Ghasemizade and Onaolapo *EPJ Data Science* (2024) 13:31
https://doi.org/10.1140/epjds/s13688-024-00470-5

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Developing a hierarchical model for unraveling conspiracy theories

Mohsen Ghasemizade[1*] and Jeremiah Onaolapo[1]

*Correspondence:
mghasemi@uvm.edu
[1]Department of Computer Science, University of Vermont, Burlington, VT, USA

Abstract

A conspiracy theory (CT) suggests covert groups or powerful individuals secretly manipulate events. Not knowing about existing conspiracy theories could make one more likely to believe them, so this work aims to compile a list of CTs shaped as a tree that is as comprehensive as possible. We began with a manually curated 'tree' of CTs from academic papers and Wikipedia. Next, we examined 1769 CT-related articles from four fact-checking websites, focusing on their core content, and used a technique called Keyphrase Extraction to label the documents. This process yielded 769 identified conspiracies, each assigned a label and a family name. The second goal of this project was to detect whether an article is a conspiracy theory, so we built a binary classifier with our labeled dataset. This model uses a transformer-based machine learning technique and is pre-trained on a large corpus called RoBERTa, resulting in an F1 score of 87%. This model helps to identify potential conspiracy theories in new articles. We used a combination of clustering (HDBSCAN) and a dimension reduction technique (UMAP) to assign a label from the tree to these new articles detected as conspiracy theories. We then labeled these groups accordingly to help us match them to the tree. These can lead us to detect new conspiracy theories and expand the tree using computational methods. We successfully generated a tree of conspiracy theories and built a pipeline to detect and categorize conspiracy theories within any text corpora. This pipeline gives us valuable insights through any databases formatted as text.

Keywords: Conspiracy Theory; NLP; Text Classification; BERT; RoBERTa; Tree; Hierarchical

1 Introduction

A conspiracy theory (CT) can have varying definitions in different disciplines. In this paper, we adopt two definitions, first from Douglas et al. [1] that defines a CT as an effort to explain or influence events for the benefit of a specific group or hidden powerful actors behind the scenes. Some elements of a CT may be true, but the truth is often misinterpreted for other purposes. According to Barkun [2], the second definition of a CT, which is very close to the first one, is the belief that a group or organization, composed of individuals, is engaging in secret activities to achieve some evil purpose. It involves the belief that malevolent forces, which are powerful and hidden, have control over human destinies.

Springer

To understand a higher topology of conspiracies, adopted from Barkun [2], first, we differentiate between a conspiracy and a CT. A conspiracy involves actual secret planning and activities, whereas a CT is a speculative belief or framework about a supposed conspiracy without proof of its existence. A CT does not accurately describe an actual conspiracy. Secondly, Barkun [2] distinguishes CTs into three categories: event, systemic, and super conspiracy. Event conspiracies account for particular events, such as the Kennedy murder [3]. Systemic conspiracies have wide-reaching aims to dominate nations or regions, such as those blamed on Jews, Masons, or capitalists. Super conspiracies link multiple conspiracies in a hierarchical order, with an allegedly secret and powerful evil force at the apex. QAnon, discussed later, is an excellent example of this type.

It is also important to note the distinction between misinformation and CTs. Misinformation involves specific false claims that can often be corrected with evidence. Conversely, CTs are broader narratives attempting to explain events in terms of secret plots that persist despite a lack of evidence and mainly appeal to believers [4]. CTs can lead to real-world incidents, such as the Pizzagate shooting [5]. In 2016, a CT emerged that Hillary Clinton, then a presidential candidate in the US election, was involved in a child sex trafficking ring based in the basement of a pizza shop in Washington, DC. This CT led a person who believed in this theory to go to the pizza shop and start firing at people, only to find no evidence of the sex ring. Fortunately, no one was injured in the incident.

However, could this tragedy have been prevented if the shooter had been aware of the CT, even just by its name? Knowing about a CT, even just its title, can help people make better judgments while reading news or articles and prevent harmful incidents. According to Ecker et al. [6], giving people a chance to deliberate can enhance their judgments. For instance, if a quick assessment of a headline is followed by an opportunity to re-evaluate, belief in false news—but not genuine news—is reduced. Encouraging individuals to think like fact-checkers results in greater reliance on previous knowledge rather than shortcuts. This indicates that being aware of the presence of misinformation, such as CTs, may assist in better assessing fresh information.

The purpose of this work is threefold:

1. To develop a scientifically structured 'tree' of CTs using publicly available online data, which will categorize and elucidate the relationships among different CTs for enhanced community understanding.
2. To build a classifier that differentiates CTs from non-CTs by labeling articles and long-text documents.
3. To enable the identification of new CTs computationally from which to expand the tree using any new text dataset.

A tree can be viewed as a roadmap used to identify various CTs, categorize them, and determine their themes and claims based on their branch. trees, as opposed to lists, highlight relationships and hierarchical structures, provide historical context, offer visual representation, and aid in recognizing patterns. This method could help the research community identify and understand CTs more easily. Proverbially, the tree's leaves represent CTs, and the branches represent the main topics to which they belong. This tree-like model enables a user to see how CTs are related by topic without implying that there are direct links from one CT to another.

To build a classifier, we first created a dataset with articles scraped from four fact-checking websites. Articles were labeled using the Natural Language Processing (NLP)

tool named Keyphrase Extraction to extract important phrases from each article, allowing us to label documents by reading only lists of words instead of whole web pages. We later built a binary text classifier using this dataset based on RoBERTa [7].

In this paper, we present a pipeline that performs classification, clustering, labeling, and position tagging to address our third research goal. First, the classifier detects CT documents within any new datasets. Such documents are passed to a clustering algorithm. The user then labels and extracts named entities from each of the clusters, which provides us with insights into the new dataset. This process allows the user to assign labels from known CTs on the tree to new documents and makes the path easier to identify new CTs that do not exist. Our clustering technique combines the UMAP dimension reduction tool [8] and the HDBSCAN clustering tool [9], leveraging their advanced computational efficiencies and robustness to facilitate the classification of documents. To label the clusters, we extract the most common action-object pairs from each cluster and assign each cluster a label. Finally, we use a Named Entity Recognition (NER) tool to perform position tagging and extract entities within each cluster.

The remainder of the paper describes the procedure for compiling the dataset, labeling it with keyphrases, and analyzing labeled data. Subsequently, we evaluate several approaches to constructing a binary classifier for identifying CT articles and outlining how to categorize the detected CT articles using clustering and dimension reduction techniques on a fresh dataset. Afterward, we assign labels to the clusters and extract named entities from each cluster to display how the entire process operates on a fresh dataset.

2 Literature review

CTs have become an increasingly prevalent part of public discourse in recent years, fueled largely by the rise of social media platforms. As false narratives can negatively impact civic discourse and public health decisions, there is a pressing need to develop automated solutions for detecting conspiratorial content at scale. This literature review surveys previous research on communication and the spread of CTs to contextualize our current work on CT detection.

2.1 Communication and spread of CTs

According to Franks et al. [10], there are three primary dimensions involved with evaluating and describing CTs: stickiness, spread, and action. The 'stickiness' of a theory refers to how appealing it is to individuals and how passionately believers espouse its notions. A successful 'spread' involves targeting receptive audiences and preempting criticisms. The term 'action' refers to how believers in CTs organize collectively to take action against those they suspect of being part of a conspiracy. Nefes [11] showed how a prominent Turkish politician exploited these dimensions for his benefit by propagating a CT among supporters to undermine protesters. Previous research has shown that simple, emotionally evocative messages that align with preexisting views are more likely to spread on social media, where malicious and unwitting actors leverage hashtags, bots, and deceptive websites to maximize dissemination [12]. These findings provide important insights into how CTs propagate online conceptually and practically.

2.2 Data collection for CT detection

Several studies have assembled datasets used to analyze CT content across various online sources. Common collection methods involve searching corpora using hashtags, key-

words, or accounts associated with specific known CTs. Gerts et al. [13] used text collected from Twitter to categorize four COVID-19 CTs using random forest classification, providing context for each. Shahsavari et al. [14] developed an automated method to analyze COVID-19 stories from news and other online sources, identifying key elements and relationships to help sort out real news from misleading ones. In the analysis of the Pizzagate CT, Leal et al. [15] utilized social network analysis to identify various roles and positions users of that site took during different time periods. Mahl et al. [16] compared the ten most frequently shared CTs on Twitter and the corresponding communities that unite them. Ahmed et al. [17] conducted a social network and content analysis of Twitter data during a 7-day period when the hashtag #5GCoronavirus was trending. De Zeeuw et al. [18] analyzed the emergence of QAnon on online platforms starting in 2017 and found that the movement had an incubation period on the imageboard website 4chan [19] before migrating to YouTube and Reddit. Pogorelov et al. [20] manually selected more than 10,000 tweets related to common targets for CTs, COVID-19, and 5G, and categorized them into three groups: tweets propagating misinformation specifically about 5G, tweets spreading other CTs, and tweets not involved in the spread of any CT. This allows the identification of discussions related to a particular CT for further classification or network analysis.

2.3 NLP techniques for analyzing CT content

In exploring the realm of CTs and their narratives, several studies have employed NLP, computational analysis, and machine learning techniques to gain deeper insights. The works of Garry et al. [21], Boberg et al. [22], and Sha et al. [23] are pivotal in this regard, focusing on understanding CT communities and narratives through diverse social media platforms, including Gab, Telegram, Facebook, and Twitter. These studies have significantly contributed to grasping the dynamics and reach of CTs, particularly in the context of QAnon followers, alternative news media's output during the early COVID-19 crisis, and Twitter narratives concerning the U.S. response to the pandemic.

Complementing these approaches, advanced machine learning classifiers have been instrumental in detecting CTs on social media. Gerts et al. [13] utilized random forest classification to categorize COVID-19 misinformation in tweets, demonstrating the effectiveness of this technique. Similarly, Peskine et al. [24] leveraged transformer-based models like BERT, achieving state-of-the-art performance in tweet classifications. These advancements highlight the evolving technological prowess in identifying and analyzing CTs.

Furthermore, the research conducted by Savinainen et al. [25], and Tangherlini [26] reveals the power of computational methods in identifying misinformation and understanding narrative structures of CTs. Savinainen et al. focused on language usage patterns related to COVID-19 CTs on social media, while Tangherlini developed an approach to analyzing the narrative structures of widely circulated CTs, including Pizzagate [5] and Bridgegate [27]. These studies underscore the role of computational content analysis in deciphering the underlying patterns and structures of CT narratives.

Moreover, the work by Smith et al. [28] and Faddoul et al. [29] offers insights into the characteristics and influence of CTs. Smith et al. analyzed anti-vaccination discourses on Facebook, employing social network analysis and generative statistical models, whereas Faddoul et al. developed a classifier for identifying conspiratorial content on YouTube, also examining YouTube's recommendation algorithm changes. These contributions are crucial in understanding the propagation and digital footprint of CTs across various platforms.

