Tackling Multilinguality and Internationality in Fake News

Andrey Tagarev
Research Unit, Sirma AI
79 Nikola Gabrovski St
Sofia, Bulgaria
andrey.tagarev@ontotext.com

Ivelina Nikolova-Koleva
Bulgarian Academy of Sciences and Sirma AI
79 Nikola Gabrovski St
Sofia, Bulgaria
ivelina.nikolova@ontotext.com

Krasimira Bozhanova
Semantic Analytics Solutions, Sirma AI
79 Nikola Gabrovski St
Sofia, Bulgaria
krasimira.bozhanova@ontotext.com

Ivan Ivanov
Life Science Solutions, Sirma AI
79 Nikola Gabrovski St
Sofia, Bulgaria
ivan.b.ivanov@ontotext.com

Abstract

The last several years have seen a massive increase in the quantity and influence of disinformation being spread online. Various approaches have been developed to target the process at different stages from identifying sources to tracking distribution in social media to providing follow up debunks to people who have encountered the disinformation.

One common conclusion in each of these approaches is that disinformation is too nuanced and subjective a topic for fully automated solutions to work but the quantity of data to process and cross-reference is too high for humans to handle unassisted. Ultimately, the problem calls for a hybrid approach of human experts with technological assistance.

In this paper we will demonstrate the application of certain state-of-the-art NLP techniques in assisting expert debunkers and fact checkers as well as the role of these NLP algorithms within a more holistic approach to analyzing and countering the spread of disinformation. We will present a multilingual corpus of disinformation and debunks which contains text, concept tags, images and videos as well as various methods for searching and leveraging the content.

1 Introduction

The topic of fake news and intentional spread of disinformation has been gaining increasing prominence over the last decade. The distinction of terminology and attempts to classify various erroneous or misleading statements online is constantly evolving but disinformation, as shown in Fallis (2015), has recently settled as the commonly accepted term to describe the intentional and systematic spread of incorrect information.

The spread of this incorrect information is strongly reliant on social media, causing a strong emotional reaction and quick propagation before its inaccuracy can be effectively exposed. This means that in most cases minimal effort is put into crafting the disinformation, instead relying on speed, volume and reuse of slightly modified pre-existing materials. There are, of course, always new disinformation materials popping up but they are in the minority and should they gain traction, they will very quickly get picked up, modified slightly and reused.

A very stark example of this kind of interaction was provided in the early months of the Covid pandemic as reliable scientifically-tested information was still rather scarce and the void was filled by a wide variety of rapidly-spreading fake and unsupported claims. This can be viewed both from the perspective of journalists mobilizing to counteract the spread[1] and from that of researchers looking into assisting their efforts e.g. in identifying spread of disinformation that has already been debunked as in Singh et al. (2021) or in tracking the comparative effect of disinformation and debunking tweets as in Jiang et al. (2021).

This means that combating the spread of disinformation can happen on a variety of levels. One option is to identify the creator and limit their reach - a replacement will pop up eventually but rebuilding a presence in the social network requires time and resources. Alternatively, it is possible to identify a piece of disinformation early in its spread and expose it to the people who interact with it before it really gains traction. Finally, it is possible to monitor social media for trending topics and work on creating convincing well-supported debunks to new disinformation that has gained popularity. An ideal approach would combine all three aspects in some manner.

Finally, the reality is that while disinformation might typically involve minor or simple modifi-

[1] https://weverify.eu/blog/speeding-up-the-debunking-process/

Proceedings of Recent Advances in Natural Language Processing, pages 1380–1386
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_154

cations, it is still intentionally crafted to be misleading and is constantly evolving and adapting. This makes completely automated approaches to combating it impractical. Meanwhile, the sheer volume of information that needs to be tracked, analysed and correlated makes a completely manual approach equally impractical. The solution will inevitably then involve a hybrid approach.

To that end, we present a data set based on a collection of fact-checker created debunks of pieces of disinformation that has been extended with additional metadata. Several forms of advanced search functionality have been developed on top of it in order to make discovering relevant content in the data set as straight-forward as possible. This will allow fact-checking experts to easily check for previous work on disinformation they encounter, point to previous instances of it being used and react quickly to its spread in order to counter it early on.

