Climate Disinformation Tracker:

An open-source tool for tracing climate denial narratives on social media

Report of the Joint Interdisciplinary Project (JIP), in collaboration with the National Police

Cesar Hernando (6285937), Kasper Trouwee (6103820),

Maria Paula Jimenez Moreno (6060684),

Manya Atul Narkar (5082919),

Shirley Li (5026555) and Vincent van Vliet (5281318)

Coach: Amir Niknam

October 31, 2025

Contents

1 Abstract

The spread of climate change mis- and disinformation poses a big threat to society; addressing this is the goal of the Joint Interdisciplinary Project (JIP) 6.1.1, in collaboration with the National Police. This report details the development of the Climate Disinformation Tracker, an open-source proof-of-concept tool designed to trace the earliest online occurrence of climate denial narratives on Platform X and provide insightful visualizations of related tweets. The methodology, adapted from the DisTrack architecture, utilizes *KeyBERT* for keyword extraction and a custom scraping pipeline relying on the Nitter front-end for data retrieval, followed by *mDeBERTa-v3-base-mnli-xnli* for natural language inference (NLI) to classify posts as entailing, neutral, or contradictory to a user-provided claim. Validation testing demonstrated that the tool correctly identified the source tweet in 72% of claims when incorporating the synonym component, thus validating the potential of this approach for misinformation tracking. The primary constraints identified are the dependence on non-deterministic Nitter scraping, which introduces operational instability and a 500-character query limit, and the accuracy ceiling of the alignment model. Despite these limitations, the tool validates a functional approach for empowering the public and investigative journalists with traceable context.

2 Introduction

Climate change refers to long-term shifts in temperatures and weather patterns. These shifts can be natural and have occurred multiple times in the Earth's history due to changes in the Sun's activity or large volcanic eruptions. However, since the Industrial Revolution in the 19^{th} century, human activities have been the main driver of climate change, primarily due to the burning of fossil fuels. This stems from the fact that burning fossil fuels releases greenhouse gases that trap the Sun's heat, raising temperatures. The connection between the increase in greenhouse gases and the rise in average global temperatures has been extensively proved by scientists across the world (Nations, 2025). Furthermore, several environmental consequences have been predicted and are already evident, including the melting of polar ice sheets, rising sea levels, extreme rainfall and droughts, and severe heat waves. These environmental changes are expected to result in social, economic and territorial challenges (eur, 2025).

In order to prevent or at least diminish these problems, governments and organizations around the globe are working on regulations and policies to fight climate change. These often negatively affect some sectors, such as oil companies. To protect their interests, some of them have opted to spread disinformation about climate change to undermine trust in science, shape public opinion, and slow down new laws and regulations (Milman, 2023) (Boston University, 2025).

This coordinated spread of disinformation not only leads to an environmental but an information crisis as well. Disinformation, however, is not a criminal offense in the Netherlands. This means the Dutch National Police do not police this kind of speech directly. One approach that separates the Dutch National Police from the rest of the world is that instead of feeling primarily responsible for policing different domains, they encourage each sector - and the public - to strengthen their own defenses against misleading narratives. This is where our tool fits in.

We utilise publicly available data through Nitter - an open source alternative to X - to provide traceable context around circulating climate claims. The tool allows for the input of a claim from the user, tracing back to the earliest observable occurrence of a tweet that aligns with said claim. It does so using several key components, namely, a keyBERT-based keyword extraction, boolean query generation, a tweet scraper, and an alignment model. A visualisation dashboard also allows the user to see not only the spread of tweets that align with the claim - along with their engagement - but also those that are neutral and/or contradictory. Furthermore, to target potential misclassifications, users are also given the option to be able to see the first 'k' tweets that have been classified as either neutral and/or contradictory.

2.1 Value Proposition

Used by the general public - schools, NGOs, local teams - the tool promotes awareness, resilience and safety. One of the key incentives behind this work is to empower individuals to trace the origins of misleading climate narratives and to reconsider their trust when faced with disinformation cascades that can be verified. This approach aligns with the Dutch sector self-defence approach that we previously mentioned. Furthermore, when a subset of this general public - such as investigative journalists - takes it further, this evidence can be used for lawful investigations. The Dutch National Police have observed that organisations spreading disinformation often engage in other prosecutable offenses, such as fraud. Therefore, by investigating such prosecutable offenses, the Dutch National Police can lawfully intervene and take action against these organisations - consequently reducing if not mitigating the spread of climate misinformation by them.

2.2 Problem Statement

Considering the above, we are able to finalise our problem statement. Despite the growing awareness of climate disinformation and the existence of tools aimed at detecting misleading content, there remains a gap in being able to *trace* the origin and spread of such claims. This lack of traceability limits the ability of investigative journalists and researchers to understand the actors responsible for the creation

and spread of disinformation campaigns. Addressing this gap requires an accessible, transparent and automated method to identify the earliest traceable source of a disinformation claim and visualisation of its spread. Our project aims to target this gap by providing a tool that does the same, fine-tuneable by multiple optional advanced-search parameters, supporting investigative journalism, enhancing public awareness and consequently allowing room for the Dutch Police to not only enable sectoral self defence but also intervene as soon as legally viable.

2.3 Sustainability Goals

Our tools contribution can be further understood through three pillars of sustainability:

- People: Sustainability begins with a resilient society. Misinformation erodes public trust in science and polarizes communities. By creating a transparent way to trace misinformation cascades online, our tool helps restore trust, critical analysis and awareness in public communities.

- Planet: Climate action relies on public understanding and effort. Misinformation can delay or derail this process. Therefore our tool aims to remove potential barriers to environmental progress by allowing the public, media, policymakers and others to act based on available facts.

- Profit: When polluting industries or lobbysists spread disinformation, they offer cheap alternatives without exposing their environmental and societal costs. This can often pose as unfair competition for sustainable companies, whose products then seem overpriced, for example, due to these false narratives. Our tool further helps foster integrity within the market and allow companies genuinely committed to sustainability to thrive.

3 Background Information

The modern-day expansion of online platforms has affected how information is shared and consumed. Social media networks such as X (formerly Twitter), Facebook, and Reddit enable fast and large-scale spreading of opinions and claims. While this accessibility allows for easy communication and expression of freedom of speech, it also facilitates the spread of misinformation and disinformation. In the context of climate change, disinformation campaigns have been shown to distort public understanding and delay effective environmental policies (Franta, 2021). Fossil fuel companies, for example, have historically contributed to climate scepticism by funding misleading studies and online narratives that question scientific consensus (Heffernan, 2024).

3.1 Climate Disinformation

Climate disinformation refers to the intentional spread of false or misleading information about climate science, causes, or mitigation strategies. This differs from misinformation, which may be unintentionally false (Fallis, 2014). Online disinformation campaigns often follow recognisable patterns, originating from a small number of accounts or organisations and then amplified through reposts, quotes, and algorithmic recommendation systems (Aïmeur et al., 2023). These coordinated activities are difficult to trace due to account or post deletions, reposting dynamics, and limitations in accessing social media data. These climate disinformation campaigns often employed rhetorical strategies such as exaggerating scientific uncertainty, promoting false balance in media coverage, and framing climate policies as economically harmful or politically motivated (Lewandowsky, 2021). As social media platforms became the primary channel for public discourse, these narratives adapted to the new digital environment, spreading (mis)information more rapidly (Del Vicario et al., 2016).

3.2 Existing methods for misinformation detection

Recent research has proposed several frameworks for tracking misinformation online, combining natural language processing (NLP) and social network analysis.

The *Truth Tracker* framework (Sadasivam et al., 2024) surveys and evaluates current rumour detection methods across social media. It highlights the limitations of traditional feature-engineering approaches, which rely on lexical, syntactic, and user-based features, and emphasizes the shift toward deep learning for more generalizable and scalable detection. Transformer-based architectures, such as *CE-BERT*, outperform earlier models by capturing both linguistic and contextual information, achieving higher accuracy and earlier detection of misinformation. These advances demonstrate the potential of large language models for real-time misinformation monitoring.

Hunt et al. (Hunt et al., 2022) propose a machine learning framework for monitoring misinformation on Twitter during crisis events such as the Boston Marathon bombing and Hurricane Harvey. Their approach automatically predicts the alignment of tweets as *true*, *false*, or *neutral* using n-gram[1] and TF–IDF[2] textual features, user metadata, and traditional classifiers such as Random Forests and Support Vector Machines. Achieving F1 scores[3] above 87%, their system effectively tracks the diffusion and correction of misinformation, offering practical tools for agencies and researchers to improve situational awareness and crisis communication. Our solution makes use of similar tweet classification, as will be discussed in section 6.

Recent studies consistently recognize that the spread of misinformation on social media remains a persistent and complex challenge. Ongoing research explores alternative strategies, including improved visualization of information networks (Shao et al., 2016) and interdisciplinary approaches that model misinformation diffusion using methods from epidemiology (Bonnevie et al., 2021). While a definitive solution has yet to be found, continuous efforts are advancing the development of more reliable and

[1] sequence of 'n' adjacent items from a given text
[2] Term Frequency-Inverse Document Frequency is a statistic used to evaluate how important a word is to a document
[3] Metric used in information retrieval to evaluate the performance as the harmonic mean of precision and recall

effective misinformation detection techniques.

3.3 DisTrack

One particularly relevant work is *DisTrack: A New Tool for Semi-Automatic Misinformation Tracking in Online Social Networks* (Villar-Rodríguez et al., 2025), which served as a foundation for our tool's architecture and model selection.

DisTrack was designed to identify and trace the spread of misinformation across social media platforms by integrating both textual and temporal signals. Its architecture consists of three main components: data retrieval, semantic analysis, and visualisation. The system begins by collecting posts from Twitter's API, which are filtered based on relevance to a specific claim or topic. The textual component relies on large language models (LLMs) to extract and compare semantic information between the original claim and retrieved posts.

For this purpose, *DisTrack* employs *KeyBERT*[4] for keyword extraction and *AIDA-UPM/xlm-roberta-large-snli_mnli_xnli_fever_r1_r2_r3*[5] for alignment classification. *KeyBERT* identifies the most representative terms within a claim using contextual embeddings derived from transformer models, enabling the construction of precise search queries. The alignment detection model, fine-tuned for multilingual natural language inference (NLI), determines whether a post entails, contradicts, or is neutral toward the input claim. This combination allows the system to filter only those posts that semantically align with the original claim.

In our project, we adopted this architecture as a baseline. The same *KeyBERT* model was used for keyword extraction for keyword extraction, due to its demonstrated effectiveness in extraction. However, while *DisTrack* relied on Twitter's official API, our implementation diverges by using Nitter, an alternate frontend for Twitter, for scraping for data collection, since the official API was not accessible for this project.

[4] https://huggingface.co/AIDA-UPM/mstsb-paraphrase-multilingual-mpnet-base-v2
[5] https://huggingface.co/AIDA-UPM/xlm-roberta-large-snli_mnli_xnli_fever_r1_r2_r3

4 Stakeholders & Business Aspect

At the moment, there is no easy way for people (like journalists) to trace back misinformation and disinformation to the source. This leads to companies such as Exxon being able to spread disinformation for 27 years (Revkin, 2015). So how do companies like Exxon spread mis- and disinformation? They fund companies such as Clintel, (Joosten and Keizer, 2020), these companies in turn make articles such as "New book on the status of the climate issue – the world's biggest problem or just a hoax?" (Foundation, 2025), which then get shared on social media (Balthus23Air, 2025). These companies also have ties to politicians and can influence their decisions, as seen in the United Kingdom with their new protest laws (Gayle, 2024).

The spreading of mis- and disinformation in turn polarises society (Azzimonti and Fernandes, 2023). The spreading of mis- and disinformation can be fought by correcting incorrect information as is done by scientists. Another way is to investigate companies that are likely to spread mis- and disinformation and find out if they did, as this is how journalists do it.

This report is about a potential third way. This approach starts at the end, this could be some tweet/claim on the internet for example. After finding the claim, the tool will search for similar claims and this can be used to determine the place where the claim started but also irregularities around the claim like multiple accounts saying the exact same thing. This approach was previous impossible as it would require an insensible amount of time of the users, but by automating the process of identifying the earliest traceable source, the tool enables users to efficiently detect patterns and narrow the source of disinformation. Making this a viable option.

The flowchart in Figure 10 within Appendix B visualises in a simple manner how the spreading of mis- and disinformation happens and also shows where certain stakeholders (can) take action. Based on the information provided by the coach, articles, papers, etc. Alongside the flowchart a power interest grid is made (see Figure 9) (Reddi, 2023).

The following stakeholders with a short explanation are visualised within the power interest grid and flowchart. These stakeholders are selected because they are important (or can be) in the industry of spreading disinformation.

- Politicians - This refers to all politicians, so those influenced by lobbyists but also those who starkly oppose them. The tool helps reveal how such narratives emerge and spread - supporting more informed policymaking.
- Social media platforms - ll social media platforms specifically for this project X since they host and moderate much of the misinformation analysed.
- Advertisement industry - Companies such as Clintel that spread incorrect information about climate change. The tools insights could increase scrutiny of such campaigns.
- Fossil fuel industry - These include companies that spread disinformation, such as Shell, BP, Exxon, etc.The tool may expose these campaigns, affecting their public image.
- Police - The Dutch National Police are the commissioning stakeholder and an indirect end user - stepping in as a result of the direct use of investigative journalists, prosecuting organisations for prosecutable offenses, inadvertently reducing the spread of climate disinformation.
- Journalists - Direct end users that can then use their network to investigate potential disinformation campaigns further.
- Society - People who get affected by climate mis- and disinformation on X. They are also a direct end-user, promoting a more informed public.

- Scientists - People who research climate change and are recognised as experts. Their work is often misrepresented and the tool benefits them by helping others trace such distortions.