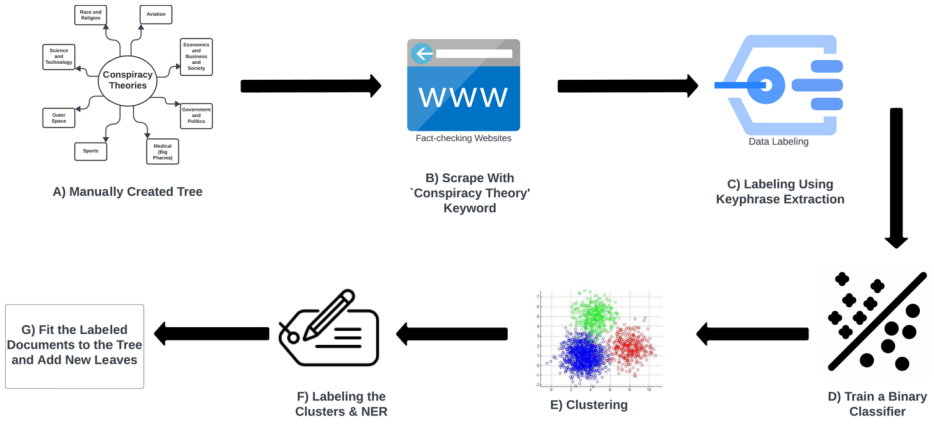


Figure 1 The pipeline

Lastly, Samory et al. [30], and Klein et al. [31] provide a unique perspective on user engagement with CTs. Their analyses of Reddit discussions in the r/conspiracy community and the factors driving individuals to engage with CTs, respectively, add valuable dimensions to our understanding of user interactions and motivations in CT contexts.

Collectively, these studies lay a foundational understanding of CT narratives, their digital dissemination, and detection methods. They collectively inform our project's approach to developing a comprehensive 'tree' of CTs, offering methodological guidance and thematic insights that are instrumental in achieving a nuanced understanding of CTs in the digital era.

2.4 Datasets and performance metrics

Prior work has often focused on limited-scope datasets like tweets or a single CT source. Through the MediaEval challenge of 2020 and 2021, participants were given a labeled dataset of tweets to build a classifier to detect non-CTs, given the topic CT and other CTs. They used the Matthews Correlation Coefficient (MCC) to evaluate their models. MCC measures the difference between predicted and actual values using all four classes of the confusion matrix. The best reported achieved score is for 2021, with an MCC score of 0.775 on the test set using a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) transformer-based model [24]. Our approach involves analyzing large text corpora from multiple fact-checking websites.

Many existing approaches center around single platforms, CT topics, or constrained data sources, which underscores the need for generalizable, scalable CT detection. The pipeline presented in the current work aims to fulfill this need.

3 Methodology

This section explains the methodology used to achieve our research goals and the pipeline we created. Figure 1 depicts the pipeline, and we will explain each part in the following sections.

3.1 (A) Manually created tree

The current authors manually identified 60 CTs from articles found through a Google Scholar search using the keyword 'conspiracy theory'. These articles encompassed CTs

on issues and fields such as climate change [32], online environments and social media [16, 33, 34], and medical domains [35, 36]. After merging the CTs from a Wikipedia article [37], we rearranged them into our tree, placing them in specific branches based on the authors' decisions.

The resulting tree has eight main branches and over 120 CTs, categorized based on the Wikipedia article and our sources. The CTs are placed on the tree according to their historical and relational context. The first layer of the tree, coming from the Wikipedia article, contains the main categories: Aviation, Economics and Business and Society, Race and Religion, Government and Politics, Medical (Big Pharma), Science and Technology, Outer Space, and Sports. The second layer contains CTs that are directly related to their categories or contain CTs in further layers having a common theme, such as '9/11', 'Malcolm X', or 'Deaths'. The third and fourth layers contain CTs that are derived from or associated with one in the previous layer, such as 'Loose Change Films', which is a CT that alleges that the US government orchestrated the 9/11 attacks to justify wars and increase its power [38]. Therefore, this CT is placed on the third layer and linked to the '9/11' CT. The same logic applies to the subsequent layers of the tree. Details of the tree and its branches are outlined in Fig. 2.

The following steps outline the process taken to generate the initial CT tree manually:

1. *Literature Review*
   - Conducted search on Google Scholar using keyword 'conspiracy theory'
   - Identified 60 initial CTs from relevant academic articles
2. *Data Collection*
   - Referenced Wikipedia article on the list of conspiracy theories
   - Combined CTs from literature review and Wikipedia
3. *Tree Structure Creation*
   - Defined 8 top-level branches based on common CT categories
   - Placed initial 60 CTs into branches based on subject matter
4. *Relationship Mapping*
   - Arranged further layers to show the origin/association between CTs
   - Grouped secondary CTs under primary in the hierarchy
5. *Manual Organization*
   - Authors arranged CTs within branches based on judgment
   - Took context like timeline into account for positioning
6. *Documentation*
   - Outlined tree structure with descriptions of branches
   - Captured process and tree overview in the paper
7. *Future Expansion*
   - Pipeline allows detecting new CTs from text corpora
   - New findings can extend tree over time

Now, we describe the 8 main branches of the tree.

Aviation: This small group contains CTs around stories and mysterious incidents that happened to aircraft in the sky. The most popular of these theories is Chemtrails, which refers to the trails airplanes leave in the sky and a false claim that these trails are chemically toxic and have been positioned as biological or chemical agents and dangerous for people and the environment [39]. Figure 3 provides a more detailed view of the family.

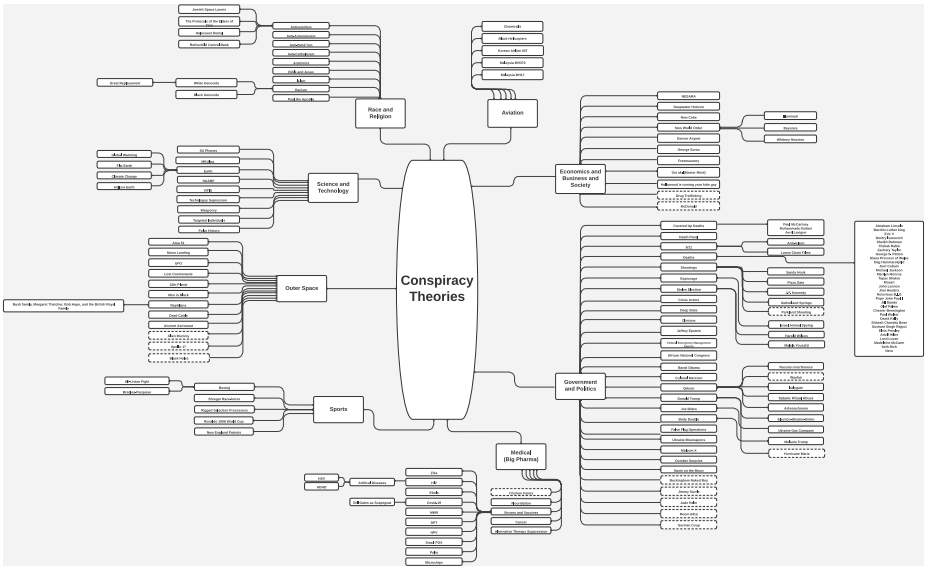
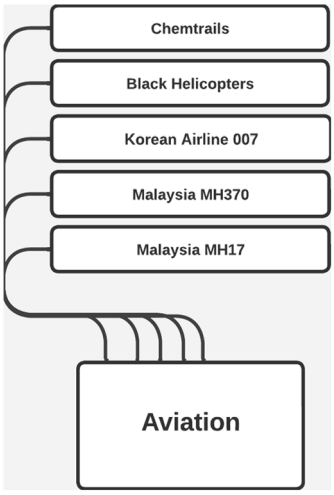


Figure 2 The final tree. Including 14 new ones found in the labeling process, highlighted in dash-lined boxes

Figure 3 Aviation family



Economics, Business, and Society: These groups have relevant topics, subjects, and connections. The resulting branch contains a variety of subjects, of which the most commonly discussed is the New World Order, which asserts that a secretive power with a globalist agenda is conspiring to eventually achieve world domination and rule the world through a one-world government [40]. Figure 4 offers a detailed examination of the family.

Race and Religion: This family includes many topics of racism and anti-religion. The most prominent topic, with ongoing branches, is antisemitism—hostility, prejudice, or discrimination against Jews [41]. A closer look at the family is presented in Fig. 5.

Government and Politics: This family has a large group with 30 branches, in which the discussions mainly revolve around political figures and events related to the American government. However, many discussions are based on information that lacks credible evidence. The two major topics of discussion in this group are the deaths of political figures

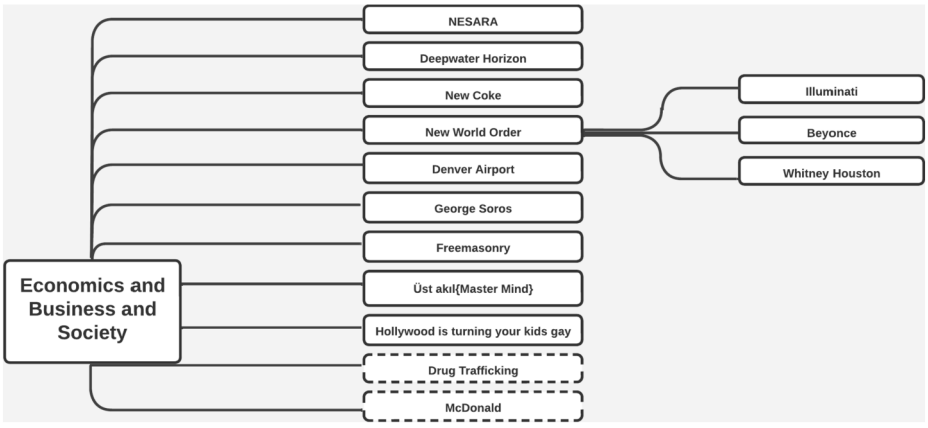


Figure 4 Economic, Business, and Society family. Including 2 new ones found in the labeling process, highlighted in dash-lined boxes

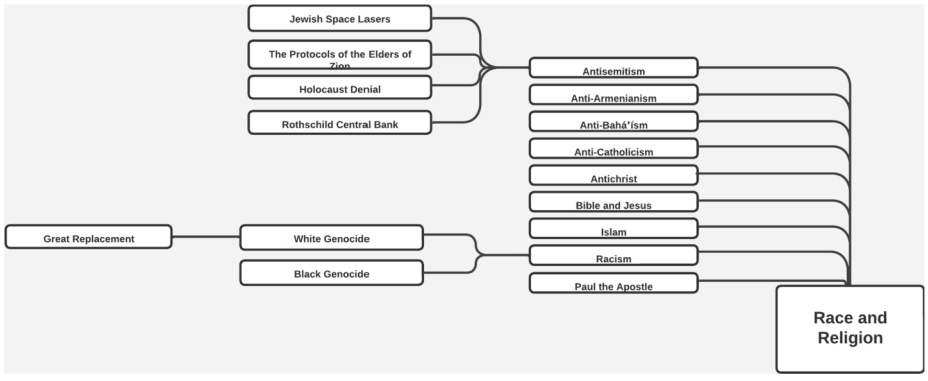


Figure 5 Race and Religion family

and other celebrities, as well as rumors and conspiracies surrounding their deaths, and QAnon. The latter of these originated from anonymous posts by a user named 'Q' on Internet forums in 2017. According to this theory, a cabal of pedophiles and deep-state actors is allegedly plotting against former U.S. President Donald Trump [42]. The family is depicted in greater detail in Fig. 6.

