Tracking Sources of Online Disinformation

Interim Report

Cand.No. 183708

November 2020

Contents

1 Introduction

The fabrication and subsequent spread of false information with intent to deceive is a tactic used by many, from authoritarian regimes and intelligence services [1] to political parties [2]. This tactic is now more commonly referred to as "Disinformation", a loan translation from the Russian "dezinfromatsiya" [3] which refers to the old KGB "black propaganda" department [4]. The definition this report takes for disinformation - as given by Wardle and Derakhshan [5] - is "information that is false and deliberately created to harm a person, social group, organization or country". A key distinction between disinformation and the often associated misinformation is that whilst both involve the spread of inaccurate or false information, disinformation carries with it mal-intent. This relationship is shown in Figure 1.

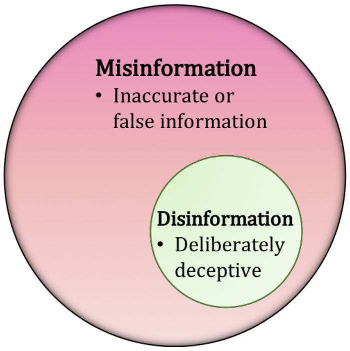


Figure 1: Disinformation as a subset of Misinformation

In an increasingly IT-centric society, especially in regards to social medias platforms, exposure to disinformation can be unavoidable. Recent large scale analyses have shown that on such platforms, false information can diffuse faster through the social media network than the truth [6, 7]. Twitter is one such social networking website vulnerable to disinformation, due to the casual formatting of the messages, or "tweets", and its key mechanism of "retweeting", which creates a loose structure in the relationship between nodes in Twitter's network [8]. This project aims to track these connections within Twitter in an attempt to accurately map sources of disinformation on that platform and investigate their origins where possible.

1.1 Project Aims and Objectives

To achieve the aims of this project, we will make use of fact-checking sites such as FactCheck.org, Snopes.com and Ground.news among others as a source for discovering potentially false claims. By accessing these reported instances of disinformation, queries can be built in Python, formed by breaking down article headlines into keywords, or potentially using a more direct link-comparison if a link to the website where the news originated is given in the article. These queries will then be run against an index of recent tweets in an attempt to match a false claim with a tweet. If a match is made that tweet is recorded and then in the instances where discovery of the tweet came from keyword-search, the matched tweet will be checked to see if it contains a link. The link will be examined to determine if it is the provenance of the purported claim, the result of which will be logged. From there we can use the accompanying metadata of these matched tweets to examine points of interest such as geographic data, characteristics of the disinformation author and the timing of near-identical posts. This will allow for tracing the spread of false information and mapping a potential network of collaborators, for the goal of monitoring and tracking these sources of online disinformation.

1.1.1 Primary Objectives

- Investigate and document potential/probable topics targeted by disinformation and the breakdown of accurate information versus false information for those topics on Twitter.
  - This will ensure that interesting subject areas (areas that provide a sufficient range for analysis) are chosen prior to implementation and perhaps help to understand which topics, people or organizations may be targeted in the future.
- Use fact checking sites as a source for building queries of false claims.
- Use Twitter's search API to match tweets containing these false claims.
  - If a tweet contains a link which redirects to an alternative news source, examine the linked page further to establish whether it contains the false claim.
- Maintain a topic-specific collection of disinformation authors and the reach of their content.
  - Found by examining "Retweet" and "Quote Tweet" metadata for the root-level object containing the screen name of the original tweeter.
  - Once an disinformation author is found, we may return collections of accounts that retweet that content and map these nodes to measure spread.
- Provide analytics for the data collected above.

1.1.2 Extension Objectives

- Most Likely: Create a user-friendly interface for end product. Potential features include:
  - Selecting false claim articles from various fact checking sites, to be used for disinformation query building.

    - Option to manually build queries for use on Twitter's Search API.
    - Data visualization tools which show spread of a disinformation piece over time and geographic distribution.

- Likely: Investigate alternative methods for discovering the source of a particular piece of disinformation on Twitter and implement systems which make use of those methods.
    - Compare performance of traditional system (fact checking site query building) versus alternatives, in the task of discovering sources of disinformation on Twitter.
    - One study looked for disparity in retweeting behaviour by measuring the inequality of the distribution of retweets. Their true positive rate for correctly identifying sources of false information was over 90% [9].

1.1.3 Expected Outcome

The expected outcome of this project is to have a created a system that can quickly scan for disinformation on Twitter, log those occurrences and their origin webpages if a relevant link is provided. Using that collected data it will then go on to provide analysis and visualizations based on these sources of disinformation. This should help an interested party more quickly and more easily gain insights on these online sources of disinformation.

1.2 Project Relevance

This project is interesting as online disinformation on social media is a relatively new area of study and so there is a freedom in the methodology for tackling this problem. This project will allow me to put in practice skills learnt from various programming modules and natural language engineering modules, whilst also allowing me to freely explore and develop data analysis skills of my own. This flexibility will hopefully provide grounds to further my creativity and lateral thinking skills as well. The potential for building a user interface means that concepts learnt from my Human-Computer Interaction module can too be put into practice. Cementing these skills is key for leaving education in good standing for the professional world and this project will also serve as a portfolio piece that can be shown to potential employers.

1.3 Project Problem Area and Motivations

With online disinformation growing as a risk to information order, I am becoming increasingly encouraged by the cautious realization of the potential scale of damage that can be caused by disinformation and the gradual adoption of fact checking tips - such as checking primary and secondary sources, being conscious of the reputation of the source and general skepticism towards outlandish stories. As such this project is of a personal interest to me too.

2 Professional and Ethical Considerations

This project makes use of public-facing data fetched from Twitter. There are no direct participants and presumably all Twitter users have previously agreed to Twitter's Terms of Service when they joined the site. However as those who's tweets which will be monitored and logged by the systems built during this project are unable to give informed consent, extra care must be taken to ensure the project conforms to professional and ethical guidelines.

2.1 BCS Code of Conduct

Below expands on the key areas stated within the BCS Code of Conduct. These are professional standards that are required by the BCS as a condition of membership and are good practices to follow for later professional life.

2.1.1 Public Interest

a. - have due regard for public health, privacy, security and wellbeing of others and the environment.

- The project shall always hold the needs mentioned above in mind. Particularly as this is a project with a focus on social media, privacy will be respected when there is no need to un-anonymize a particular Twitter user or disinformation author.

b. - have due regard for the legitimate rights of third parties.

