University of Sussex

Assessing the influence of sockpuppets on the Transgender Twitter discourse

Ethan Pannell

Computer Science and Artificial Intelligence BSc

Department of Informatics

University of Sussex

Supervisor: Luc Berthouze

May 2022

Statement of Originality

This report is submitted as part requirement for the degree of Computer Science and Artificial Intelligence BSc at the University of Sussex. It is the product of my own labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged. I hereby give permission for a copy of this report to be loaned out to students in future years.

Signature:

Ethan Pannell

Acknowledgements

I would like to thank my friends, family (including our kitty) and disability mentor for the emotional support as well as Luc Berthouze for all his support throughout this project.

Summary

Sockpuppets appear online everywhere without us knowing, usually aiming to spread disinformation and hate; sockpuppets are bots or humans who create accounts for such purposes. Discourse which is highly polarised attracts these users which further polarise it into more fringe group ideas; some of which are bigoted who then get banned and create new accounts or want to produce hate at a larger scale. One discourse which has been exploding is around trans peoples existence. We theorised that sockpuppets may be influencing the discourse to go to a certain direction depending on their view. We want to assess the influence of these sockpuppets within the twitter transgender discourse. Influence would be a measure for a particular users influence over a network, one measure is Retweet Influence which finds how influential a user is based off of the amount of unique retweeters retweeting their tweets. We used snowball sampling to map the discourse community through selecting 30 seed users which were collected via finding the most influential users who were replying under trans-related phrases near the time JK Rowling released her essay "TERF Wars". We snowball sampled with a two hops, collecting 1.8 million tweets from seven thousand users with a total of fourteen thousand users stored. We had to create a set of queries to capture the discourse successfully which was split into two types: neutral terms and non-neutral terms. Using the collected data we aimed to find suspicious users who could be sockpuppets, we used several methods such as: repeated text, replying on the same minute as many others do, and tweeting something which got more retweets than likes. We stored the data into sqlite databases and encrypted fields or the whole file using self created libraries. A linear SVM was used with extracted features which are indicative of a particular user through called authorship attribution. The final data implied that there were no sockpuppets within the discourse when we know this is extremely unlikely. The data collection for relevant data around the discourse was successful, but was difficult logistically to do this in a timely manner which we mitigated through experimenting with how we snowball sample. The features used for machine learning resulted in high accuracy but low F1 score. It was difficult differentiating between users who were heavily invested in the discourse and potential sockpuppets which were overt and produced large volumes of tweets like an highly active user as well as possibly not going deep enough into the discourse network to find the more overt sockpuppets. There could be more subtle sockpuppets within the data, we were focused on overt patterns.

Contents

1 Introduction

On Twitter, a user can easily create an account using an email address or phone number, and this opens up the issues of users exploiting this to create malicious user accounts which are either controlled automatically via the Twitter API or a human user who has another account which they use as an anonymous pseudonym to hide behind and express their true views online [1]. Both can be labeled under the category of a sockpuppet. Sockpuppets have a puppet master, the human who owns the accounts credentials for the account and can orchestrate multiple of them [2].

Sockpuppets can affect online social media platforms, such as Twitter, to spread disinformation; these sock puppets could just be random users or more organised such as state-sponsored [3]. Twitter in 2019 had 290.5 million users [4]; this is a huge attack vector for puppet masters to use. In 2017, it is thought that between 9 and 15 percent of Twitter accounts are bots and that bots use retweeting strategies to target specific communities of people [5]. Not all bots are bad; some are neutral or designed to be helpful, however, they have been used to infiltrate political discourse and manipulate them [6][7]. It was estimated that $\frac{2}{3}$ of tweeted links were posted by these automated accounts [7]. Past twitter data show similar margins to polls collected during the election period in 2016 from three million tweets [8]. During the COVID-19 pandemic, many conspiracies were propagating on Twitter (and as of writing there still is); there exists a set of bots that mainly posts conspiracy theories of the political type around the COVID-19 topic [9].

Given social media affect people and society as a whole, it is particularly interesting to understand how ideas propagate within an online discourse. There is a rise in reported hate crimes in the UK. Race, sexual orientation, disability, and transgender hate crime categories have risen over the year. Race was the highest at a 12% increase from just this year alone (76,158 to 85,268). Over the span of five years, there has been a 120% (in 2016/2017 1,195, in 2020/2021 2,630) increase in reported transphobic hate crimes [10]. This is only the reported hate crimes. Transphobia online, as well as any other kind of bigotry, is abundant, with collected tweets which span from 2016 to 2019 discussing transgender people on Twitter where 12% ($\frac{789615}{5547445}$) are being abusive [11].

