Examining Misinformation via Search Directives

Stanford CS224N Custom Project
With External Collaborator Ronald Robertson (ronalder@stanford.edu), Stanford Internet Observatory

Amy Dunphy
Department of Electrical Engineering
Stanford University
adunphy@stanford.edu

Michal Adamkiewicz
Department of Electrical Engineering
Stanford University
mikadam@cs.stanford.edu

Abstract

Web searches are a common mechanism by which people find information on the internet; however, the relationship between web searches and misinformation spread has generally been understudied. One way people may encounter misleading web search results is through social media posts called search directives, which encourage users to search for potentially dubious queries. For this project, we seek to develop 1) a classifier, which can identify whether or not any given post is a search directive, and 2) a query extractor, which, given a search directive, can extract the query being suggested.

We collected a labelled dataset of 2,811 examples from across four social media platforms. We fine-tuned a pretrained BERT classifier on our dataset, and were able to identify search directives with 88% accuracy. We then fine-tuned HuggingFace's T5 model to extract the queries from a set of search directives, reaching 74% accuracy. These models could be used in order to generate a large dataset of search directives, which then could be studied to better understand what sorts of web searches may spread misinformation.

1 Introduction

1.1 Data voids

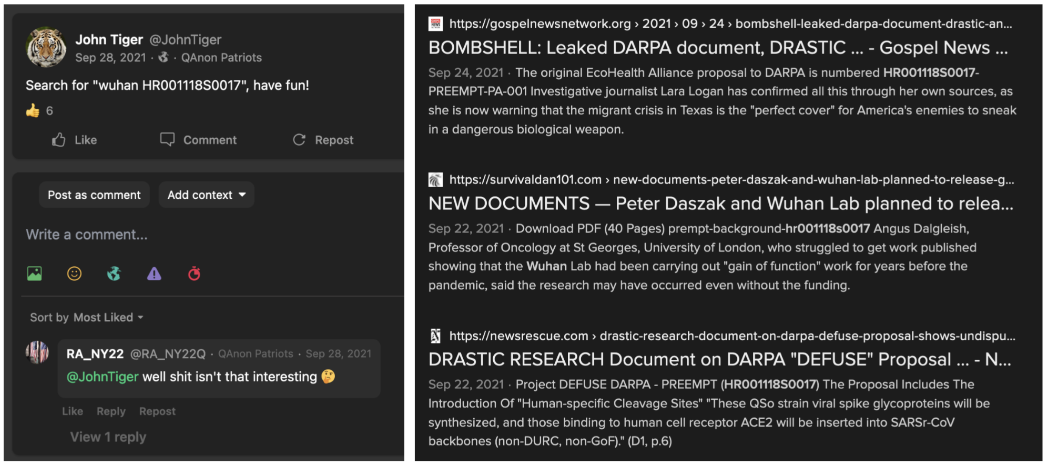
Web searches are a common mechanism by which people find information on the internet. However, the relevance of web searches to misinformation spread has generally been understudied.

Search terms which yield mostly bad or misleading information are often referred to as data voids. They are often unique words or word combinations which ordinarily would not have many search results (Golebiewski and Boyd, 2018). This allows the search results to be populated instead with misinformation. For example, searching "wuhan coronavirus" will return relatively reliable sources such as CNN, wikipedia, and the WHO website. Meanwhile, searching "wuhan HR001118S0017" will return many results declaring that COVID-19 was a biological weapon developed by America's enemies. Data voids are particularly severe on alternative search engines such as DuckDuckGo or Yandex, which are favored by conspiracy theorists due to the perception that Google down-ranks conspiratorial results.

1.2 Search directives

While data voids can lead people to misinformation, they are necessarily very specific – otherwise the conspiratorial content would be drowned out by the much larger volume of news, factchecks, and other internet content. They often take the form of names ("Dr Andreas Noack") or combinations of words that would not usually appear together ("Disney clone lab"). Their specificity makes it very difficult for anyone to come across them in the first place, likely limiting their impact.

One mechanism through which people can find data voids is through search directives, which we define as a social media post that could move a reader to conduct an online search. When users read these posts, some will search them online, thereby encountering the data void. Figure 1 shows an example search directive post instructing people to make a search, along with some of the top search results that appear.