4.1 Business Aspect

While the stakeholders represent those affected by or involved in the spread of false climate related information, we now discuss how this tool can be sustainably deployed and provide value in practice.

From an operational perspective, the project is envisioned as a non-profit, open-access tool developed in collaboration with the Dutch National Police. Ideally, it would be hosted on a secure server - allowing users to perform source-finding analyses through an accessible web interface. Since it serves as a public-interest project, maintenance and hosting costs would be covered through either public funding or taxation, as the product in and of itself does not generate revenue.

Currently, after discussion with our representative at the Dutch National Police, the deployment and infrastructure aspects fall outside the scope of this project. The tool currently exists as an open-source prototype on GitHub, promoting transparency and replicability for the project.

This can be understood in terms of a "pain" vs "gain" tradeoff which is expanded on below. Pain:

- The project does not generate any financial profit, meaning long-term hosting and maintenance would rely on institutional support or public funding.
- Operating the tool requires ongoing computational resources and server upkeep
- Adoption may not be very immediate and broad as the tool targets a niche but crucial use case, therefore, depending on general user awareness and a specific subset of users such as investigative journalists.

The gain in turn can be recognised as the following:

- The tool significantly reduces the time and effort required to trace the origins of online climate mis/disinformation on X.
- When used, it provides high social value, improving the investigative capacity for journalists as well as awareness for the general public.
- Being open source ensures transparency and accessibility, hopefully restoring trust in digital investigations.
- Establishes a scalable framework that can then be expanded to other social media platforms and domains.

Therefore, although there is a "pain" aspect to the tool, its gain in terms of improved information integrity as well as reduced climate disinformation online makes it a valuable long-term investment for an institution responsible for public good.

5 Possible Solutions

During the development of our tool, we explored several alternative solutions before finalising our current implementation. The goal of the exploration was to be able to identify the most suitable approach for reliably tracing the source of a climate disinformation cascade online. Broadly, we explored three domains:

- Data gathering: Deciding whether to obtain posts through real-time platform queries or through existing datasets, and if the former, then what platforms to use.
- Search strategy: Determining how to efficiently locate the earliest post within the retrieved data.
- Synonym generation: Accounting for linguistically varying posts of semantically the same claim.

Each alternative posed its own set of pros and cons. The following sections consist of our findings.

5.1 Dataset of tweets

Our project was initially inspired by prior research (Villar-Rodríguez et al., 2024) that used the Twitter Academic Tier API. This seemed like the most convenient option, however, it was no longer available after its acquisition by X. The closest equivalent in terms of the X API (X, 2025a) was the Enterprise version, which exceeded the budget caps for the scope of the project. The remaining API tiers were either still financially prohibitive or heavily rate-limited - proving themselves unfit for the project. Additionally, using a scraper to scrape X in and of itself is against its Terms of Service (X, 2025b). Given this constraint, we explored the option of using publicly available datasets of tweets. However, the problem with this approach was that datasets containing a small sample of tweets limited our scope of successfully finding a 'source', and datasets containing the full archive were extremely large - spanning around 6.8TB just for tweets from 2012-2023 (Archive Team, 2012), making them difficult to process. The larger files we found still had several months' worth of data missing within. Additionally, it was difficult to find datasets with complete data related to climate change. Due to these reasons, using a dataset seemed unsuitable for reliably tracing the earliest source of diverse claims.

5.2 Platform selection

Since using pre-existing datasets was deemed unfit, we shifted our focus to retrieving data directly from social media platforms in real time. Since X previously proved to be an unviable platform, we decided to explore alternative platforms. The following are some we considered, evaluating each of them based on accessibility and functionality:

5.2.1 Threads

One platform that we took into consideration was Threads. The platform provides an API that can be accessed for legitimate reasons such as academic research. However, this required an application and approval process that would exceed our ideal project timeline. In addition, the Threads API had several technical limitations as well. Each query was capped at retrieving 100 posts, making finding a 'source' unreliable. Moreover, the Meta Content Library - which serves as a shared archive for platforms such as Facebook, Instagram and Threads - restricts data retrieval to posts shared by public profiles with 1000 or more followers (Met, 2025). Additionally, Meta uses automated systems and third-party fact-checkers to exclude posts flagged as sensitive or misinformation from search, so queries could return empty arrays (Meta Platforms, Inc., 2025), making it impossible to find sources of certain topics.

5.2.2 Facebook

Facebook has two APIs: GraphAPI and Meta Content Library API. Graph API is commonly used for content management and analytics as it mainly supports retrieving data from specific objects, ie., pages, users or posts that a user has explicit access to. It does not support open keyword-based search

across public content. The Meta Content Library API, on the other hand, provides searchable access to public posts across Facebook, Instagram and Threads; however, this interface, as previously discussed, is only available to approved research institutions - the application process for which was far beyond the scope of our project. Additionally, the available data is subject to visibility restrictions, making it difficult to find sources for Facebook as well.

5.2.3 Reddit

We also considered Reddit, given its large volume of discussions. Reddit, we found, provides two types of API access: personal and commercial use. The personal use API prohibits use on behalf of organisations, deeming this unfit for our project. The commercial API, on the other hand, also requires an application and approval process that could take up to several weeks. Additionally, the Reddit API lacks full support for fixed date-range searches, ie., the time filters are relative ("a month ago", "a year ago") to the date of searching. This means that queries made on different days would yield shifting time windows, preventing consistent replication of searches. These constraints made Reddit impractical for our project as well.

5.2.4 Nitter

After evaluating available social media platforms, we found that none were ideal for our use case. They either imposed strict access limitations or required an approval process that proved to be too lengthy for the duration of our project. To overcome these constraints, we decided to develop our own data collection pipeline - a custom scraper - instead of relying on APIs. We further decided to use this scraper on Nitter (Zedeus, 2025; Zedeus and contributors, 2025) - a free and open-source alternative front-end for X. Admittedly, the usage of Nitter is a grey area, but one loophole we found was that since the scraper interacts only with Nitter's publicly available data, this approach does not technically violate X's Terms of Service. This solution provided us with a practical balance between accessibility and functionality - enabling a relatively more reliable retrieval process for our tool.

5.3 Search strategy

With some additional time on our hands, we decided to explore two different search strategies for the tool: linear and binary. In theory, we would assume binary to be faster - ideally having time complexity O(log n) compared to linear O(n). However, this is based on the assumption that each search has a constant time lookup - for example, an in-memory access. This assumption does not hold in our implementation as we do scraping and running the alignment model to classify tweets.

Every probe to decide which half of the timeline to explore next requires a network call to Nitter, followed by page rendering - each of which is expensive. As a result, the theoretical advantage does not directly translate to real-time execution.

One noticeable tradeoff in both is that of batch size vs probe frequency. The linear approach retrieves tweets in larger annual batches, running fewer probes overall, while the binary method works on much smaller windows - retrieving tweets for around 45 days, but using multiple probes and peeks to determine which half of the timeline to explore. As a result, the efficiency of these algorithms depends on the distribution of tweets related to a particular claim across time. If relevant tweets are sparse and widely spread, binary search could skip high-volume tweet retrieval per year, and in theory, be faster. However, when tweets are clustered within a few active periods, linear scanning seems to be faster since it makes fewer probes and retrieves fewer tweets.

To gain an empirical advantage for selecting one method over the other, we tested both searches on fifteen climate-related claims. Linear search, for the most part, completed within a hundred seconds, while binary search, for the most part, took a minimum of a hundred seconds, reaching up to several hundred seconds sometimes. It should be noted, however, that these results could be specific to our dataset. We selected linear search for the proof of concept as it seemed to be the quickest on the claims we tested. For a more general conclusion on which method is better, a broader and more thorough

evaluation could be conducted, ensuring claims that support all types of tweet distributions are used. This is further discussed in the future work section.

5.4 Synonyms

The search query uses multiple keywords to search, for example, the claim *"electric cars are worse for the environment"* gives the following keywords `electric, cars, worse, environment`. But the claim *"electric vehicles are worse for the environment"* means the same, but will not be found since it does not have "car" in the search query.

This problem can be resolved by adding synonyms so the query gets transformed from:

```
electric AND cars AND worse AND environment
```

Into:

`electric AND (cars OR vehicles) AND worse AND environment`,

This way, both "cars" and "vehicles" will be included in the search query.

To retrieve synonyms, a REST API (**Re**presentational **S**tate **T**ransfer **A**pplication **P**rogrammable **I**nterface) is used (Amazon Web Services (AWS), 2025). The specific REST API is from Datamuse, for an example see `https://api.datamuse.com/words?rel_syn=car` or see Appendix C.

As the example in Figure 11 shows, The REST API gives some good suggestions for "car", but in the context of the previously mentioned claims, some are relevant, like "auto", while others like "gondola" are not.

Another problem with Datamuse is that sometimes it gives empty results even if there are synonyms. To fix this problem, the NLTK Python package is used in combination with Wordnet. Wordnet is a big database of the English language that is easy to navigate (Princeton University, 2010). NLTK allows the easy use of Wordnet, but also in combination with OMW, which now enables support for multiple languages (Bond and Paik, 2012).

Yet this still gives similar results for "car" as the REST API. To fix that, the context is used to get the top results and this is done by using Spacy (`https://spacy.io/`). For the project, `en_core_web_sm` is used, this is an English-only model, but there is also a multi-language model option.

The pipeline is as follows: first, Wordnet is used to get synonyms for each keyword. Then the original claim is inputted into Spacy, after which the synonyms are ranked (value between 0 and 1) on how well they match with the keyword within the context. Synonyms above the threshold 0.1 and the top 5 ranked synonyms are returned.

6 System design

Having explained the different possibilities that we explored during this project and why they were discarded, we will now give a detailed explanation of the workflow of the latest version of the Climate Disinformation Tracker tool. As mentioned in the previous section, we opted to focus on misinformation spread on X, and due to the limitations of its API, we decided to use Nitter, a front-end alternative for X. Nitter allows retrieval of tweets of the entire history of X (formerly Twitter) and gives the user the choice of excluding and filtering certain types of posts, such as replies, retweets, videos or news. However, similarly to X, it has the limitation of having a keyword-based search. Thus, it does not take into account the meaning of sentences and consequently, it does not retrieve tweets that are phrased in a different way but have a very similar meaning. This drawback impeded us to utilize advanced search methods that make use of transformer models, such as dense retrieval. Furthermore, it shaped our system design.

Our tool consists of three main steps or components, which follow after the user's input. The input includes a claim, which could be a hoax related to climate change but not necessarily, together with a date range and some desired filters. In the first step, we employ an LLM named *KeyBERT* (Grootendorst, 2020) to extract the relevant keywords, and then we construct a boolean query. This enables the search of combinations of keywords, retrieving tweets that may not include all keywords. The tool also permits including synonyms in the query, which will be explained in detail later on. Secondly, we scrape Nitter and retrieve tweets. Finally, we use another LLM named *mDeBERTa-v3-base-mnli-xnli* (Laurer et al., 2024) to determine whether the retrieved tweets entail the original claim, contradict it or are neutral towards it.

By combining these components smartly we can find the earliest entailing tweet to the user claim, which ideally is the source of misinformation in that topic. Additionally, we can retrieve *all* the tweets within the given date range to perform more sophisticated data analysis and return insightful visualizations. These two functionalities will be explained in section 7. In the following subsections, the different components of the tool will be described in depth, as well as the synonym augmentation add-on. In the last subsection, we will provide details on our frontend user interface and visualization dashboard.

6.1 Keyword extraction and boolean query construction

Extracting the keywords of a claim manually can be tedious for the user, especially for long sentences. To make the user experience smoother, we opted to use an off-the-shelf model to extract the most representative words, *KeyBERT*. This transformer-based keyword extraction technique leverages contextual embeddings generated by BERT (Bidirectional Encoder Representations from Transformers) or any of its derivatives to find keywords that are semantically close to the overall meaning of the text. Instead of using the original BERT model, we used the following model: *mstsb-paraphrase-multilingual-mpnet-base-v2* (AIDA-UPM, 2024). This model was introduced in a study (Villar-Rodríguez et al., 2025) which also analysed misinformation in Twitter, but via the Twitter API and about other topics. The model was fine-tuned for semantic textual similarity with multilingual data (15 languages) and can thereby be used to analyze misinformation in several languages.

One of the parameters of the model is the maximum number of keywords extracted (*max_keywords*). For claims between 10 and 20 words we found that setting $max\_keywords = 5$ was enough to summarize the content of the sentence. However, for longer claims, more keywords may be needed, so we give the user the possibility to choose this parameter.

Inspired by the *DisTrack* paper, after extracting the keywords we construct a boolean query. As they mention, and we tested ourselves, including all the important keywords in these generated queries may not lead to the expected results because there can be tweets with some of the keywords from the claim but not necessarily all of them. Therefore, we generate boolean queries so that the retrieved tweets contain any combination of k keywords out of all the keywords [6]. After several tests, we discovered

[6] Internally, X does $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ parallel searches, corresponding to each of the clauses between parentheses

that for 5 keywords, using $k = 4$ was a good balance between flexibility and accuracy in the search. However, we give the user the option to select the parameter *n_keywords_dropped*, which in this case would be $5 - 4 = 1$.

For example, if the claim is: *Electric vehicles are actually worse for environment than gas cars*; the keywords extracted are: *electric, gas, worse, environment* and *cars*; and we use *n_keywords_dropped* = 1, the boolean query will be:

(*electric* AND *gas* AND *worse* AND *environment*) OR (*electric* AND *gas* AND *worse* AND *cars*)
(*electric* AND *gas* AND *environment* AND *cars*) OR (*electric* AND *environment* AND *worse* AND *cars*)
OR (*environment* AND *gas* AND *worse* AND *cars*)

This methodology makes our search more flexible and it indirectly tackles the problem of tweets using synonyms, given that the tweet includes less or the same number of synonyms as the number of keywords dropped. However, increasing *n_keywords_dropped* would lead to retrieving tweets that do not longer revolve around the topic of the original claim. In the following section, we will describe an add-on of our tool that tackles the synonyms challenge.