- The project makes only limited use of third party tools and content, all of which are free to use and licensing terms will not be breached. Where used the creators will be referenced and due credit given.

c. - conduct your professional activities without discrimination on the grounds of sex, sexual orientation, marital status, nationality, colour, race, ethnic origin, religion, age or disability, or of any other condition or requirement.

- There is a potential for personal data (of a Twitter user) such as sex/ age/ nationality/ location to be collected for data analysis towards the end of the project. As mentioned in response to *a.*, privacy will be respected if personal details are of no relevance to the project. If that data is however relevant then the project will take steps to ensure that there is no discrimination or bias against any particular group.

d. - promote equal access to the benefits of IT and seek to promote the inclusion of all sectors in society wherever opportunities arise.

- The system/tools developed in this project will be free and available for all to use.

2.1.2 Professional Competence and Integrity

a. - only undertake to do work or provide a service that is within your professional competence.
b. - NOT claim any level of competence that you do not possess.
c. - develop your professional knowledge, skills and competence on a continuing basis, maintaining awareness of technological developments, procedures, and standards that are relevant to your field.

- The project and its potential scope has been discussed with my technical supervisor and has been deemed to be within my ability to complete to a high standard and within the time frame set out by the examining board. An open-minded approach will be taken so to stay flexible to any new technologies or other developments that are relevant to this project.

d. - ensure that you have the knowledge and understanding of Legislation and that you comply with such Legislation, in carrying out your professional responsibilities.*

- Research will be conducted to determine if there are any potential breaches of legislation that could arise during this project's lifetime and if so those breaches will be avoided. And so all applicable laws, statutes and regulations will be upheld throughout this project.

e. - respect and value alternative viewpoints and, seek, accept and offer honest criticisms of work.

- Progress will be continuously shared with my technical supervisor and when possible, and not at risk of academic misconduct, ideas and progress will shared with peer group to seek criticisms and advice.

f. - avoid injuring others, their property, reputation, or employment by false or malicious or negligent action or inaction.
g. - reject and will not make any offer of bribery or unethical inducement.

- Throughout this project caution will be exercised in areas that could lead to the issues raised in *f.* and I declare I will not accept any offer of bribery or unethical inducement and will report any such instances.

2.1.3 Duty to Relevant Authority

a. carry out your professional responsibilities with due care and diligence in accordance with the Relevant Authority's requirements whilst exercising your professional judgement at all times.

- By meeting all set deadlines and producing work of a high standard I shall meet the requirements of my relevant authority (University of Sussex).

b. seek to avoid any situation that may give rise to a conflict of interest between you and your Relevant Authority.

- I shall seek to avoid any such situation, as stated previously this project has no intentions of compromising the relationship I have with my relevant authority. In the event of a conflict of interest coming about I would immediately seek advice from my technical and academic supervisors.

c. accept professional responsibility for your work and for the work of colleagues who are defined in a given context as working under your supervision.

- I accept all responsibility for the work I produce and in the event of colleagues working under my supervision I shall accept professional responsibility for them also.

d. NOT disclose or authorise to be disclosed, or use for personal gain or to benefit a third party, confidential information except with the permission of your Relevant Authority, or as required by Legislation.

- I will not disclose or make use of any confidential information that comes about during this project. Should a situation arise where I am being asked to provide confidential information to a person outside of the project I shall decline to comment and report the incident.

e. NOT misrepresent or withhold information on the performance of products, systems or services (unless lawfully bound by a duty of confidentiality not to disclose such information), or take advantage of the lack of relevant knowledge or inexperience of others.

- I will not take advantage of those less informed than I and I will only put forward information without any intent to mislead or any prejudice.

2.1.4 Duty to the Profession

a. accept your personal duty to uphold the reputation of the profession and not take any action which could bring the profession into disrepute.
b. seek to improve professional standards through participation in their development, use and enforcement.
c. uphold the reputation and good standing of BCS, the Chartered Institute for IT.
d. act with integrity and respect in your professional relationships with all members of BCS and with members of other professions with whom you work in a professional capacity.

- I take upon myself to uphold the standards mentioned above and by doing so hope to further the goal of work being done in a professional and considerate manner. I will treat all BCS members with the same respect they afforded me and remain in accordance with their Code of Conduct.

e. notify BCS if convicted of a criminal offence or upon becoming bankrupt or disqualified as a Company Director and in each case give details of the relevant jurisdiction.

- If such a situation arose and I was in a relevant position I would notify BCS with these details at once.

f. encourage and support fellow members in their professional development

- I will support and further the collective goal of helping those who seek help out.

2.2 Potential Ethical Pitfalls

As previously mentioned this project has no direct participants however will make use of public-facing social media data. Due to the nature of this research it becomes a grey area in how this project complies with the points listed on the University of Sussex's ethical compliance form for undergraduate projects (School of Engineering and Informatics). The novelty of using social media data for research brings forth new contextual challenges that may not be encompassed by traditional ethical frameworks and as such this project will be applying for ethical review. Listed below are some of the common key ethical issues [10, 11] facing social media research, expansions on the reasons as to why the project will be seeking ethical approval, and the considerations that will be taken by this project to avoid any ethical transgressions:

2.2.1 Issue 1: Public versus Private Space

This issue relates to whether social media data, e.g. Twitter data, should be considered as public or as private data. All Twitter users agree to the Terms of Service detailing how their data can be used by Twitter and by third parties. Public data would imply that the content used on the platform could be used much more freely and with little consideration to the user - a determination taken by some researchers in regards to social media data [12]. This begs the question: "Should a dataset's accessibility be tantamount to ethical justification?"

2.2.2 Issue 2: Anonymity

Whilst the Terms of Service of a social media platform might state that the content is freely available, collecting personal data that is outside the scope of the research in question can only increase the risk of harm to the user. An argument could be made that by initially posting content online the user opens themselves up to public judgement, but by publishing that content, without anonymizing it, the researcher then takes an active role in overexposing the user. Some academics have suggested that increasing the risk of harm also increases the researcher's own responsibility [13], a sentiment that this project agrees with.

2.2.3 Issue 3: Informed Consent

Informed consent relates to the explicit agreement of a person before they disclose personal information. In research this is usually a process built into the design of the experiment, for example a check-box at the start of a form and a signature. Social media confuses this issue as while explicit agreement is implied in Terms of Service, the reality is many do not read these legal documents in whole [14] and so users are unaware of how their data might be used. For example in traditional experiments a key aspect of informed consent is the ability to withdraw from the study at any time. In regards to social media research where data is unknowingly collected there is no clear option for this.

