A Thesis Proposal
ClaimInspector Framework: A Hybrid Approach to Data Annotation using Fact-Checked Claims and LLMs

Başak Bozkurt
Oxford Internet Institute, University of Oxford
basak.bozkurt@oii.ox.ac.uk

Abstract

This thesis explores the challenges and limitations encountered in automated fact-checking processes, with a specific emphasis on data annotation in the context of misinformation. Despite the widespread presence of misinformation in multiple formats and across various channels, current efforts concentrate narrowly on textual claims sourced mainly from Twitter, resulting in datasets with considerably limited scope. Furthermore, the absence of automated control measures, coupled with the reliance on human annotation, which is very limited, increases the risk of noisy data within these datasets. This thesis proposal examines the existing methods, elucidates their limitations and explores the potential integration of claim detection subtasks and Large Language Models (LLMs) to mitigate these issues. It introduces ClaimInspector, a novel framework designed for a systemic collection of multimodal data from the internet. By implementing this framework, this thesis will propose a dataset comprising fact-checks alongside the corresponding claims made by politicians. Overall, this thesis aims to enhance the accuracy and efficiency of annotation processes, thereby contributing to automated fact-checking efforts.

1 Introduction

The initial step in researching misinformation necessitates a set of criteria to determine the accuracy of a claim. Due to the impracticality of manually scrutinising each piece of information, researchers often rely on the evaluations of fact-checking organisations. They construct datasets that consist of claims that have previously been fact-checked.

However, these datasets also come with a set of limitations. Although a wealth of fact-checking resources exists to document the infiltration of misinformation across various channels, including political ads, politicians' websites and newspapers, the majority of current efforts concentrate on analysing textual claims from a single source, with Twitter being the predominant platform for claim collection. In addition, due to the methods applied in claim matching and the lack of additional controls, the datasets generated often carry a high risk of containing a considerable amount of noisy data. Efforts have been made to mitigate this issue through human annotation; however, limited resources allow such annotation to be performed on only a limited portion of the data (Kazemi et al., 2022; Shahi et al., 2021a; Vo and Lee, 2020). As a result, all these limitations may pose a risk of reduced efficacy in detecting misinformation, since claim detection and fake news detection models may be trained on this limited – and potentially noisy – subset of data.

Therefore, this thesis proposal is centred on addressing the limitations of this process. Informed by these challenges, the main objective of this thesis is to answer these questions:

- RQ1: What are the limitations of current data annotation methods for identifying misinformation?
- RQ2: How can the use of methods for the detection of previously fact-checked claims mitigate these limitations?
- RQ3: To what extent can LLMs be utilised in claim matching during data annotation to address these limitations?

This thesis considers the multimodality of misinformation across various channels. It aims to refine the matching process by drawing on automated fact-checking literature and seeks to establish a more efficient annotation process by incorporating LLMs into the annotation workflow.

By improving this process, this study seeks to not only contribute to the automated fact-checking process, but also to provide support to fact-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 215–224
March 21-22, 2024 ©2024 Association for Computational Linguistics

checkers. Manual fact-checking demands both rigorous attention to detail and a significant investment of time. In this regard, identifying claims that have previously been fact-checked can offer a substantial time-saving advantage for fact-checkers, as it eliminates the need for the redundant verification of claims that have already undergone scrutiny (Shaar et al., 2020; Shaar et al., 2022). Moreover, it can enable swift intervention, which can limit the dissemination of false claims (Nakov et al., 2021).

The remainder of this paper is organised as follows: Section 2 introduces the related work and discusses the limitations. Following this, Section 3 provides information on the proposed method and describes the ClaimInspector framework. Section 4 details a case study that applies the framework for building a dataset of claims made by politicians. Section 5 presents a preliminary plan for experiments. Lastly, Section 6 states the conclusion.

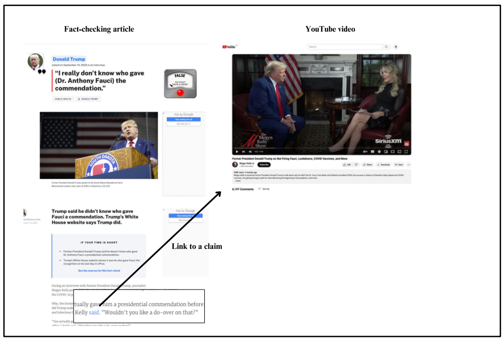
2 Related Work

In this section, I review the literature on data annotation and claim detection and discuss the limitations.