2 The Data Set

The data set used for our experiments is based on a snapshot of the Database of Known Fakes[2] (DBKF). It is a collection of debunking content from highly respected fact-checking organizations around the world extended with additional metadata related to said debunks in order to enable the advanced search and correlation functionality we present here.

Figure 1 shows an overview of the major types of objects contained within the DBKF and the connections between them. At its core, the data model of the data set is based around the Claim[3] and ClaimReview[4] format defined within schema.org which is already familiar to and used by many fact checking organizations.

The core objects defined within the schema.org specification are a Claim (a short statement summarizing the target of the debunk) and ClaimReview (a typically article-length debunk of the claim being discussed). As can be seen in Figure 1, these two objects are extended with additional explicit objects. The two most important additions are Appearances and Evidences - the former are links to posts where a specific Claim is being made and the latter are external content supporting the explanation and reasoning contained within a debunk.

Appearances are automatically expanded to include additional metadata available at the external website (more on that in Subsection 3.1) and a number of state-of-the-art systems are used to enrich the objects further (more on that in Subsections 3.3 and 4.2). Some of this metadata is more detailed and contained within specialized objects such as Image, Video, Concept and Annotation (which is an instance of a concept at a specific location in a document's text).

At the time of writing, the data set contains

- **Claims**: 32,138
- **Debunks**: 32,220
- **Appearances**: 74,099
- **Evidences**: 348,100
- **Concepts**: 110,158
- **Annotations**: 359,032
- **Images**: 9,774 links to an image of which 9,217 unique image urls
- **Videos**: 9,866 links to a video of which 9,745 unique video urls

but the numbers are always growing as new input is being collected from many fact checking and debunking organizations daily.

Figure 2 provides a specific example of a debunk that might be retrieved by the system (pre-advanced enrichment steps). The *claim* is that an immigrant destroyed a statue in Italy, three *appearances* provide links to three tweets that made that claim, the *debunk* disproves the claim and explains what actually happened while the *evidences* support that explanation.

2.1 Sources

The contents of the data set are sourced from organizations in the Facebook third-party fact-checking program[5]. At the time of writing the data set contained data from 17 separate organizations in at least 13 languages and based on disinformation encountered in over 20 different countries.

Due to the variety in article formatting and institutional approaches to writing the debunks, there is some significant variety in the details of the retrieved debunks but the objects in the high level model presented in Section 2 are present for all

[2] https://weverify-demo.ontotext.com/
[3] https://schema.org/Claim
[4] https://schema.org/ClaimReview
[5] https://www.facebook.com/journalismproject/programs/third-party-fact-checking

