Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims

Ivan Srba
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
ivan.srba@kinit.sk

Branislav Pecher*
Faculty of Information Technology, Brno University of Technology
Brno, Czech Republic
branislav.pecher@kinit.sk

Matus Tomlein
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
matus.tomlein@kinit.sk

Robert Moro
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
robert.moro@kinit.sk

Elena Stefancova
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
elena.stefancova@kinit.sk

Jakub Simko
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
jakub.simko@kinit.sk

Maria Bielikova†
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
maria.bielikova@kinit.sk

ABSTRACT

False information has a significant negative influence on individuals as well as on the whole society. Especially in the current COVID-19 era, we witness an unprecedented growth of medical misinformation. To help tackle this problem with machine learning approaches, we are publishing a feature-rich dataset of approx. 317k medical news/blog articles and 3.5k fact-checked claims. It also contains 573 manual and more than 51k predicted labels mapping the claims to the articles. They represent claim presence, i.e., whether a claim is contained in the given article, and article stance towards the claim. We provide several baselines for these two tasks and evaluate them on the manually labelled part of the dataset. The dataset enables a number of additional tasks related to medical misinformation, such as misinformation characterization studies or studies of misinformation diffusion between sources.

CCS CONCEPTS

• **Information systems** → *Web mining*; *Document representation*; • **Computing methodologies** → **Natural language processing**; **Machine learning**.

KEYWORDS

medical misinformation, dataset, fact-checking, Monant platform

*Also with Kempelen Institute of Intelligent Technologies.
†Also with slovak.AI.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGIR '22, July 11–15, 2022, Madrid, Spain
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-X/18/06.
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Ivan Srba, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION AND RELATED WORKS

False information on the Web has been a widely researched phenomenon in computer science for the past few years, as evidenced by many recent surveys, e.g., [1, 10, 21, 29, 35–37]. The main focus was – until recently – on political fake news; however, with the arrival of COVID-19 pandemic, it shifted towards the medical domain. We are witnessing an *infodemic* – a surge of new misinformation[1] related to the COVID-19 disease, such as "Drinking bleach or pure alcohol can cure the coronavirus infections"[2], "5G installations would be spreading the virus"[3], or that the "Virus was engineered in clandestine US biological laboratories in Ukraine"[4]. The consequences are quite alarming, since there are many cases when people refuse a scientifically-proven medical treatment or vaccination; or take substances that are non-functional or even dangerous to their health. Misinformative claims and disinformation campaigns can be – according to the World Health Organization, United Nations and other international organizations – "harmful to people's physical and mental health; increase stigmatization; threaten precious health gains; and lead to poor observance of public health measures, thus

[1] We use the term *misinformation* to describe false or misleading information that is spread regardless of an intention to deceive, in contrast to *disinformation*, which refers specifically to false information created and spread deliberately.
[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters
[3] https://ec.europa.eu/info/live-work-travel-eu/health/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en
[4] https://euvsdisinfo.eu/eeas-special-report-update-short-assessment-of-narratives-and-disinformation-around-the-covid19-pandemic-updated-23-april-18-may/

reducing their effectiveness and endangering countries' ability to stop the pandemic"[5].

Motivated by significant negative consequences, a number of approaches based on information retrieval and machine learning have been proposed to counter false information. Although rarely, some specific research addressing medical misinformation exists. For example, the characteristics of medical misinformation were examined in [6]. Apart from medical-specific approaches, general state-of-the-art solutions for addressing false information are commonly used in the medical domain as well.

Countering false information is, without any doubts, a challenging task and existing solutions (either general or medical-specific ones) still pose a number of shortcomings. We emphasize particularly two prevalent open problems.

At first, the majority of existing works addressing the task of misinformation detection rely on indirect features derived from content style and context (e.g., text style or users' reactions, references of the sources, etc.). This kind of approach has many advantages, e.g., it allows to detect new false narratives early (since new narratives typically share the similar characteristics with previous cases of false information). However, at the same time, these indirect features *do not consider the actual content veracity*. Existing methods also usually provide only a *limited binary prediction* (i.e., a news article/blog is/is not fake news), which, in combination with indirect features, may not be sufficient for an *explainable detection*. In addition, methods utilizing such features are prone to *suffer from domain shift* (when domain characteristics change) and may be vulnerable to adversarial attacks. This will become even worse as machine generated texts, or texts created partially by a machine and partially by a human, become more prevalent thanks to the availability of advanced language models such as GPT-3 [2]. These shortcomings of current misinformation detection methods are even more eminent in the medical domain that (from its inherent characteristics) requires accurate, easily explainable and robust approaches for misinformation detection.

Secondly, another significant open problem preventing advances in the area of misinformation detection is the *lack of suitable content-rich and benchmark datasets*. The existing datasets contain various forms of misinformative content (e.g., social media posts or news articles). Among them, datasets providing news articles and blogs are less common. In the medical domain specifically, a significant amount of medical misinformation is spread by news articles and blogs (72% of adult internet users search online for health-related issues [7], what commonly directs them to various reliable or unreliable portals publishing this form of the content). In addition, the existing works utilize either expertly annotated, but very small datasets; or larger datasets annotated only by some simple heuristics (e.g., the veracity of articles is determined by the credibility of their source). However, the simple heuristics do not necessarily capture the real veracity of the articles (e.g., articles published in reliable sources may sometimes contain misinformative content and vice versa) and therefore should be used only as weak labels. This kind of simplification is especially undesirable in the case of medical misinformation.

Substantial research and industry efforts in combating false information on the Web resulted also in emergence of various datasets [27]. Nevertheless, we see a number of issues with finding a suitable dataset for comparing the performance of different detection methods on the same data.

First, the ambiguous terminology translates into a variety of different types of datasets and does not help the researchers to locate the appropriate ones. For example, the term *fake news* sometimes refers to a social media post, sometimes to a false claim and sometimes to a news article. Majority of existing datasets contain social media posts (from Twitter [19], Facebook[6], Reddit [20]) or claims (from fact-checking sites [31, 32]), while news articles and blogs remain in the minority.

In the case of datasets containing news articles, they are often labeled based on some simple heuristics, such as the credibility of their source[7] [12–14]. Understandably, such labels are easier to get, e.g., from OpenSources[8], but they also introduce a significant noise, as they may not sufficiently capture the veracity of articles (for example, misinformative articles can appear at reliable sources). Contrary to that, datasets with manually created labels reflecting the actual content veracity remain small and often not fully annotated (e.g., just by its title [34]). Yet overall, the availability of datasets gradually improves (e.g., FakeNewsNet [28] or Deception Detection Fake News [22]).