2.2.4 Avoiding these Pitfalls

The choice of using Twitter's API to collect information rather than Facebook or other alternatives was a deliberate one. One reason for this choice was due to the difference between Twitter and other social medias as shown through their default privacy settings in both profile and posts. For example Facebook is considered more private, with the default user requirement being to select friends who can view the user's personal page. Twitter on the other hand initializes with complete public visibility. And so by removing the link between data and user we are left with an assumption that the user is happy with the content being online in some capacity. As this project has no intention of collecting irrelevant personal information and plans on anonymizing relevant information, the issues covered above should be mitigated and the potential risk of harm to the user reduced.

3 Related Research

3.1 Previous Works

Identifying and subsequently tracking misinformation and disinformation on Twitter became an urgent subject of research when it was shown how politically-motivated organizations could create a network of accounts to spread false information and simulate an appearance of support for a particular candidate [15]. An early instance of this was a machine-learned approach, targeted at astroturf campaigns [15], where the researchers described a need for automation in regards to monitoring Twitter. In later works the same team designed a system named Truthy, which relied on network analysis techniques [16]; It looked at the topology and crowd-sourced features of information diffusion networks on Twitter and which was considered a top platform of its time [17]. An alternative was the TweetCred system [18] which used a semi-supervised ranking model to classify a user's timeline of tweets in terms of their credibility. Available as a browser extension each tweet would then be accompanied by a credibility score, TweetCred was one of the first to develop a real-time system for determining credibility on Twitter [18].

Other similar systems [19, 20, 21] have been designed for the purpose of rumour detection, and range in varying degrees of automation. These systems all require a user input to build a query in which they search for possible instances of mis/disinformation, rather than monitor Twitter or another social media's stream automatically. However a commonality between most of the aforementioned systems is that they include a tool or browser-based interactive dashboard for monitoring the sources or the propagation of a particular piece of false information. These dashboards give a mixed display of interactive visualization elements and analytical information and allow for sorting of the information based on values such as number of retweets. Truthy also allowed a user to select a particular piece of tracked false information, inspect its diffusion through Twitter in the form of mapped geo-locations, and examine temporal data shown as a timeline of frequency. These features allow the system to be accessible to the interested layman who does not share the same time or potential reward for investment, as a professional journalist or researcher would.

Most similar to the system proposed in this project is Hoaxy [17]. Their focus was on the automatic tracking of news sharing, which used social media API's and web-scrapers to collect data and left the process of fact-checking to independent sources who's credibility has already been established. To demonstrate the capabilities of this type of system they analysed a dataset of tweets with a focus on two aspects. The first was to examine the temporal relationship in the propagation of misinformation and fact-checking. The second aspect was to look at differences in the ways this information was shared by users, a similar idea to the one employed in the earlier mentioned cognitive psychology detection system, where researchers looked at disparity in retweeting behaviour by measuring the inequality of a distribution [9]. A key finding of this demonstration was the larger scale at which misinformation is produced compared to fact-checking content.

3.2 Contribution to Knowledge

This project will expand on key aspects of each of the above tools, such as automation and an user-friendly interactive dashboard for displaying the analysis. It will also aim to cover a larger set of false information stories than previous systems. The uniqueness of this project however, lies in its focus on sources of disinformation. Whilst most the previously mentioned works target misinformation, not as much research has been done into the subcategory of deliberately false stories and their sources. This project aims to provide analysis on the types of accounts classified as sources and the network of actors these accounts make up. This will aid future research into this area and map potential future sources at the centre of disinformation diffusion on Twitter.

4 Requirements Analysis

This project sets out to create a system of data collection and analysis that ultimately will track sources of disinformation on Twitter. Due to the relative novelty of research into disinformation in social media it is useful to provide a breakdown of the potentially interested parties, their requirements from such a project, their satisfaction with the current system and how an ideal system could be better suited.

4.1 Academic Researchers

Their needs:

- A system that can easily be placed in their own disinformation analysis pipeline.
- Easily accessible data and analysis, for their own research and to entice previously uninterested academics to pursue research in disinformation.
- A project with a focus on tracking sources of disinformation which also highlights it's relevance will aid in the justification of future related works.

Disinformation is a problem best confronted with a multi-disciplinary approach, drawing from a range of ideas from contrasting backgrounds. To realize this and for researchers to be interested in this work and potentially incorporate the project into future disinformation analysis, the report and the code will need to be easily accessible. Whilst this project is a final year assessment for an Informatics degree, the subject matter is of interest to those in different fields, such as media studies as well. This means that over-complicating the report with excessively technical language could be to the detriment of its own accessibility. Increasing the project's ease of access might help to bridge any disconnect that exists between differing fields of study. In terms of the code, all features shall be implemented in Python. This is due to Python being one of the most used languages worldwide and also its status as a "Top language" and a "Most used topic" on the Twitter developer GitHub - a sign that it is a popular choice for working with the Twitter API. The project's Python implementations will be created using Google Colaboratory as this is a free Jupyter notebook environment that is run in the cloud and allows for anyone to instantly engage with the code.

4.2 Fact Checking Websites/Organizations

Their needs:

- A system which will aid in identifying disinformation authors and allow for a head-start in tackling a misleading story.
- Attractive and explanatory visualizations which track the sources of disinformation online and potentially the spread of a particular piece.

- Analysis that provides support to the organization's viewer base, increasing their own capacity for potential sources of disinformation.

Most fact check sites use reader interest to determine their coverage. Snopes.com, a site established in 1994 that has done with for Facebook in a fact-checking partnership effort [22], will determine reader interest based on search engine queries, reader submissions, comments posted to Snopes.com social medias and trending stories on Google or social media. This project could aid in the determination of relevant stories to cover/fact-check in multiple ways. Firstly once disinformation authors have been identified, any co-collaborators can be logged and their stories monitored, allowing for potentially earlier identification of other sources of disinformation. Secondly data collected regarding the rate of spread of a story or its magnitude can be used as a metric for determining reader interest. This data can be used within the graphical user interface (GUI) to create visualizations to accompany the story and help people realise the scale of deliberately misleading information on social medias.

5 Project Plan

The project is currently on track, research has been done into existing related works and potential methods of tracking online sources of disinformation. A basic data collection system has also been implemented as to start an early process of iterative improvement to the design. Figure 2 gives an idea as to the estimated timeline for completion of the key tasks that make up this project.