6.2 Synonym-aware search

Different words can convey the same meaning—for instance, "humans" and "people." We introduced synonym functionality to improve search query by ensuring a query retrieves more relevant tweets regardless of the specific equivalent term used. The synonyms were already mentioned in subsection 5.4.

The final solution is as follows: first the tool gets synonyms for the keywords with the use of Wordnet, which is simply a database of the English language that is easy to navigate (Princeton University, 2010). and is combined with Open Multilingual Wordnet (OMW), which enables multilingual support (Bond and Paik, 2012).

As OMW builds on top of Wordnet and the research is primarily focused on English, both were added since there are no significant overhead costs.

After the synonyms are extracted, Spacy, a natural language model processing Python package (Explosion, 2025), is used to score the synonyms relevance based on the context. As the context can change the meaning of a word. For example, in the sentences *"the light is bright"* and *"the student is bright"*, "bright" means something different. Synonyms are scored on a scale of 0.0 to 1.0. The top 5 results are returned as long as they score above the threshold of 0.1. The threshold (0.1) and top N amount (5) were selected following an iterative tuning process to maximize list quality, while also accounting for the user's ability to manually refine the final set.

The Spacy model used is *en_core_web_md* as that is the easiest to implement but it only supports English. Other models exist as well such as *xx_ent_wiki_md* that support multiple languages, however these have not yet been tested due to time constraints.

While testing the tool, we realized that the extracted synonyms were sometimes inaccurate, as described in subsection 5.4. In this context, inaccurate means that the word would not be replaceable by the synonym in the context of the claim, as a consequence of setting a low threshold for the Spacy model. To address this issue, we decided to treat the extracted synonyms as suggestions for the user. The user can then select the synonyms they prefer and add their own. By incorporating a human in the loop, we shift some work to the user but also improve the tool's performance, as investigative journalists, in particular, might have better intuition about which synonyms to use.

6.3 Tweet retrieval via Nitter scraping

After selecting the claim, date range and search filters, we retrieve tweets by scraping Nitter. Web scraping is the process of programmatically retrieving and parsing the HTML of webpages to extract specific datafields for storage, analysis, or further automated processing. Technically, scraping does

not include automatically visiting multiple URLs. This process is called web crawling. In our tool we combine both processes to harvest tweets.

Nitter allows the user to search tweets without logging in, but it internally uses tokens to access X. As a measure to protect itself from running out of tokens or X blocking it, Nitter has several domains that work in the same way but with different URLs. Through the duration of our project, only two domains were available: https://nitter.tiekoetter.com/ and https://nitter.poast.org. Our tool relies completely on the availability of these domains, working as long as one of the domains are accessible.

In order to crawl Nitter, we use the *Playwright async API* (Microsoft, nd). The URL that we request to access depends on the claim, dates and filters. The following example showcases how Nitter forms the URL:

https://nitter.tiekoetter.com/search?f=tweets&q=Electric+vehicles+are+actually+worse+for+environment+than+gas+cars&e-nativeretweets=on&e-replies=on&since=2007-01-27&until=2025-10-27&near=

If fetching this URL is satisfactory, that is, the status code is 200, we proceed to parse the tweets from the HTML. This is done using the Python package *BeatufilSoup* (Richardson, nd). Each page contains a list of tweets, each with the following datafields: username, tweet content, date, X post link, replying to (list of users that the tweet is replying to), number of comments, retweets, quotes and likes. It is important to note that Nitter displays tweets in reverse chronological order, showing the most recent ones first. This feature influenced our strategy to find the earliest entailing tweet, which will be explained in the next section. Nitter only shows a few tweets per page, and includes a *Load More* button in the bottom. Inspecting the page source we noticed that *Load More* contains a link which adds a cursor field to the original URL. Therefore, in order to retrieve all the tweets we iterate over the pages, modifying the cursor and progressively storing the tweets, until no more tweets are found.

Nonetheless, during testing we encountered different errors that we had to handle. One of them occurs when Playwright cannot access a Nitter domain due to its unavailability, or when the Nitter instance has no authentication tokens or is fully rate limited. To circumvent these errors, we switch to another domain. Furthermore, we had to handle the error when the query is too long, which occurs when the boolean query itself (excluding dates and filters) has more than 500 characters. In this case, we just output an error message to the user.

6.4 Prediction of tweet alignment with the original claim

When searching with keywords in Nitter, the retrieved tweets can fall in one of the following three categories. The tweet could entail with the original claim, meaning that it supports it. However, the tweet could also contradict it. Finally, the tweet could contain the search keywords but neither entail nor contradict the original claim, falling in the neutral category. This language task is called Natural Language Inference (NLI) in the literature. It is defined as the task of determining whether a natural language hypothesis h can justifiably be inferred from a natural language premise p (MacCartney, 2009).

In order to classify the tweets according to their alignment to the original claim, we used *mDeBERTa-v3-base-mnli-xnli* (Laurer et al., 2024), a multilingual large language model available on Hugging Face. It was pre-trained by Microsoft on the *CC100* multilingual dataset (Conneau et al., 2019), and then fine-tuned on the *Multi-Genre Natural Language Inference (MNLI)* (Williams et al., 2017) and *Cross-lingual Natural Language Inference (XNLI)* (Conneau et al., 2018) datasets. This model was evaluated on the *XNLI* test set, and it is reported have almost 90% accuracy on English and around 80 % on average over the 15 languages used for fine-tuning.

The workflow to classify tweets is as follows. First, both the original claim and retrieved tweet are tokenized[7] and transformed into numerical embeddings that the model can process. The model then

[7]Tokenization is the process of breaking down text into smaller units called tokens so that it can be processed by a machine learning or natural language processing model (Grefenstette, 1999).

performs forward inference to compute logits, which are converted into probabilities using a softmax function. The relation with the highest probability is taken as the predicted alignment label.

Since we do not retrieve only one tweet but several (hundreds or thousands), we perform this process in batches, leveraging the parallelization capabilities of transformer models. Ideally, the tokenization and evaluation of NLI by the model is performed using a GPU, but it is not necessary. For our laptops, we found that the alignment prediction was fastest when using batches of 4 when running on CPU, whereas one of our laptops with GPU had better performance when using batches of 16.

If we are using the *find first related tweet* functionality, in which we are interested in the earliest entailing tweet, the tweets are filtered based on alignment, selecting only the entailing ones. In contrast, in the *find all* functionality, we just label the tweets, keeping all of them for different analyses and visualizations. These functionalities will be described in detail in the next section, where all the components will be integrated.

6.5 Frontend and User Interface

To make the Climate Disinformation Tracker accessible to non-technical users such as investigative journalists and researchers, we developed an intuitive frontend and dashboard to visualize the data using the Dash framework (Plotly, 2025a). Dash was selected because it allows for the simple creation of an interactive dashboard.

Upon launching the application, users are presented with a user interface for configuring all the necessary parameters for the system, which can be seen in Figure 1 below. The main parameter, the mode, allows users to switch between finding the first tweet that semantically aligns with the user-provided claim and finding all related tweets to be further viewed in a visualization dashboard. Both of these modes are further explained in section 7. Clicking on the information button on the left corner opens a pop-up with instructions on how to use the tool and what all the parameters mean, which can be seen in Figure 2.

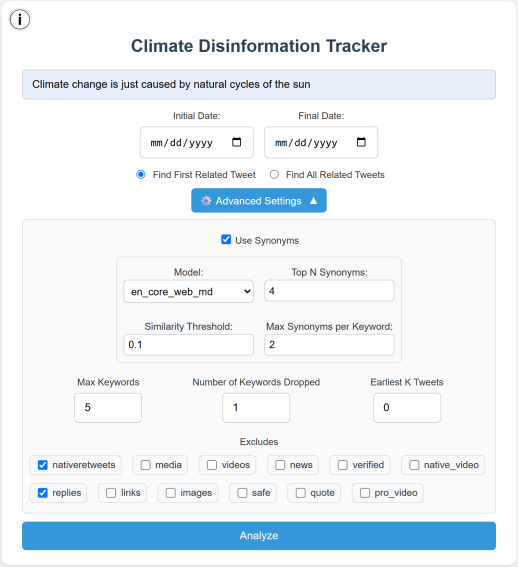


Figure 1: Frontend UI with advanced settings opened

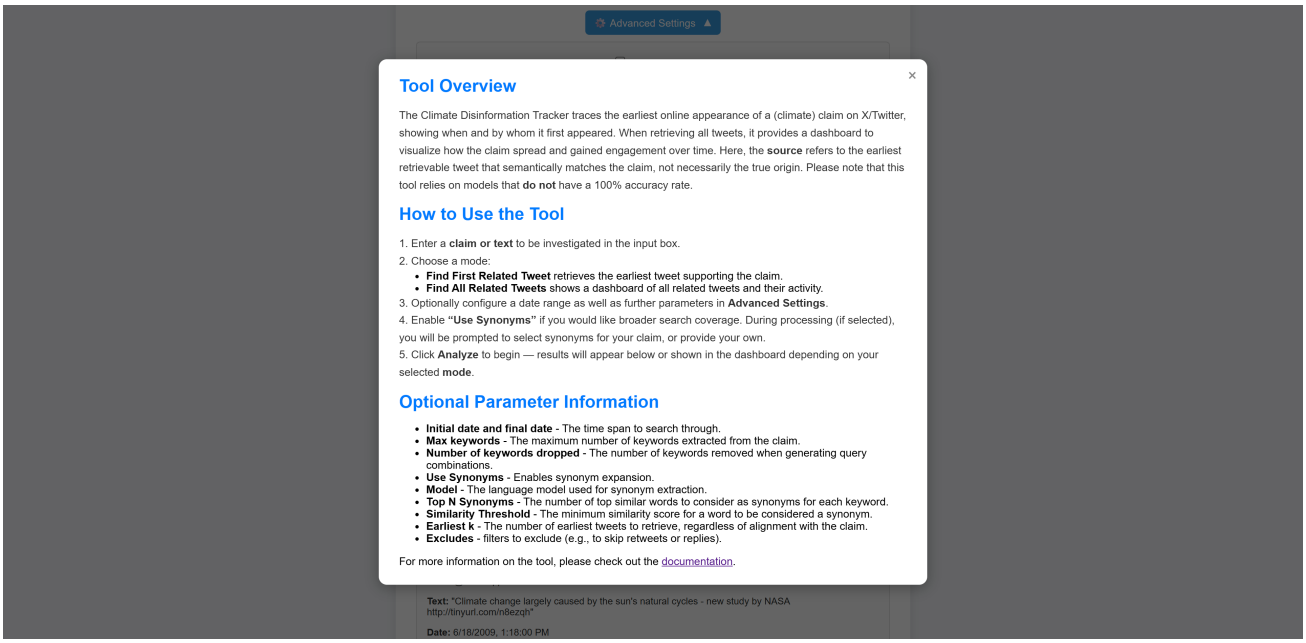


Figure 2: Tool instructions

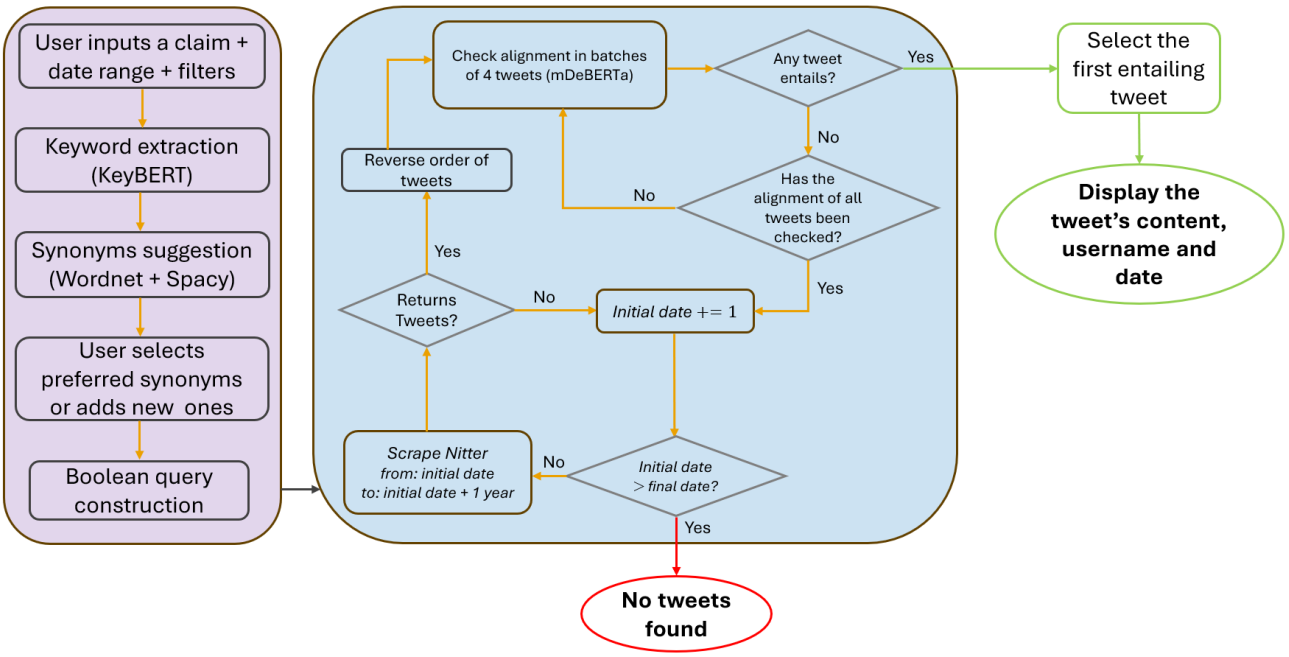


Figure 3: Pipeline of the *find first related tweet* functionality. After the user inputs a claim, keywords are extracted and a boolean query is constructed based on the synonyms selected by the user. Then, a linear search strategy is used to find the earliest entailing tweet, which consists of scraping Nitter in windows of one year starting from the initial date, and checking the alignment of tweets using the model *mDeBERTa*.