The lack of suitable datasets is even more compelling in the medical domain or for the purpose of claim-based detection. Kinsora et al. [16] presented a labeled dataset of misinformative and non-misinformative comments developed over posted questions and comments on a health discussion forum. Ghenai and Mejova [9] created a medical dataset of 139 unproven cancer treatments, which they used to search for tweets discussing them and identify users prone to propagate such misinformation. FakeHealth dataset introduced in [5] contains expertly annotated news stories published at HealthNewsReview.org[9] together with their social engagements on Twitter, but it does not map to a list of fact-checked claims. In [33], the authors created a large manually-annotated dataset (covering different domains). They mapped fact-checking articles to relevant documents containing the fact-checked claims along with stance of the documents. Unfortunately, this dataset is not public.

Recently, two datasets specifically addressing COVID-19 misinformation were published. Both were created by means of fact-checking articles which debunk COVID-19 claims. In [26], authors extracted 5 182 fact-checking articles circulated in 105 countries from 92 fact-checkers. The dataset, however, does not contain information about original news articles spreading the fact-checked claims. The second dataset named CoAID [4] provides the mapping of claims to news articles, videos or social media posts, as the fact-checking articles sometimes link the original source of the debunked information. The number of news articles covered by the dataset is, however, quite small.

We can conclude that a publicly available, feature-rich and large enough dataset containing medical news articles/blogs with labeled

[5] https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

[6] https://github.com/BuzzFeedNews/2016-10-facebook-fact-check

[7] https://github.com/several27/FakeNewsCorpus, https://www.kaggle.com/mrisdal/fake-news

[8] https://github.com/BigMcLargeHuge/opensources

[9] https://www.healthnewsreview.org/

mappings between articles and fact-checked claims is still missing. In contrast to the described datasets, our work specifically focuses on creation of the dataset containing news articles/blogs only. Focusing on one content type allows us to extract a rich set of metadata (e.g., articles' authors, sources, categories). To achieve a large set of labeled data, we do not rely on inconsistent links between fact-checking articles and news articles/blogs. Rather, we provide both manual human-created and automatic predicted annotations of claim presence, article stance and final article-claim pair veracities.

In order to address the first of the stated open problems, we argue that it is important to perform actual *fact-checking* of the content to detect medical misinformation. The fact-checking is traditionally done for short claims by evaluating their veracity against some kind of knowledge bases (e.g., scientific articles). In our solution, we employ a different approach – we *perform fact-checking of news articles or blogs against previously fact-checked claims* (this task is sometimes denoted also as *claim matching* [18]). Luckily, many misinformative articles in the medical domain are reusing claims, which have already been expertly fact-checked (and debunked), what makes the use of existing databases of fact-checks (e.g., FullFact.org or Snopes.com) feasible.

In order to address the second stated open problem, we would like to emphasize that utilization of the previously fact-checked claims can also make annotation process less expert-demanding and less time-consuming, since the most expensive part of the annotation process – the fact-checking of the claims – is already done. There are already some approaches relying on previously fact-checked claims, e.g. [24], yet there is still a lack of suitably annotated content-rich datasets containing news articles and fact-checked claims for the medical domain.

In this paper, we are introducing a novel *medical misinformation dataset* containing approx. 317k news articles and blog posts on medical topics from a total of 207 reliable and unreliable sources. The dataset contains full-texts of the articles, their original source URL and other extracted metadata. If a source has a credibility score available (e.g., from Media Bias/Fact Check), it is also included in the form of annotation. Besides the articles, the dataset contains around 3.5k fact-checks and extracted verified medical claims with their unified veracity ratings published by fact-checking organisations such as Snopes or FullFact. Lastly and most importantly, the dataset contains 573 manual and more than 51k predicted labels (annotations) mapping verified claims to the articles; they represent claim presence (i.e., whether a claim is contained in the given article) and article stance (i.e., whether the given article supports or rejects the claim or provides both sides of the argument).

The dataset is primarily intended to be used as a training and evaluation set for machine learning methods for claim presence detection and article stance classification, but it enables a range of other misinformation related tasks, such as misinformation characterisation, analyses of misinformation spreading or classification of source reliability. Its novelty and our main contributions lie in (1) focus on medical news article and blog posts as opposed to social media posts or political discussions; (2) providing multiple modalities (beside full-texts of the articles, there are also images and videos), thus enabling research of multimodal approaches; (3) mapping of the articles to the fact-checked claims (with manual as well as predicted labels); (4) providing source credibility labels for 95% of all articles and other potential sources of weak labels that can be mined from the articles' content and metadata.

The dataset has been collected with our universal and extensible platform Monant [30], which was designed to monitor, detect and mitigate false information. We are publishing a static dump of the Monant data.[10] However, the dataset in Monant is being continuously updated with latest articles and fact-checked claims from medical and other domains (e.g., general news) and also in languages other than English (currently in Slovak and Czech). To access the live version of the dataset, the Monant platform provides an easy-to-use access by the means of a REST API.[11]

2 METHODOLOGY

2.1 Data collection methodology

To create a medical misinformation dataset of news articles/blogs and fact-checked claims (and to continuously obtain new data), we used our research platform Monant [30]. Scraping of the relevant web content and extraction of metadata is implemented by the means of so called *monitors* and *data providers*. Data providers implement the scraping functionality. General parsers (from RSS feeds, WordPress sites, Google Fact Check Tool[12], etc.) as well as custom crawlers and parsers were implemented (e.g., for the fact-checking site Snopes). All data is stored in the unified format in a central data storage. Monitors define which data providers should be used, their scheduling (i.e., frequency of extractions), parameters setup (e.g., a list of RSS feed URLs used as an input to the RSS feed parser), and data provider chaining (if additional data providers should be chained, e.g., when a new article is found).

To compile a list of medical English news sites/blogs, we used expertly-curated lists of reliable and unreliable sites (e.g., Media Bias/Fact Check[13] or OpenSources[14]) and previous related works (e.g., [6]). We added additional sources of unknown credibility that were often referenced (linked) by the sources in the initial list. Next, we checked for each source, whether it still existed and how the data could be obtained from it (e.g., using a WordPress or RSS feed parser or if it required a custom parser). We ended up with a list of 207 medical sources in English; we have a credibility (reliability) score for 70 of them. Examples of reliable (credible) sources include healthline.com, or who.int; examples of sources marked by the listings as unreliable are naturalnews.com, or healthimpactnews.com.

Next, we searched for fact-checking sources that also perform fact-checking of medical claims; we compiled a list of 7 of them (Snopes.com, MetaFact.io, FactCheck.org, Politifact.com, FullFact.org, HealthFeedback.org, and ScienceFeedback.co). Since fact-checked claims are explicitly stated by the fact-checkers, it was possible to automatically extract claims from the fact-checking articles. Additional claims were supplemented from the list of unproven cancer

[10] A sample of the data together with accompanying documentation and analyses in Jupyter notebooks is available at https://github.com/kinit-sk/medical-misinformation-dataset/ The full static dump is available at TODO://Zenodo/link

[11] Instructions how to get access to the Monant API, its detailed documentation, and a guide with examples of API calls and the structure of the data can be found at the GitHub link above.