John Tiger @JohnTiger
Sep 28, 2021 · QAnon Patriots

Search for "wuhan HR001118S0017", have fun!

RA_NY22 @RA_NY22Q · QAnon Patriots · Sep 28, 2021
@JohnTiger well shit isn't that interesting 🤔

https://gospelnewsnetwork.org › 2021 › 09 › 24 › bombshell-leaked-darpa-document-drastic-an...
BOMBSHELL: Leaked DARPA document, DRASTIC ... - Gospel News ...
Sep 24, 2021 · The original EcoHealth Alliance proposal to DARPA is numbered **HR001118S0017**-PREEMPT-PA-001 Investigative journalist Lara Logan has confirmed all this through her own sources, as she is now warning that the migrant crisis in Texas is the "perfect cover" for America's enemies to sneak in a dangerous biological weapon.

https://survivaldan101.com › new-documents-peter-daszak-and-wuhan-lab-planned-to-release-g...
NEW DOCUMENTS — Peter Daszak and Wuhan Lab planned to relea...
Sep 22, 2021 · Download PDF (40 Pages) prempt-background-**hr001118s0017** Angus Dalgleish, Professor of Oncology at St Georges, University of London, who struggled to get work published showing that the **Wuhan** Lab had been carrying out "gain of function" work for years before the pandemic, said the research may have occurred even without the funding.

https://newsrescue.com › drastic-research-document-on-darpa-defuse-proposal-shows-undispu...
DRASTIC RESEARCH Document on DARPA "DEFUSE" Proposal ... - N...
Sep 22, 2021 · Project DEFUSE DARPA - PREEMPT (**HR001118S0017**) The Proposal Includes The Introduction Of "Human-specific Cleavage Sites" "These QSo strain viral spike glycoproteins will be synthesized, and those binding to human cell receptor ACE2 will be inserted into SARSr-CoV backbones (non-DURC, non-GoF)." (D1, p.6)

Figure 1: Search directive with query "wuhan HR001118S0017" and DuckDuckGo search results.

Our goal in this paper is to create models that enable us to detect then extract the queries mentioned in social media posts, since these are potential data voids. Using this query extractor, we could compile a large number of potential data voids for further study, yielding some insight into what sorts of web searches people are using to discover misinformation.

1.3 Search directive taxonomy

A very wide variety of posts could be described as search directives. We have defined four primary classifications of search directives, in decreasing degree of how directly they may motivate an online search. Table 1 shows examples of each of the four classes.

Type	Definition	Example
Linked	Includes a link to a search result page	Check out: https://www.google.com/search?q=black+sun
Suggested	Tells the reader to conduct a search	Do your research, using #DuckDuckGo and search for 'dangers of 5G'.
Modeled	Tells the reader about conducting a search	So weird I can't find this story when I search "Jackie Gordon"... What's up with that @Google @googlenews?
Mentioned	Tells the reader about a search	"2014 obama coup ukraine" is trending across google search.

Table 1: Examples of four types of search directive.

The most explicit form of search directive is a direct link to a web search, called "linked" search directives. We do not examine linked directives in this paper, since it is trivial to identify and extract queries from them using classical methods (the pattern in the link is always predicable).

The next most explicit form is "suggested", which are directives which directly instruct the user to carry out a search. Then, we have "modeled", where the post author talks about a search they personally carried out. Last is "mentioned", which describes any post which specifically mentions a specific search.

2 Related Work

Digital trace studies have found that news engagement is typically higher via search engines than via social media (Guess et al., 2020). Recent surveys have also shown that users trust search engines more than social media as a source for general news and information (Edelman, 2021). This could potentially be related to the “IKEA effect”, which states that people overvalue the things they have personally built (Norton et al., 2012). An “IKEA effect of misinformation” could imply that people are overconfident in information they feel they have discovered themselves, such as via an online search (Tripodi, 2022).

These patterns of usage and trust in search engines can be cause for concern, because search results vary widely depending on the query used. Politicans and pundits are known to actively exploit data voids, using new terms, phrases, or names in speeches in order to direct users to searches filled with poor information (Tripodi, 2019).