7 Tool functionalities

As mentioned previously, the main goal of this tool is to help the police and investigative journalists trace back the origin of disinformation about different topics on social media. In the context of climate change, this tool could expose organisations' efforts to spread climate denial narratives for their benefit. Hypothesising that these organisations might have fabricated false climate claims, our first approach was to build a functionality to find, for a given claim, the earliest entailing tweet in Twitter/X.

After testing this functionality, we found that the earliest entailing tweet did not yield particularly valuable insights for the claims we examined, as the corresponding accounts were typically not affiliated with organisations and generally had a low number of followers. Nevertheless, investigative journalists could experiment with more strategically selected claims and potentially uncover more informative results. The pipeline for identifying the earliest entailing tweet is described in subsection 7.1.

Since we were aware of the fact that organisations, such as oil companies, could use bots to spread disinformation for them, and that they could have used other social media platforms to propagate misleading narratives in the first place, we opted to make our tool more valuable by implementing another functionality. This functionality consists of retrieving every tweet in the desired date range and analysing the temporal evolution of tweet activity through a bubble chart, while also identifying the accounts with the highest posting frequency on the topic. Furthermore, a network analysis component was implemented to examine how information and disinformation are propagated on X and to potentially identify echo chambers. This functionality is described in detail in subsection 7.2.

7.1 Finding the earliest entailing tweet as a potential source

In this section, we will integrate the components explained in section 6 to find the earliest entailing tweet to an input claim. The whole pipeline is shown in Figure 3. First of all, the user inputs a claim, some custom filters (to exclude replies and retweets, for example), the initial and final date of the search, whether they want to include synonyms, and other hyperparameters such as *max_keywords* or *n_keywords_dropped* (explained in subsection 6.1). Then, keywords are extracted from the claim, and

the boolean query is constructed, which can include custom synonyms. Next, starting from the selected initial date, we scrape Nitter to retrieve tweets in windows of one year. If tweets are found, we reverse their order to take into account that Nitter returns the latest tweets first. After this, we predict the alignment of these tweets in batches of four tweets, and if we encounter one or more entailing tweets in a batch, we return the first one as the earliest entailing tweet. If no tweets are found in the window of one year, or if none of the tweets found entail, we move to the next year. Furthermore, if one of the Nitter domains is not available, we catch the error and move to a different one.

Apart from the potential failure encountered when all the domains are unavailable or fully rate-limited, the functionality could also fail if earlier tweets to the found source are misclassified, or if the source itself is misclassified. This can happen, as the alignment model used has an accuracy of 90%. To overcome this drawback, we added the option of returning a list of candidate sources together with the earliest entailing tweet found by the tool. To support this, we introduce an additional parameter: k, which determines how many of the earliest tweets are stored alongside and the identified source. For any value $k > 0$, we scan chronologically, collecting up to k of the earliest tweets, regardless of their alignment label. This ensures that the stored set includes the earliest occurrences of tweets containing claim-related keywords, some of which may have been misclassified as neutral or contradictory. Including this list allows users to manually review all early mentions of the claim and verify whether the automatically detected entailing source is indeed the earliest relevant post. When $k = 0$, which is the default setting, only the single earliest entailing tweet is returned. Figure 4 shows the frontend when $k = 5$.

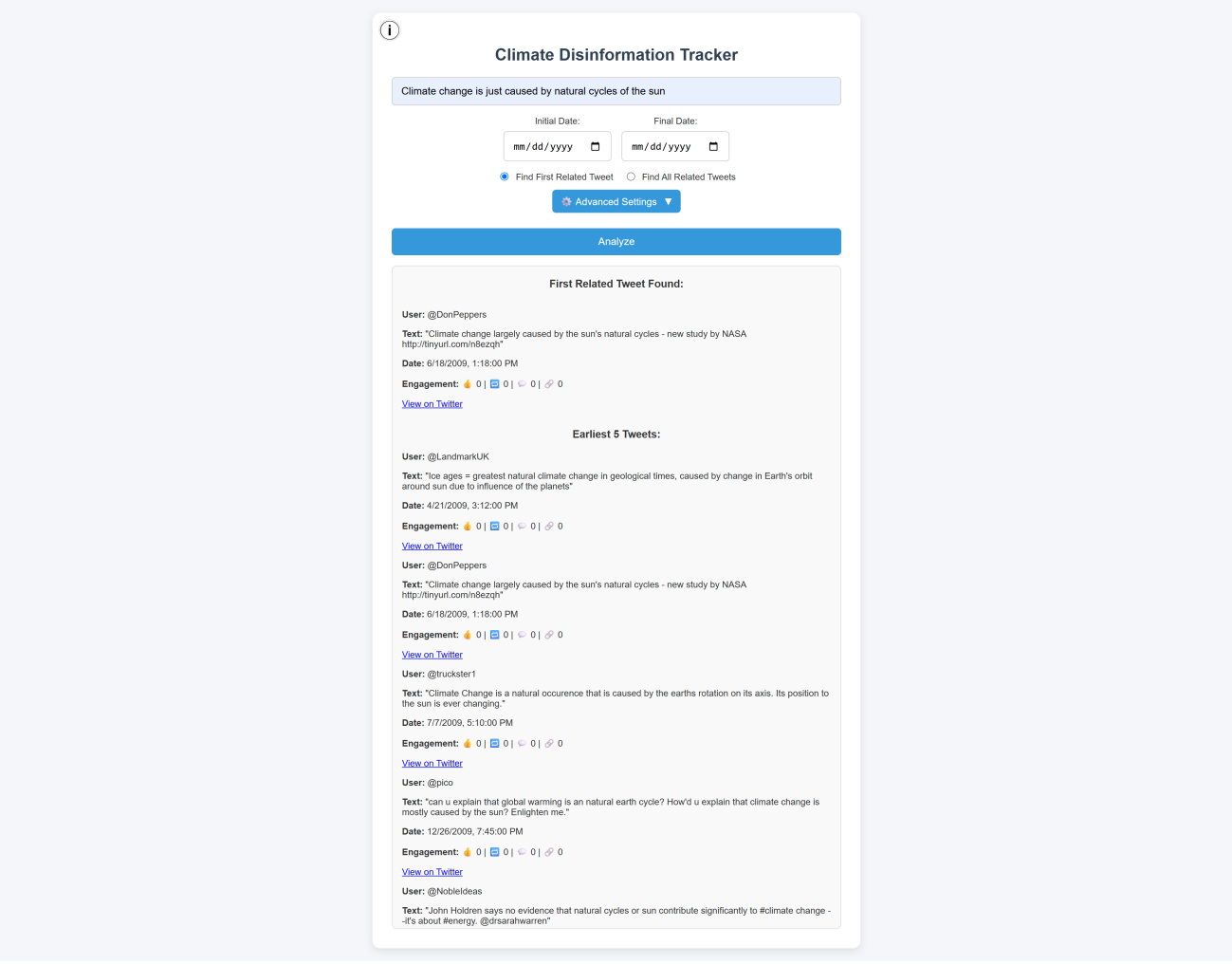


Figure 4: Frontend result with $k = 5$ and input claim *"Climate change is just caused by natural cycles of the sun"*

7.2 Data analysis and visualization

Next to finding the earliest aligning tweet of the user-inputted claim, it is also possible to find all related tweets and visualize them. The tool then extracts the keywords from the claim and retrieves all tweets using the generated query from the given date range (or all time, if no date range is given). After retrieving all tweets, the alignment model is used to predict the alignment for each of the tweets and the tweets with alignment result are saved into a csv file. This file is then used to generate figures with the Plotly (Plotly, 2025c) library, a Python framework for creating interactive visualizations. The results are displayed in the front-end using Dash (Plotly, 2025a), a web application framework built on top of Plotly that enables the creation of interactive dashboards in Python.

7.2.1 Overview

When the tweet retrieval and alignment steps are finished in the back-end, the user is redirected to the visualization overview page. This page can be seen in Figure 5, and the different elements are described below.

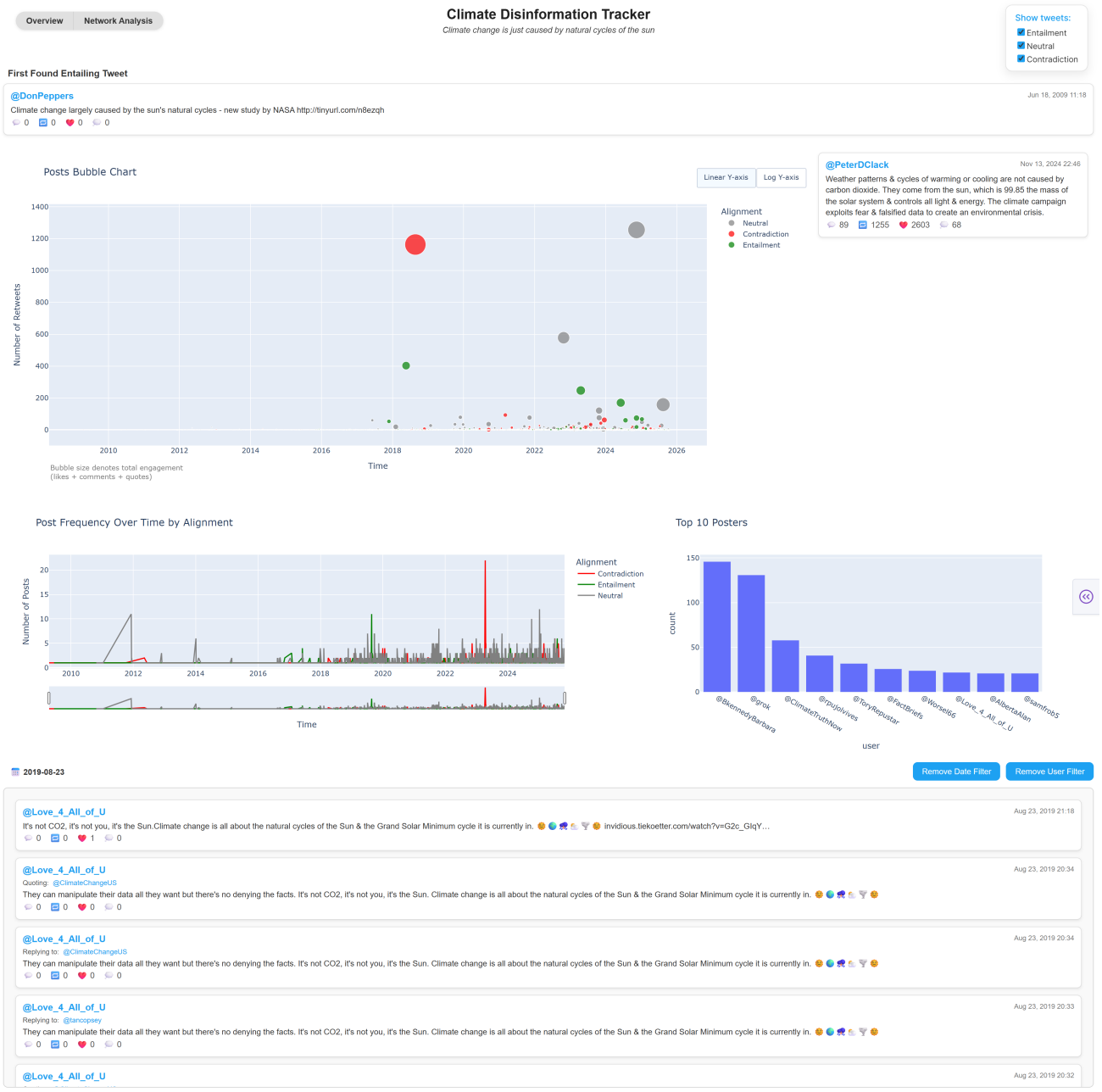
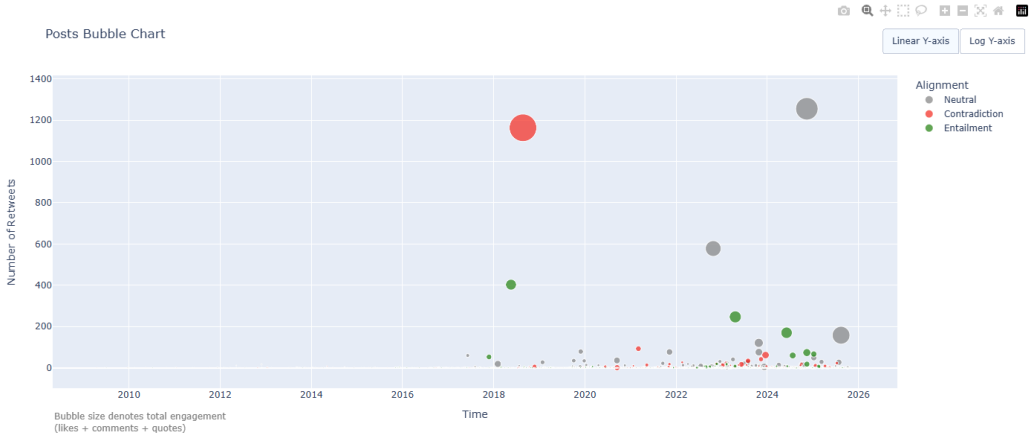


Figure 5: Visualization overview page

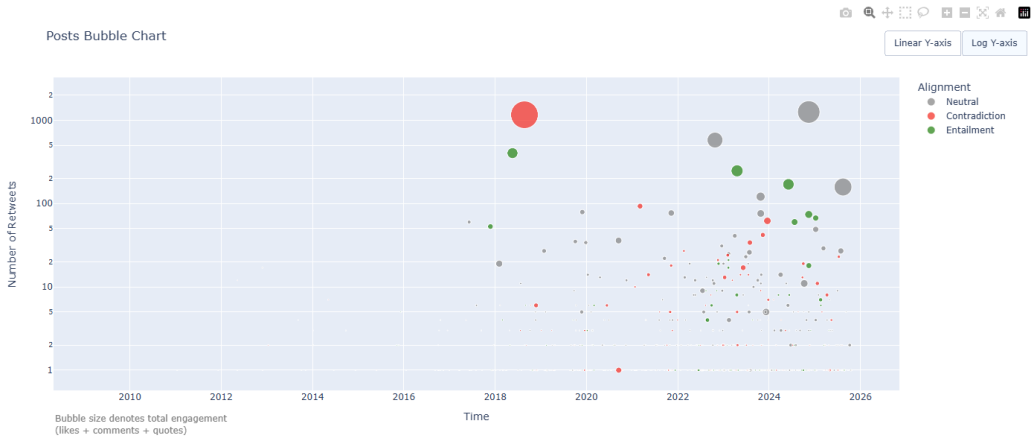
First entailing tweet. On top of the page is the first found entailing tweet. This is the first tweet that is found with the extracted keywords and aligns with the inputted claim by the user.