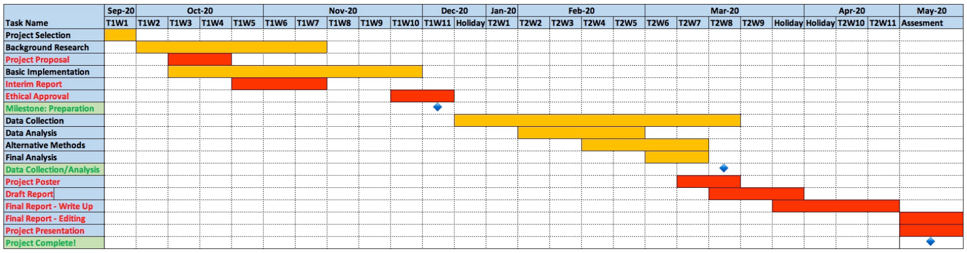


Figure 2: Gantt Chart Breakdown of Task Areas

5.1 Task Breakdown

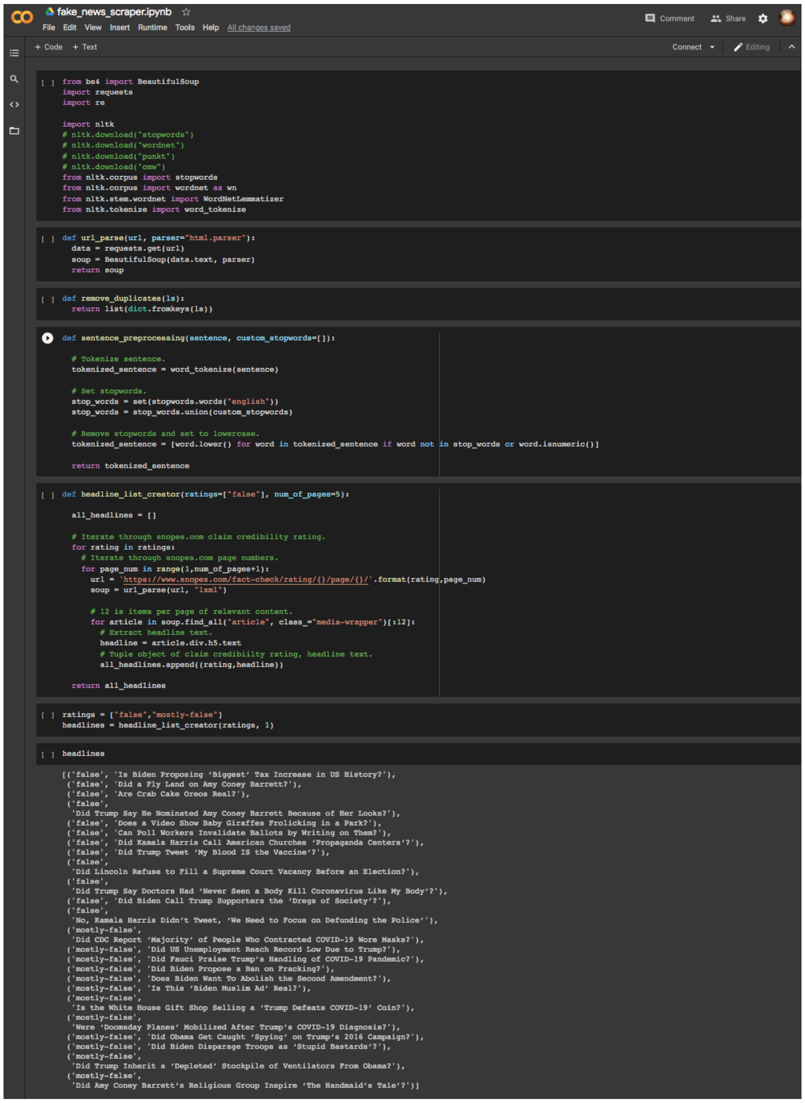
- Project Selection: Accepted by supervisor to work on project titled: "Tracking Sources of Online Disinformation".
- Background Research: This task is one of the earliest due to the need to have some form of a knowledge base before any other work can begin. It is ongoing as other related works can only offer more insight to the potential routes this project could go down.
- Project Proposal: A brief description of the project and the work it entails. Attached in the Appendix.
- Basic Implementation: A complete data collection model - from query building to Twitter search - that can be used to refine later models.
- Interim Report: An assessed piece of work that expands on the Project Proposal and goes on to cover the sections as seen earlier on in this document.
- Ethical Approval: Confirmation that this project is within ethical guidelines.
- Milestone 1! Preparation Complete: Once ethical approval is granted, the project will be set up entirely to begin social media data collection and the remaining tasks.
- Data Collection: Collection of Twitter data for use of tracking sources of disinformation online. This is the longest task due to limitations within

Twitter's free subscription API access: Only Tweets from the previous seven days are accessible and so it is in the projects best interest to begin this task as early as possible.

- Data Analysis: An initial analysis of the data collected so far.
- Alternative Methods: Implementation of alternative data collection methods to run in parallel with main method. Comparisons will be drawn between the two or more sets of data collected to determine which best suits the project.
- Final Analysis: Using all data collected draw conclusions relevant to the problem area. In this time, if all else is on track, a simple GUI can be implemented based on the data collection and analysis pipeline.
- Milestone 2! Data Collection and Analysis Complete:
- Project Poster: An assessed piece of work that showcases key areas of this project.
- Draft Report: A working version of the final report, sent directly to the supervisor for review.
- Final Report - Write Up: Expand on Draft Report using supervisor feedback as reference.
- Final Report - Editing: Last stage of Final Report process, ensure all chapters are relevant and concise.
- Project Presentation: 20 minute talk demonstrating knowledge in the project's subject area and offer an overview of the project and its achievements.
- Milestone 3! Project Complete: All final year project tasks finished!