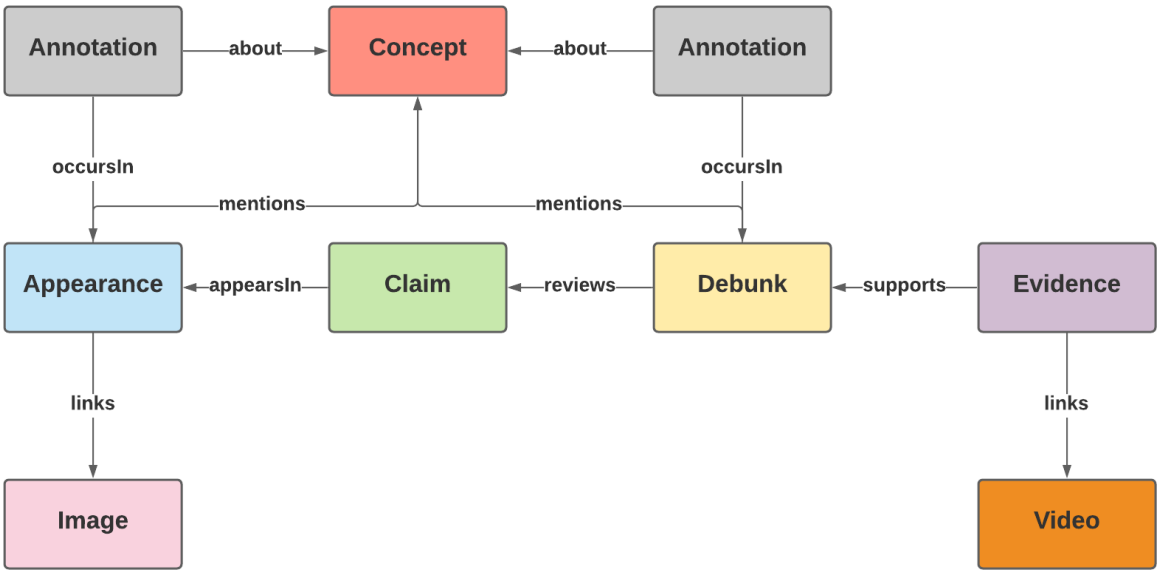

Figure 1: An overview of the major objects in the Database of Known Fakes

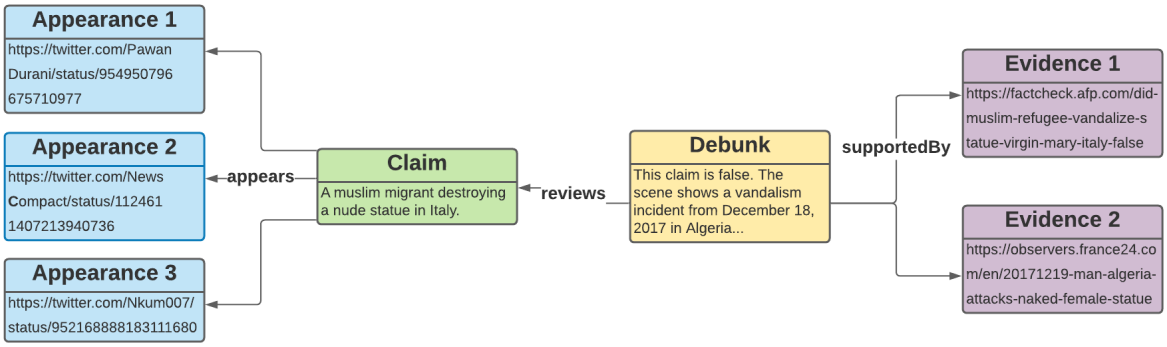

Figure 2: An example of the major objects associated with a specific debunk

sources. The differences mostly concern details like quantity of appearances, types of evidence provided, etc.

3 Enrichment

This section will go over the additional enrichment steps that are used to take the DBKF beyond a simple collection of existing debunks and unlock the ability to do advanced searching and correlation by expanding the metadata available on the debunks. This contains the additional metadata added to Appearance and Evidence objects, language tagging of all texts and named entity recognition.

3.1 Appearances and Evidence

When originally extracted from the debunking articles, appearances and evidences are just plain urls. These objects are expanded to include additional metadata, whenever possible. This has multiple goals:

1. Extract text, images, videos, author information, publication time, etc. to be used in searching, filtering and analysis

2. Archive the link so that it is still accessible should the original be removed (often the case with disinformation that gets debunked)

3. If the link is already an archive, extract the original url for domain analysis purposes

The retrieval of this metadata is, of course, not always possible. Aside from the common case of a link being removed or simply inaccessible, there is also no guarantee how the target website will be formatted. To that end we have chosen

to focus metadata expansion on a few social media websites that are particularly common (Twitter, Facebook, YouTube), popular archive websites (perma.cc, archive.org, archive.is, etc.) and websites that follow the Google guidelines for publishing news articles with properly tagged metadata. This allows us to collect at least some metadata for over 90% of appearances.

3.2 Language Recognition

Language detection is a simple task but an important first building block for some of the more complex enrichment steps presented later such as for selecting the appropriate NER pipeline in subsection 3.3. For this reason we ran the texts of claim, debunks and appearances through a well-tested language detection algorithm- Shuyo (2010).

It is worth noting that while language detection is not an especially difficult task, the contents of the data set are quite varied in a number of ways - quite a variety of languages, mixed-language texts, various lengths (from few words to multi-page articles). This all means that some amount of errors will inevitably be introduced at this step of the process.