Researchers divide the fact-checking pipeline into four main subtasks: (1) the assessment of checkworthiness, (2) the detection of previously fact-checked claims, (3) the retrieval of evidence and (4) the verification of the factuality of the claim (Shaar et al., 2020). Data annotation, while not listed among these subtasks, can be considered a preliminary task (0). This foundational step is crucial, as it involves labelling data, which supports both the preparation of data for the entire fact-checking pipeline and the training of algorithms. Claim detection, another integral part of this sequence, is closely linked to the assessment of checkworthiness and the detection of previously fact-checked claims.

2.1 Data Annotation

The scarcity of annotated datasets for training and benchmarking has constituted a substantial obstacle in NLP research (Chapman et al., 2011). Recruiting an annotator with specialised expertise is financially expensive, and providing the necessary training to non-experts is time-consuming (Shahi and Majchrzak, 2022). This challenge is particularly pronounced in areas such as misinformation research, where domain-specific knowledge and a deep understanding of context are essential. For instance, when annotating information related to COVID-19, proficiency in medical terminology and scientific context is required.



Donald Trump

"I really don't know who gave (Dr. Anthony Fauci) the commendation."

Trump said he didn't know who gave Fauci a commendation. Trump's White House website says Trump did.

IF YOUR TIME IS SHORT

- Former President Donald Trump said he doesn't know who gave Dr. Anthony Fauci a presidential commendation.
- Trump's White House website shows it was he who gave Fauci the recognition on his last day in office.

See the sources for this fact-check

tually gave him a presidential commendation before Kelly said. "Wouldn't you like a do-over on that?"

Former President Donald Trump on Not Firing Fauci, Lockdowns, COVID Vaccines, and More

Figure 1: The overview of the extraction of URLs from a fact-checking article.

To overcome these challenges, researchers have directed their focus towards domain-specific information sources, which manually verify each claim. Fact-checking websites, in particular, have proven invaluable for large-scale annotation tasks. Within a fact-checking article, fact-checkers typically cite the source of a claim, providing links to news articles, video platforms or campaign websites. Researchers extract these links and assign labels, such as true, partially false or false, based on the verdict provided within the fact-checking article. Figure 1 illustrates the annotation of a claim source (e.g. YouTube video) using the labels given in the fact-checking article (e.g. PolitiFact article).

This approach has seen widespread application across various domains, particularly when retrieving claims made on social media platforms (e.g. Kazemi et al., 2022; Shahi et al., 2021). The AMUSED framework (2022) thoroughly details the stages of this approach for claims made on social media platforms. These stages include searching for anchor tags <a>, which indicate hyperlinks in fact-checking articles. Subsequently, hyperlinks are filtered to identify those leading to social media posts. Following this, corresponding social media data is collected and labelled based on the ruling assigned to the news articles by a fact-checker. The final stage includes human annotation to verify the assigned label.

In particular, studies have adopted the AMUSED framework to extract claim URLs. However, extracting claim URLs is not a straightforward task. There were some efforts to make fact-checking websites structured in order to obtain data, such as creating a JSON format to use the ClaimReview-type specified by Schema.org (RAND, 2015). However, metadata is not always complete for claims from those websites (Shahi et al., 2021b; Quelle et al., 2023). In addition, fact-checkers often present the source of a claim along with various links that support their judgement on the claim. Therefore, in most cases, it is difficult to pinpoint the exact location of a source URL among the others. As the AMUSED framework searches for all anchor tags leading to social media platforms, it may fetch unrelated URLs, potentially leading to mismatches. For instance, a fact-checking article may refer to a subsequent tweet debunking the misinformation or an earlier tweet sharing accurate information that was later repurposed for spreading misinformation (Shahi et al., 2021a).