5.2 Progress so Far

The project currently has already been approved for developer access by Twitter. This means, pending approval from the ethics board, we can progress with starting to collect data from the Twitter platform. Figure 3 shows a simple implementation of the a false claim scraper, written in the Python language. This implementation scrapes articles based on their authenticity rating from the fact checking website Snopes.com. This process is the first step in the data collection process with plans for the false claims to eventually be used for potential disinformation queries.



```
from bs4 import BeautifulSoup
import requests
import re

import nltk
# nltk.download("stopwords")
# nltk.download("wordnet")
# nltk.download("punkt")
# nltk.download("omw")
from nltk.corpus import stopwords
from nltk.corpus import wordnet as wn
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import word_tokenize
```

```
def url_parse(url, parser="html.parser"):
  data = requests.get(url)
  soup = BeautifulSoup(data.text, parser)
  return soup
```

```
def remove_duplicates(ls):
  return list(dict.fromkeys(ls))
```

```
def sentence_preprocessing(sentence, custom_stopwords=[]):

  # Tokenize sentence.
  tokenized_sentence = word_tokenize(sentence)

  # Set stopwords.
  stop_words = set(stopwords.words("english"))
  stop_words = stop_words.union(custom_stopwords)

  # Remove stopwords and set to lowercase.
  tokenized_sentence = [word.lower() for word in tokenized_sentence if word not in stop_words or word.isnumeric()]

  return tokenized_sentence
```

```
def headline_list_creator(ratings=["false"], num_of_pages=5):

  all_headlines = []

  # Iterate through snopes.com claim credibility rating.
  for rating in ratings:
    # Iterate through snopes.com page numbers.
    for page_num in range(1,num_of_pages+1):
      url = 'https://www.snopes.com/fact-check/rating/{}/page/{}/'.format(rating,page_num)
      soup = url_parse(url, "lxml")

      # 12 is items per page of relevant content.
      for article in soup.find_all("article", class_="media-wrapper")[:12]:
        # Extract headline text.
        headline = article.div.h5.text
        # Tuple object of claim credibiilty rating, headline text.
        all_headlines.append((rating,headline))

  return all_headlines
```

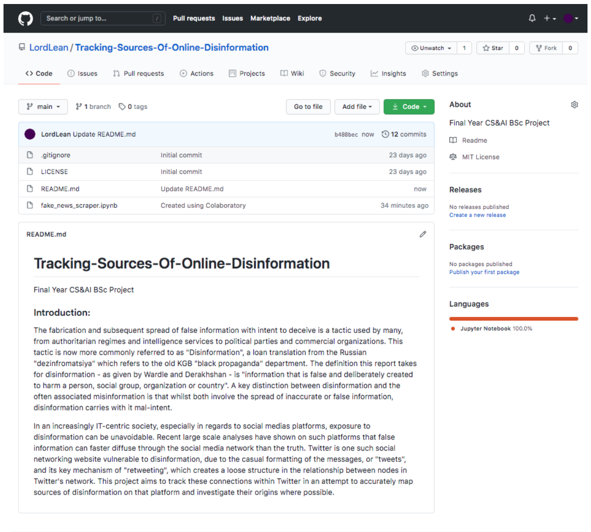
```
ratings = ["false","mostly-false"]
headlines = headline_list_creator(ratings, 1)
```

```
headlines
```

```
[('false', 'Is Biden Proposing ‘Biggest’ Tax Increase in US History?'),
 ('false', 'Did a Fly Land on Amy Coney Barrett?'),
 ('false', 'Are Crab Cake Oreos Real?'),
 ('false',
  'Did Trump Say He Nominated Amy Coney Barrett Because of Her Looks?'),
 ('false', 'Does a Video Show Baby Giraffes Frolicking in a Park?'),
 ('false', 'Can Poll Workers Invalidate Ballots by Writing on Them?'),
 ('false', 'Did Kamala Harris Call American Churches ‘Propaganda Centers’?'),
 ('false', 'Did Trump Tweet ‘My Blood IS the Vaccine’?'),
 ('false',
  'Did Lincoln Refuse to Fill a Supreme Court Vacancy Before an Election?'),
 ('false',
  'Did Trump Say Doctors Had ‘Never Seen a Body Kill Coronavirus Like My Body’?'),
 ('false', 'Did Biden Call Trump Supporters the ‘Dregs of Society’?'),
 ('false',
  'No, Kamala Harris Didn’t Tweet, ‘We Need to Focus on Defunding the Police’'),
 ('mostly-false',
  'Did CDC Report ‘Majority’ of People Who Contracted COVID-19 Wore Masks?'),
 ('mostly-false', 'Did US Unemployment Reach Record Low Due to Trump?'),
 ('mostly-false', 'Did Fauci Praise Trump’s Handling of COVID-19 Pandemic?'),
 ('mostly-false', 'Did Biden Propose a Ban on Fracking?'),
 ('mostly-false', 'Does Biden Want To Abolish the Second Amendment?'),
 ('mostly-false', 'Is This ‘Biden Muslim Ad’ Real?'),
 ('mostly-false',
  'Is the White House Gift Shop Selling a ‘Trump Defeats COVID-19’ Coin?'),
 ('mostly-false',
  'Were ‘Doomsday Planes’ Mobilized After Trump’s COVID-19 Diagnosis?'),
 ('mostly-false', 'Did Obama Get Caught ‘Spying’ on Trump’s 2016 Campaign?'),
 ('mostly-false', 'Did Biden Disparage Troops as ‘Stupid Bastards’?'),
 ('mostly-false',
  'Did Trump Inherit a ‘Depleted’ Stockpile of Ventilators From Obama?'),
 ('mostly-false',
  'Did Amy Coney Barrett’s Religious Group Inspire ‘The Handmaid’s Tale’?')]
```

Figure 3: Colab Notebook: False Claim Scraper

A private repository has been set up on GitHub (Figure 4) for maintaining version control of any code written. This also acts as a simple way of pulling the data collected straight to our Python notebook and vice versa. GitHub was selected due to familiarity with the ecosystem, having previously used the service for other projects.



LordLean Update README.md b488bec now 12 commits

.gitignore	Initial commit	23 days ago
LICENSE	Initial commit	23 days ago
README.md	Update README.md	now
fake_news_scraper.ipynb	Created using Colaboratory	34 minutes ago