[12] https://toolbox.google.com/factcheck/explorer

[13] https://mediabiasfactcheck.com/conspiracy/

[14] https://github.com/BigMcLargeHuge/opensources

treatments published by [9]. As veracity ratings can differ between fact-checkers, we unified them into a scale of 6 values: false, mostly false, true, mostly true, mixture and unknown (meaning a veracity of the claim could not be evaluated by a fact-checker or experts' consensus has not been reached yet). The latter originates mostly from the MetaFact.io site, where the experts' evaluations are crowdsourced (in comparison with other fact-checking portals where the fact-checking process is done by one expert only) and the claim veracity is determined only when the evaluation of a sufficient number of experts is available.

2.2 Data labelling methodology

Our aim was to obtain manual ground-truth labels of *claim presence*, i.e., whether a given verified (fact-checked) claim is present in an article, and of *article stance*, i.e., what the stance of the article is towards the matched claim. Our proposed data labelling methodology was inspired by the work of Wang et al. [33]. The labelling is performed in four steps: First, we identify possible article-claim pairs to label. Second, the pairs are distributed to annotators in batches guaranteeing that one pair is given to multiple annotators to minimize possible mistakes in the labelling process and that the same annotator never sees the same pairs multiple times, even across batches. Next, the pairs are annotated by the annotators. Lastly, the labels from all annotators are aggregated into a single claim presence and article stance label for each labelled article-claim pair.

A total number of 28 annotators participated in the labelling process, including the authors of this paper, master students, and other researchers. To prevent potential subjectivity and low-quality labels, a match of at least two annotators had to be achieved for the label to be included into the dataset.

2.2.1 Labels and their aggregation. To annotate claim presence, the annotators could select one of four possible labels:

1. **Yes** – when the annotator can find a part of the article (a sentence or a paragraph) that literally or semantically contains the claim.
2. **Suggestive** – when the article relates to the claim, but the annotator cannot identify any specific part of the article that contains it (e.g., an article discusses the flu vaccine efficacy and suggests that they are ineffective or even harmful by providing anecdotal evidence but never explicitly makes that claim).
3. **No** – when the claim is not present in the article.
4. **Can't tell** – when the annotator cannot, for some reason, choose any of the options above.

When the annotators selected one that claim is present in the article ("Yes" or "Suggestive" labels), they were further asked to label the stance of the article towards the identified claim, by selecting one of four possible labels:

1. **Yes** – when the article supports the claim (directly or indirectly from its context).
2. **No** – when the article contradicts the claim (directly or indirectly from its context).
3. **Both** – when the article presents arguments both *for* and *against* the claim.
4. **Can't tell** – when the annotator cannot, for some reason, choose any of the options above.

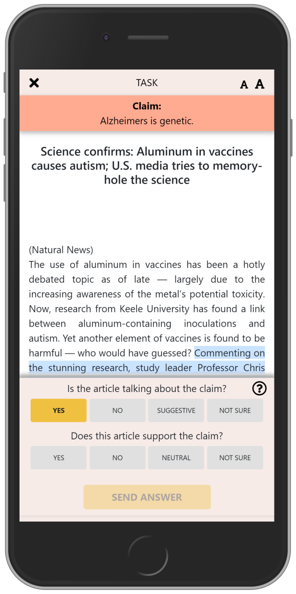
The individual article-claim pair labels are aggregated as follows: First, we filter out all "Can't tell" labels. Next, if any of the remaining claim presence or article stance labels was chosen by two or more annotators for a given article-claim pair, this label is assigned as the final aggregated one. In case of no match in claim presence labels, we lower the requirement by joining the "Yes" and "Suggestive" labels into one and check again for a match. If a match is found, we assign a "Suggestive" label as the final aggregate claim presence label. When there is no match of at least two annotators in claim presence or article stance labels, the article-claim pair is assigned to new annotators to collect more labels. It is also worth noting that article stance labels can be evaluated only when a given claim is present in the article. As a result, there is a lower number of article stance labels compared to the number of claim presence ones.

2.2.2 Selection of article-claim pairs for labelling. The number of all possible article-claim pairs is equal to the number of claims times the number of articles, which is far too many to label. Moreover, most of them would be irrelevant, i.e., it would consist of claims completely unrelated to the articles. To deal with this problem, we select for labelling only a subset of pairs with a high possibility to be relevant. We used two selection methods during our labelling.

At first, we used ElasticSearch to select a subset of the article-claim pairs. More specifically, we used each claim in turn as a query to find matching articles. This returned a large set of articles along with an ElasticSearch score for each article. We kept only articles with the score higher than the $\frac{2}{3}$ of the maximum score, i.e., the score associated with the first matched article. We then shuffled the resulting set of article-claim pairs and sampled two batches, each with 100 random pairs, i.e., 200 pairs in total. We split them among six annotators so that each pair was assigned to three annotators. The annotations were collected using spreadsheets: each annotator was assigned one sheet per batch, with each row describing a single article-claim pair. For each article-claim pair, the annotators were presented with the title of the article, the claim, article URL and the claim URL for information.

However, this selection method led to a significant class imbalance. Out of 197 article-claim pairs, where there was an agreement between the annotators, the claims were labelled as present only in ~10% of cases, which also limited the number of stance annotations. We also observed a relatively large number of "Can't tell" labels which were caused by several claims. These mostly too generic claims (e.g., "There are more doctors") were mistakenly matched with many articles. To mitigate the latter, we manually filtered out these problematic claims from further labelling. The former was addressed by using our proposed claim presence detection baseline (cf. Section 4.1) instead of the simple querying in ElasticSearch.

We also switched from spreadsheets to a custom-made web-based annotation application, suitable also for mobile devices, which enabled us to reach to a wider range of annotators. The application streamlined the annotation process and the article-claim pairs distribution to the annotators. The article-claim pairs were served to annotators until a match of at least two annotators was achieved in the values of claim presence as well as the article stance. Pairs with at least one label but where no consensus had been achieved


✕
TASK
A A
Is the article talking about the claim?
?
YES
NO
SUGGESTIVE
NOT SURE
Does this article support the claim?
YES
NO
NEUTRAL
NOT SURE
SEND ANSWER

Claim:
Alzheimers is genetic.

Science confirms: Aluminum in vaccines causes autism; U.S. media tries to memory-hole the science

(Natural News)
The use of aluminum in vaccines has been a hotly debated topic as of late — largely due to the increasing awareness of the metal’s potential toxicity. Now, research from Keele University has found a link between aluminum-containing inoculations and autism. Yet another element of vaccines is found to be harmful — who would have guessed? Commenting on the stunning research, study leader Professor Chris

Figure 1: The mobile interface of the annotation application used in the later stage of article-claim pairs labelling. The annotators were presented with an article, a claim in the top, highlighted most similar sentence and buttons for selecting claim presence and article stance labels.

yet were served to the annotators with a higher priority to keep the “unfinished” pair labels to a minimum.