The aim is to detect sockpuppets within the twitter transgender discourse and to assess the influence of the sockpuppets if they do exist. This research project is split into separate questions; we will look at the influence of users within the transgender discourse; try to detect sockpuppets within the discourse, and see how much influence they have on the discourse. As an extension, we will then assess the particular topics the users discuss within the discourse.

2 Professional Considerations and Ethics

2.1 BCS Code of Conduct

We comply with BCS Code of Conduct; it can be seen in [12]. We shall describe the relevant points to the project.

2.1.1 Public Interest

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

We will be collecting data on a sensitive discourse topic without the direct consent of those users, what we will do is pseudo-anonymise the data and only publish aggregate and/or paraphrased data so no content can be reversed searched to target a particular individual. The sensitive data collected will be stored on secure servers hosted by the University of Sussex.

b. have due regard for the legitimate rights of Third Parties*.

Yes, we will follow the Twitter developer policy while using their API.

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

We will not discriminate on any basis, we are looking to protect transgender people computationally through detecting sockpuppets within the discourse, not to further promote disinformation or discrimination.

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

Yes, we want to include all. If this research goes well, this could be used to detect potential sockpuppets who may harm the transgender community or people within the transgender discourse who are trying to harm others. The data collected will not parse out particular identities to discriminate on background, however, users' data collected on all will have to have internet access to create a Twitter account.

2.1.2 Professional Competence and Integrity

a. only undertake to do work or provide a service that is within your professional competence.

This project will be implementing already existing algorithms but to a new problem area or even the same problem area for instance influence measures for Twitter. Our professor/supervisor has said this is challenging but believes we can implement this.

b. NOT claim any level of competence that you do not possess.

We shall not claim competence in any area we do not possess, the course we have taken touches on each area within the research questions to give a good foundation as well as my research into the separate areas gives us an acceptable amount of competence.

f. avoid injuring others, their property, reputation, or employment by false or malicious or negligent action or inaction.

We shall avoid damaging others' reputation, property, employment by false or negligent action or inaction. No data displayed to readers will be reversible to a specific user.

2.1.3 Duty to Relevant Authority

The relevant authority for this research is the informatics department at the University of Sussex.

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.

We will not reveal any third party or confidential information, especially due to the sensitive group and community members within which will be pseudo-anonymised. We shall only discuss such information with those associated with the project.

2.2 Ethical Considerations

This project was subject to ethical application reference number ER/EP396/1 submitted to the Sciences & Technology C-REC on 26/11/2021 and was approved on 13/01/2022. See the whole application in 9.1.

Ethical considerations within this project can be summarised to three points:

1. Collecting data from users who do not know
2. A large number of tweets being collected
3. The data collected potentially contains data of a particularly sensitive community, transgender individuals

For the first point: All of our figures show only aggregate or paraphrased data so users will not have their IDs leaked via this paper or any of the tweet IDs. If we display text, this text is pseudo-anonymised users and is on a large enough scale which doesn't attribute to a specific user. We have accepted Twitters terms of service and developer agreement and applied to their developer platform under the academic research API, the application was accepted for academic use by Twitter staff. The users of twitter has accepted the Twitter ToS whole registering for their accounts granting us permission to use their tweets.

For the second point: We pseudo-anonymised the data via encrypting the user id column and using a new generated id to reference to the user instead. For data which was all sensitive information where we could not generate secure pseudo-anonymised fields, we encrypted the whole file and when in use briefly decrypt and load into memory if it was a file, if a database we decrypt it and work with the data then re-encrypt it. We implemented a small library for this process using Fernet encryption provided by the python Cryptology library [13], for per file/database we used age [14] which was integrated into python through the library we made by our self. We believe we tried our absolute best to keep this data secure while working with it.

For the final point: During data collection, we favoured methods to reduce the data-set size collected, in total we stored 1873093 tweets which were encrypted. Using this with the aforementioned implemented procedures carried out during this project will protect these individuals from their data being exploited or exposed to/by the public.

We completed all of the sensitive parts of the project on the Unix university server, and used Google Colab for training the machine learning algorithm with the anonymised feature set database due to the lack of RAM on the university server.

Great attention was given to creating a workflow that maximises the security of the data and minimises the likelihood of data leakage.

3 Background research

3.1 Twitter

Twitter is an online social network, specifically a micro-blogging platform. Users can sign up with an email or phone number. Each user has their own display name and username. A username is a unique string identifier for a user

which is attached to a Twitter user id. A display name is a public-facing name that acts like a nickname for a user, for example, you could have the username as "@HelloWorld" and have a display name of "foo bar".

