towards developing systems that cater to fact-checkers. More systems are now designed to *efficiently* identify claims originating from *various types* of sources (e.g., news articles, broadcast, and social media). Moreover, the fact-checker is now becoming a part of the system by providing feedback, rather than just being a consumer of its output. Finally, we see an increase in systems' transparency by providing explainable decisions, thus making them more an assistive tool rather than a replacement for the fact-checker. However, there are several challenges left to tackle, as we present in the next sections.

4 Lessons Learned

The main lesson from our analysis is that there is a partial disconnection between what fact-checkers want and what technology has to offer. We provide more detail below.

1. Over time, many tools have been developed, either to automatically fact-check claims or to provide facilities to the fact-checkers to support their manual fact-checking process. However, there are still limitations in both automated and manual processes: (*i*) credibility issue for automated systems, as they do not provide supporting evidence, and (*ii*) scalability issue for manual fact-checking.
2. Automated fact-checking systems can help fact-checkers in different ways: (*i*) to find claims worth fact-checking, (*ii*) to find relevant previously fact-checked claims; (*iii*) to find supporting evidence (in the form of text, audio or video), translating (for multilingual content) and summarising relevant posts, articles and documents if needed, and (*iv*) to detect claims that are spreading faster to slow them down.
3. There is a lack of collaboration between researchers and practitioners in terms of defining tasks and developing datasets to develop automated systems. In general, a human-in-the-loop can be an ideal setting for fact-checking, which is currently not fully explored.

5 Challenges and Future Forecasting

Below we discuss some major challenges and we forecast some promising research directions:

5.1 Major Challenges

- **Leveraging multi-lingual resources:** The same claim, with slightly different variants, often spreads over different regions of the world at almost the same or at different time periods. These may be "international claims" such as medical claims about COVID-19, or stories that are presented as local, but with varied, false locations. Those claims might be fact-checked in one language, but not in others. Moreover, resources in English are abundant, but in low-resource languages, such as Arabic, they are clearly lacking. Aligning and coordinating the verification resources and leveraging them across different languages to improve fact-checking is a challenge.
- **Ambiguity in the claims:** Another reason why automatic fact-checking is challenging is related to the fact that often a claim has multiple interpretations. An example is "*The COVID death rate is rising.*" Is this about mortality or about fatality rate? Does it refer to today/yesterday or to the last week/month? Does it refer to the entire world or to a specific area? In such cases, knowledge about the context is necessary in order to properly frame the claim and to filter out unlikely interpretations. After that, all remaining interpretations should be analyzed, which would further slow down the work of fact-checkers. One system that proposes a solution to this problem is CoronaCheck.[23]

[20] http://idir.uta.edu/claimportal/

[21] http://reporterslab.org/squash-report-card-improvements-during-state-of-the-union-and-how-humans-will-make-our-ai-smarter/

[22] http://fullfact.org/blog/2020/jul/afc-global/

[23] http://coronacheck.eurecom.fr

- **System bias:** The majority of existing systems are trained using datasets curated by a small group of people and often annotated by non-experts. This in turn results in systems biased towards how the system developers perceive factuality and how the annotation task was described to the annotators. The dangers of bias in large language models is becoming increasingly obvious [Bender *et al.*, 2021], and should not be ignored just because the purpose of the system is benevolent.

- **Contextual information:** The current state-of-the-art for automated fact-checking makes limited use of contextual information, e.g., reader's comments, linked sources of news articles, social network data for social media posts. Such information can provide useful signals for enriching the current models.

- **Multimodality:** Information is typically disseminated through multiple modalities such as text, image, speech, video, temporal, user profile, and network structure. Addressing the problem based on a single modality can be a step towards failure. For example, it might be difficult to detect fake news pieces that are automatically generated using deep fakes and/or GPT-3-style text generation. To avoid such issues, multimodal approaches would be one way to go, if evidence can be gathered from multiple types of sources at the same time. This in turn requires multimodal datasets to develop suitable models.

5.2 Future Forecasting

- **Close collaboration between fact-checking platforms and researchers:** We envision closer collaboration between professionals from fact-checking platforms alongside researchers in the domain to discuss common interests, existing solutions, and future directions, has been a challenge.

- **Integrated solutions:** We also envision unified and open-source initiatives to develop resources for system development and benchmarking.