Each article-claim pair was presented in the application as shown in Figure 1. The claim was presented at the top, visually separated from the rest of the presented content. Underneath the claim, the title of the article, followed by its formatted body, was presented to the annotator. On the bottom, the annotators were presented with buttons for assigning the claim presence label and—if the annotator chose that the claim is present in the article—also the article stance label. As the articles were long and often dealt with multiple claims at the same time, we used a supportive text highlighting feature: the application highlighted sentences in the article that were most similar to the claim. The similarity was determined by cosine similarity between a sentence embedding representation of the given claim and the sentences of the article. Using this approach, we collected additional 376 article-claim pair labels from 28 annotators.

The collection of labels was also distributed in time. First 439 article-claim pairs (denoted as *Sample 1* in sections below) were annotated in 2019 and early 2020; since this was before the onset of the COVID-19 pandemic, this sample does not contain any claims or articles pertinent to it. The remaining 134 pairs (denoted as *Sample 2* in sections below) were annotated in June 2021, thus capturing also narratives spread in that time.

3 DATASET DESCRIPTION

3.1 Descriptive analysis of raw data

The dataset consists of medical news/blog articles and fact-checked claims in English language. However, the Monant platform, which was used to collect the dataset and makes it accessible via an API endpoint, also collects articles from other domains (e.g., politics or general news) and in other languages (currently mostly in Slovak and Czech). Out of all 885,403 unique news/blog articles from 256 sources, there are 316,832 English medical articles from 207 sources.[15] Out of all extracted 9,633 fact-checking articles from 17 fact-checking sites, there are 3,423 fact-checked medical claims from 7 fact-checking sites.

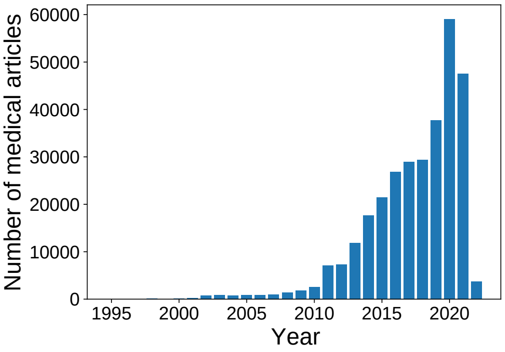


Figure 2: Number of collected medical articles in our dataset according to their publication year.

The dataset provides a rich set of features about each article. Besides article’s URL, title, textual body, and attached multimedia, it also contains information about an article’s authors, category, tags, and references. In addition, we collect (in regular intervals) the users’ feedback on Facebook (i.e., the number of likes or shares) for each news article. In some cases, the posts from the attached discussions are available as well (there are 778,947 discussion posts related to 47,849 articles).

For 70 sources, we have an explicit source reliability (credibility) label (cf. Section 2.1 for more details): 22 sources are considered to be reliable sources, 48 sources are considered to be unreliable. Out of all medical articles, 38.89% were collected from reliable sources, 55.93% from unreliable sources, and only 5.18% articles are from the sources without any reliability label.

Where possible, we collected all articles published by a given source. Consequently, some of the articles in the dataset were published in 1995. Nevertheless, the majority of the collected news articles were published between years 2010–2021 as shown in Figure 2. We can see an increasing trend in the number of medical news articles, with a significant increase in the last three years (the extreme rise in year 2020 can be explained by the onset of the COVID-19 pandemic).

Figure 3 shows the distribution of veracity ratings of the fact-checked medical claims contained in the dataset. 983 were evaluated as false, 60 as mostly-false, 100 as mixture, 39 as mostly-true, and 259 as true. The rating of a significant number of claims (originating mostly from MetaFact.io, cf. Section 2.1) is currently unknown.

[15] The content of this section is based on the dataset’s descriptive analysis published at: https://github.com/kinit-sk/medical-misinformation-dataset/. To make the analysis replicable, it uses a “freeze time” set to February 1, 2022. As a result, only those data, that were present in the Monant platform up to this date, are considered.

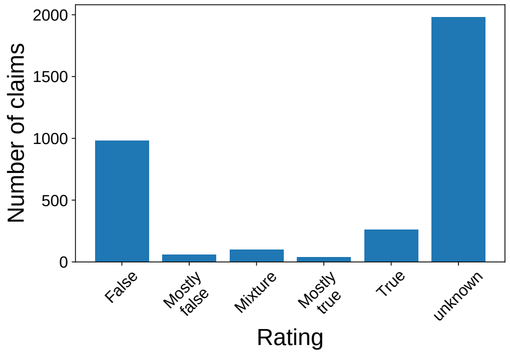


Figure 3: Number of medical claims in our dataset according to their veracity rating.

Table 1: Distribution of claim presence and article stance labels in the dataset. "Supporting" label means that an article supports the matched claim, "Contradicting" that an article contradicts or rejects the claim and "Neutral" that it discusses both sides of the argument (corresponding to "Both" label selected by the annotators).

	Sample 1	*Sample 2*	Overall
Claim presence			
Present (incl. Suggestive)	222 (51%)	101 (75%)	323 (56%)
Not present	217 (49%)	33 (25%)	250 (44%)
Total	439	134	573
Article stance			
Supporting	129 (61%)	74 (75%)	203 (66%)
Contradicting	62 (30%)	24 (24%)	87 (28%)
Neutral	19 (9%)	1 (1%)	20 (6%)
Total	210	99	309

3.2 Labelled dataset

The dataset contains 573 labels from human annotators. It contains 323 positive claim presence labels overall and out of these, there are 309 article stance labels, where there was an agreement between the annotators. The overall distribution of the claim presence and article stance labels is shown in Table 1. It also shows distributions for individual Samples 1 and 2. As we can see, while there is a balance between present and not present labels in *Sample 1* as well as overall, *Sample 2* is skewed towards present labels. As to the article stance, most articles support the matched claims. There is a lack of "Neutral" stance labels in our dataset, i.e., of articles that would present both sides of the argument. This can make it difficult for models trained on this data to correctly classify this stance class.

Besides the labels from human annotators, the dataset also contains approx. 51k *predicted* labels using our proposed baselines. Their analysis is provided in Section 4.3.

3.3 Downstream tasks

The collected dataset can support a range of fact-checking and misinformation-related tasks. Its main intended use is for training and evaluation of machine learning methods for the tasks of *claim presence detection* and *article stance classification*. The former can be considered a claim-oriented document retrieval problem, i.e., given a fact-checked claim, all documents, where it is present, are retrieved, or alternatively as previously fact-checked claims detection, i.e., given an unverified piece of text or claim, all relevant previously fact-checked claims are retrieved [25]. The latter is a classification problem; the aim is to detect stance (position) of the author of an input piece of text towards a specified target [17].