Medical: Medical science and the pharmaceutical industry are crucial in developing vaccines and medications to treat various illnesses that affect people worldwide. However, some cultural and religious beliefs promote false information about these cures, leading people to refuse them. For instance, some people believe that vaccines are satanic or that powerful entities like Bill Gates aim to use them to control the world by implanting microchips in people who receive the COVID-19 vaccine [43]. Figure 7 illustrates the family with enhanced detail.

Science and Technology: The CTs in this tree are interconnected. The only outwardly branching CT, Earth, includes false narratives related to climate change. Believers of such CTs claim that scientists have fabricated the evidence behind global warming and climate change for personal or financial gain [44]. A comprehensive depiction of the family is shown in Fig. 8.

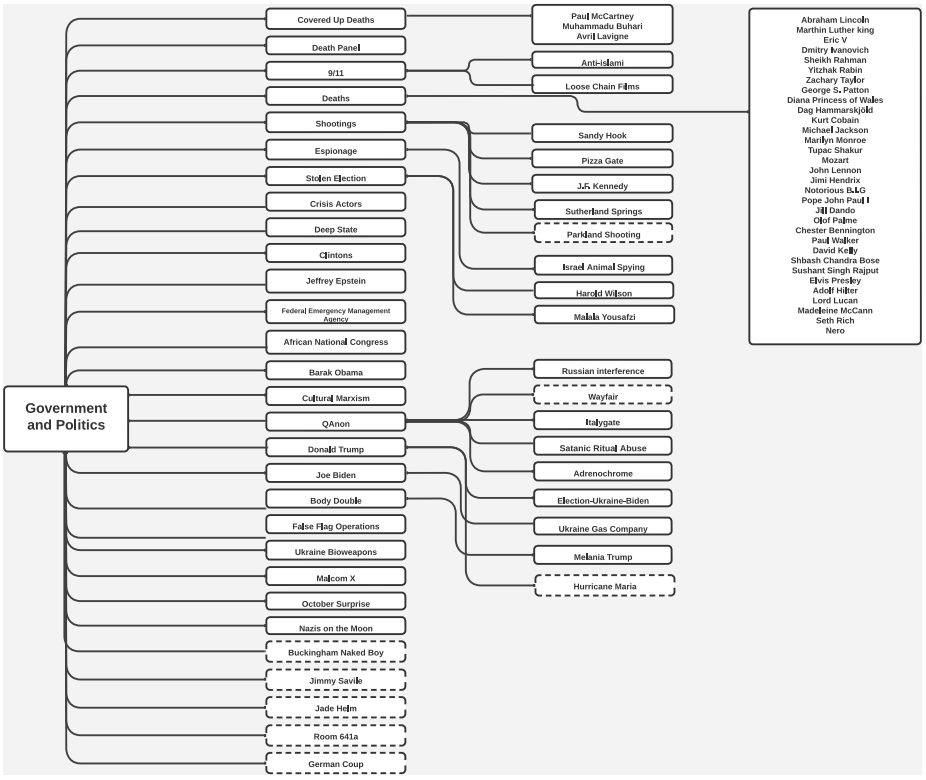


Figure 6 Government and Politics family. Including 8 new ones found in the labeling process, highlighted in dash-lined boxes

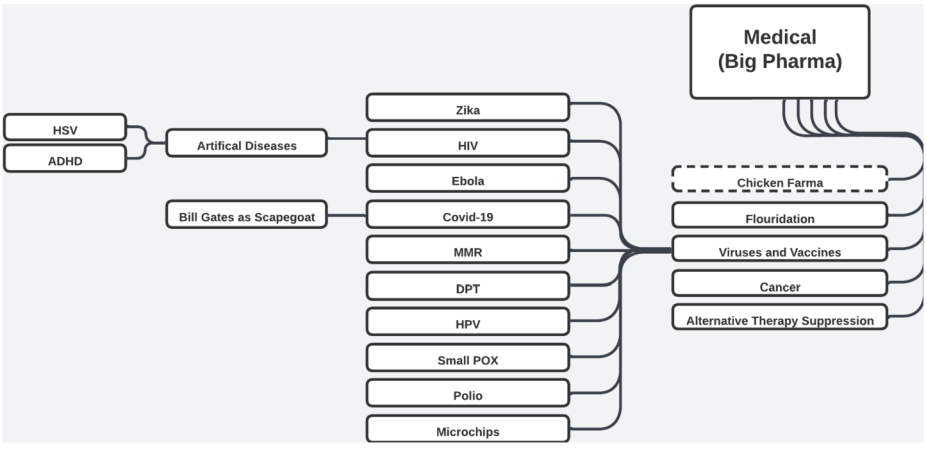


Figure 7 Medical family. Including 1 new one found in the labeling process, highlighted in a dash-lined box

Outer Space: CTs in this family attempt to justify unusual occurrences by associating them with outer space. One of the most well-known theories is about the moon landing. It suggests that NASA, potentially with assistance from others, faked the Apollo program and moon landings [45]. Figure 9 shows a closer view of the family.

Sports: The second smallest family, Sports, contains CTs about events and entities involved in some sports. For example, the mystery of Ronaldo's 1998 World Cup Final. After

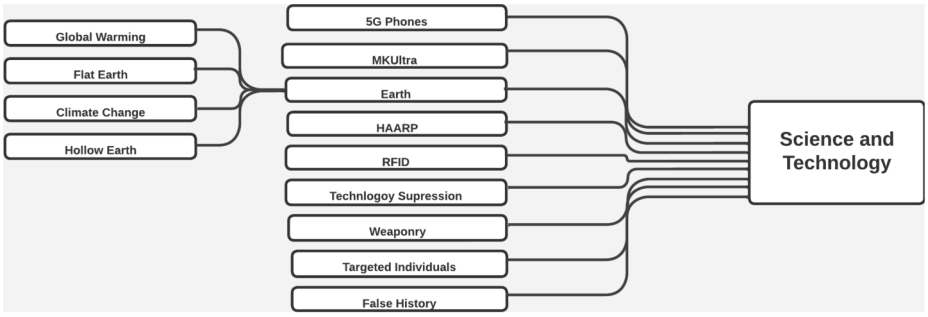


Figure 8 Science and Technology family

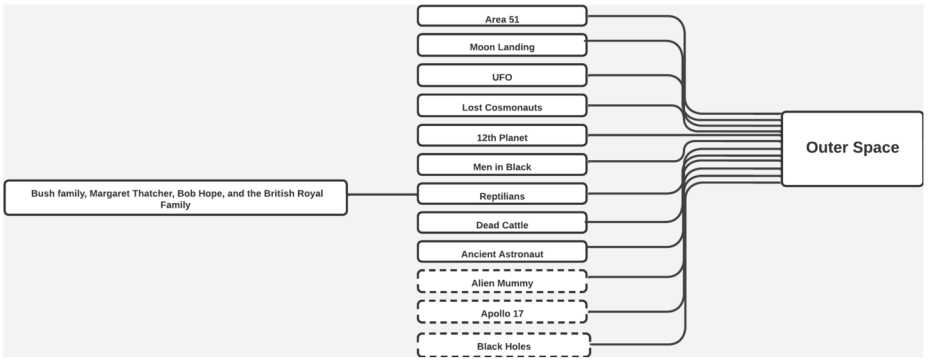


Figure 9 Outer Space family. Including 3 new ones found in the labeling process, highlighted in dash-lined boxes

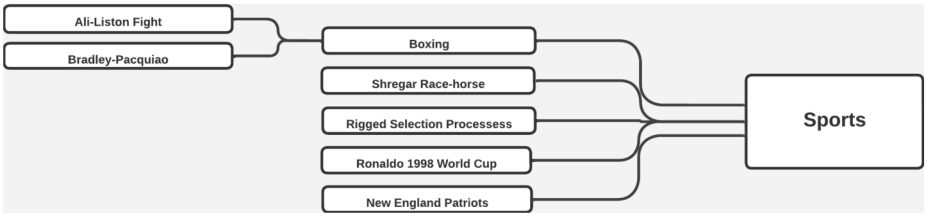


Figure 10 Sports family

suffering a seizure, Ronaldo was removed from the starting line-up but was restored at the last minute [46]. Figure 10 provides a more detailed view of the family.

The largest family is Government and Politics, with almost 60 CTs if all theories related to Deaths are included. Most (although not all) CTs about Deaths involve political figures, which is why they all have been categorized in the Government and Politics. The smallest group belongs to Aviation, with only 5 children, where 3 of them are about airplane crashes.

3.2 (B) Scraping and creation of the original dataset

We selected data from fact-checking websites to create the largest possible tree. Using these websites to identify and label CTs has several advantages over relying on social me-

dia platforms. Firstly, fact-checking websites usually have a rigorous vetting process for the information they publish, which helps to ensure that the information is accurate and reliable. In contrast, social media platforms are often unregulated and may contain substantial false or misleading information. Secondly, fact-checking websites typically provide detailed explanations and evidence to support their claims, which can help increase the labeling process's transparency and credibility. In contrast, social media platforms often provide limited context and may not provide any supporting evidence for claims. Finally, fact-checking websites may have a broader range of topics and sources than social media platforms, which can help to ensure that a wider range of CTs are identified and labeled. Additionally, social media platforms may need to be more expansive in the scope and depth of the discussed CTs.

Four fact-checking websites, Fullfact [47], Factcheck [48], Politifact [49], and Snopes [50], were scraped with the Python tools, Selenium [51] and BeautifulSoup [52], all using the keyword 'conspiracy theory' for the search. Our database contains raw text files from the main body of articles on these websites. In the research studies done earlier to find conspiracies, almost all picked one or a few online social media platforms like Reddit and Twitter as their seed repository [16, 23, 24, 30].

3.2.1 Snopes

This fact-checking website covers various topics, including politics, current events, and social media. It is dedicated to analyzing news stories and videos and labeling them according to their accuracy. To investigate the prevalence of CTs, we scraped 1015 articles related to this topic from this website. The oldest article was published in January 1998.

3.2.2 Politifact

This website focuses mainly on American politics, and it features statements from social media platforms like Facebook and Twitter. Among the articles available on this website, we found a total of 566 related to CTs. These articles were published between December 2017 and July 2022.