A user can post tweets on their account; tweets are 280 characters max length text, which can have additional attachments such as up to four images, or one video, or one audio recording. Tweets can link to external web pages such as news articles, but this takes up characters. Users can follow other users to have their tweets appear on their timelines. The count of the users following and the followers are publicly viewable. Each user has their timeline, where tweets are displayed from the users they follow, this also includes retweets.

A user has several actions they can perform onto a tweet: like, reply, and retweet. Likes are a way of showing one likes the content of the tweet, and replies allow one to comment on a tweet, and a reply in itself is a tweet that can be replied to as well, creating long threads of them. Retweets are when you amplify a tweet by showing their tweet as a part of your collection of tweets, users can also quote retweet, which allows you to reference a tweet and add your own 280 character max length text to it too, usually adding on top of the tweet. For each type of action, a numerical value is attached showing the amount of that action that has been performed, one for likes, tweets, retweets, and quote retweets.

Within tweets, a user can mention a hashtag; hashtags are keywords or phrases which have a suffix of #. They are used to help with searching for a particular topic, such as *#javascript* for tweets around the JavaScript programming language which a user has tagged it as.

Twitter supplies three types of APIs. API stands for (A)pplication (P)rogramming (I)nterface. APIs allow a programmer to interface and exchange data with another service, such as a third-party service like Twitter [15]. The three API types are Standard, Academic Research, and Business. We will be using the Academic Research API, as this provides a higher tweet cap per month, 10 million tweets a month for academic research API while the standard only provides 500,000 a month. It allows access to "real-time and historical public data with additional features and functionality that support collecting more precise, complete, and unbiased datasets" [4].

The API allows us to collect data on the number of likes, retweets (both normal and quote retweet), and replies for each tweet. For user accounts, the user creation time, bio, followers, following, and more are available. We can use the historical API to get old tweets from given users or a search term or use the real-time stream of tweets which can then be filtered through for specific keywords.

3.2 Sockpuppets in Social Media

Estimates show that between 9% and 15% of active Twitter accounts are bots [5]. Not all are malicious, some are even helpful tools for people, however, there are still many which are harmful which use tactics such as to try and manipulate users without them knowing as well as exploiting the Twitter algorithm [6].

Social media has been exploited by people to push certain ideas to the platform as a whole, such as to push users into groups such as ISIS sympathisers in the Syrian revolution, the alt-right, and the activists of the Euromaidan movement. They exploit twitters algorithms to gain more awareness for their groups; this all occurred on Twitter [2]. Governments have also exploited social media networks using cyber troops[3] to try to skew the general populations' opinion such as for election periods. Cyber troops are government military or political party groups to manipulate public opinion via social media such as Twitter.

Data collected after the 2016 election showed that one can get the consensus of the population's political leaning as the data showed had similar margins to polls which commenced during that election period [8]. This shows the potentiality of the network having an effect in the real world.

3.3 Transgender Twitter discourse

The *Transgender community* is made up of trans individuals who include binary trans individuals and nonbinary gender identities[16].

The *Twitter Transgender community* includes the *Transgender community* as a subset of this community; additionally, it includes cisgender allies who either could be a part of the LGBTQ+ community or are heterosexual; in general, they are people who are affected by or interested in transgender issues.

Twitter Transgender discourse includes any discourse related to the topic of transgenderism. This includes The Transgender community, the Twitter Transgender Community, people generally discussing topical related content and can also contain transphobes / bigots.

The Transgender community is an online social network but an in-person social network too, while the Twitter Transgender community is predominately online due to it being based around Twitter as a platform, however, this does not mean that users do not know each other in-person.

Relationships between trans people as a support group helped with the shared experience of the psychological stress of being Transgender [17]. Trans youth online use unique strategies for moving through a binary gendered online world, creating their own communities [18].

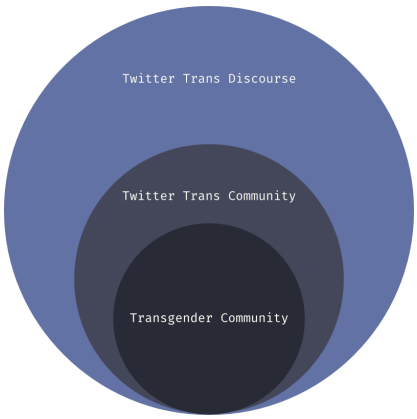


Figure 1: Community Subsets

3.4 Community Detection

Our definition of community is not to be interpreted as the term within network science. In network science, a community is defined as a subset of nodes within the graph where connections between the nodes are denser than connections with the rest of the network, this is not just a computational problem but a social and biological one too [19]. We just want to find one community, the Twitter Transgender Discourse.