There is fairly little existing computational research into the impact of search engines on misinformation spread. Makhortykh et al. (2020) carried out a cross-comparison of the amount of misinformation found across six search engines in three languages, confirming that sites such as Yandex (a Russian search engine favored by conspiracy theorists) do in fact host more alternative media than competitors. Zade et al. (2022) examined political headlines on Google related to the 2020 election, with a particular focus on differences in search results based on query or location of search. However, there has yet to be a comprehensive study across a range of topics of what sorts of keywords are most likely to become data voids full of bad information, or what their distribution looks like. We hope that our model will allow for the compilation of a large number of potential data voids, which can then be examined for misinformation content in a follow-up study.

3 Approach

3.1 Search directive classifier

Our search directive classifier was developed by fine-tuning a BERT model on our dataset. BERT, which stands for Bidirectional Encoder Representation from Transformers was first introduced in Devlin et al. (2019). It uses a now standard self-attention transformer proposed in Vaswani et al. (2017). Before the main model, the inputs are lower-cased and tokenised using WordPiece, a tokeniser similar to the byte pair encoding we discussed in class but with a different pair-selection criteria.

For our purposes, we used the "bert-base-uncased" model from HuggingFace, which was pre-trained on Wikipedia and the book-corpus datasets using masked language modelling and next sentence prediction (HuggingFace). It uses a total of 110M parameters. The version hosted on HuggingFace is prepared for use in a classification task by taking the hidden state corresponding to the [CLS] token and passing it into a linear classifier. We then fine-tune the model for the search directive classification task on our dataset of 2,811 examples.

3.2 Query extractor

Our query extractor model is developed by fine-tuning Huggingface’s T5 model. T5 is a model first introduced by google in Raffel et al. (2020), specifically to explore the limits of text to text transfer learning. It was first pre-trained on the common crawl dataset, before being fine-tuned on 7 NLP tasks including question answering, paraphrasing and sentence completion.

Architecturally, the T5 model is similar to the self-attention transformer proposed in Vaswani et al. (2017). Just like that paper, it uses fully-visible masking in the encoder fulled by causal masking in the decoder, however it uses a simplified positional embedding.

We have taken this pre-trained and fine-tuned model (specifically the "t5-small" checkpoint with 60 million parameters) and further fine-tuned it on our dataset of 875 search directive-query pairs.

3.3 Baselines

To create a baseline for the classifier, we looked for the strings ’search:’, ’search for’, ’search "’, ’search ”, ’search term’, and ’search bar’, all of which are common in directives and uncommon elsewhere. If post contained one of the strings, we marked it as a directive. This baseline had an accuracy of 69% across our entire dataset.

Creating a baseline for the query was more complicated. We defined a set of start tokens, which generally occur before directives: "search:", "search for", "search "", "search ’", "search term", """, and "’". We also defined a set of stop tokens which indicate that the query is over: ".", """, "’", and newline. We took the query to be any text between the start and stop tokens, with whitespace and quotes removed. This baseline had an accuracy of 23% across our full dataset.

4 Data

Our dataset contained the text of 2811 social media posts (including both posts themselves and comments – we will use "post" as a shorthand to describe both) from across four different social media sites: twitter, gab, reddit, and gettr. Every post was labelled by hand, with labels containing 1) whether or not the post was a directive, 2) if directive, what class of directive it was, and 3) if directive, what the query was (sometimes none). 775 of them were search directives. Figure 2 shows a summary of the statistics of our dataset. Table 2 shows a sample of the types of posts we encountered.