3.2.3 Factcheck

The main focus of this website is to scrutinize the accuracy of statements made by major U.S. politicians, including their speeches, debates, TV ads, interviews, and news releases. We used two methods to search for articles related to CTs on this website: one based on relevance and another based on date. We gathered a total of 104 articles, with the oldest one published in July 2009.

3.2.4 Fullfact

The primary goal of this UK-based independent fact-checking website is to enhance accuracy and transparency in public debate. We utilized the Selenium tool to crawl through 84 articles relevant to our keyword, with the oldest dating back to April 2019.

The dataset contains 1769 documents, with 769 labeled as CT and the rest as non-CT.

The NLP tools that we used in this work are all based on BERT [53]. Except for removing unnecessary white spaces, no other text preprocessing (e.g., removing stop words, lemmatization) was conducted because BERT comprehends the context of words in sentences through its bidirectional approach, which adds vital context. BERT uses WordPiece tokenization to break words into subwords, effectively handling out-of-vocabulary words. As

BERT is trained on complete sentences, it understands the relationships between words and sentences. Removing stop words could interfere with this understanding. Additionally, BERT is robust to noise, such as irrelevant words or misspellings, and can learn to ignore or leverage such elements.

3.3 (C) Labeling using keyphrase extraction

To construct an accurate database, we required a labeled dataset. However, labeling 1769 web pages manually would have been prohibitively labor-intensive. To simplify the process, we utilized an NLP tool named Keyphrase Extraction. The purpose is to extract a few phrases that can help readers comprehend the content of a text quickly. Keyphrases are made up of multiple words, unlike keywords, which consist of single words. For instance, in the sentence 'The food was delicious, and the staff was wonderful', the primary topics returned by Keyphrase Extraction would be 'food' and 'wonderful staff'. The Keyphrase Extraction process is performed in two stages: candidate generation, where all possible keywords are identified from the input text, and keyphrase ranking, where the candidate keywords are ranked in order of importance. We applied this method to the entire document collection in the dataset and identified the documents that discussed a CT. This resulted in a labeled dataset that we used subsequently to train our classifier.

We adopted the `keyphrase-extraction-kbir-inspec` tool from the Hugging Face platform [54], a repository of NLP models and datasets. This tool achieved the highest F1 score of 62% on the Inspec dataset [55], which we used as our benchmark. The tool extracts keyphrases, which are concise and informative expressions that capture the main topics of a document. The tool is based on Keyphrase Boundary Infilling with Replacement (KBIR) [56], a pre-trained model that leverages a multi-task learning framework consisting of three components: Masked Language Modeling (MLM), which predicts masked words in a document; Keyphrase Boundary Infilling (KBI), which infers the boundaries of keyphrases; and Keyphrase Replacement Classification (KRC), which determines whether a keyphrase can be replaced by another word. By combining these components, the model learns to identify and generate keyphrases from any document. The model is further fine-tuned on the Inspec dataset, a corpus of 2,000 computer science papers annotated with keyphrases by experts.

The comparatively low F1 score could be due to several factors. These include the intricacy of Keyphrase Extraction, discrepancies between predicted and actual keyphrases, challenges in applying methods across different domains, and the limitations of current NLP techniques. Nevertheless, this score is one of the best on the dataset and indicates a promising strategy [56].

We evaluated the usefulness of our model by analyzing 90 articles and examining the keyphrases generated for them. We checked if the keyphrases contained the CT label needed for our labeling task. We found that over 75% of the generated keyphrases mentioned the exact CT label, the labels we already have on the tree as the leaves. We used context and Google search for the remaining ones to determine their CT label. Below is a summarized sample of one of the articles from Snopes [57], where the extracted keyphrases have the exact CT label in them:

> On 26 March 2016, a YouTube user 'RussianVids' posted a video attempting to prove that the 1969 'moon landing' was a hoax. The video's main piece of 'evidence' is a clip

> from a season four episode of 'Arrested Development' titled 'The B. Team', featuring actors 'Ron Howard' and Jason Bateman.
> In the clip, Howard suggests that the 1969 moon landing was faked on a sound stage, while the 1971 landing was real. The video's narrator claims that Howard is deliberately 'mixing truth with lies' to convince viewers that humans have never been to the moon.
> The choice of 'Ron Howard' in this conspiracy theory is peculiar, given Howard's well-documented interest in space exploration and his involvement in space-related projects. Notably, Howard directed 'Apollo 13' and contributed to a documentary about the Apollo missions called 'In the Shadow of the Moon'. The documentary includes interviews with Apollo astronauts refuting claims that the moon landings were faked.
> This video is among numerous attempts by conspiracy theorists to prove the 1969 moon landing was a hoax, including a debunked claim that 'Stanley Kubrick' admitted to helping fake the moon landings. Despite these theories, the overwhelming consensus among scientists and experts is that the moon landings did occur as reported.

The generated keyphrases, as seen in the last one, mention the CT label 'moon landing', which was previously part of the tree:

> ['Apollo 13', 'Arrested Development', 'Ron Howard', 'Stanley Kubrick', 'landing', 'moon landing']

Here is an example of where the CT label was not mentioned in the keyphrases, but we could still understand the label from the semantics of other keyphrases. This article was published in Fullfact [58], and a summary of it is presented below.

> An Instagram post has claimed that 'World Trade Center 7' collapsed on 9/11 'for no apparent reason'. In reality, the collapse of World Trade Center 7 has been thoroughly investigated by the 'National Institute of Standards and Technology', among other organizations.
> On the day of the attacks, two planes were flown into the 'Twin Towers' (the North and South towers of the World Trade Center complex). The 'North Tower' collapse resulted in debris which ignited fires on at least ten floors of World Trade Center 7. The 'automatic sprinkler system' for some floors of World Trade Center 7 failed due to damage to the 'water lines' caused by the Twin Towers' collapse, allowing some fires to burn out of control.
> The intense heat from these fires caused 'steel support beams' in World Trade Center 7 to expand, leading to the collapse of several floors. This triggered other 'structural failures' within the building, leading to the entire building's collapse. While theories of an explosion causing the collapse were investigated, no evidence was found to support them.
> NIST lead investigator, Shyam Sunder, acknowledged that the collapse of 'tall buildings' due to fire is rare, stating that this is the first known instance of a building over 16 stories tall collapsing primarily due to fires. The Instagram post's claim disregards these findings and feeds into 'conspiracy theories' about the 9/11 attacks, which have been previously debunked.

And the resulting keyphrases:

Table 1 Found conspiracies while labeling the dataset

#	Conspiracy	#	Conspiracy
1	The Jimmy Savile	8	McDonald
2	Black Holes	9	Apollo 17
3	Alien Mummy	10	Buckingham Naked Boy
4	Jade Helm	11	Parkland Shooting
5	Wayfair	12	Room 641a
6	Chicken Farma	13	German Coup
7	Drug Trafficking	14	Hurricane Maria

['National Institute of Standards and Technology', 'North Tower', 'Twin Towers', 'World Trade Center 7', 'automatic sprinkler system', 'conspiracy theories', 'steel support beams', 'structural failures', 'tall buildings', 'towers', 'water lines']

We see phrases like 'Twin Tower', 'World Trade Center 7', and 'conspiracy theories' in the list of keyphrases that clearly allude to the September 11, 2001 attacks and its associated CT label in the tree, which is '9/11'.

Labeling required a significant amount of background knowledge about existing CTs. The authors acquired this knowledge by reading academic papers and the Wikipedia article mentioned earlier while building the tree. Without this prior knowledge, phrases such as 'moon landing' may go unnoticed.

To label the whole dataset, we used keyphrases to indicate whether an article contained a CT. We have a label of 1 for any article with a keyphrase related to any CT in the tree and a 0 for the others. For some articles with a label of 1, we also assigned specific CT names from the tree based on the key phrases and the context. However, there were scenarios where keyphrases pointed out more than one CT label. The authors assigned as many labels as the keyphrases identified in such cases. We found 14 new CTs not previously part of the tree during this labeling process by looking at their keyphrases and a Google search. The newly added CTs in the tree were not previously mentioned in the Wikipedia articles or academic papers reviewed earlier. However, they have now been included in the tree using dash-lined boxes, as presented in Fig. 2 and Table 1. The only difference between these and the rest of the tree is their previous absence from the articles and papers.

3.4 (D) The binary classifier

This work focuses on a dataset of documents from fact-checking websites that share a common theme of analyzing one or more CTs in each article. The classification task aims to identify which articles discuss a conspiracy and, for future research, to determine the specific conspiracy or conspiracies they address. The language of the articles is relatively uniform, as they only describe the CTs without endorsing, advertising, or using other complex language scenarios. Therefore, the classifier performs a simple function of detecting which article is related to a CT.

The classifier operates on the document level, taking one article as a unit of analysis and assigning it a label. The design rationale for this choice, rather than the sentence level, is threefold: first, the dataset consists of long text articles; second, the goal is to quickly ascertain the main topic of the entire document, not just a single sentence; and third, the Large Language Model (LLM) used, has limitations in processing small inputs, such as one sentence, and require more context for better comprehension [59].

Table 2 Precision, Recall, and F1 score for different models. The sorting of the results was based on the F1 score in descending order. The numbers are in percentage

Model	Precision	Recall	F1
RoBERTa	86.8	87.2	**87**
BERT	84	84.5	**84.2**
Random Forest	77	74	**75.5**
SVM	76	74	**75**
Naive Bayes	72	70	**71**
KNN	70	70	**70**
CNN	55.1	58.3	**56.5**

Given a new dataset, we want to find the documents discussing CTs, and to do that, we need a binary classifier. The classifier assigns a value of 1 when it identifies a CT article and 0 when it does not. We divided our dataset into 80% for training and 20% for testing the classifier. In this work, we trained our classifier for the long text documents with a set of machine learning algorithms: Support Vector Machine (SVM), Naive Bayes Classification, Random Forest Classification, K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), BERT, and Robustly Optimized BERT Pretraining Approach (RoBERTa). We performed a grid search for each model to find the best hyperparameters representing the best F1 score. The results for each model are presented in Table 2.

We used different types of inputs for the traditional machine learning models (SVM, Naive Bayes, Random Forest, KNN) and the neural network models (CNN, BERT, RoBERTa). For the former, we extracted TF-IDF features from the text as inputs. For the latter, we used padded and embedded text sequences as inputs for CNN. Using their respective tokenizers, we tokenized and padded text sequences as inputs for BERT and RoBERTa.

TF-IDF is an acronym for the term frequency-inverse document frequency. It is a metric that determines the significance of a word in a document within a group of documents. This is based on how frequently the word appears in the document and how uncommon it is in the collection [60].

When dealing with an unbalanced dataset, there is a difference in the number of one label compared to the other. In our case, where there are more 0s than 1s, accuracy alone is not an informative metric. This is because simply predicting the majority class would result in high accuracy. To overcome this problem, we need better measurements. Precision measures the fraction of correct positive predictions, while recall measures the fraction of actual positives that were correctly predicted. F1 score combines precision and recall by taking their harmonic mean and is a more reliable metric to evaluate the performance of models trained on unbalanced datasets. Based on the results shown in Table 1, both BERT and RoBERTa models significantly outperformed all other models. Additionally, RoBERTa had a slightly higher score, 87%.