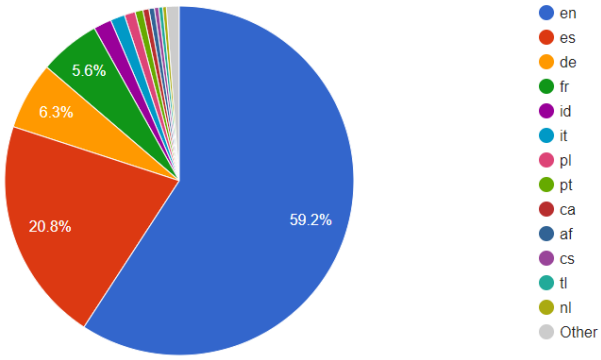

Figure 3: A pie chart of languages distribution in the data set

In Figure 3 we can see the language distribution produced by the algorithm. The distribution of languages corresponds to what we expect to see based on the fact-checking sources present in the data set. English, Spanish, German and French are currently the major languages and there is a long tail of languages where the total number of available debunks is much lower.

It is worth keeping in mind that the distribution is both a reflection of currently active fact-checking organizations and the specific sources that are processed and ingested in the system. This is to say that the situation is quite fluid and more languages can become relevant in the future.

	Detected	P (strict)	P (partial)
CES	156	0.74	0.95
spaCy	171	0.66	0.83
Google Cloud	190	0.59	0.81

Table 1: A comparison of the precision of general concepts over English text between three systems

3.3 Locations and Concepts

The language tagging of all text in the data set allows the final step in metadata enrichment - named entity recognition carried out over the corpus. This task was further divided in two, based on the needs of the users and analysis of available algorithms. After reviewing the literature on comparative analysis of available algorithms Schmitt et al. (2019), we ran some additional comparisons of different approaches since the diverse nature of our data set and unconventional target concepts make comparison over standard data sets less feasible.

Table 1 shows a comparison between an in-house developed NER pipeline targeted at the publishing domain, which is based on GATE[6] using a subset of the Wikidata[7] data set (referred to as CES - Concept Extraction Service), the default spaCy pipeline[8] and the NER functionality of Google Cloud[9]. The comparison was carried out over a variety of language although the CES algorithm was only used on English texts.

The conclusion is that CES, while limited to only English, has a notable performance advantage. Meanwhile the spaCy pipelines and Google Cloud offering are roughly on par but both support a larger variety of languages.

For that reason, CES was used to do NER of location mentions over the data set. It was decided that the better performance in correctly identifying location mentions offsets the limiting of that enrichment to only English. Conversely, the general "concept" tags rather typical POL entities are more useful when applied to as many of the texts as practical.

A word of caution on the applicability of simple numerical comparison in the case of general con-

[6]https://gate.ac.uk/
[7]https://www.wikidata.org/
[8]https://spacy.io/usage/models/
[9]https://cloud.google.com/natural-language

cept recognition. The case of disinformation spread is a very challenging one for the NER task since the uniquely identifying concepts associated with given misinformation are quite varied and much less well-defined than the typical POL formulation of the task. Words like ”immigrant”, ”hospital”, ”lemon”, ”vaccine”, ”5G” and many others have actually proven quite important but their popularity and usefulness is actually quite limited in time. To that end we placed particular importance on the system's ability to identify such concepts when they first begin to gain prominence.

4 Search

The final step in unlocking the full potential of the data set is to enable powerful search functionality that can make discovering existing debunk information and locating similar cases of disinformation in the past as quick and easy as possible. This kind of searching should utilize the full capability of the collected and enriched data set and can be used as a stepping stone to semi-automated and automated systems such as early detection of disinformation and chat bots. It is also a first step to being able to detect larger trends within the data such as tracking the spread of a particular disinformation claim across countries or watching a particular piece of disinformation change and evolve in response to fact-checker debunks.

There are several aspects of this search functionality that build on each other. Firstly, we will describe the full-text and faceted search then near-duplicate detection based on visual similarity and multilingual search that uses latest neural machine translation. Subsection 5.1 will also briefly discuss what we envision these search capabilities building to in the future.