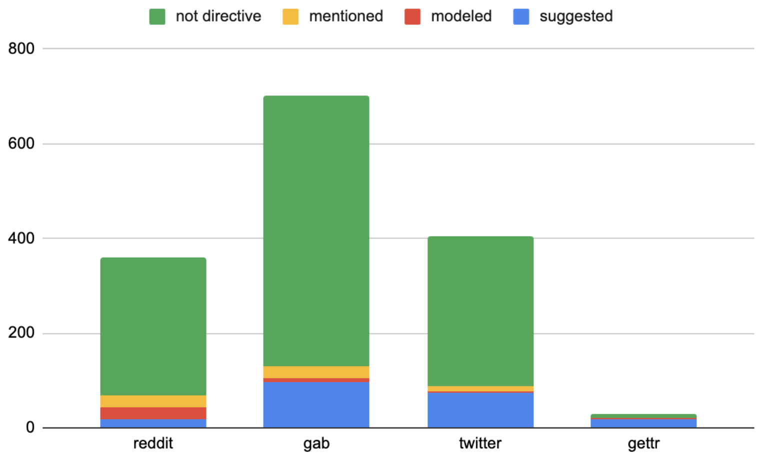


Figure 2: Breakdown of dataset by source and directive type

The posts were collected using a variety of methods. Due to the enormous volume of existing social media posts, the comparatively small percentage of them that are search directives, and the limitations of human labelling, we took several steps to maximize the probability that any post we looked at was a directive.

4.1 Reddit

Our 364 Reddit posts were collected using the Reddit API with the PRAW python wrapper (PRAW, 2022). Our data included both post text and comments. At first, we gathered several thousand posts from across the site (any post containing the words "search", "look" (as in "look up"), and "google" from the last year). However, it quickly became clear that the vast majority of these posts were not directives, and we did not have the time to label them all.

Since search directives are much more common (and relevant to our interest in misinformation-focused data voids) in conspiracy-oriented content, we focused on posts in the subreddit r/conspiracy from the last year. We also found that directives containing the keywords "look up" or "google" but not "search" were very rare, so we focused on posts specifically containing the word "search".

Content	Directive?	Directive Type	Query
Ocak, a search and rescue dog, became a hero in Turkiye’s Kahramanmaras province, the epicentre of 7.6- and 7.7-magnitude earthquakes last month. #RescueDog #Earthquake	0		
@JH dare you go show your search history.	0		
Use the Brave search engine.	0		
Vaccine reactions mimic COVID reactions in many cases. Search “vaccine tinnitus” and maybe acknowledge that real people are suffering from these shots. #realnotrare	1	suggested	vaccine tinnitus
Died Suddenly top search on Google for android phone in Canada	1	mentioned	Died Suddenly
I searched "theatre queer" to find a more articulate definition of the term for a friend than the one I could come up with in the moment…	1	modeled	theatre queer

Table 2: Samples of posts.

4.2 Gab & Gettr

Gab and Gettr are two American alt-right-focused social media platforms, both quite similar in form to twitter. Conspiracy theories and misinformation are rampant across both. 702 of our 1,207 Gab posts and all 28 Gettr posts were collected using the Social Media Analysis Toolkit (SMAT) 3rd party API (SMAT, 2022). SMAT allows users to collect all posts containing a certain word; here, we looked at posts containing the word "search" and labelled a subset of them.

4.3 Twitter & Gab

All 1,253 twitter posts and 505 of our gab posts were collected using DGAP, Stanford Internet Observatory’s data gathering tool. We began with around 500,000 Gab posts and 10,000 Tweets containing the word "search". They were filtered down significantly by removing advertisements and duplicates (eg. multiple users re-posting the same comment). We also removed all posts where "search" was only part of a link: while some of these count as linked directives (discussed above), they are easy to retrieve using classical methods and therefore were not useful for our model training data.

We had already trained an early version of our search directive classifier on the SMAT-generated GabGettr dataset described above. On its training dataset, it correctly classified 84% of posts as either directive or not directive. We ran all remaining Twitter and Gab data through this classifier. We then labelled a random subset of the posts that the classifier marked as directives.

4.4 Data augmentation

Despite the hundreds of thousands of posts we began with, we ended up with only 624 examples with specific queries for the query extractor to train on. In order to improve the volume of the training data, we decided to augment the dataset using Google Translate.

With the Google Cloud Translate API, we translated examples with queries into Chinese and then back into English. Chinese was chosen as a language very different than English, which would therefore perturb the grammar and structure of the post enough for the translation to be distinct from the original. Table 3 shows some examples of augmented data compared to the original versions. We did have to relabel the augmented data by hand, due to small changes in the search terms themselves (especially cases of singular vs. plural, which translate poorly between English and Chinese).