Since the dataset contains articles from a number of reliable and unreliable sources, it could be used for the *misinformation characterisation* task, i.e., for analyses of characteristics of articles (how they are written) similar to [6]: what topics they cover and how these topics evolve over time. The mapping of articles to fact-checked claims provides a straightforward grouping of the articles based on the misinformation they are related to.

The misinformation sources often create inter-connected networks which spread and amplify the false information [15]. Since the dataset contains full-texts of the articles, it supports the task of *misinformation spreading analysis*; it is possible to analyse linking patterns between the sources, search for content that is similar or even taken over from other sources, etc. Misinformation can spread between countries and across languages. Since the data available in Monant via an API endpoint also contain non-English sources (at this moment Slovak and Czech), it can be used to develop and test multilingual methods and analyse spreading patterns from English-language sources to other languages.

Besides text, the dataset contains other modalities, such as images, article and source metadata, etc. These can be all utilised to develop *multimodal detection methods*. Lastly, the dataset can also be used for the task of *source credibility identification* by utilising the existing source credibility labels and extracting a range of credibility indicators from the articles and available metadata, such as polarity of the articles, use of references, use of authors, etc.

3.4 Ethical considerations

The dataset was collected and is published for research purposes only. We collected only publicly available content of news/blog articles. The dataset contains identities of authors of the articles if they were stated in the original source; we left this information, since the presence of an author's name can be a strong credibility indicator. However, we anonymised the identities of the authors of discussion posts included in the dataset.

The main identified ethical issue related to the presented dataset lies in the risk of mislabelling of an article as supporting a false fact-checked claim and, to a lesser extent, in mislabelling an article as not containing a false claim or not supporting it when it actually does. To minimise these risks, we developed our labelling methodology as described in Section 2.2 and require an agreement of at least two independent annotators to assign a claim presence or article stance label to an article. It is also worth noting that we do not label an article as a whole as false or true. Nevertheless, we provide partial article-claim pair veracities based on the combination of claim presence and article stance labels (cf. Section 4.3).

As to the veracity labels of the fact-checked claims and the credibility (reliability) labels of the articles' sources, we take these from the fact-checking sites and external listings such as Media Bias/Fact Check as they are and refer to their methodologies for more details on how they were established.

Lastly, the dataset also contains automatically predicted labels of claim presence and article stance using our baselines described in the next section. These methods have their limitations and work with certain accuracy as reported in this paper. This should be taken into account when interpreting them.

The means for reporting mistakes and possible redress in both the manual as well as the predicted labels are described in the accompanying repository[16].

4 CLAIM PRESENCE AND ARTICLE STANCE BASELINES AND ANALYSIS

Except for sentence tokenization, we did not employ any additional preprocessing on the articles and claims before creating sentence embedding representations. For obtaining the sentence embeddings, we use Universal Sentence Encoder [3] (model version 4). In case of claims, we apply it on the whole statements, while considering them to be single sentences. In case of articles, we apply it on whole titles (regardless of possible multiple sentences there), and also on article bodies, tokenized to sentences by punctuation tokenizer.

4.1 Evaluation of claim presence baselines

The evaluation of claim presence detection method utilizes both *Sample 1* and *Sample 2*, while using only 2 classes: 1) *not-present*, which maps exactly to the original class; and 2) *present*, which is a result of aggregation of the three stance classes, as they indicate that the claim is present in the article.

To evaluate the IRSE method for detection of claims in articles presented in Section **??**, we compared it with baseline methods that stem from commonly used approaches for similar problems and build on best performing approaches analysed in Section **??**. The method itself is a combination of two baseline methods introduced below – one based on information retrieval (IR method), and one based on sentence embedding similarity (SE method).

4.1.1 Information retrieval (IR method). This IR method follows similar steps to the IRSE method presented in Section **??** except that it does not use sentence embeddings to score the similarity of article sentences and claim statements. Instead, only TF-IDF weights of n-grams are used to calculate mapping scores without the additional similarity score. Thus, it is purely based on information retrieval techniques.

4.1.2 Sentence embedding similarity (SE method). This method calculates a mapping score based on sentence embeddings extracted from article sentences and a claim. The mapping score is an average of two similarity comparisons: (1) cosine similarity of article title and a claim, and (2) cosine similarity of 5 most similar article sentences and a claim. The decision that a claim is present in an article is made based on whether the mapping score is above a set threshold.

4.1.3 Performance Comparison For Claim Presence Detection. This evaluation compares the performance of our method (*IRSE*) with two baseline methods proposed by us (*IR* and *SE* method) and an existing method for detecting textual entailment implemented in the AllenNLP libraries [8] (*AllenNLP TE* method).

The *IRSE* method also contains a prefiltering step (Section **??**) in addition to the main algorithm. Our experimentation showed that setting the threshold for the prefiltering step to *0.25* enabled it to discard a large number of potential mappings without affecting the overall performance of the method. Having not impacted the performance, we omit the prefiltering step from further evaluation and comparison of the *IRSE* method with other methods.

The *IR*, *SE*, and *IRSE* methods required a choice of threshold in order to make the claim presence decisions based on their scores. We chose the thresholds such that recall of the methods on the positive class (i.e., claim present in article) would roughly match the recall of the *AllenNLP TE* baseline method (around 0.4). By setting the common level for recall, we could compare the methods working under the same requirement for the proportion of relevant items to be selected. We used the following thresholds for the *IR*, *SE*, and *IRSE* methods respectively: 0.5, 0.5, 0.45.

The performance results of the compared methods are shown in Table 2. The *AllenNLP TE* method achieves the lowest score for both precision and recall on both the positive and negative class. This results in the lowest F1-score from the compared methods. Compared to the other methods we see a relatively larger number of *false positives* predicted by the *AllenNLP TE* method (22.3% false positives compared to 2–6.5% for the other methods).

The *IRSE* method achieved higher precision and recall than the *IR*, and *SE* methods. Figure 4 illustrates a relation between true positive rate and false positive rate of these methods using ROC curve (receiver operating characteristic curve). The *IRSE* method retains lower false positive rate with increasing true positive rate than both baseline methods. Out of the two baseline methods, the *IR* method performs better with lower false positive rate than the *SE* method.

Accuracy of the methods is shown in Table 3 separately for *Sample 1* and *Sample 2* datasets. Although the *IRSE* method retains the highest accuracy, the accuracy drops for all methods except for *AllenNLP TE* in *Sample 2* compared to *Sample 1*. Manual inspection of the errors made by the IRSE method revealed that the decrease cannot be explained by a domain shift due to COVID-19. Majority of the errors were not related to COVID-19 articles and claims. Most commonly, the errors were due to the claim presence method neglecting some information in claims and mapping them to articles that were related but did not discuss that specific case. For instance, for claim "Do Omega-3 fatty acids decrease triglycerides?", we observed results that discussed other effects of Omega-3 fatty acids that did not relate to triglycerides. To handle such cases, a more strict threshold for the method could be used.