4.1 Facets

The basic functionality needed by expert fact-checkers is the ability to quickly look for debunks related to the claim they are currently investigating using keywords or even phrases. To this end, we have implemented a full-text search based on the Elasticsearch[10] engine. Results from the full-text search are presented, based on the user preference, either by the relevance score returned by Elasticsearch, or by date of publication. In order to fully exploit the information collected from the sources, as well as the metadata and enrichments created while populating DBKF, we have implemented faceted search. The facets currently supports filtering and slicing DBKF content or search results by: language, author, debunk publisher (source), time of publication, locations and concepts. For example, using the full-text search with the ”5G” keyword, together with facets can help debunkers quickly find the false claims that were circulating in different languages on the subject during a specified timeframe. Similarly, fact-checkers can use facets to quickly check what locations are mentioned in ”vaccine” (selected from the concept facet) related disinformation. The ability to search DBKF with the help of facets can be beneficial not only to verification professionals but also to researchers in the field of disinformation, social scientists and policy makers.

4.2 Visual Similarity

Searching based on visual similarity relies on the research performed by Kordopatis-Zilos et al. (2019) and is carried out by using the near duplicate detection (NDD) service[11]. This service supports indexing and searching for both images and videos based on visual similarity between the contents. This means that every image or video discovered within appearance and evidence objects is automatically indexed within the NDD service in order to be available for visual similarity searches.

When a visual search is initiated by the user, the image or video they provide is also indexed into the NDD service and then all returned results are tied back to their corresponding debunks within the DBKF. This effectively enables us to discover debunks that contain similar images and videos even if they have been reuploaded or slightly modified which are the typical way bad faith actors reuse them for spreading disinformation.

It is worth noting that once indexed, an image or video does not need to be stored in its original form and, in fact, due to concerns about storage and distribution rights of digital content, they are usually not stored locally. Instead, the final representation of the visual object is a single vector in a highly-dimensional space which cannot be used to recreate the original digital object. In practice, this means that the visual similarity service can discover connections to similar content but cannot show that content to the user.

As a practical step to combat the frequent disap-

[10] https://www.elastic.co/elasticsearch/

[11] http://ndd.iti.gr/

pearance of content tied to disinformation, we work with internet archiving websites to preserve any Appearance when we initially encounter it. These websites have a procedure for the content owner to have the a specific archived item removed but in practice while the social media post often disappears within months, the archiving organizations are rarely contacted to have the archived content removed. So as a response to a visual similarity search, we provide the similar item, a link to its original url and a url to an archiving website making it quite likely but not guaranteed that the user can view the original image or video.

There are other image and video processing services available for integration with the contents of the DBKF such as automated image forensic analysis Zampoglou et al. (2016) and deep fake detection Charitidis et al. (2020) but those are more suited to producing evidence to support debunks than for searching the data set. That said, it is possible to extend the metadata associated with debunks to reflect the kind of visual manipulations encountered within a piece of disinformation and make that available for search as another facet similar to the ones described in Subsection 4.1.

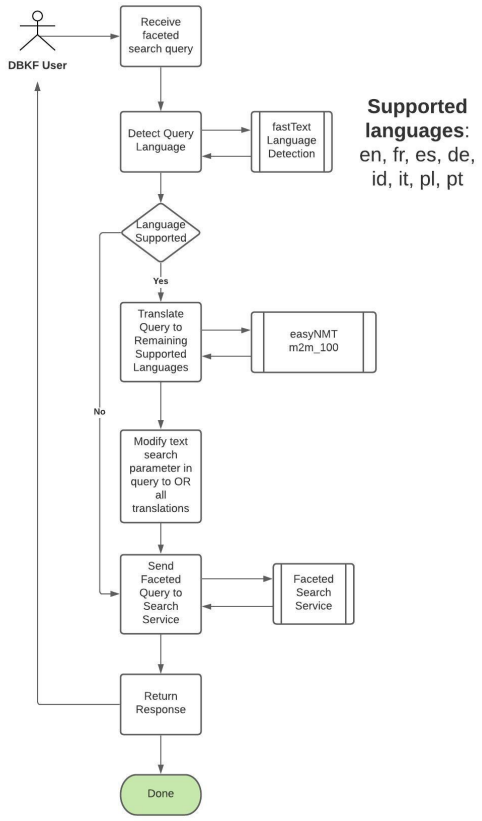

Figure 4: The workflow of multilingual search

4.3 Multilingual

The newest kind of search functionality enabled in the DBKF focuses on the multilingual aspect of the data set. Its intent is to vastly improve the ability to track the spread of disinformation in international situations.