Original	Translate-Augmented
@Paprwiz Search Disney "Clone Lab" Very frightening	@Paprwiz Searching for Disney's Clone Lab is scary
Do an internet search for "PANDA EYES"...........	Search for "PANDA EYES" on the internet...
@JaneDoe1976 Do an internet search for 'Dov Zakheim'.	@JaneDoe1976 Search the internet for "Dov Zakheim".
@LouisianaBull search Small aeroplanes crashes .	@LouisianaBull searches for small plane crash.

Table 3: Examples of augmented data.

5 Experiments

5.1 Classifier

5.1.1 Experimental details

We fine-tuned a BERT classifier on the entire dataset of 2811 posts. A random 15% of the dataset was set aside as a testing dataset. The classifier was instructed to label each piece of text either 0, for not directive, or 1, for directive. We trained our classifier for 150 epochs with a learning rate of 2e-5, then evaluated it against the labelled testing dataset. The classifier output was expected to be an exact match for human labels.

5.1.2 Results

The output accuracy on the testing dataset was 88%, compared with a baseline accuracy of 69% (discussed above). Figure 3 shows the confusion matrices for both the model and our baseline.

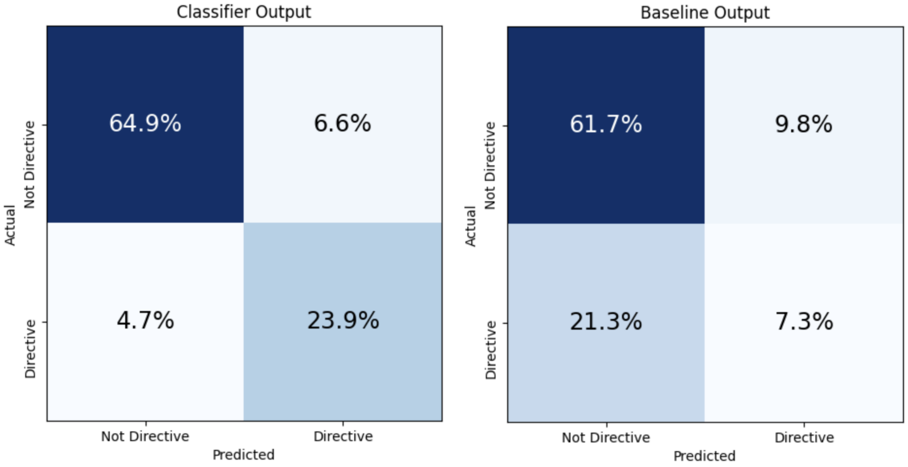


Figure 3: Confusion matrices for classifier model and baseline.

5.2 Query extractor

5.2.1 Experimental details

We fine-tuned Huggingface's T5-small model on a dataset of 775 search directive posts, mixed with 100 more pieces of Google-translate-augmented data. 15% of the 775 posts were set aside as a testing set. Both the augmented data and the posts that we augmented data from were included exclusively in the training set, in order to prevent situations where near-identical posts were in the training and test sets.

The model's loss plateaued around 460 epochs. Drawing from other people's advice online, we chose to use a learning rate of 2e-4, slightly higher than default (Patil, 2020).

In early versions of the model, we encountered issues where the model would misspell queries (eg. "zoomsday bunker" in place of "doomsday bunker"). As an experiment, we modified beam search so that the model would only consider tokens in the input (always including None as an option). We hoped that this would reduce spelling errors; however, while it worked in specific cases, it decreased the overall accuracy by several percent. We eventually decided to forgo this step.

5.2.2 Evaluation method

To evaluate our accuracy, we set both the human-labeled queries and the model output queries to lower case, stripped front and back whitespace, and removed all quotes (including some non-ascii quotes found in the dataset). We counted the model as correct only where this processed model output was an exact match for the human label.

5.2.3 Results

Overall, we were able to achieve a 74% accuracy on the full test dataset, compared with a baseline accuracy of 23%. We examined the accuracy percentages by classification of the search directive: it was highest on the most explicit type, suggested directives, with an accuracy of 81%. The model struggled the most on modeled directives, with an accuracy of 50%, which makes sense – the modeled directives were the rarest by far in our dataset. Importantly, the suggested directives which performed best are suspected to be the most relevant to misinformation spread. Figure 4 shows the breakdown of accuracy per directive type.

We found that the addition the augmented data increased our final evaluation accuracy by just under 10%. We also experimented with using T5-base instead of T5-small, however, the increase in accuracy was minimal (around 2%) and we determined it was not worth the increase in training time.