Another approach in data annotation (e.g. Vo and Lee, 2020) is to search for links to fact-checking articles among responses to social media posts. If a fact-checking link is found, then a pair of a social media post and its corresponding fact-checking article link matched. This approach operates under the assumption that these links signify fact-checking interventions relevant to the post being responded to. For instance, if user A responds to user B's tweet by sharing a link from PolitiFact, a researcher detects B's tweet by searching for links that include the PolitiFact hostname among its direct replies. Then, they annotate B's tweet with a fact-checking rating, assuming that the verification of A is relevant to the claim posted by B.

Although this approach has only found a limited application in automated fact-checking research, researchers have widely used this approach in researching the spread of misinformation on social media (e.g. Vosoughi et al., 2018; Bond and Garrett, 2023; Friggeri et al., 2014). However, this approach also has several limitations. First, posts shared on social media that have not yet received a reply containing a fact-check link elude the researcher's scrutiny. The absence of such links does not necessarily indicate the absence of misinformation. For instance, research has shown that partisan communities avoid using fact-checking and, in some cases, they have moderation policies that delete fact-check links automatically (Parekh et al., 2020). This means that researchers are likely to miss these posts in their data. Second, the link shared may be unrelated and did not fact-check the content of the social media post. Moreover, there may be instances where fact-checking articles, despite addressing similar topics, may concentrate on different aspects (Vo and Lee, 2020).

These methods create uncertainty about whether the link extracted from the fact-check article represents the original source disseminating misinformation. This situation underscores the importance of additional checks on claim URL-fact-check pairs. While the AMUSED framework proposes a labelling step by human annotators to ensure that the pairs are matched correctly, studies often perform this task on only a subset of claims, such as randomly selecting 100 pairs (e.g. Kazemi et al., 2022), or do not perform it at all (e.g. Shahi et al., 2021a).

Overall, these limitations raise concerns about the potential noise in datasets. There is a need for solutions that can use more automation. At this juncture, it appears that claim detection methods and LLMs could offer solutions that support human augmentation in addressing these challenges, which are key objectives of this thesis.

Furthermore, this thesis broadens its scope to encompass not only social media content but also news articles and video platforms. Previous studies have primarily focused on claims originating from social media, with a particular emphasis on Twitter. Apart from a small number of studies that explored multimodal claims (e.g. Vo and Lee, 2020; Shahi and Majchrzak, 2022), the majority of these works were predominantly focused on analysing text-based content. This limited focus inevitably results in selection bias, capturing only a fragment of the information landscape. By expanding its scope, this thesis aims to provide a more comprehensive analysis of misinformation, ensuring a thorough examination across diverse media sources.

2.2 Claim Detection

Claim detection is an integral step in the subtasks for assessing checkworthiness and detecting previously fact-checked claims. This study will specifically concentrate on its role in the second subtask. There is no need to focus on checkworthiness here, as claims have already been extracted from fact-checks.

Detecting previously fact-checked claims can be defined as follows: "Given a check-worthy input claim and a set of verified claims, rank the previously verified claims in order of usefulness to fact-check the input claim" (Nakov et al., 2022). Most of the prior works have mainly focused on the retrieval and ranking of fact-checks based on their relevance to a given tweet or a political statement (e.g. Shaar et al., 2020; Nakov et al., 2022; Kazemi et al., 2022). These works measured token similarity and semantic similarity between a given tweet/political statement and previously fact-checked claims. They used classical lexical retrieval models, such as BM25 (Robertson and Zaragoza, 2009), and transformer-based models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019).

Another line of research approached claim detection with a reverse formulation (e.g. Hossain et al., 2020). Given a database of verified claims, they identified social media posts that make similar claims. In addition to using common semantic similarity models for information retrieval, Hossain et al. (2020) detected the stance of tweets, whether the tweets agreed, disagreed or no stance was taken, and demonstrated that most models do not perform well in the agree and disagree classes. However, when they first identified whether the fact-check-tweet pair was relevant using BERTScore (Zhang et al., 2019), and then only relevant pairs were further classified based on their stance using Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019), the model performed well.

Recently, researchers have focused on the use of LLMs in automated fact-checking. LLMs have a high potential to assist in pinpointing portions of documents that reiterate a claim that was previously verified or express a claim with a similar meaning to one that has already been confirmed (Augenstein et al., 2023). A recent study (Choi and Ferrara, 2023) has demonstrated that fine-tuned LLMs can assist in evaluating the textual entailment between social media posts and verified claims. Fine-tuned LLMs (GPT-3.5-Turbo, Llama-13b-chat-hf, Llama-7b-chat-hf) surpassed the performance of pre-trained LLMs in claim detection.