Bubble chart. The first figure is a bubble chart displaying the retrieved tweets, with the date of the tweet on the x-axis and the number of retweets on the y-axis. The color depicts the outcome of the alignment model for the tweet: green for entailment, gray for neutral and red for contradiction. Lastly, the size denotes the total engagement: number of likes + comments + quotes. Using this figure, users can quickly get an overview of tweets which had a large reach and when they were posted. In the figure, it is possible to switch between a linear (Figure 6a) and a logarithmic (Figure 6b) y-axis. This functionality is implemented because there can be large differences in the number of retweets between the tweets. A linear scale is more intuitive and better for seeing absolute differences and detecting outliers, while a logarithmic scale is more beneficial for relative comparisons for tweets with less retweets when the outliers would otherwise dominate the chart. Lastly, when the user clicks on a bubble in the chart, the specific tweet is displayed on the right side of the figure.

It can be noted that the number of followers is not used in this figure, while it would be a good addition for the engagement and reach of a tweet. The reason this statistic is not shown in the figure is that it would require another call per user to Nitter to retrieve the user data. This means that per tweet, which could be thousands, another call would have to be made to Nitter and the page would have to be scraped to retrieve the number of followers, increasing the processing time even further. Therefore, to minimize the processing time, we opted not to include this statistic in the current version of the tool.



(a) Bubble chart with linear y-axis.



(b) Bubble chart with logarithmic y-axis.

Figure 6: Comparison of bubble charts showing posts over time, with a linear and a logarithmic y-axis for the number of retweets.

Timeline. Below the bubble chart is a timeline with the date on the x-axis and the number of tweets per day on the y-axis. There are three lines in this graph, one for each alignment prediction: green for entailment, gray for neutral and red for contradiction. With this figure, users can see when there were a lot of tweets posted about their input claim, providing an idea of when the spreading of the information happened.

Top posters. On the right side of the timeline graph, there is a bar chart showing the top 10 posters about the claim. This allows users to see which accounts posted the most about the specific topic.

Tweet display. Clicking on a day or a user in the timeline and top posters charts results in the display of the tweets on the specified day by the selected user. This way, the tool allows users to see what kind of tweets were posted which is for example useful for detecting spam accounts that post a lot of the same tweets. It is possible to filter by both date and user, or by one of the two.

Alignment filter. On the top right corner of the page, there is a box to select which kind of tweets to show in terms of their alignment. This allows users to filter tweets based on whether they agree with their input claim or not. This is useful for seeing when certain kinds of tweets were posted and by whom. The filter is applied to all the figures in the visualization, including the network analysis, which is described in the subsection below.

7.2.2 Network analysis

In addition to the figures on the overview page, the visualization part also offers a network analysis of the reply/quote interaction between users. This page is accessible by clicking on the 'Network Analysis' button in the menu in the top left corner of the page and can be seen in Figure 7. To get the full functionality of this analysis, it is important that the replies are not excluded from the search in the advanced settings.

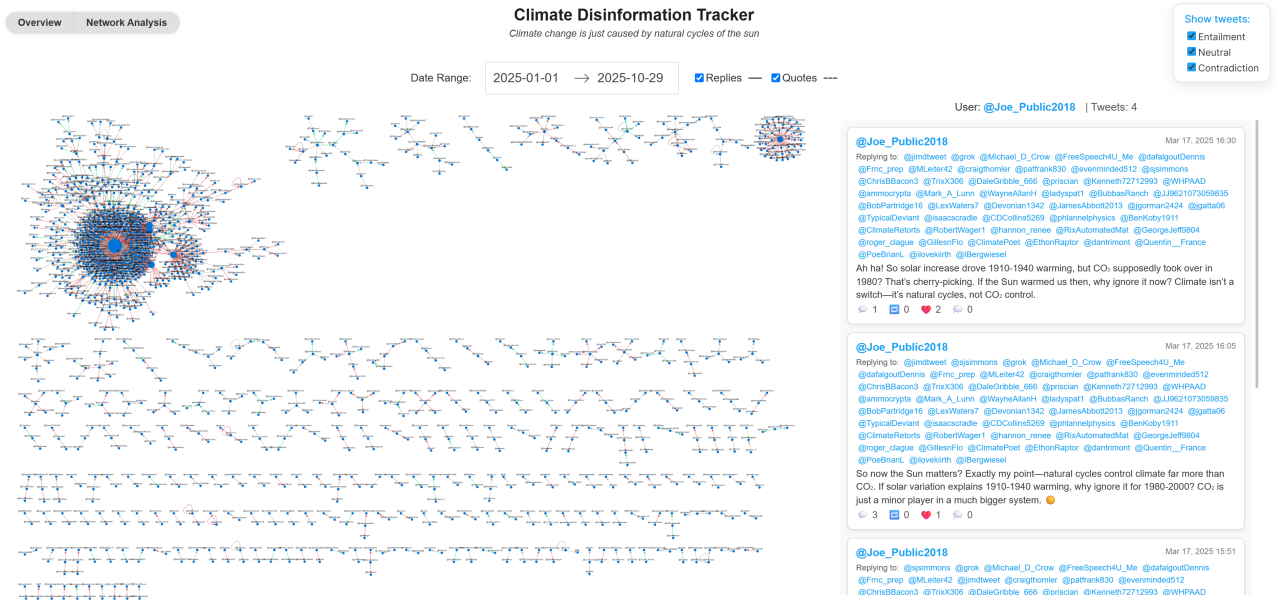


Figure 7: Network analysis page

To visualize the reply/quote interaction between users, a directed graph is constructed using the NetworkX (NetworkX, 2025) library, where each node represents a unique user and edges represent replies/quotes between users. For example, the edge $A \rightarrow B$ means that user A replied to or quoted user B. This graph is then converted into a Dash Cytoscape[8] (Plotly, 2025b) compatible format to be

[8]Dash Cytoscape is a graph visualization component that integrates the Cytoscape.js library into Dash, enabling interactive and easily customizable network visualizations.

displayed using Dash. To distinguish between reply and quote interactions, different arrow styles are used: a solid line for replies and a dashed line for quotes. Next to that, the graph also differentiates between the three alignment classifications by coloring the arrows accordingly: green for entailment, gray for neutral and red for contradiction. To visualize the relative influence of each user, the nodes are sized according to their degree centrality, which is the normalized amount of incoming and outgoing edges of that node. Lastly, the cose layout option in Cytoscape is used to position the nodes based on physics. This produces a visually intuitive layout where clusters become clearly visible, as can be seen in Figure 7. A downside of this layout is that it is computationally heavy to generate this layout. Therefore, if there are thousands of tweets, it might take a very long time to load the graph, or it might even crash.

Similar to the figures on the overview page, clicking on a node reveals the tweets of that user on the right side of the screen and as mentioned before, it is possible to filter on the alignment classifications. Next to that, it is also possible to filter on date range and interaction type.

The reason we opted for a reply/quote interaction network, instead of including retweets as well, is because it was not possible to get the username of the account the made the retweet with the way we are currently scraping Nitter. Only the display name is shown, which can be seen in Figure 8, making it impossible to extract useable data for the network graph since this uses the distinct usernames.

Figure 8: Retweeted tweet in Nitter

8 Testing and Results

Once we had the design for our final solution and its implementation, we proceeded to validate the tool we built for tracking the source of false climate information.

One of the first things we noticed was the difficulty of getting test cases, knowing the source is not publicly available information (otherwise there is no need for this tool) and the source also needs to be within the social media platform. Therefore, we created our own test cases by using tweets that we know should be the source within a confined time range. Since we can define the time range in which to search for the source, our tool identifies the earliest source within that specified period. This means we may not be able to confirm if the tool finds the very first source overall, but we are able to find the first one within the chosen time range.

In this way the procedure to find test cases for our tool comprised the following steps:

1. In the X platform find a random tweet about climate change or even about any other topic (as the tool is not topic dependent and the source tracking method can be applied to different topics).

2. From the original tweet, make a very short *new claim* that is broader in scope and has roughly the same keywords as the original tweet.

3. Input the start date in the interface of the tool as the date that the tweet was published. Therefore by reverse engineering, this guarantees said tweet to be the source within the range.

It is worth noting that given that we are doing search through exact keyword matching, it was important to create test claims that had the same word forms/inflected forms, i.e. there is a need to use the exact version of a word form to find the desired tweet.

By using this methodology we have at our disposal the tweet which should be found by the tool (the original tweet) and the date at which it should be found (the date in which the original tweet was posted).

As parameters for our tool testing we used the input claim (built by us and referred to as *new claim*), the starting date of the search being the date in which the original tweet was posted, a maximum of 5 keywords extracted, and dropping 1 keyword in the query construction.

We first tested without using the synonyms component of our tool, that is, without including in the query search the synonyms provided by the user nor the synonyms model. Later on we did a second round of testing by considering synonyms, this meant that the user could add its own synonyms or use the ones provided by the synonyms model.

For the purpose of our project we did not consider testing with just the synonyms from the synonym model because this would mean testing the accuracy with which the model provides synonyms. But since the user is also able to input its own synonyms it is not critical for this proof of concept to determine how accurate the synonym model works within the tool.

The main aim of the testing of our tool was to check the accuracy, the number of times in which the tool correctly finds the source. This means that the tweet is retrieved *and* is correctly classified as aligned with the *new claim*. We knew that that underlying models (the keyword model and the alignment model) had their own accuracy and that they were trained with generalized data while the intended use of this tool is for climate-related topics. For that reason, our goal was to evaluate how the models work when interconnected with all the other components of our tool and for this specific use case.

There was a total of 31 claims without synonyms for which 68% were found correctly (see appendix D). Most of the times the ones that failed were because the *new claim* contained synonyms and therefore when doing the search scraping, the keywords would not match the original tweet and hence not be

retrieved.

An example of such a test claim that uses as input for the tool words that are synonyms from the original tweet is the following. The input claim is "A vegan diet is more environmentally friendly than a meat diet" while the original tweet states "Oxford scientists found that even the least sustainable plant-based diet was still more environmentally friendly than the most sustainable meat eater's diet - backing up decades of research saying the exact same thing.". This shows that by replacing plant-based with vegan at the input, the tool fails to retrieve the original post.

Given the findings of our first testing, we decided to incorporate synonyms into the implementation of our tool. For this case, 25 claims were tested with the synonym functionality achieving a slightly higher accuracy of 72% (see Appendix E). This demonstrates that the inclusion of synonyms can effectively enhance the tool's ability to identify relevant sources, showing good cohesion between the synonym-based retrieval and the alignment model in most cases. The main reason why the tool fails to find the source tweet in these set of cases is due to the alignment model, as the claims were incorrectly classified as "neutral" or "contradiction". This means that in these cases, thanks to the inclusion of the synonyms the post was able to be retrieved by the scraper but the alignment model still compares the original tweet to the *new claim* which does not use the synonyms added for searching.

An example of this wrong classification is seen in the test case that is classified as a "contradiction" when it uses as input claim: "Moths are actually spies developed by the U.S. military." while the original tweet says "The U.S. military is advancing technology to control moths in flight and use them as spies.".

Additionally, when testing with synonyms there were a couple of cases in which the original tweet was not retrieved even though the keywords match and the characters in the query also do not exceed 500 characters. Searching manually with the same query on Nitter as the one built by the tool does retrieve it, leading us to think that depending on the Nitter domain that is used by the tool there might be different data available.

While testing we noticed a couple of things about the tool:

- The tool is not deterministic due to Nitter scraping: different domains retrieve different data, so there is not always the same output for the exact same input.
- Contractions (and other words with apostrophes) get tokenized as different keywords. For example, the keyword model for the word "don't" extracts it as the keyword "don". In order to overcome this we included in our implementation a step that expands contractions.
- The keyword search is sensitive to hyphens. For instance, when using "worst-case" and "low-meat" in the input claim to the tool, even though this is the correct way to write them, the original tweet would not be found given that it used the same words but without the hyphens.
- The tool is sensitive to word forms/inflected forms.
- The posts in Nitter are stored with dates that are in coordinated universal time but the user query includes their own time zone. As a consequence, there are conflicts with a couple of tweets that might be stored in the database in a different date from the perspective of the user. For these cases, adding as starting point a day earlier than the original post solved the issue.