6 Analysis

6.1 Classifier

The classifier accuracy was 88%, which is only a few percent lower than the consistency between individual human raters: due to edge cases of what is considered a search directive, there can be considerable variation between even human raters. The examples the classifier struggled most on were cases where web searches were discussed, but the post itself was not a directive – for example, cases where posters discussed their search history on google. This makes sense; this sort of post is often difficult for human raters as well. Altogether, the classifier performed quite well, and will be very useful for our future data collection efforts.

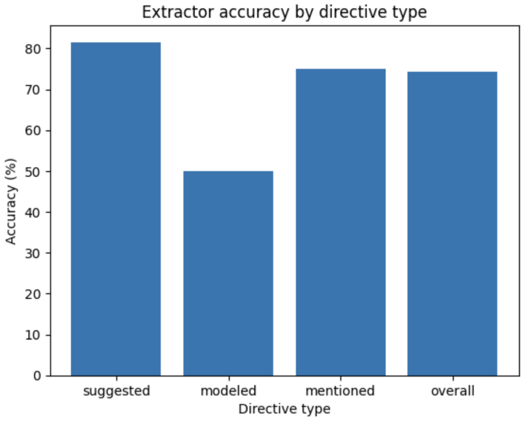
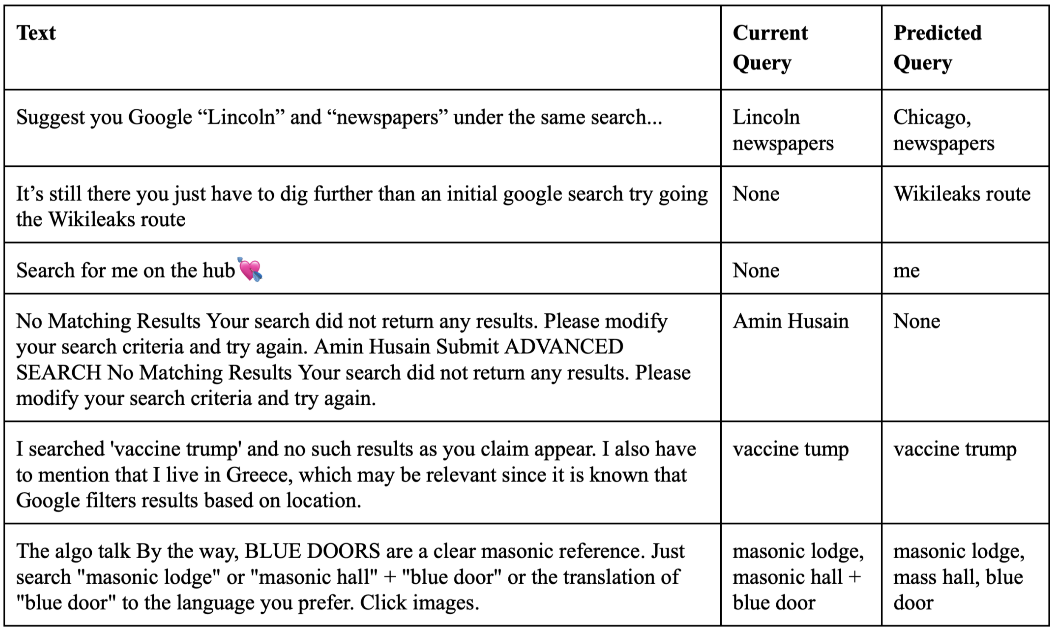


Figure 4: Query extractor performance by classification of search directive

6.2 Query extractor

The query extractor failed in several very interesting ways: Table shows some examples of posts which it incorrectly categorized. In general, many of its errors were fairly small. Mispellings, plurals, and verb conjugations were frequent issues.

It also struggled with cases where context would be required to understand the directive. For example, a user posted "search for me", presumably indicating that readers should look for their username. However, the extractor simply returned "me" as the query. Similarly, it was not able to parse the search directive "If you write "Illuminati" backwards on google you get the NSA page as first result.", returning "Illuminati" instead of "itanimullI".