4.2 Evaluation of article stance baselines

In case of the stance classification method, we utilize the *Sample 1* as training set, with only 210 pairs, and the *Sample 2* as testing set, with 100 samples. In both cases, we dropped the *not-present* class, as it is not relevant for the method. By removing the *not-present* class, we are left with following distribution of the stance

[16] https://github.com/kinit-sk/medical-misinformation-dataset/

Table 2: Precision, recall and F1-score statistics for the evaluated methods for detection of claim presence in an article calculated on *Sample 1* and *Sample 2* datasets together. The results show that the IRSE method outperforms baseline methods.

	Present			Not-present		
	Precision	Recall	F1-score	Precision	Recall	F1-score
AllenNLP TE (baseline)	0.49	0.38	0.43	0.39	0.5	0.44
SE method (our baseline)	0.79	0.43	0.56	0.55	0.86	0.67
IR method (our baseline)	0.81	0.40	0.53	0.54	0.89	0.67
IRSE method (ours)	**0.91**	**0.45**	**0.6**	**0.58**	**0.95**	**0.72**

Table 3: Accuracy of evaluated claim presence detection methods on the *Sample 1* and *Sample 2* datasets collected and annotated in 2019 and in 2021 respectively. Overall accuracy is calculated across both datasets.

	S1 Acc.	*S2* Acc.	Overall Acc.
AllenNLP TE (baseline)	0.41	0.50	0.43
SE method (our baseline)	0.65	0.53	0.62
IR method (our baseline)	0.66	0.46	0.62
IRSE method (ours)	**0.71**	**0.56**	**0.67**

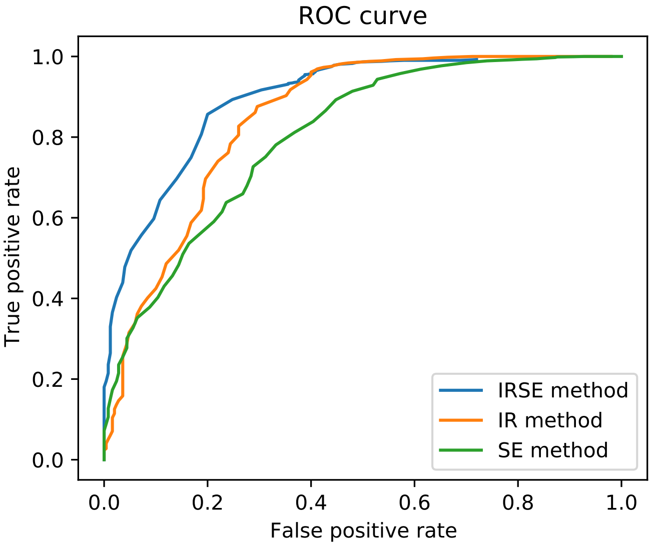


Figure 4: ROC curve showing relation between true positive rate and false positive rate of the evaluated methods. Our *IRSE* outperforms the *SE* and *IR* methods by achieving lower false positive rate at most evaluated true positive rates.

classes in training/testing sets respectively: 61.4%/74.0% *supporting*; 29.5%/25.0% *contradicting* and 9.1%/1.0% *neutral*.

In addition to the manually labeled data from Monant, we also utilize the Fake News Challenge dataset in the evaluation of the stance classification method. Similarly, we drop the class denoting that the article is unrelated to the claim. This leaves us with ~20450 samples, with following distribution: 27.24% *supporting*; 7.5% *contradicting*; and 65.26% *neutral*.

To evaluate the performance of our proposed stance classification method, we use the best models from the FNC as baselines. These models were independently evaluated by the competition and so can be considered as models that perform good on task of determining the stance towards claims. This includes following approaches utilizing both hand-crafted, but also automatically extracted features:

- *Talos*[17] – an ensemble of decision tree and convolutional neural network, where the final decision is obtained by simple 50/50 voting. This approach uses both hand-crafted features in the decision tree and the word embeddings in both the decision tree and the neural network.
- *Athene* [11] – an ensemble of multiple multi-layer perceptrons, where the final decision is obtained by hard voting between them. All models use the same set of hand-crafted features, with only difference being their random initialisation.
- *Athene-ext* [11] – an extension of *Athene* approach, developed after analysis of various models in the challenge, designed to overcome the observed problems. A single stacked LSTM is used with the best subset of the hand-crafted features, as determined by ablation study.
- *UCL* [23] – a simple multi-layer perceptron which uses TF-IDF scores of the claim and the article and the similarity between them.

When evaluating the various models, we apply a slightly different methodology for each of our two datasets. In case of FNC data, we have enough data and therefore we evaluated the models using a test subset of the dataset, as it was originally released for the competition. In case of manually labeled Monant data, we perform 2 evaluations. First, we perform a 5-fold cross-validation on the training set (represented by *Sample 1*) and report the mean performance of the model, which is determined by running the cross-validation 10 times. Then, to determine the possible effect of a concept drift, we evaluate models trained on our training data using the testing set (represented by *Sample 2*).

4.2.1 Performance Comparisons For Article Stance Classification.

When evaluating our proposed approach, we investigate the various variants presented in Section **??**, which build on the best performing approaches from the Section **??**. The best performing models from the investigated approaches are the following:

[17] https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html

- *All Sentences CNN* - a model that uses CNN for obtaining high-level representations. As input, the claim followed by the first 100 sentences of the article, without any detection of relevance for these sentences, is used. The articles with higher number of sentences are clipped and those with lower number of sentences are padded with zero vector. This network is meant for comparison purposes, to determine the effect of sentence relevance detection.
- *Attention LSTM* - a model that uses the LSTM network for obtaining high-level representations for both the claim and the article body. An attention mechanism is applied on the high-level representations to identify the important parts of the article. Another LSTM layer is applied on the output of the attention layer. A dropout with rate of 0.4 is applied to prevent overfitting, followed by a dense layer and a softmax layer for classification.
- *Similarity CNN* - a model that uses CNN for obtaining high-level representations. As input, we use three most similar sentences, along with one previous and one following sentence. We use three different convolutional layers, the output of which is concatenated together. A dropout of rate 0.25 is applied before the convolutional layers and one with rate of 0.5 is applied on the concatenated output of the convolutions, to prevent overfitting. Finally, we apply a dense layer and a softmax layer for classification.

For our two proposed models, the *Similarity CNN* and the *Attention LSTM*, we also employ the transfer learning approach. We first train a general model using the FNC data and fine-tune it using the manually labeled training data afterwards, without freezing their weights.