To conclude our testing, we found that the tool performed reliably in most cases, achieving an improved accuracy of 72% when using the synonym functionality. This shows that the addition of synonyms effectively enhanced the retrieval process and allowed the tool to identify relevant source posts more consistently. However, due to the nature of searching by keywords, the tool is highly sensitive to anything that modifies the words that are used for scraping; whether that is contractions, hyphens, word forms, or synonyms. In addition, one of the biggest bottlenecks remains to be the alignment

model, as it can misclassify certain claims even when the correct posts are retrieved. Overall, these results demonstrate the potential of our approach while also highlighting areas for further refinement in classification accuracy and robustness.

9 Risk Assessment

A tool designed to identify the source of misinformation can introduce several risks, both for individual users and for society as a whole. These risks primarily concern the accuracy of the tool and the potential to misuse the tool.

9.1 Incorrectly identifying source of information

The ultimate goal of the tool is to help people find the source of certain claims on platform X. However, the first retrieved post is not necessarily the true source of the claim. Posts and/or accounts can be deleted, posts can be misclassified by the alignment model, or they could originate on different platforms entirely. Therefore, the tool may incorrectly identify a tweet as the first entailing post, potentially leading to misleading conclusions.

This inaccuracy poses risks for both society and users of the tool. For society, the risk is that innocent people can be falsely blamed for spreading misinformation. Being falsely accused is unwanted and can have negative impact on those individuals. A less harmful, but still important risk is for the users of the tool. They can be misdirected to look a certain way for the source, while the true source remains hidden.

To mitigate these risks, three strategies have been implemented. Firstly, next to only returning the first found entailing tweet, it is also possible to return the earliest k tweets (explained in subsection 7.1), regardless of their alignment classification. This allows the user to also see tweets that might have been misclassified by the alignment model. Secondly, the tool also offers the option to create a visualization of tweets (explained in subsection 7.2), allowing the user to not only see the first found entailing tweet, but also other tweets about the same topic. This way, users can also see tweets that are not classified as entailing by the alignment model and see how much reach certain tweets have (by the amount of retweets/likes/comments). With this extra information, we hope to give extra context and inform users more about the spread of misinformation claims, mitigating the risk of drawing incorrect conclusions about the source. Next to that, the tool has disclaimers about the usage and accuracy (which can be seen in Figure 2), and we do not say that the first found tweet is the source, but we just display it as the 'first found entailing tweet'. With these measures, users of the tool are informed about the limitations of the tool and are encouraged to draw conclusions critically and responsibly.

9.2 Misuse of tool

Another significant risk of the tool is that it can be misused. Because the tool is open source and publicly available, it may not only be used by individuals who want to use it for good reasons but also by individuals or organizations with malicious intent, such as the spreaders of disinformation. This again has several risks for both society and users of the tool.

There are two identified misuse scenarios that are worth mentioning. Firstly, people or organizations can use it to find the spreaders of claims that do not align with their ideas and harass or intimidate them to stop spreading this information. Another way that the tool could be misused is that malicious actors can orchestrate the spread of a certain claim in such a way that it will not be traceable back to the true source.

Both of these cases undermine the goal of creating transparency and understanding within society; therefore, we have implemented several mitigation strategies. Firstly, as mentioned in the previous subsection, we have disclaimers in our tool about the usage and accuracy of the tool, and we encourage users to do further research about the users/posts before drawing conclusions. Next to that, the open-source nature of our tool also allows users to understand the limitations and thus anticipate the misuse possibilities of our tool.

With these measures we cannot prevent the tool from being misused, but we aim to encourage users of the tool to interpret the results critically and perform their own research into their findings.

10 Ethical Considerations

During the development of the disinformation tracker tool, several ethical considerations have been taken into account. The goal of the tool is to promote transparency and provide insight into how information spreads online, but this goal must be balanced against individuals' rights to privacy and freedom of speech.

10.1 Privacy vs. transparency

One of the main considerations is the balance between creating transparency and respecting privacy. The tool aims to increase transparency by identifying where specific claims come from and how they spread across platform X. However, this analysis relies on data generated by individuals, which introduces privacy concerns. While the tool makes use of publicly available data from X, it processes it and presents new insights to users of the tool.

By using X, users agree to the Terms of Service (X, 2025b) and the privacy policy (X (formerly Twitter), 2024), which describes how user data is collected and shared. Section 3 in the privacy policy specifies how the user data is shared; in section 3.1 it is mentioned that the information is available to the general public (unless a user has a private account), and in section 3.2 it is stated that the data is shared through the X API, which can be used for analysis purposes. However, our tool does not gather the data through the API, so we would not fall under the conditions of section 3.2, but section 3.1. This places our tool in a gray area: the data is public, but the way it is collected and processed differs from X's intended usages as described in their privacy policy.

Next to the concerns about respecting X's privacy policy, we also took GDPR compliance into account. Since this is a complex topic, our coach connected us to a legal professional who could provide us with advice. They assured us that there is no problem since we are using publicly available tweets. However, if the tool would be deployed on a server in the future, it is important to reevaluate the GDPR compliance since it would then process data on this server.

Despite the privacy concerns, we think the value of creating transparency and understanding weighs up against the ambiguity of respecting X's privacy policy. The tool has a valuable role in creating more transparency about the spread of online (mis)information, helping society with verifying spreaders of certain claims and becoming resilient against disinformation. Nevertheless, privacy remains an important topic; therefore, we do not process any data that is not needed for the analysis or not publicly available, and the tool is made open-source, so individuals can see how the data is being used.

10.2 Freedom of speech

Another concern is that people might feel a suppression of their freedom of speech due to the existence of our tool. The knowledge of the existence of a tool to trace the spread of information can create a fear of being falsely accused of spreading disinformation. This idea can hold people back from publicly posting their opinions or even asking questions, which is an undesired effect. Therefore, it is important to note that the tool does not aim to differentiate between mis- and disinformation, nor does it detect whether a tweet is false or true. The tool is created to be as neutral as possible, without drawing conclusions about posts and users. This is clearly stated in the documentation and instructions of the tool to avoid misinterpretation and prevent a feeling of censorship.

10.3 Usage of Nitter

The last ethical concern is about the usage of Nitter (Zedeus, 2025). As mentioned in section 5, it is against X's Terms of Service (X, 2025b) to scrape the platform to retrieve data. Due to this and the high prices of using X's API, the decision was made to scrape Nitter. While this does make the tool technically adhere to X's Terms of Service, it is debatable whether this is ethical, since Nitter is doing the 'dirty work' for us and we are making use of this. Next to this, the privacy concern arises due to

the usage of Nitter. If the tool made use of the X API, it would fall under the provisions of X's privacy policy, making it easier to justify the processing of user data.

Next to the ethical concerns of Nitter, there are also practical implications of the alternative front-end, which are described in section 11. However, since the development of this tool is still in the early stages and considering the budget limitation, using Nitter for this project is still the best option despite its limitations.

11 Limitations

The development of the Climate Disinformation Tracker tool is a proof-of-concept. This means while working as expected the tool has some key limitations.

11.1 Technical and Data-Sourcing Constraints

- **Dependence on Nitter for Data Scraping**: The tool relies entirely on the availability and functionality of Nitter, an unofficial alternative front-end for X. This introduces several critical limitations:
  - **Non-Deterministic** - As mentioned in section 8, the tool is not deterministic because of Nitter. This means different Nitter domains retrieve different data, so the same input on domain A could result in different output for Domain B.
  - **Rate Limit and Availability** - The tool is depended on Nitter which in turn is depended on X. This means that Nitter can be (temporarily) blocked by X or the tool can be (temporarily) blocked by Nitter, which means the tool won't work (for a period of time).
  - **Grey Area** - The tool operates in a grey area, this is because it is against X terms of service to scrape but the tool scrapes from Nitter which gets their data from X.
- **Accuracy Bottlenecks of Core Models**: The overall accuracy of the tool is limited by the performance of the underlying models:
  - **Alignment Model Error** - The `mDeBERTa-v3-base-mnli-xnli model`, while robust, operates with an approximate 90% accuracy. A single misclassification (labeling an "entailing" tweet as "neutral" or "contradiction") can lead the tool to miss the true source and return a later post as the "earliest entailing tweet."
  - **Keyword Sensitivity** - The initial keyword search remains highly sensitive to linguistic variations, including hyphens, contractions, and inflected word forms. Despite the synonym component, the effectiveness of retrieval relies heavily on the quality of the user's initial claim or the synonym model's context-matching.
  - These limitations are partly covered by allowing the user to interact at each step in the process, so users can also add synonyms to decrease the keyword sensitivity for example.
- **Query and Character Limits**: The boolean search query constructed for Nitter scraping is constrained by a 500-character limit. For complex or highly detailed claims, this limitation can prevent the creation of comprehensive search queries that would otherwise improve retrieval success.

11.2 Operational and Scope Limitations

- **Single-Platform Scope**: The tool is limited exclusively to platform X. Disinformation often originates or is simultaneously amplified on other platforms (e.g., Facebook, Reddit, forums), creating a gap between the earliest entailing tweet found by the tool and the actual, global source of the claim.
- **Time Zone Discrepancies**: Posts in the Nitter data are stored using Coordinated Universal Time (UTC), while the user inputs their local time zone. This conflict introduces ambiguity around the exact date and time of posts, particularly those at the beginning or end of the search window, potentially leading to the misidentification of the true earliest source.
- **Network Analysis**: The network analysis uses a computationally heavy layout to generate, this

means that it takes time to load the entire graph.

12 Conclusion

The Climate Disinformation Tracker represents a step toward improving transparency in the digital information landscape, particularly regarding climate change disinformation. By using natural language processing and machine learning models, the tool enables users to trace the earliest verifiable online occurrence of climate-related claims and draw their own conclusions via the data visualisation dashboard. Its architecture demonstrates the feasibility of automating claim tracing even under restrictive data-access conditions, integrating keyword extraction, synonym-aware query construction, Nitter scraping, and claim alignment classification.

The results of our evaluation indicate that while the tool performs effectively in retrieving via synonym expansion and classifying tweets that align with given claims, the true accuracy of the tool of 72% is constrained by the limitations of both Nitter data retrieval and the accuracy of the underlying models. Moreover, the tool's single-platform focus and minor inconsistencies related to time zones limit its ability to capture the broader, cross-platform dynamics of disinformation spread. Nonetheless, the achieved performance validates the potential of our approach as a proof of concept for scalable misinformation tracking systems. The tool should serve to provide valuable insights for investigative journalists and organisations seeking to strengthen their resilience against disinformation campaigns, thereby indirectly aiding authorities in addressing one of humanity's most urgent global challenges.

13 Recommendations

As with any proof of concept system, this work opens doors to a multitude of avenues for future development. Following we distinguish between research extensions and technical enhancements, as well as concrete actions the Dutch National Police can take to translate this proof of concept into a deployable solution.

13.1 Research and Technical Extensions

Due to the short duration of the project, we limited ourselves to the spread of climate disinformation on X. We developed a tool that serves as a proof of concept of a climate disinformation tracker. However, in order to make the tool even more valuable for investigative journalists or the police, misinformation on different social media platforms and even the internet should be considered. This would reduce the gap between the earliest entailing post and the actual source.

With regard to potential future improvements within the Nitter implementation, a more advanced model could be employed to generate synonym suggestions, thereby reducing the time and effort required by users. It is important to note that more powerful models would be heavier to install on a laptop and could increase computation time. If budget is not a constraint, integrating the OpenAI API should also be considered. Despite these suggestions, we believe that investigative journalists possess a better intuition for selecting appropriate synonyms, which is why synonym suggestion is not particularly crucial.

Moreover, further research could be done regarding the optimal search strategy of the earliest entailing tweet. We tested a linear and binary approach on a limited dataset of claims, for which the linear approach was proven to be generally faster. However, testing both strategies for a larger dataset with a greater diversity of tweets distribution would allow drawing more general conclusions, and perhaps show cases where the binary approach is beneficial. Additionally, other search strategies could be explored, such as adaptative methods which could modify the search window depending on the volume of tweets.

In relation to data analysis and visualizations, a potential enhancement would be to incorporate the amount of followers into both the bubble chart and the network analysis. This addition could substantially improve the user's ability to extract meaningful insights. Additionally, integrating retweets into the network analysis would provide a more comprehensive representation of how information propagates on X through time.

Furthermore, claims addressing different aspects of climate change could be analyzed collectively to identify common accounts across the datasets. This approach could potentially serve as a means to uncover organizations that spread disinformation for their own benefit. We hypothesize that an individual may unintentionally share misinformation about a specific aspect of climate change. However, if an account repeatedly spreads misinformation across multiple topics, it is likely that an organized entity is operating behind it, seeking to manipulate public opinion, undermine trust in science, and delay the implementation of policies aimed at mitigating global warming.

13.2 Logistical and Organisational Steps

In addition to extending the tool through the previous section, following are logistical steps we have identified that can aid the Dutch National Police to extend our proof of concept into a fully operational tool:

The tool is currently available in a public Github repository[9], which can be cloned to run locally. Although we do include clear instructions on how to install the necessary packages and launch the app, this process might not be straightforward for every end-user. Ideally, to enhance usability and accessibility, the tool should be hosted on a server so that users do not have to install the requirements

[9]Link of the Github repository: https://github.com/vincentvvliet/climate-disinformation-detector.

themselves. While we have developed the front end for the tool, it has not yet been hosted due to the scope of the project, but it may be in the future.

We also recommend developing a sustainable funding and maintenance model to ensure the tool remains operationally up to date. This can include aspects such as public funding, institutional partnerships, as well as a governance framework that covers maintenance, compliance with frameworks, and handling bug reports of the tool.

Finally, pilot sessions and workshops can be conducted with individuals using our system that are closely aligned with the Dutch National Police, such as investigative journalists, police analysts in order to familiarise them with the tool as well as test the tool in real-world scenarios with real end-users. We strongly believe that this is the most effective way to improve the tool. Such an approach would help identify technical bottlenecks and ensure future releases align further with the needs of the stakeholders.