The search utilizes the latest advances in neural machine translation and the translation is based on M2M-100 - a many-to-many multilingual translation model presented in Fan et al. (2020). It supports bidirectional translation between any pair of over 100 languages and shows a marked improvement in translation between non-English languages when compared to English-centric model. English is often not the first language in which disinformation appears so this is a very useful feature for our use case.

As shown in Figure 4, we have chosen to focus on the eight major languages of the data set. This is a reflection of the analysis shown in Figure 3 about the current distribution of data in the data set. A major advantage of the M2M-100 model is that it allows seamless adaptation to changes in the available data.

To quickly summarize the workflow presented, the search first identifies the language of the query. If it is a supported language, it translates it into all other supported languages and sends off a multi-expression query to be processed. Otherwise it forwards it without modifications. The search returns the results in order of relevance without regard to which translation they have matched.

The decision to not translate queries in unsupported language is a reflection of the limitation of the search. If the search query is too short or ambiguous (a not unlikely situation), the language tag will be unreliable and the translations will likely be of equally low quality.

5 Conclusion and Future Work

In conclusion, the DBKF already contains a large amount of data extended with useful metadata and powerful search capability. This can make it a powerful tool in the arsenal of fact-checkers and also allows its incorporation in counter-disinformation campaigns where people are targeted with evidence of a claim's falsehood before they spread it unknowingly. The contents of the database are also constantly growing with the automatic ingestion of new content. Future developments can include

the addition of new fact-checking sources, support for metadata-expansion of more types of social media posts and further building on the modeling and search functionality.

5.1 Multimodal Search

One improvement of particular interest is the option to enable a true multimodal search over the data set. As discussed in Section 4, we already have full-text, faceted, image and video search so the next step would be to combine them into a single endpoint. This would enable to effortlessly search for debunks relevant to a social media post, essentially the automatic ability to ask "*Is this post repeating known disinformation?*"

The challenge is actually combining the various results in a meaningful way. The various modalities operate on completely different scales, not to mention that they are all optional and possibly multi-valued e.g. how do you compare a post with a sentence of text, three concepts, a location mention and five images to a debunk that has three pages of text, seventy concepts, no images and two videos? It is by no means an insurmountable obstacle but extensive experimentation and careful fine-tuning will be required to produce intuitive and helpful results.

5.2 Model Extension

There are various ideas for adding additional aspects to the data model. One idea briefly mention in Subsection 4.2 is tagging debunks based on the type of disinformation techniques they represent e.g. deep fake videos, out-of-context images, etc. Work has began on building a vocabulary for disinformation techniques but we are consulting with fact-checking experts to align it to their expectations and needs.

The more interesting but complex direction of expansion would be to incorporate deeper understanding and tracking of disinformation campaigns into the model. This would allow to explicitly connect individual debunks into the larger trends they are coming up against.

Acknowledgments

This work is funded by the EU H2020 WeVerify (grant agreement: 825297) project.

References

Polychronis Charitidis, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2020. Investigating the impact of pre-processing and prediction aggregation on the deepfake detection task.

Don Fallis. 2015. What is disinformation? *Library Trends*, 63(3):401–426.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Ye Jiang, Xingyi Song, Carolina Scarton, Ahmet Aker, and Kalina Bontcheva. 2021. Categorising fine-to-coarse grained misinformation: An empirical study of COVID-19 infodemic. *CoRR*, abs/2106.11702.

Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. 2019. Visil: Fine-grained spatio-temporal video similarity learning.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343.

Nakatani Shuyo. 2010. Language detection library for java.

Iknoor Singh, Kalina Bontcheva, and Carolina Scarton. 2021. The false COVID-19 narratives that keep being debunked: A spatiotemporal analysis. *CoRR*, abs/2107.12303.

Markos Zampoglou, Symeon Papadopoulos, Yiannis Kompatsiaris, Ruben Bouwmeester, and Jochen Spangenberg. 2016. Web and social media image forensics for news professionals. *Proceedings of the International AAAI Conference on Web and Social Media*, 10.