README.md

Tracking-Sources-Of-Online-Disinformation

Final Year CS&AI BSc Project

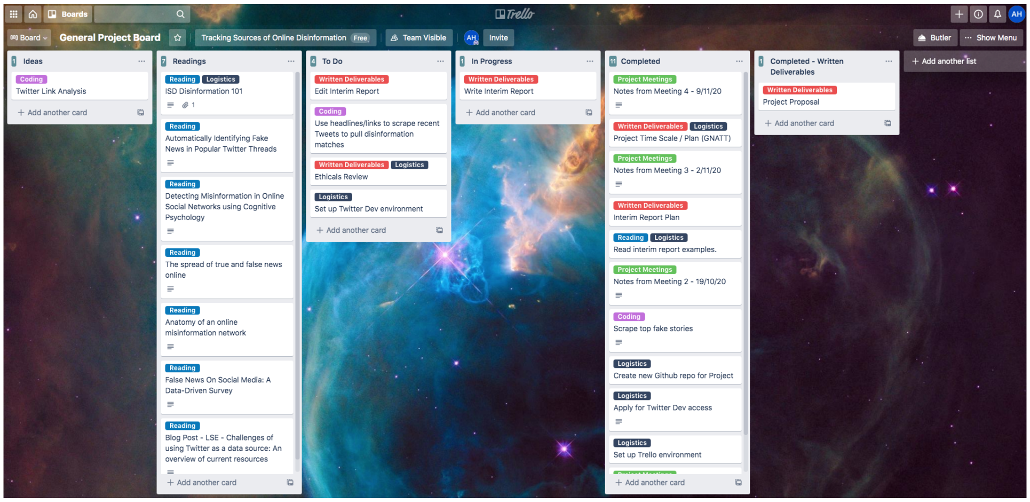
Introduction:

The fabrication and subsequent spread of false information with intent to deceive is a tactic used by many, from authoritarian regimes and intelligence services to political parties and commercial organizations. This tactic is now more commonly referred to as "Disinformation", a loan translation from the Russian "dezinfromatsiya" which refers to the old KGB "black propaganda" department. The definition this report takes for disinformation - as given by Wardle and Derakhshan - is "information that is false and deliberately created to harm a person, social group, organization or country". A key distinction between disinformation and the often associated misinformation is that whilst both involve the spread of inaccurate or false information, disinformation carries with it mal-intent.

In an increasingly IT-centric society, especially in regards to social medias platforms, exposure to disinformation can be unavoidable. Recent large scale analyses have shown on such platforms that false information can faster diffuse through the social media network than the truth. Twitter is one such social networking website vulnerable to disinformation, due to the casual formatting of the messages, or "tweets", and its key mechanism of "retweeting", which creates a loose structure in the relationship between nodes in Twitter's network. This project aims to track these connections within Twitter in an attempt to accurately map sources of disinformation on that platform and investigate their origins where possible.

Figure 4: GitHub Repository: Tracking-Sources-of-Online-Disinformation

For project management Trello (Figure 5) was selected due to the simplicity of its layout and accessibility as a free service. It uses a three tier information hierarchy of Boards, Lists and Cards which are intuitive to use and highly visual. Trello also features an optional AI assistant for organizational optimization. One current application is the automatic labelling of cards depending on which list they appear in, saving time amending these all by hand.



Ideas (1)
- Coding — Twitter Link Analysis

Readings (7)
- Reading, Logistics — ISD Disinformation 101 (1)
- Reading — Automatically Identifying Fake News in Popular Twitter Threads
- Reading — Detecting Misinformation in Online Social Networks using Cognitive Psychology
- Reading — The spread of true and false news online
- Reading — Anatomy of an online misinformation network
- Reading — False News On Social Media: A Data-Driven Survey
- Reading — Blog Post - LSE - Challenges of using Twitter as a data source: An overview of current resources

To Do (4)
- Written Deliverables — Edit Interim Report
- Coding — Use headlines/links to scrape recent Tweets to pull disinformation matches
- Written Deliverables, Logistics — Ethicals Review
- Logistics — Set up Twitter Dev environment

In Progress (1)
- Written Deliverables — Write Interim Report

Completed (11)
- Project Meetings — Notes from Meeting 4 - 9/11/20
- Written Deliverables, Logistics — Project Time Scale / Plan (GNATT)
- Project Meetings — Notes from Meeting 3 - 2/11/20
- Written Deliverables — Interim Report Plan
- Reading, Logistics — Read interim report examples.
- Project Meetings — Notes from Meeting 2 - 19/10/20
- Coding — Scrape top fake stories
- Logistics — Create new Github repo for Project
- Logistics — Apply for Twitter Dev access
- Logistics — Set up Trello environment

Completed - Written Deliverables (1)
- Written Deliverables — Project Proposal

Figure 5: Project Management: Trello

Bibliography

[1] J. Goldman, *Words of intelligence : a dictionary.* Lanham, Md: Scarecrow Press, 2006.

[2] Y. Benkler, R. Faris, and H. Roberts, *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press, 2018.

[3] L. Bittman, *The KGB and Soviet disinformation : an insider's view.* Washington: Pergamon-Brassey's, 1985.

[4] G. Jowett, *Propaganda and persuasion.* Thousand Oaks, Calif: Sage, 2006.

[5] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe report*, vol. 27, 2017.

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[7] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Anatomy of an online misinformation network," *PloS one*, vol. 13, no. 4, p. e0196087, 2018.

[8] P. Chamberlain, "Twitter as a vector for disinformation," 2009.

[9] K. K. Kumar and G. Geethakumari, "Detecting misinformation in online social networks using cognitive psychology," *Human-centric Computing and Information Sciences*, vol. 4, no. 1, pp. 1–22, 2014.

[10] L. Townsend and C. Wallace, "Social media research: A guide to ethics," *University of Aberdeen*, vol. 1, p. 16, 2016.

[11] W. Ahmed, P. Bath, and G. Demartini, "Chapter 4 using twitter as a data source: An overview of ethical, legal, and methodological challenges," in *The Ethics of Online Research*, ser. Advances in Research Ethics and Integrity, K. Woodfield, Ed. Emerald, December 2017, no. 2, pp. 79–107. [Online]. Available: http://eprints.whiterose.ac.uk/126729/

[12] danah boyd and K. Crawford, "Critical questions for big data," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, 2012. [Online]. Available: https://doi.org/10.1080/1369118X.2012.678878

[13] A. Markham and E. Buchanan, "Ethical decision-making and internet research: Version 2.0. recommendations from the aoir ethics working committee," *Available online: aoir. org/reports/ethics2. pdf*, 2012.

[14] M. L. Williams, P. Burnap, and L. Sloan, "Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation," *Sociology*, vol. 51, no. 6, pp. 1149–1168, 2017.

[15] J. Ratkiewicz, M. D. Conover, M. Meiss, A. Flammini, and F. Menczer, “Detecting and tracking political abuse in social media,” in *In Proceedings of the 5th AAAI International Conference on Weblogs and Social Media (ICWSM’11*, 2011.

[16] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, “Truthy: mapping the spread of astroturf in microblog streams,” in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 249–252.

[17] C. Shao, G. Ciampaglia, A. Flammini, and F. Menczer, “Hoaxy: A platform for tracking online misinformation,” 03 2016.