14 Team Reflection

We started this project on the 1st of September. In the first and second week we immediately were able to connect and we even met after the workshops. This really helped with the team bonding which made the project a lot more fun and progress extremely smooth. Everyone was always helpful and actively engaged in discussions and nobody felt scared to share their perspective.

In the second and third week, we started with the preparations, which included reading relevant sources, setting up the project, etc. Our overall workflow was really flexible, we met every week and set goals for each week, basically a simple sprint implementation that fit the project. During these weeks we also had the problem statement presentation. It went well, and we received a lot of feedback. Unfortunatey, we never it received digitally (via BrightSpace), so we only remembered and wrote down parts of the received feedback. During these weeks we also designed the system architecture, and consequently found it aligning with the third learning objective (LO3) of JIP: Being able to research, design (and eventually model) technical approaches for the problem proposed. Furthermore, the research led us to discovering some ethical and societal consequences of our project, helping us align with the second learning objective (LO2) as well.

In the fourth and fifth week we started with the actual development of the tool, and we quickly encountered a big problem: X's API is really expensive, so we needed to determine alternative options as discussed in subsection 5.2. Overall, we reacted quickly and calm to the situation, we divided the work equally, everybody researched an alternative solution and by the next day we already solved the problem. We then decided to redesign our architecture - where we ran into differences in opinion. This was ideal as we were able to offer perspectives as honed by our own studies. Some offered technical insights, some offered management insights, some offered logistical insights. Eventually, we were able to successfully reach an architecture that everyone agreed on. Observed during multiple weeks, we were able to achieve the fourth learning objective (LO4) in this manner. During the work split, most of the project was related to software, however, we were all able to explore and work on aspects we hadn't before. Luckily, most of us had an undergraduate background in computer science, and a masters in differing fields allowing a mutual understanding of most software related aspects, while accounting for differing perspectives where needed as honed by our masters. In relation to LO4, although we attempted to set up a few meetings with stakeholders, we were unfortunately unable to schedule them successfully due to unavailability. Therefore, although limited on that front, we were able to get the professional opinion of different stakeholders on not only the usability of the product, but also for example certain aspects of the project - such as legal ones.

Week six and seven were about finishing the main features of the project and preparing for the midterm presentation and report. It was during this time that the project workload was intensive - keeping everyone busy with developing specific features. We started on time and were able to deliver everything without problems. One thing we changed compared to the problem statement presentation is that not everybody was presenting because that felt unnecessary, and in this way somebody would be taking notes during the presentation and question round. Therefore, this time we did have all the feedback written down and were able to use it.

In the last three weeks we finalised our project, worked on optimisation and added some new features such as synonym incorporation and visualisation in the frontend which elevated the project to the next level. These implementations were based majorly in feedback we received during the midterm presentations. During this week was when we were able to see our interdisciplinary aspect the most. Work during this week was divided based on efficiency, and each member was able to work on what they had experience with - contributing their skillset to our project, unlike what we did in the beginning - where some of our members explored with domains they had not worked with before. In this manner, throughout the project - we were not only able to work on what we were confident in, but also tried out new skillsets, helping us achieve the first learning objective (LO1) as well.

The team blogs every other week helped keep all of us up to date with the progress. The BuddyCheck

also served as a moment for us to express our feelings - good and bad - towards our teammates. This was helpful as this proved to be a reflection reminder every now and then. The personal blogs added further introspection to this, allowing each team member to be aware of and work on (if need be) certain skills. This helped with awareness and reassurance. In this manner, we were also able to achieve the fifth learning objective (LO5) of JIP.

Overall, we are very pleased with the final product and our approach. We worked really well because we trusted each other and everybody communicated clearly and on time, so we could handle problems on time. Besides this, everybody was very open-minded, willing to learn and open for discussions. However, this can also sometimes be a double edged sword as we sometimes discussed for too long. For example, we had a one-hour meeting to distribute who would check which section of the final report. During this meeting, in the first 20 minutes we distributed the work and the other 40 was discussing parts of the report. While this was very valuable, the time spent was quite a lot since we had a planned deadline to finish proofreading the report.

For a timeline of the project see Appendix F.

References

(2025). Consequences of climate change — climate.ec.europa.eu. https://climate.ec.europa.eu/climate-change/consequences-climate-change_en. [Accessed 08-09-2025].

(2025). Content library - meta content library and api - documentation - meta for developers. [Accessed 08-09-2025].

AIDA-UPM (2024). mstsb-paraphrase-multilingual-mpnet-base-v2: Pre-trained model. Hugging Face model hub. Retrieved from https://huggingface.co/AIDA-UPM/mstsb-paraphrase-multilingual-mpnet-base-v2.

Aïmeur, E., Amri, S., and Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Amazon Web Services (AWS) (2025). What is RESTful API?

Archive Team (2012). Archive team: The twitter stream grab. Internet Archive dataset; JSON "spritzer" sample of global Twitter stream.

Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76:102256.

Balthus23Air (2025). New book on the status of the climate issue – the world's biggest problem or just a hoax? Accessed: 2025-09-08.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71. Open Multilingual Wordnet.

Bonnevie, E., Sittig, J., and Smyser, J. (2021). The case for tracking misinformation the way we track disease. *Big Data & Society*, 8(1):20539517211013867.

Boston University (2025). Research identifies origins of climate misinformation crisis: Fossil fuel companies. https://www.bu.edu/met/news/research-identifies-origins-of-climate-misinformation-crisis-fossil-fuel-companies/. [Accessed 08-09-2025].

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.

Explosion (2025). spaCy 101: Everything you need to know. https://spacy.io/usage/spacy-101. Accessed: 2025-10-27.

Fallis, D. (2014). The varieties of disinformation. *The philosophy of information quality*, pages 135–161.

Foundation, C. (2025). New book on the status of the climate issue – the world's biggest problem or just a hoax? Accessed: 2025-09-08.

Franta, B. (2021). Early oil industry disinformation on global warming. *Environmental politics*, 30(4):663–668.

Gayle, D. (2024). Uk government adviser on disruptive protest accused of conflict of interest. *The Guardian*. Accessed: 2025-09-08.

Grefenstette, G. (1999). Tokenization. In *Syntactic wordclass tagging*, pages 117–133. Springer.

Grootendorst, M. (2020). Keybert: Minimal keyword extraction with bert.

Heffernan, A. (2024). Countering fossil-fuelled climate disinformation to save democracy.

Hunt, K., Agarwal, P., and Zhuang, J. (2022). Monitoring misinformation on twitter during crisis events: a machine learning approach. *Risk analysis*, 42(8):1728–1748.

Joosten, T. and Keizer, P. (2020). Klimaatsceptisch nederland profiteert nog altijd van netwerk en geld uit fossiele industrie. *Follow the Money*.

Laurer, M., Van Atteveldt, W., Casas, A., and Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Lewandowsky, S. (2021). Climate change disinformation and how to combat it. *Annual review of public health*, 42(1):1–21.

MacCartney, B. (2009). *Natural language inference*. Stanford University.

Meta Platforms, Inc. (2025). Threads api: Keyword search. Developer documentation.

Microsoft (n.d.). Playwright python api reference: Playwright class. https://playwright.dev/python/docs/api/class-playwright. Accessed: 2025-10-29.

Milman, O. (2023). Revealed: Exxon made 'breathtakingly' accurate climate predictions in 1970s and 80s. https://www.theguardian.com/business/2023/jan/12/exxon-climate-change-global-warming-research. [Accessed 08-09-2025].

Nations, U. (2025). What Is Climate Change? — United Nations — un.org. https://www.un.org/en/climatechange/what-is-climate-change. [Accessed 08-09-2025].

NetworkX (2025). Networkx: Software for complex networks. Accessed: 2025-10-29.

Plotly (2025a). Dash: A web application framework for python. Accessed: 2025-10-29.

Plotly (2025b). Dash cytoscape: A graph visualization component for dash. Accessed: 2025-10-29.

Plotly (2025c). Plotly: Python graphing library. Accessed: 2025-10-29.

Princeton University (2010). About WordNet.

Reddi, L. T. (2023). Stakeholder analysis using the power interest grid. Accessed: 2025-09-08.

Revkin, A. (2015). Exxon knew about climate change in 1981, email shows. *The Guardian*.

Richardson, L. (n.d.). Beautiful soup 4 documentation: Pulling data out of html and xml files. https://beautiful-soup-4.readthedocs.io/en/latest/. Accessed: 2025-10-29.

Sadasivam, V., Udhaya, T., Saravanan, K., and Keerthana, S. (2024). Truth tracker: Unveiling fact from fiction in social media. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–4. IEEE.

Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.

Villar-Rodríguez, G., Huertas-García, Á., Martín, A., Huertas-Tato, J., and Camacho, D. (2025). Distrack: A new tool for semi-automatic misinformation tracking in online social networks. *Cognitive Computation*, 17(1):12.

Villar-Rodríguez, G., Álvaro Huertas-García, Martín, A., Huertas-Tato, J., and Camacho, D. (2024). Distrack: a new tool for semi-automatic misinformation tracking in online social networks.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

X (2025a). X api v2 introduction. Developer documentation.

X (2025b). X terms of service. Terms of Service documentation.

X (formerly Twitter) (2024). X Privacy Policy.

Zedeus (2025). Nitter: Alternative twitter front-end. Accessed: 2025-09-25.

Zedeus and contributors (2025). Nitter: Alternative twitter front-end. Accessed: 2025-09-25.

A Appendix Codebase

The entire codebase can be found on GitHub and is publicly available.