We first evaluate the performance of our proposed stance classification approach using accuracy metric and compare it with other baselines on both the FNC, as well as the manually labeled data, both on training data using cross-validation and testing data. The results of this evaluation are presented in Table 4. The results show us that the models that utilize the simple hand-crafted features struggle when dealing with a different dataset. This is evident in the *Athene* and its extension. We can presume that the hand-crafted features used are too specific for the FNC data, causing overfitting in the models that use them. On the other hand, the models with automatic feature extraction, which includes UCL baseline model, *Talos*, and our proposed models, show a better performance and better generalization. In addition, we can see that the effect of concept drift is minimal in models with automatic feature extraction, allowing them to retain their accuracy.

Furthermore, we explored the performance gain of our settings of the models for stance classification in the context of misinformation detection presented in Section **??**. The results from the comparison suggest, that the identification of relevant parts of the articles is necessary when dealing with longer articles. In case of FNC data, where the average length of article is ~16 sentences, the performance increase is not as evident. This may be due to the specificity of the shorter articles, which mostly deal with a single claim, and therefore can be considered relevant as a whole for the classification. However, when investigating the articles from Monant, where the average article length is ~55 sentences, the increase in performance is noticeable. In such articles, the extraction of features from the whole article results in a lot of unnecessary information, or noise, which causes problems for the classification.

Table 4: Performance comparison of proposed stance classification models and baselines from the Fake News Challenge. The reported metric is accuracy, calculated from test subset in case of FNC. In case of manually labeled data from Monant, we report both the accuracy on training samples (*Sample 1*), as a mean of 10 runs of 5-fold cross-validation, as well as on testing samples (*Sample 2*) capturing a potential concept drift. Our approach using similarity CNN with transfer learning shows the best performance in comparison with evaluated models and baselines.

Model / Performance on dataset	FNC	*S1*	*S2*
Talos	66.93	42.57	48.00
Athene	67.81	14.36	15.00
Athene-ext	69.00	19.31	10.00
UCL	65.76	37.13	47.00
All Sentences CNN	64.91	40.54	57.00
Attention LSTM	63.19	43.78	40.00
Similarity CNN	65.57	56.76	63.00
Attention LSTM - transfer	64.79	61.83	65.00
Similarity CNN - transfer	**71.86**	**74.23**	**73.00**

When comparing attention mechanism with the similar sentences extracted using cosine similarity, we found out that the attention mechanism struggled to identify relevant parts of the article for classification. The main struggle was stemming from the characteristics of the articles we use. Specifically, the arguments regarding the claims were not contained within most similar sentences (those usually mentioned the claim in its basic form), but by the surrounding sentences instead. Even though the attention mechanism deemed the most similar sentence as the most relevant, its surroundings were not considered as relevant, which misled the decision process.

The use of transfer learning attributed to a significant increase of performance on the data from Monant platform, even though the discrepancy in the distribution of classes across the datasets was significant. When we were training the LSTM networks using the transfer learning, they often broke down and started predicting the most dominant class in the data. Even though the use of attention mechanism helped in this regard in the LSTM based networks, the convolutional neural networks proved to be more stable and reliable for generating good claim and article representations and therefore attained better performance.

4.3 Descriptive analysis of predicted annotations

Besides raw data and manual labels, the introduced dataset[18] also contains the predicted annotations (for claim presence, article stance and article-claim pair veracities). These annotations give interesting

[18] The content of this section is based on the dataset's descriptive analysis published at: https://github.com/kinit-sk/medical-misinformation-dataset/. In order to make the analysis replicable, it uses a "freeze time" set to February 1, 2022. As a result, only those data that were present in the Monant platform up to this date are considered.

insights into the proliferation of medical claims in articles on the Web. In addition, the annotations can be used as (weak) labels for other misinformation detection methods (based on articles' content style and context) while accepting some noise introduced by the methods' performance imperfection. Such annotations are less precise in comparison with experts' annotations, but at the same time are available for a much larger number of articles. They are also more accurate in comparison with commonly used heuristics (e.g., articles' annotations derived from reliability of their sources). This section gives a short overview of the predicted annotations (since the dataset of news articles/blogs and claims is continuously updated, the number of predicted annotations continuously increases as well).

In total, there are 50,953 article-claim mappings labeling the presence of claims in articles as true[19]. Out of 316,832 of articles stored in the platform, 34,850 (11%) articles are mapped to at least one claim. Out of total number of 3,423 medical claims, 1,193 (34.85%) claims are mapped to at least one article. The majority of claim-presence annotations are related to claims from metafact.io (33,950, 66.63%), Fullfact.org (9,418, 18.48%), Healthfeedback.org (4,878, 9.57%), the list of cancer-related claims created in [9] (1,906, 3.74%), and Snopes.com (793, 1.56%).

Out of 50,953 claim stance annotations, 40,168 (78.83%) annotations are labeled as supporting, 2,065 (4.05%) as neutral and 8,720 (17.11%) as contradicting.

The resulting article-claim pair veracity annotations (50,953 in total) show the following distribution: 10,170 (19.96%) article-claim pairs are classified as false, 45 (0.09%) as mostly false, 55 (0.11%) as mixture, 41 (0.08%) as mostly true, 8,814 (17.30%) as true; and finally 31,828 (62.47%) article-claim pairs are labeled as unknown. The high number of article-claim pairs labeled as unknown is caused by the fact that many claims – 1,982 (57.9%) medical claims – have an unknown veracity (see Section **??**).

Out of 34,850 articles mapped to at least one claim, 7,348 (21.08%) are annotated consistently only with true article-claim pairs, 7,680 (22.04) only with false article-claim pairs, and finally, 971 (2.79%) articles contain a mixture of true as well as false article-claim pairs (i.e., some pairs contribute to truthfulness of an article, while others indicate its falseness). The remaining articles are associated only with one or several unknown article-claim pairs that we omit in this consistency evaluation.

Out of 50,953 article-claim pair veracity annotations, 35,094 (68.88%) article-claim pair veracity annotations relate to articles which come from unreliable sources; out of them, 7,571 (21.57%) label article-claim pairs as false and 5,794 (16.51%) as true. 12,815 (25.15%) article-claim pair veracity annotations relate to articles which come from reliable sources; out of them, 2,186 (17.06%) label article-claim pairs as false and 2,549 (19.89%) as true. Although further investigation is needed, we can see that more veracity annotations relate to articles from unreliable sources (even when we consider the distribution of articles from un/reliable sources in our dataset). However, it also suggests that the information on the source's credibility (commonly used as a heuristic to label articles) is not sufficient and the articles need to be assessed by the claims they make.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a labelled dataset of medical articles with mappings to fact-checked claims for training and evaluation of machine learning methods supporting the fact-checking process. Besides providing a static dump of the dataset, we also provide a programmatic access to continuously updated data in our Monant platform. The platform has already been maintained for over 1.5 years, collecting, updating, and annotating new data. The main supported tasks are claim presence detection and articles stance classification, for which we provided manual labels, and which are essential for searching and checking whether a new article contains claims that have already been fact-checked. In addition, the dataset enables a range of other tasks, such as misinformation characterisation studies, studies of misinformation diffusion, source credibility classification, etc. Thus, the dataset can be useful for researchers interested in misinformation, automatized or ML-supported fact-checking as well as for NLP and IR community in general.