3 Proposed Method

The proposed method consists of two main stages. The first stage, outlined in Section 3.1, involves extracting the source link of a claim from a fact-checking article and verifying its relevance to the fact-check. The second stage concentrates on broadening the dataset's scope by retrieving relevant news articles or video content associated with a verified claim, as elaborated on Section 3.2. Subsequently, Section 3.3 introduces the ClaimInspector framework, providing a summary of the overall process.

3.1 Identifying Original Sources in Fact-checking Articles

Task: This stage is closely related to the works of Shahi et al. (2021a), Shahi and Majchrzak (2022) and Kazemi et al. (2022), which focused on finding existing fact-checks for claims made in social media posts. The objective is to perform this not only for social media posts but for all types of claim sources. This task can be divided into the following two subtasks:

- Original Source Identification: Given the URL of a fact-checking article, return the URLs of the sources that are cited as the origin of the fact-checked claim.

- Stance Detection: For each fact-check and matching source pair, predict whether the fact-check and matching source agree or disagree or whether the matching source takes no stance with respect to the fact-check.

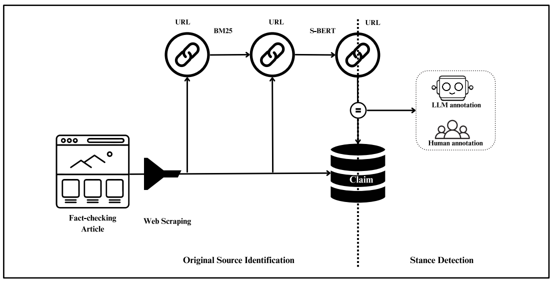


Figure 2: The workflow for identifying original sources in fact-checking articles.

Methods: Similar to the aforementioned studies, I will extract claim URLs mentioned in fact-checking articles. I will use Beautiful Soup (Richardson, n.d.), a Python library for extracting data from HTML, to retrieve the content of the fact-checking articles and prepare a list of source URLs. An illustration of the overall workflow for fetching claim sources cited in the fact-checked articles is shown in Figure 2.

However, to address the limitations discussed in Section 2, this thesis differs from the previous studies in several key aspects. First, this thesis will focus on a diverse range and types of sources cited in fact-checking articles, including, but not limited to, the official websites of politicians, campaign ads and news articles. The prior works restricted their scopes to claims that were made in social media posts, in particular, Twitter posts. In order to mitigate this selection bias, this thesis aims to explore both textual and video content.

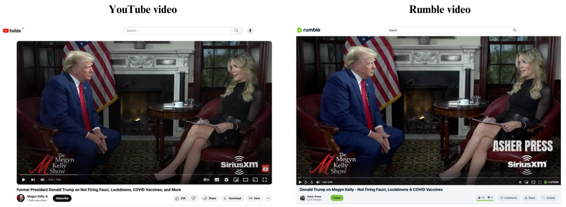
Second, for claim URL-fact-check pair validation, I will assess both token and semantic similarity – common metrics often employed in the claim detection stage. This approach is designed to bolster the robustness of the dataset. As highlighted in Section 2.1, the methods used in the previous works may lead to noisy data. To solve this issue, I will conduct an additional verification step to confirm the relevance of the identified pairs. Similar to the recent research (Choi and Ferrara, 2023), this will involve leveraging the BM25 algorithm and S-BERT to capture both token and semantic similarity between a verified claim by fact-checkers and the source of a claim. I will utilise Beautiful Soup to extract data from the source URLs.

In addition, as a final control step, this stage will include both LLM annotation and human annotation. As I automatically pair the source and fact-check through the references in fact-checking articles, I will conduct an extra step to confirm the relatedness of these pairs. Each pair of claim sources and-fact-checks will be classified into one of the following options: entailment, contradiction and neutral. If it is classified as entailment, then I will assign labels to claim sources based on the label assigned to the fact-checking article. If not, then the data will be excluded from the dataset.