[18] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, “Tweetcred: A real-time web-based system for assessing credibility of content on twitter,” *CoRR*, vol. abs/1405.5490, 2014. [Online]. Available: http://arxiv.org/abs/1405.5490

[19] N. Hassan, A. Sultana, Y. Wu, G. Zhang, C. Li, J. Yang, and C. Yu, “Data in, fact out: Automated monitoring of facts by factwatcher,” vol. 7, 09 2014.

[20] S. Finn, P. T. Metaxas, and E. Mustafaraj, “Investigating rumor propagation with twittertrails,” *CoRR*, vol. abs/1411.3550, 2014. [Online]. Available: http://arxiv.org/abs/1411.3550

[21] P. Resnick, S. Carton, S. Park, Y. Shen, and N. Zeffer, “Rumorlens: A system for analyzing the impact of rumors and corrections in social media,” 2014.

[22] D. Mikkelson, “Disclosures,” Jun 2020. [Online]. Available: https://www.snopes.com/disclosures/

Appendix

Interim Log

Notes from Meeting 1 - 12/10/20:

- Set up trello and teams infrastructure.
- Following agile framework - create end product then iteratively refine.
- Investigate fact-check-sites (FCS) for APIs / stable html (easy of scraping).
- Investigate multiple web scraping libraries and compare performance on varying FCS.
- NLP para-phrasing tools for comparing possible disinformation string from FCS with original twitter post + if exists, affiliated link.
- Will require ethical approval.
- Potential use/misuse cases.

Notes from Meeting 2 - 19/10/20

- Consider using link analysis - meta data if tweet contains link already included.
- Scrape sites for false reports containing links to source material.
- Ethics review template shared.
- User interface - Streamlit discussed (extension objective)

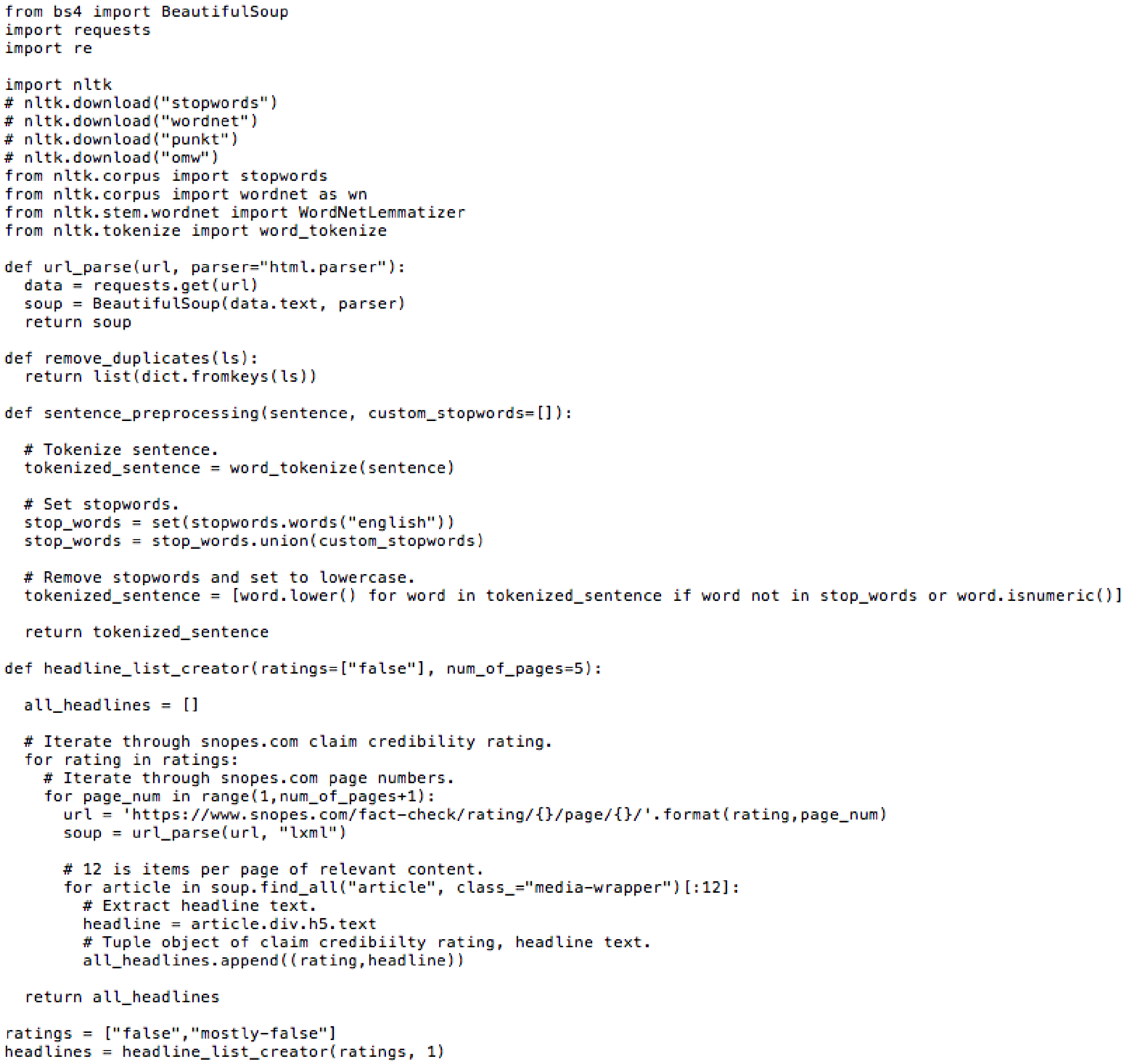
Notes from Meeting 3 - 02/11/20

- Write Interim report to be accessible to all.
- Introduction is crucial to show understanding of the project.
- Start write up of interim report if reading/background knowledge is complete.
- Tier system for objectives.

Notes from Meeting 4 - 09/11/20

- Good topic to explore: Relevance of AI and data analysis growth for social media analysis.
- Ensure consistent report structure. I.e. relevance of subsections, font selection.