- **Usability:** We further forecast more research on the system interface design, which would facilitate the adoption of AI by fact-checkers. It is important to develop systems that require minimal technical knowledge and reduce cognitive load. Such systems can help a larger number of fact-checkers and journalists in the fact-checking process.

- **Interpretability and explainability:** Models should be designed in such a way that their outcomes are explainable, unbiased, and more accountable to ethical considerations.

- **Efficient and real-time solutions:** Finally, in order to tackle the velocity of the spread of fake news there is a need to develop systems that are efficient and scalable for real-time solution. To be effective, such systems would need to be embedded within, or accessible by, social networks and other big technology companies.

6 Conclusion

We have presented a survey of the available intelligent technologies that can support the human experts in the different steps of the manual process of fact-checking claims. These include tasks such as identifying claims worth fact-checking, detecting relevant previously fact-checked claims, retrieving relevant evidence to support the manual fact-check of a claim, and actually verifying a claim. In each case, we paid attention to the challenges in future work and to the potential impact on real-world fact-checking.

We argued that there is currently only a partial overlap between what fact-checkers want and what the research community considers as a priority. We then discussed lessons learned and major challenges that need to be overcome. Finally, we suggested several research directions, which we forecast will emerge in the near future.

Acknowledgments

This work was made possible in part by grant# NPRP 7-1330-2-483 from the Qatar National Research Fund (a member of Qatar Foundation). Partial support also comes from a Google gift. This work is also part of the Tanbih mega-project, developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news," propaganda, and media bias by making users aware of what they are reading. The statements made herein are solely the responsibility of the authors.

References

[Ahmadi *et al.*, 2019] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. Explainable fact checking with probabilistic answer set programming. In *TTO*, 2019.

[Akkalyoncu Yilmaz *et al.*, 2019] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *EMNLP*, pages 3488–3494, 2019.

[Alam *et al.*, 2021] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. *arXiv/2103.12541*, 2021.

[Alshomary *et al.*, 2020] Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. Extractive snippet generation for arguments. In *SIGIR*, page 1969–1972, 2020.

[Arnold, 2020] Phoebe Arnold. The challenges of online fact checking. Technical report, Full Fact, 2020.

[Augenstein *et al.*, 2019] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP*, pages 4684–4696, 2019.

[Barrón-Cedeño *et al.*, 2020] Alberto Barrón-Cedeño, Tamer Elsayed, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, and Preslav

Nakov. CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims on social media. In *ECIR*, pages 499–507, 2020.

[Bellot *et al.*, 2013] Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric Sanjuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, Matthew Trappett, and Qiuyue Wang. Overview of INEX 2013. In *CLEF*, pages 269–281, 2013.

[Bender *et al.*, 2021] Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*, pages 610–623, 2021.

[Botnevik *et al.*, 2020] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. BRENDA: Browser extension for fake news detection. In *SIGIR*, pages 2117–2120, 2020.

[Castillo *et al.*, 2011] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on Twitter. In *WWW*, page 675–684, 2011.

[Chen *et al.*, 2020] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. TabFact: A large-scale dataset for table-based fact verification. In *ICLR*, 2020.

[Elsayed *et al.*, 2019] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Pepa Atanasova, and Giovanni Da San Martino. CheckThat! at CLEF 2019: Automatic identification and verification of claims. In *ECIR*, pages 309–315, 2019.

[Fan *et al.*, 2020] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. Generating fact checking briefs. In *EMNLP*, pages 7147–7161, 2020.

[Gad-Elrab *et al.*, 2019] Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. Tracy: Tracing facts over knowledge graphs and text. In *WWW*, pages 3516–3520, 2019.

[Gencheva *et al.*, 2017] Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. A context-aware approach for detecting worth-checking claims in political debates. In *RANLP*, pages 267–276, 2017.

[Hanselowski *et al.*, 2018] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. UKP-Athene: Multi-sentence textual entailment for claim verification. In *FEVER*, pages 103–108, 2018.

[Hasanain and Elsayed, 2020] Maram Hasanain and Tamer Elsayed. bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness. In *CLEF*, 2020.

[Hassan *et al.*, 2015] Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting check-worthy factual claims in presidential debates. In *CIKM*, pages 1835–1838, 2015.

[Hassan *et al.*, 2017] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *SIGKDD*, pages 1803–1812, 2017.

[Jaradat *et al.*, 2018] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. ClaimRank: Detecting check-worthy claims in Arabic and English. In *NAACL-HLT*, pages 26–30, 2018.