Three human annotators will be recruited through Amazon Mechanical Turk to annotate a randomly chosen sample of 100 pairs. I will employ the majority rule for human annotation to establish ground truth. This approach holds up when there is a high-level of agreement among annotators. While recognising that this may not always hold true (Plank, 2022), in the context of this thesis it is deemed appropriate. As the semantic and token similarity will already have been conducted, pairs that have reached the final phase are presumably related, thereby rendering the task less challenging. I anticipate a high level of consensus among annotators in deciding whether or not a fact-check and claim source matches. Consequently, the majority rule will be the method of choice for human annotation. Following this, similar to the prior work (Choi and Ferrara, 2023), I will compare these human annotations with those from LLMs. Overall, these measures are designed to ensure that the URLs collected are correctly matched with their corresponding fact-checks and labels, thereby enhancing the overall integrity of the dataset.

3.2 Detecting Relevant Claim Sources Containing Previously Fact-checked Claims

This thesis aims to identify content that is similar to the source of a claim cited by fact-checking organisations, acknowledging the circulation of misinformation beyond sources listed in fact-checking articles. Fact-check organisations typically focus on the source where a claim is first stated, often prioritising mainstream outlets. However, the claim may also have been circulated through other mediums. Especially in recent years, misleading information has been disseminated through algorithmically generated or 'junk news' sources that do not adhere to journalistic norms (Burton and Koehorst, 2020). Therefore, this research will detect news articles containing previously fact-checked claims.



Former President Donald Trump on Not Firing Fauci, Lockdowns, COVID Vaccines, and More

Donald Trump on Megyn Kelly - Not Firing Fauci, Lockdowns & COVID Vaccines

Figure 3: While originally published on YouTube, the claim was also spread on Rumble.

In addition, this study will consider the visual contents of fact-checked claims on non-mainstream platforms. For example, a fact-checked claim (e.g. Ramirez Uribe, 2023) presented as a video link on YouTube (e.g. Kelly, 2023) may also circulate on non-mainstream video sharing platforms, such as Rumble (e.g. Asher Press, 2023), a popular platform among conservatives and far-right communities, as shown in Figure 3. Therefore, this study will also search for claims made in video formats across popular non-mainstream video platforms, in particular, two popular alternative social media websites that focus on videos, Rumble and BitChute (Pew Research Center, 2022).

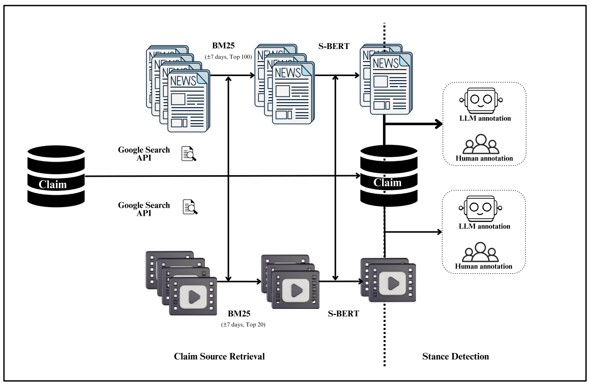


Figure 4: The workflow for detecting relevant claim sources containing previously fact-checked claims.

Task: An illustration of the overall workflow is shown in Figure 4. I formulate the task of detecting related news articles and videos as retrieving relevant content and classifying whether it is related to the fact-check. This task can be summarised into the following subtasks:

- Claim Source Retrieval: Given a fact-check, return a subset of relevant news articles and video sources.

- Stance Detection: For each fact-check and matching text/video source pair, predict whether the fact-check and matching source agree or disagree, or whether the matching source takes no stance with respect to a fact-check.

Methods: The headlines of fact-checking articles are generally written in a way that reflects the actual claim, and so they can be used to get the original news articles. Therefore, I will search the headline of the fact-checking article on Google via Google Search API, and retrieve the top 1,000 results that best match each fact-checked claim within a ± 7-day timeframe from the day the initial claim was made. This approach will allow us to identify content that is most closely related to the claim source.