Code



```
from bs4 import BeautifulSoup
import requests
import re

import nltk
# nltk.download("stopwords")
# nltk.download("wordnet")
# nltk.download("punkt")
# nltk.download("omw")
from nltk.corpus import stopwords
from nltk.corpus import wordnet as wn
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import word_tokenize

def url_parse(url, parser="html.parser"):
  data = requests.get(url)
  soup = BeautifulSoup(data.text, parser)
  return soup

def remove_duplicates(ls):
  return list(dict.fromkeys(ls))

def sentence_preprocessing(sentence, custom_stopwords=[]):

  # Tokenize sentence.
  tokenized_sentence = word_tokenize(sentence)

  # Set stopwords.
  stop_words = set(stopwords.words("english"))
  stop_words = stop_words.union(custom_stopwords)

  # Remove stopwords and set to lowercase.
  tokenized_sentence = [word.lower() for word in tokenized_sentence if word not in stop_words or word.isnumeric()]

  return tokenized_sentence

def headline_list_creator(ratings=["false"], num_of_pages=5):

  all_headlines = []

  # Iterate through snopes.com claim credibility rating.
  for rating in ratings:
    # Iterate through snopes.com page numbers.
    for page_num in range(1,num_of_pages+1):
      url = 'https://www.snopes.com/fact-check/rating/{}/page/{}/'.format(rating,page_num)
      soup = url_parse(url, "lxml")

      # 12 is items per page of relevant content.
      for article in soup.find_all("article", class_="media-wrapper")[:12]:
        # Extract headline text.
        headline = article.div.h5.text
        # Tuple object of claim credibiilty rating, headline text.
        all_headlines.append((rating,headline))

  return all_headlines

ratings = ["false","mostly-false"]
headlines = headline_list_creator(ratings, 1)
```

Tracking Sources Of Online Disinformation Project Proposal

Cand.No. 183708

November 12, 2020

With the advent of social media and wide-spread internet access, there has been a shift in the information ecosystem; Historically one might actively seek out their favorite broadsheet and possibly be able to place some trust in the journalistic process. Now, for most to have any presence online it means being exposed to information from possibly unreputable sources, pushing their own brand of news, be it suggested articles from a web browser or more commonly in the form of posts on social media. This makes it difficult to fit in an increasingly IT-centric society without at least some exposure to disinformation. Twitter is one such social media that has been plagued with false information; A 2018 study showed that inaccurate news reached more users than factually correct information, and that these falsehoods often even spread at a faster rate (Vosoughi, Roy, & Aral, 2018). As such Twitter will be a focus of this project, in its aim to track sources of disinformation online.

Projects Aim and Objectives

This project will use various fact-checking sites such as FactCheck, Snopes and GroundNews among others as a source for discovering potentially false claims. By accessing these reported instances of disinformation, queries can be built in Python which will then be run against an index of recent tweets in an attempt to match a false claim with a tweet. If a match is made then that tweet is recorded and checked to see if it contains a link. This link will then be examined to determine if it is the provenance of the purported claim, the result of which will be logged for the goal of monitoring these sources of disinformation.

Primary Objectives

- Investigate the level of disinformation versus accurate information for specific topics on Twitter.
- Research existing work on similar subjects for best practices and potentially novel methods.
- Identify the success cases and shortcomings in existing works.
- Implement a basic finished product early on to pinpoint necessary features, then refine according to the above objectives.

- Implement multiple models that target varying collections of fact-checking sites to see which combination is best suited for the task.
- Identify best way to represent and maintain collection of logged sources of disinformation.
- Investigate both the scope and scale of damage caused by disinformation on Twitter.

Extension Objectives

- Design a user-friendly interface for searching and viewing collections of false claims and their associated sources.
- Research ways of combating sources of online disinformation, making use of the tool or tools created during this project.
- Reach out to fact-checking sites to investigate the potential incorporation of the project's end-product in their fact-checking process.

Relevance

This project comes during a global pandemic and in the wake of a large increase in amount spent on online advertising campaigns for elections at high levels of government. This has meant a move toward an increased level of vigilance from independent watchdogs, governments and the technology firms who are in part responsible for hosting the "fake news"; the project sets out to assist this shift in attitude by coming up with solutions to the issue of tracking sources of disinformation as well as to research the spread of disinformation on social media and its wider impact.

As someone who believes skepticism to be a useful trait in this day and age, I am becoming increasingly encouraged by the gradual adoption of fact-checking tips (checking primary and secondary sources, being conscious of the reputation of the source and general skepticism towards outlandish stories) and the cautious realization of the potential scale of disinformation and by extension misinformation online. As such this project is of a personal interest to me too.

Required Resources

Twitter developer account(s) to be able to access their APIs, specifically Twitter's Search API to return collections of Tweets relevant to a specific disinformation related query.

Institutional account access for overleaf.com, an online LaTeX editor which will be used for the writing and editing of this project.

An account on Trello for handling project management.

A computer with a Python 3.X installation and various related natural language processing, web-scraping and Twitter packages installed, such as but not limited to:

Natural Language Processing: NLTK, spaCy, and cdQA.

Web-Scraping: Requests, Beautiful Soup 4, lxml, Selenium, Scrapy.

Twitter: Tweepy.

System Resources such as CPU, RAM etc. should be covered by most modern machines as this is not a computationally intensive project however if there is a need to scale up the Sussex University lab machines or a cloud-based solution such as Google Colab will most likely be sufficient.

Time Management

Weekly Timetable

This is an initial plan of the hours in a week that will be dedicated to this project - it marks 10 hours total currently and is subject to change depending on the needs of the project. Dedicated hours can be seen below marked in bold.

Time	Monday	Tuesday	Wednesday	Thursday	Friday
9:00		Seminar		Seminar	**Project**
10:00	**Project**		**Project**	Seminar	**Project**
11:00	**Project**	**Project**	**Project**	Lecture	
12:00					Lecture
13:00	Lecture	Lecture		Workshop	
14:00		Lecture	**Project**		
15:00	Lecture		**Project**		
16:00				**Project**	

Bibliography

Chamberlain, P. (2010). Twitter as a vector for disinformation. *Journal of Information Warfare*, *9*(1), 11–17.

Fletcher, R., Cornia, A., Graves, L., & Nielsen, R. K. (2018). Measuring the reach of "fake news" and online disinformation in europe. *Reuters institute factsheet*.

Galitsky, B. A. (2015). Detecting rumor and disinformation by web mining. In *Aaai spring symposia.*

Kumar, K., & Gopalan, G. (2014, 09). Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, *4*, 14. doi: 10.1186/s13673-014-0014-x

Pierri, F., Artoni, A., & Ceri, S. (2020). Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PloS one*, *15*(1), e0227821.

Pierri, F., Piccardi, C., & Ceri, S. (2020). *A multi-layer approach to disinformation detection on twitter.*

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. Retrieved from `https://science.sciencemag.org/content/359/6380/1146` doi: 10.1126/science.aap9559