Our BERT-based models showed better performances because an LLM model like BERT can detect based on narratives by learning the patterns and relationships between different words and phrases in the text. The model is trained on a large corpus of text data and learns to recognize the features that are most predictive of the target variable, in this case, whether the document is conspiracy-related or not. The model can then use these features to classify new documents based on their similarity to the training data.

When classifying fact-checking articles about CTs, more than a simple topic modeling approach like LDA [61] may be required. This is because LDA is an unsupervised learning

method not designed to capture the complex relationships between different elements of a CT. Although LDA can identify hidden relationships in data and discover topics using a probabilistic framework, it heavily relies on the words used in a corpus and their frequency. Therefore, better choices for accurately classifying fact-checking articles about CTs may exist.

3.5 (E) Clustering

After detecting the documents as CT with the classifier, we must group them by similarity to better understand the corpus. Our preferred grouping algorithm for clustering is the latest tool, Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) [9]. However, before proceeding with HDBSCAN, our text input must be converted into vectors, numerical representations that computers can understand. We use a RoBERTa vectorizer that matches our classifier. The RoBERTa model processes our text and outputs a vector representation from one of its layers. Since each unique word can be considered a separate dimension, each document would have thousands of dimensions, making the clustering process computationally expensive. To mitigate this problem, we reduce the dimensionality of the vectors using the Uniform Manifold Approximation and Projection (UMAP) technique [8].

We use the SRoBERTa model from Reimers et al. [62], to generate 768-dimensional dense vectors. This model modifies the pre-trained BERT network to produce sentence embeddings that can be compared with cosine similarity. It uses a siamese or triplet network structure to encode both sentences into the same space. This makes it faster and more suitable for semantic similarity search and clustering than BERT or RoBERTa, which require both sentences as input and take about 65 hours to find the most similar pair in 10,000 sentences. The proposed model can do the same task in about 5 seconds. We use the RoBERTa-based version of this model, SRoBERTa, that matches our classifier base, and we feed the embeddings to our classifier and UMAP for dimension reduction.

HDBSCAN is a clustering algorithm that differs from others as it can automatically separate and cluster our dataset. It can also detect how many clusters the dataset requires without a specific number of clusters being set initially. HDBSCAN uses a density-based clustering algorithm, particularly useful in cases with denser clusters of conspiracies like COVID-19 and less dense clusters for conspiracies like Ronaldo's 1998 World Cup Final. One important hyperparameter used in this algorithm is *min_cluster_size*, which determines the smallest grouping we consider a cluster. Another parameter, known as *min_samples*, controls the level of conservatism in the clustering. The smaller this value, the fewer points are detected as noise or outliers.

When using UMAP for dimension reduction, there are two key hyperparameters to consider: *n_components* and *n_neighbors*. n_neighbors determines the number of neighboring points used in approximating the manifold structure. A larger value of n_neighbors focuses on the bigger picture, while a smaller value emphasizes local structures. n_components controls the final dimension of the input data.

Bayesian optimization is a method to find the best hyperparameters for a machine learning model. It is better than grid search, which takes too much time and memory, and random search, which may miss some good options. Hyperopt [63] is a Python library that makes Bayesian optimization easier. We want to minimize the cost of this problem because it means how far the points are from their clusters. The lower the cost, the better

the clustering. So, we use the cost as the penalty for Hyperopt to find the best hyperparameters.

Dimensionality reduction and clustering are powerful techniques for analyzing and interpreting data. However, dimensionality reduction can lead to losing important information and selecting the appropriate dimensions can be challenging [64]. Furthermore, when dealing with large datasets, the curse of dimensionality can make computations impractical. Clustering algorithms, on the other hand, can be sensitive to noise and outliers, and the choice of distance metric can have a significant impact on the results. Furthermore, they can get stuck in local minima, which can be problematic when dealing with high-dimensional data. Finally, the curse of dimensionality can make it difficult to determine the optimal clustering of high-dimensional data [64].

3.6 (F) Labeling and named entity recognition

We now move on to the task of labeling the clusters. To achieve this, we are adopting the method outlined by Liu et al. [65]. Essentially, we label each cluster automatically by identifying the most frequently occurring action-object pair in the phrases contained within the cluster. While this method provides us with a good insight into the cluster's context, it may only sometimes generate a precise label.

In this work, a BERT-based NER tool [53] from the Hugging Face framework was utilized to extract named entities from the dataset. This model has been fine-tuned on the standard CoNL-2003 NER dataset [66]. The tool is capable of recognizing four different entity types: Location, Organization, Person, and Miscellaneous. The tool's performance was evaluated using the F1 score, which was reported as 95.1% on the development dataset and 91.3% on the test dataset [67]. Using this tool, we analyze the entities inside each cluster and also the entities inside the original dataset we scraped in the first place.

3.7 (G) Fitting to the tree and adding new leaves

This section addresses the third research objective, which is to extend the existing tree with novel CTs discovered from a new text corpus. We aim to identify and label the documents that contain CTs in the new dataset and to align them with the existing tree or add new branches if necessary. However, this task poses several challenges, especially for the newly found CTs that do not have a corresponding match on the tree. We will discuss these challenges in the following sections and propose methods to extract CT labels from documents and expand the tree accordingly.

4 Discussion and analysis

In this section, we begin by examining and analyzing the initial dataset, which consists of articles scraped from four fact-checking websites that we obtained earlier. Next, we will demonstrate how the pipeline functions on a fresh dataset that we have scraped specifically for this purpose. Finally, we will examine the pipeline's outputs and discuss how it can benefit us.

4.1 The original dataset

During the process of labeling the dataset, we came across 14 previously unknown CTs. These theories were not mentioned in any of the papers or Wikipedia articles we analyzed. However, they are still well known from other sources. The new CTs have been categorized as follows:

1. The Jimmy Savile CT refers to a disinformation campaign accusing British politician Keir Starmer of failing to prosecute the late BBC presenter Jimmy Savile. The CT about this known sex offender originated from far-right circles and was subsequently circulated in mainstream media [68]. This CT has been placed in the Government and Politics family.

2. Black Holes, categorized in the Outer Space family, propagated by a controversial men's rights activist named Roosh V, claims that the first-ever photo of a black hole, an achievement widely hailed by physicists, is almost certainly fake [69].

3. Alien Mummy CT, categorized in the Outer Space family, suggests that a video shows Russian agents in the 1960s discovering the remains of a 13,000-year-old mummified alien in an Egyptian tomb, as part of an operation codenamed Project ISIS, with the footage allegedly obtained from secret KGB archives [70].

4. Jade Helm CT asserts that a 2015 United States military training exercise was actually a cover for a planned hostile military takeover, a plot to impose martial law, confiscate firearms, invade Texas, institute population control, or prepare for an apocalyptic event such as a comet or asteroid striking the Earth [71]. This CT has been placed in the Government and Politics family.

5. Wayfair CT is a groundless claim suggesting that the U.S.-based company, which sells affordable furniture, is involved in child trafficking, with the unfounded allegations originating from the QAnon community and spreading globally [72]. This CT has been placed in the Government and Politics family.

6. Chicken Farma, refutes a CT that feed producers are intentionally altering their products to reduce egg production and inflate prices, pointing out that the real cause of decreased egg production and increased prices is the outbreak of highly pathogenic avian influenza (HPAI), which has led to the depopulation of millions of layer hens [73]. This CT has been placed in the Medical family.

7. Drug Trafficking CT suggested that the CIA and its operatives in the 1980s used crack cocaine—sold via the Los Angeles African–American community—to raise millions to support the agency's clandestine operations in Central America [74]. This CT has been put in the Economics and Business and Society family.

8. McDonald CT falsely claims that human meat was found in the freezers of an Oklahoma City McDonald's meat factory and about 90% of other McDonald's meat factories inspected a hoax that originated from the satirical blog Huzlers.com in 2014 [75]. This CT has been put in the Economics and Business and Society family.

9. Apollo 17 CT alleges that the moon landing was fake because an unidentified reflection in an astronaut's visor during the 1972 mission suggests a 'stagehand' was present, rather than another astronaut in a spacesuit [76]. This CT has been placed in the Outer Space group.

10. Buckingham Naked Boy CT claims that an image shows a naked child escaping from Buckingham Palace is false; the image is from a viral hoax designed to promote the E! TV series 'The Royals' and features an adult man, not a child [77]. This CT has been placed in the Government and Politics family.

11. Parkland Shooting CT posits that the student survivors of the tragedy, who became vocal advocates for gun control, were actually 'crisis actors' paid to propagate a false narrative, a baseless claim that rapidly spread through fringe Internet circles, social media platforms, and even some mainstream media outlets [78]. This CT has been placed in the Government and Politics family under the Shooting branch.

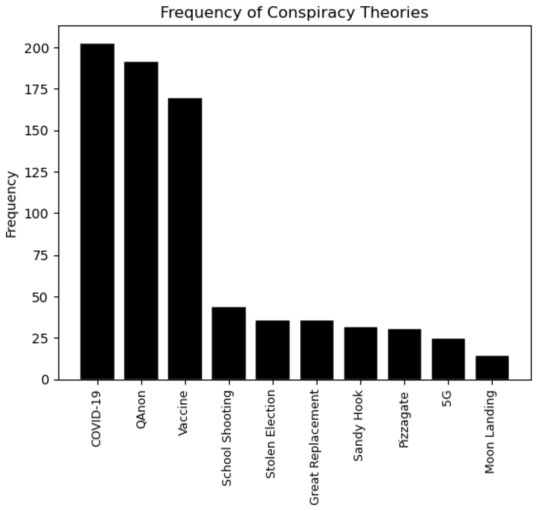


Figure 11 Frequency of CTs

12. Room 641a refers to a secretive room in a building owned by AT&T, suggesting that it was a secret surveillance hub used by the U.S. National Security Agency (NSA) to monitor and analyze Internet data and communication, potentially infringing on privacy rights and laws [79]. This actual conspiracy, which is not a mere CT, has been placed in the Government and Politics group.

13. German Coup plot involved a group associated with the extremist Reichsbürger movement, which rejected the legitimacy of the modern German state and was comprised of 25 individuals, including a special forces officer and an eccentric aristocrat, who were planning an armed insurrection with the intention of installing their government. However, the movement has become politicized and believes in the 'deep state' allegedly holding Germany captive [80]. This CT has been placed in the Government and Politics group.

14. Hurricane Maria CT emerged from Donald Trump contesting the rigorously researched death toll of 3,000 from Hurricane Maria in Puerto Rico, suggesting it was part of a Democratic CT to make him look bad, despite evidence showing that thousands of Americans died due to lack of oxygen, food, water, medical attention, and suicide, as well as from other indirect effects of the hurricane [81]. This CT has been placed in the Government and Politics group under the Donald Trump CT.