I will utilise Beautiful Soup to extract data from source URLs. To find claims that are related to fact-checked claims, similarity measures will be calculated using the BM25 algorithm. Similar to the previous work (Choi and Ferrara, 2023), these matching results will be reranked based on the cosine similarity between the sentence-BERT embeddings of each fact-checked claim and the result. This will yield a distinct set of news article-claim pairs with varying degrees of token and semantic similarity. The final step involves selecting the top results from the list. If the type of source is a video, these steps will be conducted for the metadata of matched video content using Python scraper for the BitChute video platform (bumatic, 2022) and Rumble API.

Lastly, to verify whether or not the extracted link authentically represents the source of the claim, this study will leverage LLM and human annotation for a text entailment task similar to the last step outlined in Section 3.1.

3.3 ClaimInspector Framework

This section outlines a comprehensive framework, ClaimInspector, developed for the data annotation process. The ClaimInspector leverages both web scraping techniques and NLP methods to identify, extract and verify claim sources that are mentioned in Section 3.1 and Section 3.2. Figure 5 illustrates the stages in the framework, which are discussed below.

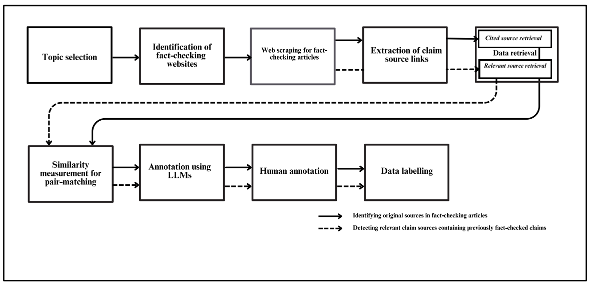

Topic selection
Identification of fact-checking websites
Web scraping for fact-checking articles
Extraction of claim source links
Cited source retrieval
Data retrieval
Relevant source retrieval
Similarity measurement for pair-matching
Annotation using LLMs
Human annotation
Data labelling
Identifying original sources in fact-checking articles
Detecting relevant claim sources containing previously fact-checked claims

Figure 5: The overview of the ClaimInspector framework.

Topic Selection: The initial phase of the framework involves researchers choosing a topic of interest. This choice may concentrate on specific areas of concern, such as COVID-19 or election-related misinformation. Alternatively, researchers may opt for a more comprehensive approach by including several types of misinformation.

Identification of Fact-checking Websites: The second step involves systematically choosing the websites of International Fact-Checking Network (IFCN)-accredited fact-checking organisations. These websites are dedicated to examining statements made in the public domain, such as in news articles, social media posts or public speeches, and assessing their accuracy. The identification process could be based on their areas of expertise, geographical focus and language.

Web Scraping for Fact-checking Articles: The third step includes utilising advanced web scraping techniques to crawl fact-checking articles related to the identified topics.

Extraction of Claim Source Links: Within these fact-checking articles, hyperlinks that lead to the original claim sources are extracted in order to trace the origin of the information.

Data Retrieval: Once the links to claim sources are collected, data from these web pages is retrieved. This process involves downloading the content and metadata for the next step, which can be referred to as "cited source retrieval." In addition, the dataset would not only include the URL sources directly cited in the fact-checking articles, but also consider other sources where the given claim appears. I refer to this process as "relevant source retrieval". To find relevant sources, the fact-checked claim is searched using the Google Search API and retrieved news articles and video contents are collected. The Beautiful Soup library is used to handle the diversity of web page structures. This library enables the parsing of HTML and XML documents, allowing for the extraction of data from a wide array of page styles. Special attention is given to alternative media platforms, such as Rumble and BitChute. Customised extraction techniques are used to handle the unique features of these platforms.

Similarity Measurement for Pair Matching: This step focuses on the measurement of similarity between the claims extracted during the search and the previously fact-checked claims with which they correspond, conducted through a two-pronged approach. First, token-based similarity is evaluated, identifying exact matches in terms and phrases. Subsequently, the analysis extends to semantic similarity, which discerns the underlying meaning beyond mere word usage. Decisions to advance to the subsequent phase are predicated on the similarity scores obtained for the pairs.

The final three stages are dedicated to ensuring the quality of the dataset. Considering the approach of automatically linking sources and fact-checks through references, an additional process to verify the relevance of these matched pairs needs to be implemented.

Annotation Using LLMs: This stage incorporates a verification step through LLM annotation, wherein each claim source is paired with a fact-checking article and categorised as either entailment, contradiction or neutral.