https://github.com/vincentvvliet/climate-disinformation-detector

B Appendix Stakeholders

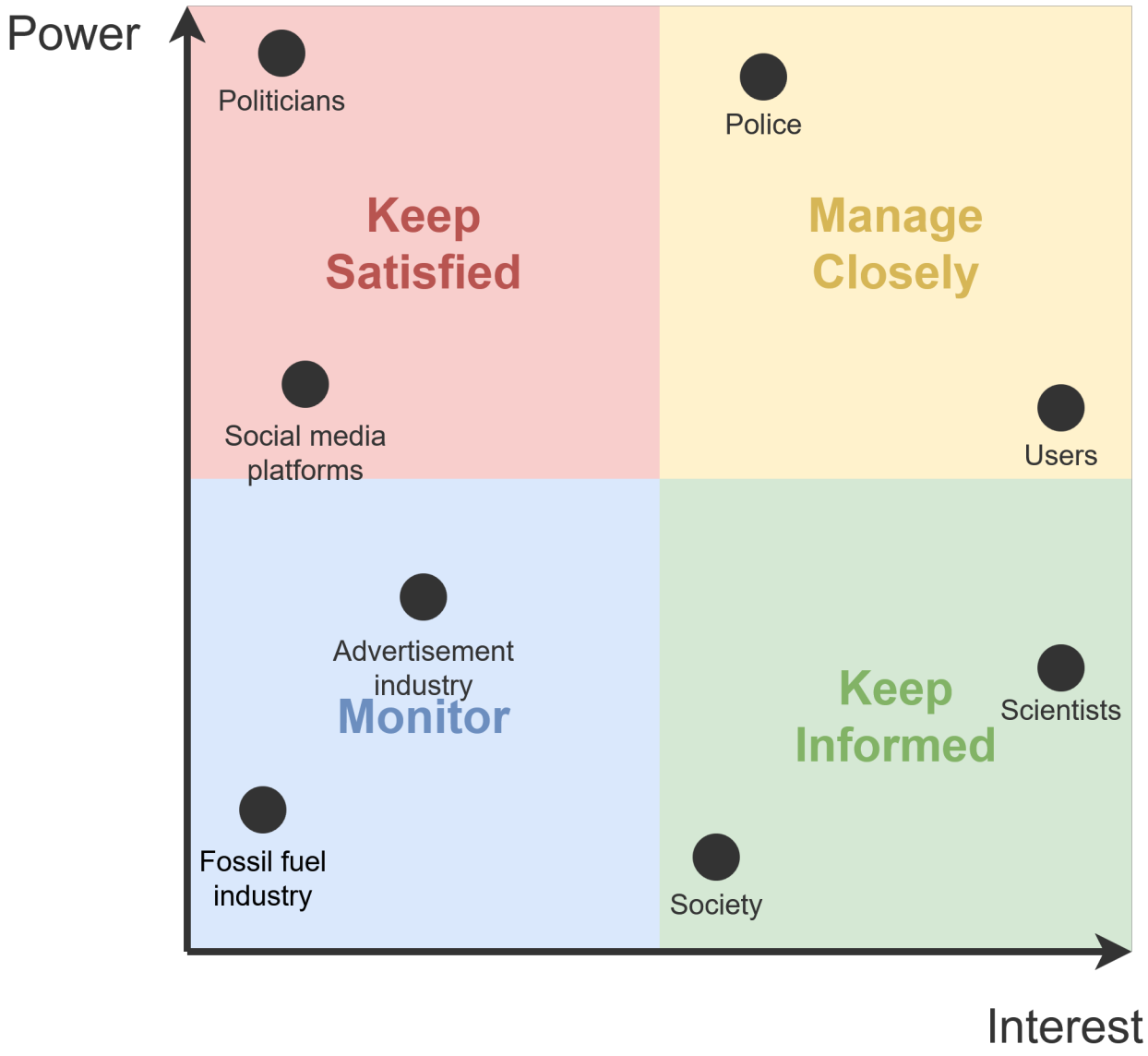


Figure 9: Power interest grid

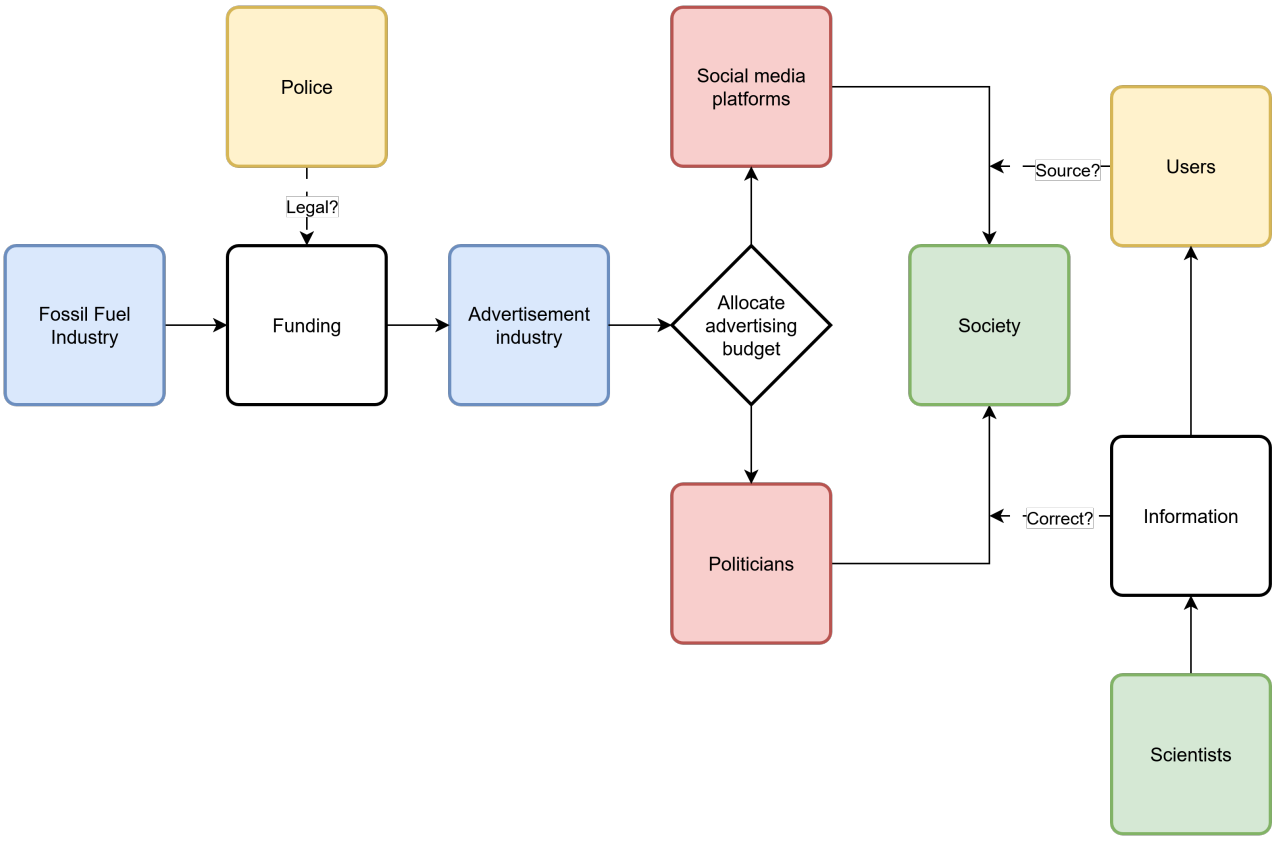


Figure 10: Flow chart of the current business model and where the tool can effect it

C Appendix Possible Solutions


```
[
    {
        "word":"automobile",
        "score":9037
    },
    {
        "word":"machine",
        "score":7053
    },
    {
        "word":"auto",
        "score":6044
    },
    {
        "word":"gondola",
        "score":1042
    },
    {
        "word":"motorcar",
        "score":1020
    },
    {
        "word":"railcar",
        "score":1018
    },
    {
        "word":"cable car",
        "score":26
    },
    {
        "word":"railroad car",
        "score":13
    },
    {
        "word":"elevator car",
        "score":10
    },
    {
        "word":"railway car"
    }
]
```


Figure 11: Output of Datamuse with ”car”

D Appendix Testing claims without synonyms

Table 1: Verification Results: Claims With Synonyms

Claim (Input to Tool)	**Found Source (0/1)**
environment models can't replicate how clouds cool the earth	1
Even in worst case scenarios, the environmental footprint of low meat diets is much, much better than high meat consuming diets.	1
An Oxford-led study found that even in worst case scenarios, vegan and low meat diets produce significantly lower environmental impacts than diets high in meat consumption.	1
In 2008, Wall Street's greed crushed the world—millions lost homes, the men who caused it made billions and walked free.	1
Wall Street and Big Banks triggered the 2008 financial crisis, but instead of holding them accountable, the government bailed out the 1%—leaving the 99% to suffer the consequences. Now, with the derivatives bubble growing again, another catastrophic crash may be inevitable.	0
The financial crisis in 2008 is all the fault of Wall Street and having that 1% bail out is so unfair	0
Wall Street's risky bets crashed the economy in 2008; the 1% got bailed out, the 99% paid—and it could happen again.	0
A plant-based diet is more environmentally friendly than a meat diet	1
A vegan diet is more environmentally friendly than a meat diet	0
Microneedle patches combine mRNA delivery with permanent quantum dot markers for use as vaccine passports.	1
The U.S. military is developing technology to remotely control moths for surveillance purposes.	0
Moths are actually spies developed by the U.S. military.	0
Moths are actually spies that the U.S. military use.	1
NASA data shows no global warming in 8+ years, despite 475 billion tons of CO2.	1
Fires are driven by weather and poor land management—not climate change or slight temperature rise.	0
Fires are not driven by climate change or slight temperature rise.	0
Western media did not report the soviets moon landing, because it was so fake-looking	1
Fox news is propaganda for the republican party	1
Covid-19 vaccines contain graphene oxide particles	1
The assassination attempt on Donald Trump was staged for publicity.	1
carbon dioxide is good for plants	0
Global warming is a hoax. In the past it was much hotter than today.	1
Elvis Presley is still alive. Pastor Bob Joyce and him are the same person.	1
Taylor Swift is a psyop asset from the Pentagon.	1
Climate change does not cause more frequent or intense hurricanes	1
Vaccines destroy immune function and can cause harm and death.	1
Higher CO2 levels are beneficial for the environment	1
Humans emit more CO2 than volcanoes	1
Tom Cruise is one of the highest ranking members of scientology	1
Climate change is a pagan religious ideology	0
Climate change narrative is about centralised control	1

E Appendix Testing claims with synonyms

Table 2: Verification Results: Claims With Synonyms

Claim (Input to Tool)	**Found Source (0/1)**
Climate simulations can't reproduce how clouds regulate the planet's temperature.	1
Climate simulations can't replicate how clouds cool the earth	1
Even in extreme case situations, the ecological footprint of low meat diets is much, much better than high meat consuming diets.	1
An Oxford-conducted research found that even in extreme case scenarios, plant based and reduced meat diets produce significantly lower ecological impacts than diets high in meat consumption.	0
A vegan diet is more environmentally friendly than a meat diet	0
A vegetarian diet is more eco-friendly than a meat-heavy diet.	0
The U.S. military is developing technology to remotely control moths for surveillance purposes.	0
Moths are actually spies developed by the U.S. military.	0
The Reptilians live among us. Experts in deception, they use mental and bodily disguises to infiltrate civilization at the top levels.	1
Phone use before sleep may be your biggest rest error. Blue light reduces melatonin and makes falling asleep tougher.	1
Western media did not report the soviets lunar landing, because it looked so staged.	1
Fox News is disinformation for the Republican organization.	1
The corona vaccine contains graphene oxide molecules	0
The murder attempt on Donald Trump was orchestrated for publicity.	1
carbon dioxide is beneficial for plants	1
Global warming is a fabrication. It has been much warmer in the history of the earth than today.	1
Elvis Presley is still living. Pastor Bob Joyce and him are identical.	1
Taylor Swift is a psychological operative from the department of defense.	0
Climate change does not lead to more frequent or powerful hurricanes.	1
Vaccines ruin immune function and can cause injury and death.	1
Satellites are balloons floating in the atmosphere not space	1
Climate consensus is debunked as propaganda	1
Electric vehicles show highest toxicity on ecosystem	1
Mt. Etna spewed more CO2 in the atmosphere than humans on the planet	1
The existence of human beings is injurious to earth	1

F Appendix Team Reflection

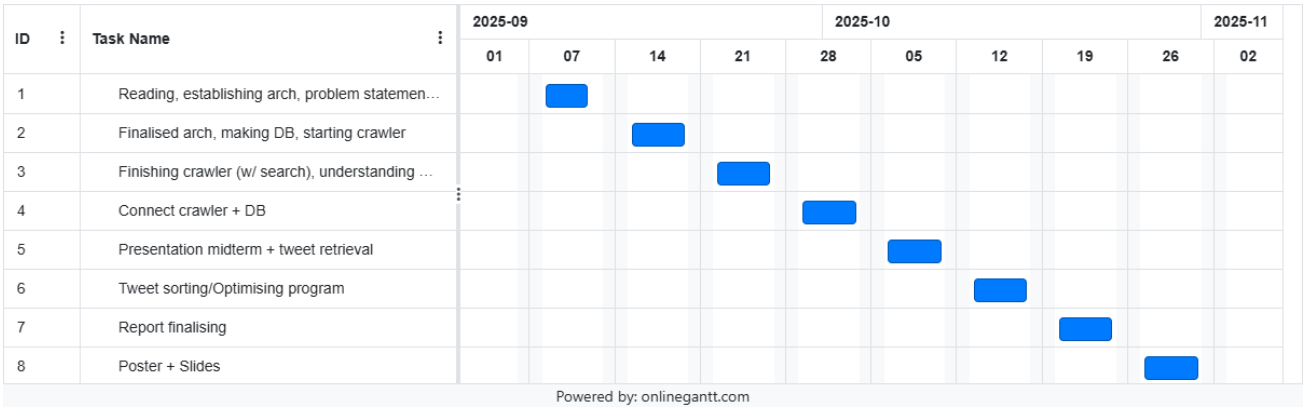


Figure 12: Timeline Problem Statement

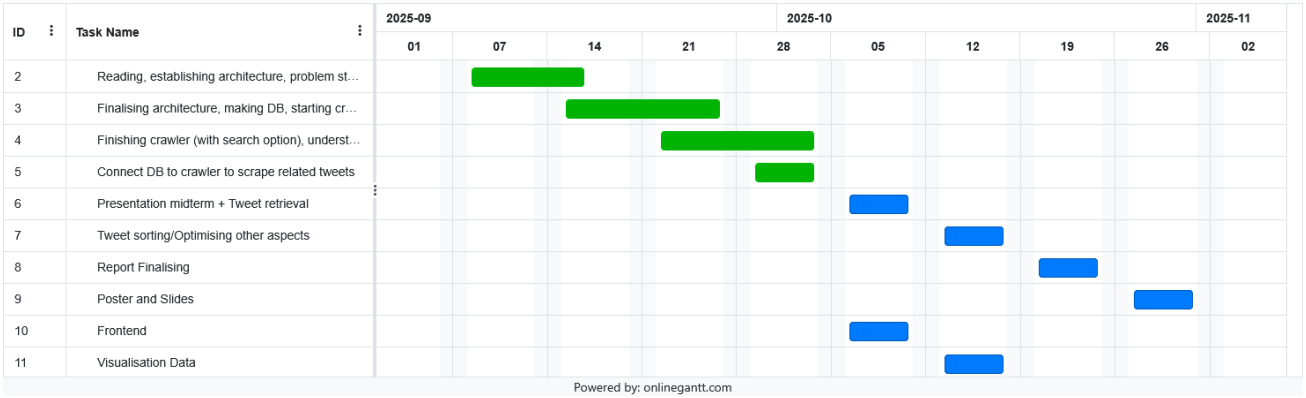


Figure 13: Timeline Midterm

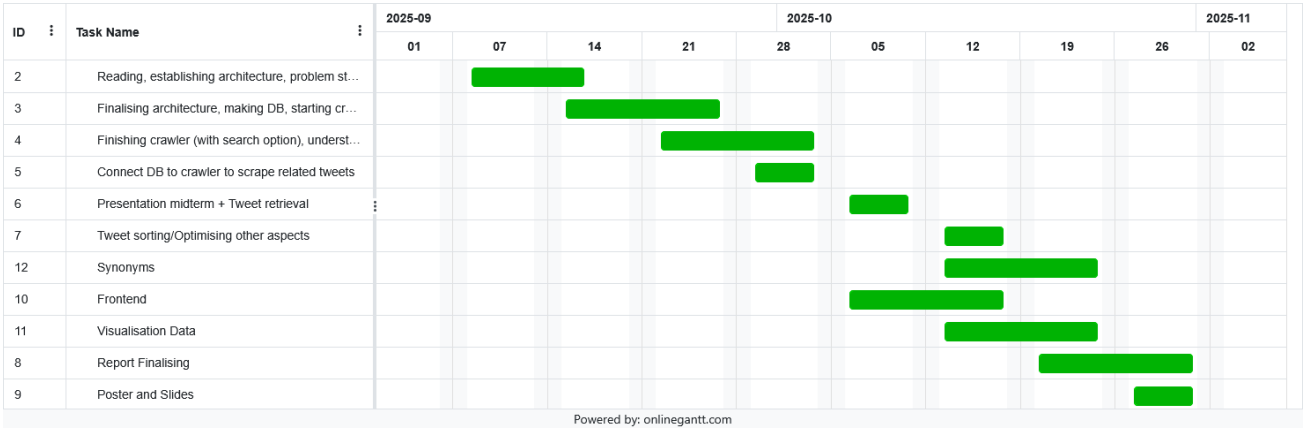


Figure 14: Timeline Final Report