The authors used the context of the keyphrases generated for each article, conducted Google searches, and read through multiple articles to determine whether the theories belonged to the tree and, if so, which category they should be placed under. Resources propagating these CTs, or to be more precise, all of the CTs, are often unreliable due to their lack of transparency, questionable methodologies, diverse origins, and tendency to propagate misinformation or 'fake news' without rigorous fact-checking or accountability, and that is how most of the CTs start spreading all over the Internet. Next, we analyze the most frequent CTs in the original dataset.

To determine which CTs are discussed the most, we did a frequency analysis by going through the dataset and seeing how many times each CT is mentioned. It can give us which CTs are the most prevalent.

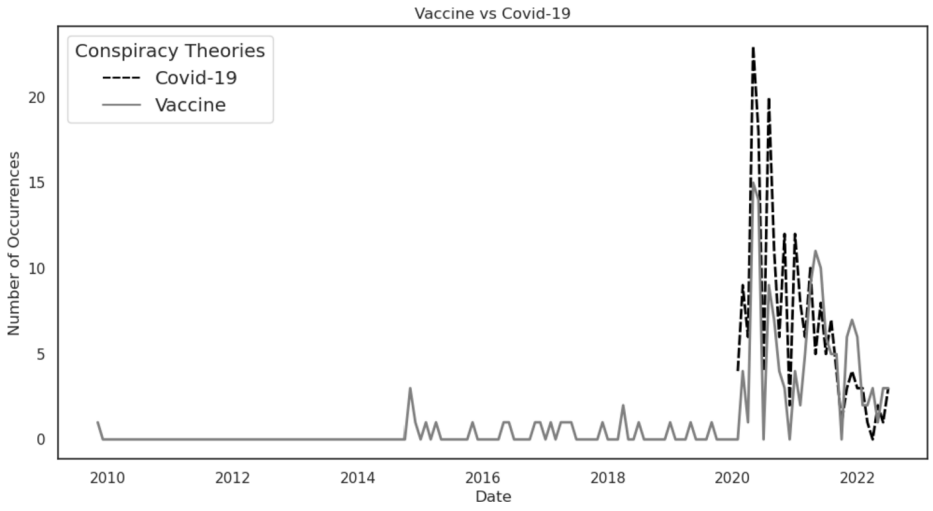


Figure 12 Vaccine and COVID-19 CTs over time

Figure 11 displays the 10 most frequently discussed CTs identified in the dataset, and the first three are COVID-19, QAnon, and Vaccines, with frequencies of 203, 192, and 170, respectively. The Vaccine CT encompasses all conspiracies related to vaccines, including those related to COVID-19 vaccines, which may explain its inclusion in the top three. We further perform a pair temporal analysis for the pairs of COVID-19-QAnon and COVID-19-Vaccine to see if there are any trends and patterns between these 3 most repeated CTs.

Figure 12 displays the frequency of CTs related to Vaccines and COVID-19 across a given time frame. Notably, the trend for Vaccine CTs increased leading up to the emergence of COVID-19 in 2019, and the two exhibited a similar pattern after that. Before the COVID-19 pandemic, Vaccine CTs often revolved around other viruses such as Ebola, HIV, and Polio. However, after the emergence of COVID-19, Vaccine CTs primarily focused on the COVID-19 vaccine.

As shown in Fig. 13, the frequency of QAnon CTs increased after the emergence of COVID-19 in late 2020, which was one year after the initial appearance of COVID-19. According to Morelock et al. [82], there is a connection between QAnon and COVID-19 CTs, as people are finding it difficult to trust the information they receive and are becoming less interested in learning from experts. These ideas are being spread on the Internet, and QAnon took advantage of the COVID-19 situation to increase its popularity.

Now, using our NER tool, we analyze the entities involved in CTs and their spread through social media platforms such as Facebook, Twitter, and YouTube. The table below presents the frequency of these entities, categorized by their types (PER—Person, ORG—Organization, LOC—Location, MISC—Miscellaneous). Table 3 displays the top 25 entities that appeared most frequently in the CT-detected articles in the original dataset.

The data indicates that individuals such as Donald Trump, Joe Biden, Marjorie Taylor Greene, Hillary Clinton, and Alex Jones are often mentioned in association with CTs. This is probably because they play crucial roles in either promoting, debunking, or being the subject of these theories.

The high frequency of organizations such as Facebook, Twitter, and YouTube indicates that these platforms act as significant sources for the spread of CTs. These social media

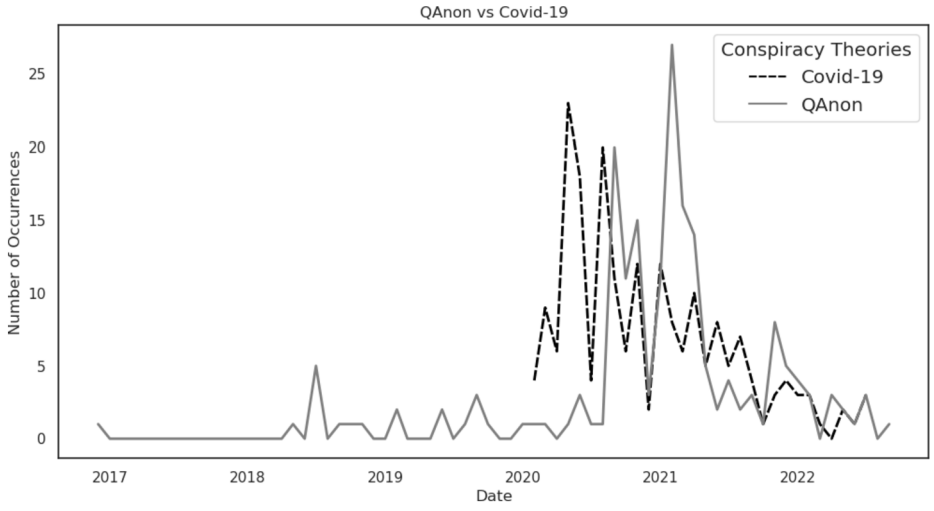


Figure 13 QAnon and COVID-19 CTs over time

Table 3 Top 25 most frequent entities

Entity	Tag	Count	Entity	Tag	Count
Trump	PER	909	Clinton	PER	95
Facebook	ORG	665	Americans	MISC	95
COVID-19	MISC	517	Florida	LOC	89
U. S.	LOC	407	Jones	PER	88
Biden	PER	240	Russian	MISC	87
CDC	ORG	148	WHO	ORG	82
Republican	MISC	113	Twitter	ORG	77
Greene	PER	113	White House	LOC	75
China	LOC	104	Republicans	MISC	72
Capitol	LOC	99	YouTube	ORG	70
Democrats	MISC	97	CNN	ORG	64
FBI	ORG	96	Texas	ORG	62
American	MISC	95			

platforms enable users to share and consume information rapidly, sometimes unverified, contributing to the proliferation of CTs.

Locations like the U.S., China, Capitol, Florida, and Texas also appear frequently, suggesting that CTs often involve geopolitical elements, regional politics, or specific events occurring in these areas.

Miscellaneous entities such as COVID-19, Republican, Russian, American, and Democrats are frequently mentioned, indicating that CTs often involve political ideologies, national identities, or global crises.

4.2 The pipeline

In this section, we subject our Classifying, Clustering, and Labeling & NER parts of the pipeline, parts D, E, and F, shown in Fig. 1, to a test with a new dataset. We will demonstrate how it works and how it can be useful. Our first step is to apply the classifier to the dataset to distinguish the CTs from the non-CTs. Next, we will apply clustering to the CTs to determine the necessary number of clusters and their corresponding labels. We then

present the first ten clusters and their labels for both methods. And lastly, we show the top 10 entities found in the CT-detected articles.

After training the classifier and making the clustering work, we need to evaluate our pipeline on a new dataset to see how the system really works and how it benefits us. We crawled the Snopes website again with the keyword 'conspiracy theory' and made sure none of the results were previously in our original dataset. The resulting dataset contains 704 articles, and we performed the slight preprocessing task mentioned before to remove unnecessary white spaces.

In order to classify documents, we need to format them as a CSV file where each document occupies one row, and the column is named 'text'. The classifier ranks each input based on the likelihood that the document is a CT or not. We can even apply a threshold to label them as 1 if the probability is higher than a specified value. The default threshold is set to 0.5, which means that a document is labeled as 1 if the probability is greater than this value. We picked a threshold of 0.6, which resulted in 391 articles being labeled as CT.

Next, the clustering algorithm runs on the ones classified as 1 and finds the optimum number of clusters based on the lowest cost function. Given a user-defined space to perform a Bayesian search for the best hyperparameter, it finds the one with the lowest cost function. The higher the cost function, the higher the number of outliers is. In this task, we want the cost function to be minimal.

The configuration of running the Bayesian optimization for a space of:

```
hspace = {
    'n_neighbors': hp.choice('n_neighbors', range(3,20)),
    'n_components': hp.choice('n_components', range(3,20)),
    'min_cluster_size': hp.choice('min_cluster_size',
        range(2,20)),
    'random_state': 42}
```

And the result is:

```
Best parameters: {'min_cluster_size': 2, 'n_components': 10,
    'n_neighbors': 3, 'random_state': 42}
Number of clusters: 50
Cost: 0.04229607250755287
```

The algorithm was able to detect 50 clusters out of the 391 CT articles, with a very low cost of just 0.04. This is an excellent result, indicating that only 4% of the documents, or about 15 of them, were not assigned to a cluster.

Using the labeling method explained earlier, we generate labels for the first 10 clusters.

```
Cluster -1: claim_theory_conspiracy_vaccine
Cluster 0: said_bus_election_conspiracy
Cluster 1: said_message_patent_conspiracy
Cluster 2: claim_death_pizza_conspiracy
Cluster 3: claim_claim_image_photograph
Cluster 4: said_people_group_time
Cluster 5: said_blood_child_conspiracy
Cluster 6: found_mask_hospital_child
Cluster 7: claim_information_vaccine_autism
Cluster 8: said_people_shooting_conspiracy
```

Table 4 The first 10 clusters with top 5 entities

Cluster	Entity	Tag	Count	Cluster	Entity	Tag	Count
–1	Twitter	ORG	12	0	Trump	PER	22
	Trump	PER	12		U.S.	LOC	15
	Flynn	PER	9		Dominion	ORG	11
	Greene	PER	8		Myanmar	LOC	10
	Americans	MISC	8		Donald Trump	PER	9
1	Malaysia Airlines	ORG	9	2	Trump	PER	11
	Flight 370	MISC	8		Donald Trump	PER	9
	CNN	ORG	5		United States	LOC	6
	Suzhou	LOC	4		U.S.	LOC	5
	China	LOC	4		Pentagon	ORG	3
3	Trump	PER	8	4	Pfizer	ORG	6
	Black	MISC	7		Peshawar	LOC	4
	BLM	ORG	6		Muslim	MISC	4
	Oprah Winfrey	PER	5		Sandy Hook	ORG	3
	King	PER	5		Pakistan	LOC	3
5	Biden	PER	29	6	CDC	ORG	10
	Joe Biden	PER	14		Julia	PER	9
	U.S.	LOC	11		Florida	LOC	6
	Democrats	MISC	8		Medicare	MISC	4
	Americans	MISC	7		WHO	ORG	4
7	CDC	ORG	13	8	Sandy Hook	LOC	10
	U.S.	LOC	10		Nazi	MISC	9
	CO 19	MISC	7		Parkland	LOC	8
	FDA	ORG	7		Flynn	PER	8
	Obama	PER	6		David Hogg	PER	7

In the given context, cluster –1 is considered to be the group of observations that lie far away from the other clusters and are called outliers. Upon analyzing the generated label, it is evident that some of the data points from this cluster can be grouped with another cluster that exhibits a predominant medical theme, namely cluster number 6. This can be attributed to the fact that the text documents being analyzed are lengthy and discuss multiple topics. The primary context of these documents may be something other than medical. The words 'claim' and 'said' appear in 9 out of 10 of the clusters. This is expected due to the language used in the original articles, and the labels may need to convey clearly what the cluster is about.