Human Annotation: A random selection of 100 claim sources and fact-check pairs is subject to human annotation to verify the relevance of these matched pairs. This human-in-the-loop approach aids in validating the annotations provided by LLMs.

Data Labelling: The data undergoes a labelling process. If the pair is categorised as entailment, the claim source will inherit the fact-checking article's label. Conversely, any data not classified as entailment will be omitted from the dataset.

4 Implementation: A Case Study on Claims Made by Politicians

While applicable for collecting and annotating data across diverse topics, this thesis will employ the framework to identify claims made by United States (US) politicians. I will scrape both PolitiFact and Snopes, which are IFCN-accredited fact-checking organisations. Gathering data from two fact-checking organisations will give us a more balanced and diverse view of fact-checked claims. PolitiFact primarily concentrates on scrutinising claims associated with politicians, and its sample of politicians is representative of the population of

Field Name	**Description**
Claim ID	A unique identifier assigned to each fact-checked claim.
Politician	The name of the politician making the claim.
Party Affiliation	The political party of the politician making the claim.
Claim Text	The claim that is being fact-checked.
Claim Category	The category of the claim (e.g. election, economy, health).
Claim Source	The origin or source of the claim (e.g. speech, TV interview, tweet).
Claim Link	The URL to the source of the claim.
Fact-check Publishing Date	The date when the fact-checking article is posted.
Fact-check Link	The URL to the fact-checking article providing evidence.
Label	The verdict assigned based on the fact-check (e.g. true, false, mostly false).

Table 1: Description of fields in the dataset.

US politicians (Bucciol, 2018). Snopes examines claims spanning a diverse range of subjects. Following the previous research (Bond and Garrett, 2023), I will collect fact-checks from Snopes' 'Politics' and 'Politicians' categories. This dataset, ClaimInspector: Politicians Edition, will include fact-checked claims made by US politicians, along with links to the claim sources. The fields of the dataset and descriptions are shown in Table 1.

5 Experiments on the ClaimInspector: Politicians Edition

In order to assess the ClaimInspector, I will conduct two sets of experiments. First, I will perform a claim detection task using BM25 and BERT-based models. As the evaluation measure, I will calculate mean reciprocal rank, mean average precision and mean average precision at k for k ∈ {1, 3, 5, 10, 20, 30}. The results from the CLEF-2022 CheckThat! Lab Task 2B will serve as the baseline for this experiment (Nakov et al., 2022).

The second set will focus on the annotation results of pre-trained LLMs. Due to the significant computational resources required for fine-tuning LLMs, this process falls outside the scope of this thesis. I plan to conduct experiments with zero-shot prompting (Kojima et al., 2022) and few-shot prompting (Brown et al., 2020). The performance of LLMs can also be significantly influenced by the prompts given. As such, I will experiment with several elements of the prompts, including the choice of words and the structure of sentences. To evaluate the efficacy of the LLMs, I will employ a range of performance indicators, including macro-level precision, recall and accuracy. The outputs from the models will be compared with benchmark annotations provided by human annotators.

6 Conclusion

This thesis proposal offered an analysis of the current annotation methods and suggested enhancements through additional controls, including similarity measures and LLM-guided annotation. It advocated for broadening the range and type of claim sources beyond mere textual content and social media. Additionally, the proposal underscored the need to identify content akin to sources cited in fact-checking articles, acknowledging the extensive reach of misinformation beyond the sources typically listed in fact-checking articles. To systematically implement these contributions, this thesis proposal introduced the ClaimInspector framework, a novel hybrid approach to data annotation. The proposal outlined the plans for applying this framework by creating a dataset called ClaimInspector: Politicians Edition. Through this dual focus on methodology improvement and dataset creation, the research intends to equip researchers and fact-checkers with reliable tools.