We also present results of claim presence and article stance baselines which are used to generate predicted labels mapping articles to fact-checked claims. Their main limitations lie in inconsistent performance of claim presence detection for short versus long claims, limited semantic understanding of matched text and claims, and in class imbalance lowering performance of the stance classification (especially with respect to the *neutral* class). Also, they currently work only for content in English language.

As future work, we plan to extend the dataset with content in other languages and develop multilingual methods of claim presence and article stance. Since the scarcity of manually labeled data will likely remain a problem, we will continue focusing on machine learning approaches that can utilize unlabeled data as is the case of semi-supervised learning. Furthermore, we will seek more efficient ways of navigating the selection of examples to label (active learning), and ways of gathering and exploiting previous experience from other tasks as is the case of transfer and meta-learning.

ACKNOWLEDGMENTS

This work is partially supported by The Ministry of Education, Science, Research and Sport of the Slovak Republic under the Contract No. 0827/2021.

REFERENCES

[1] A. Bondielli and F. Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences* 497 (2019), 38–55. https://doi.org/10.1016/j.ins.2019.05.035

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[3] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. 2018. Universal sentence encoder. arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[4] L. Cui and D. Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885 [cs.SI]

[5] E. Dai, Y. Sun, and S. Wang. 2020. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. arXiv:2002.00837 [cs.SI]

[19] Note that the dataset contains also additional 366 thousand annotations labeling claim-article pairs as not present (i.e., when the mapping score was below the set threshold, but still achieves a meaningful value) – it allows dataset users to use another (lower) threshold if they would like to increase recall at the expense of precision.

[6] S. Dhoju, M. Main Uddin Rony, M. Ashad Kabir, and N. Hassan. 2019. Differences in Health News from Reliable and Unreliable Media. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 981–987. https://doi.org/10.1145/3308560.3316741

[7] Susannah Fox. 2014. The social life of health information. https://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/

[8] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proc. of Workshop for NLP Open Source Software (NLP-OSS)*. Assoc. for Computational Linguistics, Melbourne, Australia, 1–6. https://doi.org/10.18653/v1/W18-2501

[9] A. Ghenai and Y. Mejova. 2018. Fake Cures: User-Centric Modeling of Health Misinformation in Social Media. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 58 (Nov. 2018), 20 pages. https://doi.org/10.1145/3274327

[10] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu. 2020. The Future of False Information Detection on Social Media: New Perspectives and Trends. *ACM Comput. Surv.* 53, 4, Article 68 (July 2020), 36 pages. https://doi.org/10.1145/3393880

[11] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1859–1874. https://www.aclweb.org/anthology/C18-1158

[12] M. Hardalov, I. Koychev, and P. Nakov. 2016. In Search of Credible News. In *Artificial Intelligence: Methodology, Systems, and Applications*, Christo Dichev and Gennady Agre (Eds.). Springer International Publishing, Cham, 172–180.

[13] B. D. Horne and S. Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. arXiv:1703.09398 http://arxiv.org/abs/1703.09398

[14] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar. 2020. BanFakeNews: A Dataset for Detecting Fake News in Bangla. arXiv:2004.08789 [cs.CL]

[15] Andrea Hrckova, Robert Moro, Ivan Srba, and Maria Bielikova. 2021. Quantitative and qualitative analysis of linking patterns of mainstream and partisan online news media in Central Europe. *Online Information Review* ahead-of-print, ahead-of-print (Dec. 2021). https://doi.org/10.1108/OIR-10-2020-0441

[16] A. Kinsora, K. Barron, Q. Mei, and V. G. V. Vydiswaran. 2017. Creating a Labeled Dataset for Medical Misinformation in Health Forums. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 456–461.

[17] Dilek Küçük and Fazli Can. 2021. Stance Detection: A Survey. *Comput. Surveys* 53, 1 (Jan. 2021), 1–37. https://doi.org/10.1145/3369026

[18] E. Lurie and E. Mustafaraj. 2020. 'Highly Partisan' and 'Blatantly Wrong': Analyzing News Publishers' Critiques of Google's Reviewed Claims. In *Conf. for Truth and Trust Online 2020*. https://emmalurie.github.io/docs/TTO_2020.pdf

[19] T. Mitra and E. Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *ICWSM*. 258–267.

[20] K. Nakamura, S. Levy, and W. Y. Wang. 2020. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. arXiv:1911.03854 [cs.CL]

[21] S. B Parikh and P. K Atrey. 2018. Media-rich fake news detection: A survey. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 436–441. https://doi.org/10.1109/MIPR.2018.00093

[22] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://www.aclweb.org/anthology/C18-1287

[23] B. Riedel, I. Augenstein, G. P Spithourakis, and S. Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. arXiv:1707.03264 http://arxiv.org/abs/1707.03264

[24] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3607–3618. https://doi.org/10.18653/v1/2020.acl-main.332

[25] Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates, Vol. 2936. CEUR-WS, Bucharest, Romania, 13. http://ceur-ws.org/Vol-2936/paper-29.pdf

[26] G. K. Shahi and D. 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*. http://workshop-proceedings.icwsm.org/pdf/2020_14.pdf

[27] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3 (2019), 21. https://doi.org/10.1145/3305260

[28] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (2020), 171–188. https://doi.org/10.1089/big.2020.0062 arXiv:https://doi.org/10.1089/big.2020.0062 PMID: 32491943.

[29] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600

[30] I. Srba, R. Moro, J. Simko, J. Sevcech, D. Chuda, P. Navrat, and M. Bielikova. 2019. Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. In *Workshop on Reducing Online Misinformation Exposure – ROME 2019, colocated with SIGIR 2019*. https://rome2019.github.io/papers/Srba_etal_ROME2019.pdf

[31] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[32] W. Y. 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)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[33] X. Wang, C. Yu, S. Baumgartner, and F. Korn. 2018. Relevant Document Discovery for Fact-Checking Articles. In *Companion Proc. of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, 525–533. https://doi.org/10.1145/3184558.3188723

[34] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao. 2020. Weak Supervision for Fake News Detection via Reinforcement Learning. arXiv:1912.12520 [cs.SI]

[35] X. Zhang and A. A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020), 102025.

[36] X. Zhou and R. Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5, Article 109 (sep 2020), 40 pages. https://doi.org/10.1145/3395046

[37] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* 51, 2, Article 32 (Feb. 2018), 36 pages. https://doi.org/10.1145/3161603