The last piece of the pipeline extracts the named entities. This would help us get to know and understand the cluster better. In Table 4, we can see the 5 most frequently generated entities for each cluster.

By examining the top 5 entities in cluster 6, particularly CDC (Centers for Disease Control and Prevention), Medicare, and WHO (World Health Organization), it is evident that the primary theme of this cluster is related to medical topics. On the other hand, the top 5 entities in cluster –1 include various entities such as Twitter, Donald Trump, Flynn (referring to Michael Flynn, a retired U.S. Army lieutenant general), Greene (Congresswoman Marjorie Taylor Greene), and Americans. This suggests that this cluster could be categorized as one of the clusters with a political theme. However, due to the nature of the dataset, which comprises long text articles, the main theme of the document that mentions these entities might not necessarily be focused primarily on politics.

By clustering similar documents together and labeling them, the existing CTs inside the corpus can emerge through this process. The labeling approach uses common words to highlight the discussion style. This helps us highlight CTs that may not be easily identifi-

able through keyword frequency alone. Additionally, NER provides context on what each cluster focuses on. Overall, the clustering brings together related documents, the labeling summarizes themes, and NER provides entity context. This multi-pronged approach helps detect and understand the CTs that may not be found in the corpus by keyword search alone.

Our methodology diverges from BERT-based tools like BERTopic [83] in several key ways. Unlike BERTopic, which primarily employs Sentence BERT for document embedding and further processes these embeddings using UMAP for dimensionality reduction and HDBScan for clustering, our approach leverages a unique pairing with RoBERTa. This pairing with RoBERTa is crucial for generating word embeddings in our pipeline. Additionally, our method is distinct in its optimization strategy. We have chosen to optimize clustering and hyperparameter selection separately through methods like grid search and Bayesian optimization. This approach allows for more flexibility and precision in fine-tuning each component of our pipeline, in contrast to the integrated workflow in BERTopic.

5 Limitations

This work utilized data from a fact-checking website, a Wikipedia article, and several scientific papers, which collectively only encompass some of the spectrum of CTs found online. Our study did not exhaustively include the comprehensive list of over 290 CTs documented in the 'Conspiracy Theories In American History' encyclopedia [84], due to its coverage only up to 2003. Of these, approximately 200 were not mentioned in our work, with only 90 overlapping with those identified in our study. Consequently, we categorized about 30 CTs that were not listed in the encyclopedia. However, it is important to note that these newly identified CTs might not necessarily be undiscovered; they could be known CTs that were not included in our initial sources. Future work could focus on external validation of these CTs to ascertain their novelty and broader recognition in the field of CT research.

A key feature of CTs is the interconnection of information across different topics, such as the link between 'Pizza Gate' and 'QAnon'. However, the tree structure of the data limits the representation of such information, as each node can only have one label and one parent. This prevents the capture of cross-cutting relationships among CTs.

A significant concern regards the classifier's ability to perform well on articles from websites beyond Snopes. This is due to differences in language and text formatting. Besides, the tree construction was limited due to the difficulty of detecting and scraping English articles and CTs, thus not providing an accurate representation of CTs in other languages or regions. Computational limitations prevented the use of more advanced algorithms, such as XLNet, or a broader grid search during classifier training. Furthermore, the model's robustness was limited since the articles used for training were scraped from only four selected fact-checking sites. The Keyphrase Extraction model showed moderate performance, leading to potential errors in manual labeling.

6 Future work

This work contributes to the state-of-the-art detection of CTs from long texts, but it also has some drawbacks. One of them is the classifier, which determines the pipeline's accuracy. Future work will investigate data augmentation and advanced models such as XLNet

to improve the classifier. Another solution is to employ fine-tuning techniques for LLMs, such as PEFT and LoRA, to adapt the model for this specific task.

This work could be extended to cover other languages, as CTs are not only discussed in English on the Web. A more accurate labeling of the clusters, using a combination of the labeling method and NER or using topic modeling techniques like LDA to extract the topics, can help in assigning more informative and specific labels to the clusters, indicating the exact CTs. This labeling can also help the public, as many users only read the labels and do not examine the clusters in detail. The main objective of future work would be to address the third research goal fully, to identify new CTs that are not in the tree. This could be achieved by building a multi-classifier that can also determine which family and child each document belongs to. If they do not belong to any of the families/children, meaning they are novel, we could infer the exact CTs using a combination of the labeling methods and keyphrase extraction. Another source of new CTs could be the outliers of the clustering algorithm, the ones that are not clustered with any other groups; they might contain new CTs.

This work is a first step towards compiling a comprehensive list of CTs on the Internet. To accomplish this goal, future work should scrape more websites, including sources other than fact-checking websites. Furthermore, the classifier should be enhanced to detect CTs across different types of text and genres, such as promotional, conversational, and journalistic texts.

7 Conclusion

This work compiles a tree of over 120 conspiracy theories. We created a computational pipeline that uses advanced natural language processing techniques, such as Keyphrase Extraction, BERT-based classification, clustering, and named entity recognition, to detect and classify new conspiracy theories from text automatically. To build a labeled dataset, we sourced data from four fact-checking websites. This approach allowed us to identify 14 new CTs, with our classifier achieving an F1 score of 87%.

Our computational pipeline consists of two modules—the classifier and the clustering module. These modules work together to identify potential new CTs in articles and group them for further analysis. The labeling and named entity recognition tools provide valuable insights into the content and context of these groupings. As new theories emerge online, our pipeline automatically expands the CT tree. This comprehensive taxonomy will help researchers and the general public become more aware of different CT narratives. In summary, our work establishes a strong foundation for mapping and monitoring the conspiracy theory landscape using cutting-edge NLP techniques.

Abbreviations

CT, Conspiracy Theory; NLP, Natural Language Processing; BERT, Bidirectional Encoder Representation from Transformers; KBIR, Keyphrase Boundary Infilling with Replacement; MLM, Masked Language Modeling; KBI, Keyphrase Boundary Infilling; KRC, Keyphrase Replacement Classification; NER, Named Entity Recognition; LLM, Large Language Model; RoBERTa, Robustly Optimized BERT Pretraining Approach; NSP, Next Sentence Prediction; HDBSCAN, Hierarchical Density-Based Spatial Clustering of Applications with Noise; UMAP, Uniform Manifold Approximation and Projection; PER, Person; ORG, Organization; LOC, Location; MISC, Miscellaneous; CDC, Center for Disease Control and Prevention; WHO, World Health Organization.

Acknowledgements

We thank Josh Minot for his helpful advice and Protiva Sen, Anthony Barrows, Sean P. Rogers, and Devon Boy for their proofreading assistance. We also thank the anonymous reviewers for their valuable feedback.

Author contributions
MG and JO proposed this paper's main idea. MG prepared the manuscript initially and performed all the steps of the proofs in this research. All authors read and approved the final manuscript.

Funding
This material is based upon work supported by the National Science Foundation under Grant No. 2242829. Computations were performed on the Vermont Advanced Computing Center supported in part by NSF Grant No. OAC-1827314.

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 3 October 2023 Accepted: 2 April 2024 Published online: 16 April 2024

References

1. Douglas KM, Uscinski JE, Sutton RM, Cichocka A, Nefes T, Ang CS, Deravi F (2019) Understanding conspiracy theories. Polit Psychol 40:3–35
2. Barkun M (2013) The nature of conspiracy belief 2nd edn. University of California Press, Berkeley, pp 1–14. http://www.jstor.org/stable/10.1525/j.ctt3fh35v.5. Accessed 2023-12-01
3. John F (2024) Kennedy assassination conspiracy theories. Wikipedia. https://en.wikipedia.org/wiki/John_F._Kennedy_assassination_conspiracy_theories. Accessed 2024-02-05
4. Enders AM, Uscinski JE, Seelig MI, Klofstad CA, Wuchty S, Funchion JR, Murthi MN, Premaratne K, Stoler J (2021) The relationship between social media use and beliefs in conspiracy theories and misinformation. Polit Behav, 1–24
5. Haag M, Salam M (2017) Gunman in 'Pizzagate' Shooting Is Sentenced to 4 Years in Prison. The New York Times. https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html. Accessed 2023-09-20
6. Ecker UK, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, Kendeou P, Vraga EK, Amazeen MA (2022) The psychological drivers of misinformation belief and its resistance to correction. Nat Rev Psychol 1(1):13–29
7. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv:1907.11692
8. McInnes L, Healy J, Melville J (2018) Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint. arXiv:1802.03426
9. McInnes L, Healy J, Astels S (2017) Hdbscan: hierarchical density based clustering. J Open Sour Softw 2(11):205
10. Franks B, Bangerter A, Bauer MW (2013) Conspiracy theories as quasi-religious mentality: an integrated account from cognitive science, social representations theory, and frame theory. Front Psychol 4:424
11. Nefes TS (2017) The impacts of the Turkish government's conspiratorial framing of the gezi park protests. Soc Mov Stud 16(5):610–622
12. Heath C, Heath D (2007) Made to stick: why some ideas survive and others die. Random House, New York, NY
13. Gerts D, Shelley CD, Parikh N, Pitts T, Watson Ross C, Fairchild G, Vaquera Chavez NY, Daughton AR (2021) "thought I'd share first" and other conspiracy theory tweets from the Covid-19 infodemic: exploratory study. JMIR Public Health Surveill 7(4):26527
14. Shahsavari S, Holur P, Wang T, Tangherlini TR, Roychowdhury V (2020) Conspiracy in the time of corona: automatic detection of emerging Covid-19 conspiracy theories in social media and the news. J Comput Soc Sci 3(2):279–317
15. Leal H (2020) Networked disinformation and the lifecycle of online conspiracy theories. In: Routledge handbook of conspiracy theories. Routledge, London, pp 497–511
16. Mahl D, Zeng J, Schäfer MS (2021) From "nasa lies" to "reptilian eyes": mapping communication about 10 conspiracy theories, their communities, and main propagators on Twitter. Soc Media Soc 7(2):20563051211017482
17. Ahmed W, Vidal-Alaball J, Downing J, Seguí FL (2020) Dangerous messages or satire? Analysing the conspiracy theory linking 5g to Covid-19 through social network analysis. J Med Internet Res 22(5):19458
18. De Zeeuw D, Hagen S, Peeters S, Jokubauskaite E (2020) Tracing normiefication: a cross-platform analysis of the qanon conspiracy theory. First Monday
19. 4chan (2024) 4chan. https://www.4chan.org/index.php. Accessed 2024-02-05
20. Pogorelov K, Schroeder DT, Filkuková P, Brenner S, Langguth J (2021) Wico text: a labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In: Proceedings of the 2021 workshop on open challenges in online social networks, pp 21–25
21. Garry A, Walther S, Rukaya R, Mohammed A (2021) Qanon conspiracy theory: examining its evolution and mechanisms of radicalization. J Deradical 26:152–216
22. Boberg S, Quandt T, Schatto-Eckrodt T, Frischlich L (2020) Pandemic populism: facebook pages of alternative news media and the corona crisis—a computational content analysis. arXiv preprint. arXiv:2004.02566
23. Sha H, Hasan MA, Mohler G, Brantingham PJ (2020) Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives. arXiv preprint. arXiv:2004.11692
24. Peskine Y, Alfarano G, Harrando I, Papotti P, Troncy R (2021) Detecting covid-19-related conspiracy theories in tweets. MediaEval
25. Savinainen O, Hvidbjerg Hansen T (2022) Covid-19 related conspiracy theories on social media: How to identify misinformation through patterns in language usage on social media. PhD thesis, Linköping University, Department of Computer and Information Science. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1679954