Text	Current Query	Predicted Query
Suggest you Google “Lincoln” and “newspapers” under the same search...	Lincoln newspapers	Chicago, newspapers
It’s still there you just have to dig further than an initial google search try going the Wikileaks route	None	Wikileaks route
Search for me on the hub💘	None	me
No Matching Results Your search did not return any results. Please modify your search criteria and try again. Amin Husain Submit ADVANCED SEARCH No Matching Results Your search did not return any results. Please modify your search criteria and try again.	Amin Husain	None
I searched 'vaccine trump' and no such results as you claim appear. I also have to mention that I live in Greece, which may be relevant since it is known that Google filters results based on location.	vaccine tump	vaccine trump
The algo talk By the way, BLUE DOORS are a clear masonic reference. Just search "masonic lodge" or "masonic hall" + "blue door" or the translation of "blue door" to the language you prefer. Click images.	masonic lodge, masonic hall + blue door	masonic lodge, mass hall, blue door

Table 4: Examples of failed query extraction

In addition to these errors, it picked up on several cases where the human rater had made mistakes. In one spot, the human typo-ed the directive "vaccine trump" as "vaccine tump", but the model got it correct. There were also some edge cases where it made a different decision from the human rater, but on reflection we may have agreed with the model instead of the human (though we did not change the dataset to recompute accuracy). The extractor will be a very useful tool.

7 Conclusion

7.1 Summary

We gathered and labeled a total of 2811 social media posts from twitter, gab, reddit, and gettr. For each, we labelled 1) whether or not the post was a search directive, 2) if it was a directive, what type it was, and 3) what the query was (sometimes None). Of the posts, 775 were search directives.

We then trained a binary classifier to output whether a given post was a search directive or not. The classifier output 88% accuracy on our testing dataset, compared with a baseline of 69%.

We followed up by training a search directive query extractor, which, given a search directive, was trained to output the most likely query. It was able to match the human-labelled output in 74% of the testing set, compared with a baseline of 23%. It was able to reach 81% accuracy on the suggested directives, which are the most common and most relevant to potential misinformation study.

7.2 Future work

We plan to label additional data in order to make some further improvements to both models. We will also continue with the translation augmentation, since the initial use of augmented data was very successful. In addition, we plan to work on better standardization for the data that we do have, particularly in cases with multiple queries, which were sometimes inconsistently handled in this iteration of the models.

After the models are improved, our next steps will be to generate a large dataset of search directives and examine the search results for them across different search engines. For queries with very few results (potential search directives), we will check the dubiousness of the top result domains using a news trust tool such as Media Bias/Fact Check. From there, we will be able to observe patterns and gain insight into the sorts of misinformation that can be reached from search directive-inspired web searches.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Edelman. 2021. Edelman trust barometer 2021. Technical report, Edelman.

Michael Golebiewski and Danah Boyd. 2018. Data voids: Where missing data can easily be exploited. *Data & Society*.

Andrew M. Guess, Brendan Nyhan, and Jason Reifler. 2020. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5):472–480.

HuggingFace. Text classification. `https://huggingface.co/docs/transformers/tasks/sequence_classification`.

M. Makhortykh, A. Urman, and R. Ulloa. 2020. How search engines disseminate information about covid-19 and why they should do better. *Harvard Kennedy School (HKS) Misinformation Review.*

Michael I. Norton, Daniel Mochon, and Dan Ariely. 2012. The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, 22(3):453–460.

Suraj Patil. 2020. T5 finetuning tips. *Hugging Face Forums*. Https://discuss.huggingface.co/t/t5-finetuning-tips/684.

PRAW. 2022. Praw: The python reddit api wrapper. `https://praw.readthedocs.io/en/stable/`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research.*

SMAT. 2022. Social media analysis toolkit. `https://www.smat-app.com/`.

Francesca Tripodi. 2019. Devin Nunes and the Power of Keyword Signaling | WIRED.

Francesca Bolla Tripodi. 2022. *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*. Yale University Press, New Haven.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Himanshu Zade, Morgan Wack, Yuanrui Zhang, Kate Starbird, Ryan Calo, Jason Young, and Jevin D. West. 2022. Auditing google's search headlines as a potential gateway to misleading content: Evidence from the 2020 us election. *Journal of Online Trust and Safety*, 1.