[Karadzhov *et al.*, 2017] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. Fully automated fact checking using external sources. In *RANLP*, pages 344–353, 2017.

[Karagiannis *et al.*, 2020] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *VLDB*, 13(11):2508–2521, 2020.

[Konstantinovskiy *et al.*, 2021] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2(2), April 2021.

[Kotonya and Toni, 2020] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In *COLING*, pages 5430–5443, 2020.

[Lazer *et al.*, 2018] David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[Lee *et al.*, 2020] Nayeon Lee, Belinda Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. Language models as fact checkers? In *FEVER*, pages 36–41, 2020.

[Lee *et al.*, 2021] Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. Towards few-shot fact-checking via perplexity. In *NAACL*, pages 1971–1981, 2021.

[Li *et al.*, 2016] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

[Majithia *et al.*, 2019] Sarthak Majithia, Fatma Arslan, Sumeet Lubal, Damian Jimenez, Priyank Arora, Josue Caraballo, and Chengkai Li. ClaimPortal: Integrated monitoring, searching, checking, and analytics of factual claims on Twitter. In *ACL*, pages 153–158, 2019.

[Malon, 2018] Christopher Malon. Team Papelo: Transformer networks at FEVER. In *FEVER*, pages 109–113, 2018.

[Martino *et al.*, 2020] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di

Pietro, and Preslav Nakov. A survey on computational propaganda detection. In *IJCAI*, pages 4826–4832, 2020.

[Miranda *et al.*, 2019] Sebastião Miranda, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel, and Zita Marinho. Automated fact checking in the news room. In *WWW*, pages 3579–3583, 2019.

[Nakov *et al.*, 2018] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *CLEF*, pages 372–387, 2018.

[Nakov *et al.*, 2021a] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *ECIR*, pages 639–649, 2021.

[Nakov *et al.*, 2021b] Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. A survey on predicting the factuality and the bias of news media. *arXiv/2103.12506*, 2021.

[Nguyen *et al.*, 2020] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. FANG: Leveraging social context for fake news detection using graph representation. In *CIKM*, pages 1165–1174, 2020.

[Nie *et al.*, 2019] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*, pages 6859–6866, 2019.

[Nikolov *et al.*, 2020] Alex Nikolov, Giovanni Da San Martino, Ivan Koychev, and Preslav Nakov. Team_Alex at CheckThat! 2020: Identifying check-worthy tweets with transformer models. In *CLEF*, 2020.

[Popat *et al.*, 2016] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *CIKM*, pages 2173–2178, 2016.

[Rashkin *et al.*, 2017] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, pages 2931–2937, 2017.

[Shaar *et al.*, 2020] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. That is a known lie: Detecting previously fact-checked claims. In *ACL*, pages 3607–3618, 2020.

[Shaar *et al.*, 2021] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. The role of context in detecting previously fact-checked claims. *arXiv:2104.07423*, 2021.

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

[Thorne and Vlachos, 2018] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *COLING*, pages 3346–3359, 2018.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The fact extraction and VERification (FEVER) shared task. In *FEVER*, pages 1–9, 2018.

[Vasileva *et al.*, 2019] Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *RANLP*, pages 1229–1239, 2019.

[Vlachos and Riedel, 2014] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Workshop on LT and CSS*, pages 18–22, 2014.

[Vo and Lee, 2018] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking URL recommendation to combat fake news. In *SIGIR*, pages 275–284, 2018.

[Vo and Lee, 2020] Nguyen Vo and Kyumin Lee. Where are the facts? Searching for fact-checked information to alleviate the spread of fake news. In *EMNLP*, pages 7717–7731, 2020.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[Williams *et al.*, 2020] Evan Williams, Paul Rodrigues, and Valerie Novak. Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In *CLEF*, 2020.

[Wright and Augenstein, 2020] Dustin Wright and Isabelle Augenstein. Claim check-worthiness detection as positive unlabelled learning. In *Findings of EMNLP*, pages 476–488, 2020.

[Zaman *et al.*, 2014] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. A Bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.*, 8(3):1583–1611, 2014.

[Zannettou *et al.*, 2019] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data and Information Quality*, 11(3):10:1–10:37, 2019.

[Zlatkova *et al.*, 2019] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. In *EMNLP*, pages 2099–2108, 2019.

[Zubiaga *et al.*, 2018] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2), February 2018.