26. Tangherlini TR, Shahsavari S, Shahbazi B, Ebrahimzadeh E, Roychowdhury V (2020) An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: bridgegate, pizzagate and storytelling on the web. PLoS ONE 15(6):0233879
27. Washington G (2016) Bridge Scandal: what You Need to Know. The New York Times. https://www.nytimes.com/2016/11/04/nyregion/george-washington-bridge-scandal-what-you-need-to-know.html. Accessed: 2024-02-05
28. Smith N, Graham T (2019) Mapping the anti-vaccination movement on Facebook. Inf Commun Soc 22(9):1310–1327
29. Faddoul M, Chaslot G, Farid H (2020) A longitudinal analysis of youtube's promotion of conspiracy videos. arXiv preprint. arXiv:2003.03318
30. Samory M, Mitra T (2018) 'the government spies using our webcams' the language of conspiracy theories in online discussions. In: Proceedings of the ACM on human-computer interaction 2(CSCW), pp 1–24
31. Klein C, Clutton P, Dunn AG (2019) Pathways to conspiracy: the social and linguistic precursors of involvement in reddit's conspiracy theory forum. PLoS ONE 14(11):0225098
32. Tyagi A, Carley KM (2021) Climate change conspiracy theories on social media. arXiv preprint. arXiv:2107.03318
33. Mahl D, Schäfer MS, Zeng J (2022) Conspiracy theories in online environments: an interdisciplinary literature review and agenda for future research. New Media Soc 14614448221075759
34. Kou Y, Gui X, Chen Y, Pine K (2017) Conspiracy talk on social media: collective sensemaking during a public health crisis. In: Proceedings of the ACM on human-computer interaction 1(CSCW), pp 1–21
35. Andrade G (2020) Medical conspiracy theories: cognitive science and implications for ethics. Med Health Care Philos 23(3):505–518
36. Wood MJ (2018) Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. Cyberpsychol Behav Soc Netw 21(8):485–490
37. List of conspiracy theories. https://en.wikipedia.org/wiki/List_of_conspiracy_theories. Accessed 22-August-2023
38. Loose Change (2021) Loose Change—Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Loose_Change&oldid=1057446339. [Online; accessed 6-December-2021]
39. Wikipedia (2023) Chemtrias Conspiracy Theory. https://en.wikipedia.org/wiki/Chemtrail_conspiracy_theory. Accessed: September 1, 2023
40. Wikipedia (2023) New World Order (conspiracy theory). https://en.wikipedia.org/wiki/New_World_Order_(conspiracy_theory). Accessed: September 1, 2023
41. Wikipedia (2023) Antisemitism. https://en.wikipedia.org/wiki/Antisemitism. Accessed: September 1, 2023
42. Wikipedia (2023) QAnon. https://en.wikipedia.org/wiki/QAnon. Accessed: September 1, 2023
43. Wikipedia (2023) Big Pharma conspiracy theories. https://en.wikipedia.org/wiki/Big_Pharma_conspiracy_theories. Accessed: September 1, 2023
44. Wikipedia (2023) Climate change conspiracy theory. https://en.wikipedia.org/wiki/Climate_change_conspiracy_theory. Accessed: September 1, 2023
45. Wikipedia (2023) Moon landing conspiracy theories. https://en.wikipedia.org/wiki/Moon_landing_conspiracy_theories. Accessed: September 1, 2023
46. Dreamteam FC (2023) Ronaldo and the unsolved mystery of the 1998 World Cup Final. https://www.dreamteamfc.com/c/news-gossip/165418/mystery-ronaldo-1998-world-cup/. Accessed: September 1, 2023
47. FullFact. https://fullfact.org/. Accessed 28-August-2023
48. FactCheck. https://www.factcheck.org/. Accessed 28-August-2023
49. Politifact. https://www.politifact.com/. Accessed 28-August-2023
50. Snopes. https://www.snopes.com/. Accessed 28-August-2023
51. Selenium automates browsers. That's it! https://www.selenium.dev/. Accessed 28-August-2023
52. . Beautiful Soup Documentation. https://beautiful-soup-4.readthedocs.io. Accessed 28-August-2023
53. Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, arXiv:1810.04805
54. Hugging Face—The AI community building the future. https://huggingface.co/. Accessed 28-August-2023
55. Hulth A (2003) Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 conference on empirical methods in natural language processing, pp 216–223
56. Kulkarni M, Mahata D, Arora R, Bhowmik R (2021) Learning rich representation of keyphrases from text. arXiv preprint. arXiv:2112.08547
57. Evon D (2016) Ron Howard Didn't Say That the 1969 Moon Landing Was Fake. https://www.snopes.com/fact-check/ron-howard-moon-hoax/. Accessed 31-August-2023
58. Rahman G (2021) World Trade Center 7 didn't collapse for 'no apparent reason'. https://fullfact.org/online/world-trade-center-7-9-11/. Accessed 31-August-2023
59. Tamkin A, Brundage M, Clark J, Ganguli D (2021) Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint. arXiv:2102.02503
60. Sparck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21
61. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
62. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. arXiv preprint. arXiv:1908.10084
63. Bergstra J, Yamins D, Cox D (2013) Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: International conference on machine learning, pp 115–123. PMLR
64. Chen Z, Mi C, Duo S, He J, Zhou Y (2023) Clustop: an unsupervised and integrated text clustering and topic extraction framework. arXiv preprint. arXiv:2301.00818
65. Liu P, Ning Y, Wu KK, Li K, Meng H (2021) Open intent discovery through unsupervised semantic clustering and dependency parsing. arXiv preprint. arXiv:2104.12114
66. Sang EF, De Meulder F (2003) Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint. arXiv:cs/0306050

67. Hugging Face. BERT-base NER. https://huggingface.co/dslim/bert-base-NER
68. Mulhall J (2022) Johnson's Savile slur isn't the first rightwing conspiracy to go mainstream. https://www.theguardian.com/commentisfree/2022/feb/10/boris-johnson-savile-slur-far-right-conspiracy. Accessed: September 14, 2023
69. Best S (2019) First photo of black hole is FAKE, conspiracy theorist claims. https://www.mirror.co.uk/science/first-photo-black-hole-fake-14280651. Accessed: 2023-09-14
70. Russell B (2016) Outrageous conspiracy theory claims this video shows alien remains being probed by Russian KGB agents. https://www.mirror.co.uk/news/weird-news/outrageous-conspiracy-theory-claims-video-8526077. Accessed: 2023-09-14
71. Wikipedia (2023). Jade Helm 15 conspiracy theories—Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Jade_Helm_15_conspiracy_theories. [Online; accessed 14-September-2023]
72. BBC News (2020) Wayfair: the false conspiracy about a furniture firm and child trafficking. https://www.bbc.com/news/world-53416247. [Online; accessed 14-September-2023]
73. Denkmann L, Leibovitz S (2022) Are we victims of a chicken conspiracy? https://www.kuow.org/stories/are-we-victims-of-a-chicken-conspiracy. Accessed: 2023-09-14
74. Frontline (2023) Special Reports—Cocaine, Conspiracy Theories And The Cia In Central America | Drug Wars | FRONTLINE | PBS. https://www.pbs.org/wgbh/pages/frontline/shows/drugs/special/cia.html. Accessed: 2023-09-14
75. Swenson A (2023) False claim of human meat in McDonald's factories stems from old hoax | AP News. https://apnews.com/article/archive-fact-checking-9233643423. Accessed: 2023-09-14
76. News (2022) Conspiracy theorists say they have new proof apollo 17 moon landing was hoax. Toronto Sun
77. Krishna R (2020) This is not an image of a naked child "escaping" from Buckingham Palace. https://fullfact.org/online/child-buckingham-palace/. Accessed: 2023-09-14
78. Arkin D, Popken B (2018) How the internet's conspiracy theorists turned Parkland students into 'crisis actors'. https://www.nbcnews.com/news/us-news/how-internet-s-conspiracy-theorists-turned-parkland-students-crisis-actors-n849921
79. Kaptsov O (2013) The secret of the room 641a. https://en.topwar.ru/32732-tayna-komnaty-641a.html. Accessed: 2023-09-14
80. Smith P, Brockling M (2022) What we know about German group that plotted coup. https://www.nbcnews.com/news/world/germany-coup-plot-reichsburger-sovereign-citizens-conspiracy-theories-rcna60492. Accessed: 2023-09-14
81. Campbell AF (2018) Trump just invented a cruel conspiracy theory about Hurricane Maria's death toll. https://www.vox.com/2018/9/13/17854408/trump-hurricane-maria-death-toll-tweets. Accessed: 2023-09-14
82. Morelock J, Narita FZ (2022) The nexus of qanon and Covid-19: legitimation crisis and epistemic crisis. Crit Sociol 48(6):1005–1024
83. Grootendorst M (2022) Bertopic: neural topic modeling with a class-based tf-idf procedure. arXiv preprint. arXiv:2203.05794
84. Knight P (2003) Conspiracy theories in American history [2 volumes]: an encyclopedia [2 volumes], 1st edn. vol 2. ABC-CLIO, Santa Barbara

Publisher's Note