Limitations

This thesis includes a number of limitations that may inform future research. First, it acknowledges the potential for selection bias in the data collection method, given that the scope is limited to claims that were examined by fact-checking organisations. This may result in the exclusion of less circulated and less controversial false claims. Second, studies in this domain suffer from several crawling problems, such as timeouts, unresolvable coding and access restrictions, which may also lead to data loss in this research. Future work can use more advanced web crawling techniques to overcome this constraint and ensure a more comprehensive data collection. Third, it is important to note a limitation related to the monolingual nature of this study, as it exclusively examines claims in the English language. This restriction may overlook the variations present in claims made in other languages. This highlights the need for further research that adopts a multilingual approach. Lastly, since substantial computational resources are necessary for fine-tuning LLMs, fine-tuning has not been included within the current scope of the thesis. Future research could address this gap, potentially enhancing the ClaimInspector framework with the precision that fine-tuned LLMs could offer.

Ethics Statement

This thesis proposal will undergo ethical review by the Central University Research Ethics Committee at the University of Oxford before any research activities begin. I am committed to adhering to ethical guidelines in the use of APIs and web scraping practices. I will ensure compliance with the terms of service and usage policies of all platforms from which data will be collected. Furthermore, I recognise the inherent risks associated with the use of LLMs in annotation, including the propagation of biases and the generation of inconsistent outputs due to their probabilistic nature. I believe that employing human annotation to test the outputs of LLMs can mitigate these risks, thereby enhancing the reliability of the results.

Acknowledgements

I would like to thank the anonymous reviewers for their constructive comments. Additionally, I would like to thank the 2213/A Overseas Graduate Scholarship Program, funded by the Scientific and Technological Research Council of Turkey, for their support.

References

Asher Press. 2023. Donald Trump on Megyn Kelly - Not firing Fauci, lockdowns & COVID vaccines.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. arXiv:2310.05189.

Robert M Bond and R Kelly Garrett. 2023. Engagement with fact-checked posts on Reddit. *PNAS Nexus*, 2(3):1–9.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Alessandro Bucciol. 2018. False claims in politics: Evidence from the US. *Research in Economics*, 72(2):196–210.

bumatic. 2022. Bitchute Scraper.

Anthony G. Burton and Dimitri Koehorst. 2020. Research note: The spread of political misinformation on online subcultural platforms. *Harvard Kennedy School Misinformation Review*.

Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association: JAMIA*, 18(5):540–543.

Eun Cheol Choi and Emilio Ferrara. 2023. Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. arXiv:2310.09223.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2.

Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):101–110.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ashkan Kazemi, Zehua Li, Verónica Peréz-Rosas, Scott A Hale, and Rada Mihalcea. 2022. Matching tweets with applicable fact-checks across languages. In *CEUR Workshop Proceedings*.

Megyn Kelly. 2023. Former President Donald Trump on not firing Fauci, lockdowns, COVID vaccines, and more.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimised BERT pretraining approach. arXiv:1907.11692.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. arXiv:2103.07769.

Preslav Nakov, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the CLEF-2022 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims. In *CEUR Workshop Proceedings*, Bologna, Italy.

Deven Parekh, Drew Margolin, and Derek Ruths. 2020. Comparing audience appreciation to fact-checking across political communities on Reddit. In *12th ACM Conference on Web Science*, pages 144–154, Southampton United Kingdom.

Pew Research Center. 2022. The role of alternative social media in the news and information environment. Technical report.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dorian Quelle, Calvin Cheng, Alexandre Bovet, and Scott A. Hale. 2023. Lost in translation -- multilingual misinformation and its evolution. arXiv:2310.18089.

Maria Ramirez Uribe. 2023. PolitiFact - Trump said he doesn't know who gave Fauci a commendation. Trump's White House website says Trump did.

RAND. 2015. Schema.org Claim Review.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Leonard Richardson. Beautiful Soup documentation.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2021a. An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, 22(100104):1–16.

Gautam Kishore Shahi and Tim A. Majchrzak. 2022. AMUSED: An annotation framework of multimodal social media data. In Filippo Sanfilippo, Ole-Christoffer Granmo, Sule Yildirim Yayilgan, and Imran Sarwar Bajwa, editors, *Intelligent Technologies and Applications*, volume 1616 of *Communications in Computer and Information Science*, pages 287–299. Springer International Publishing, Cham.

Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021b. Overview of the CLEF-2021 CheckThat! Lab: Task 3 on fake news detection. In *CEUR Workshop Proceedings*, Bucharest, Romania.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? Searching for fact-checked information to alleviate the spread of fake news. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. arXiv:2309.13638.