The Twitter transgender discourse can be seen as a type of discussion community around the discourse; we will aim to detect, meaning to capture, this discourse community.

Community structure detection is usually a set of techniques in network science to detect multiple communities within one large set of data from a network/graph, but these attempt to capture all of the communities within the network [20] when we only want to capture one. To be able to map the discourse we need to apply specific capturing techniques for a single community.

One method is collecting tweets around terms of interest for the detection of a community around a certain topic, which would be the discourse. In [21], they use snowball sampling [22] at first which collected 119,156 users, which then with certain keywords in tweets to limit down the results to the users affiliated with ISIS only.

Another technique is snowball sampling [22], this is where a selection of users are chosen as **seed agents**, the user's followers from these agents are then added to the selection of users. This can be iterated, called **hops**, such as finding all seed agent followers (1-hop), and then finding all followers of the seed agents followers (2-hop). This technique was used to capture three communities [2]: the alt-right, ISIS sympathizers in the Syrian revolution, and activists of the Euromaidan movement. This method was used over the first method as it may potentially make it easier to observe the behaviours of the sockpuppets. They collected 106k users and 268 million tweets on the alt-right alone. This could collect mutual users who are not a part of the discourse however we can experiment with this.

Using the Twitter stream API is useful for research in general [23], giving us all the real-time created tweets. We could potentially use the technique of the typical meaning of community detection itself, however, typical techniques are not to detect a specific community but to find all the communities within the network. It may capture the specific discourse community we want to collect data on, however, this may be difficult from the vast amount of data generated on Twitter from users.

3.5 Measuring Influence

To know how much influence a sockpuppet has, we need some way to measure influence over the network. There is no definitive definition of what an influential user in a social network is within this research area. New research papers with proposed influence measures give different perspectives and methods which define an influential user. Influential users can be split into several categories, such as opinion leaders, influencers and, discussers [24], there are many more, see [25] for survey. In general, we can define a user as influential if their actions within the network can affect the behaviour of the other users in the network [25].

Twitter has its own metric system to see how well a tweet has performed; they simply count the number of times the tweet was seen, the number of times someone engaged with tweets, and more specific metrics seen in Figure 2. Note, however, these details are not available via the Twitter Research API, which means that we cannot make influence measures using these metrics.

However, as mentioned in 3.1, the metrics available to us are the number of likes, normal retweets, quote retweets, and replies for each tweet, as well as the tweet text itself. This data can be a metric for measuring influence as this is the way a user can interact with a particular influential tweet or account [26] [27] [28] [29] .

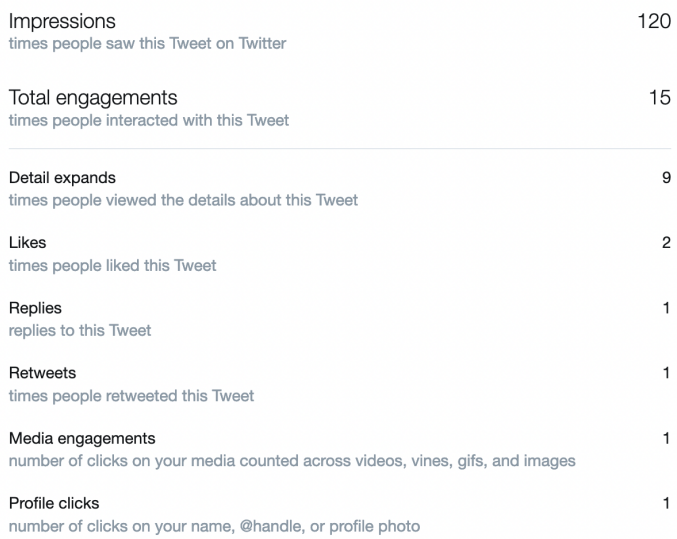


Figure 2: Official Twitter metrics

The general measures researched are indegree, retweet, and mention influence. Indegree is the number of inbound connections to a particular node. where the node would be a user and the inbound connections are the users who follow the user. Popular users who have a high indegree are not as influential as people may think for making users retweet or mention[26]; the topological measures such as indegree alone reveal very little about the influence of a user.

In [27], for the time published, there were pre-existing closed source tools to measure influence, one of which is Klout which uses more than 25 variables, they see influence as the "ability to drive people to action", which makes replies and retweets the most important factors of their measure. Another service is Twitter Grader; this measure scores out of 100 and is also closed source, but factors that contributed to its algorithm were the number of followers, Twitter Grader score of those followers, the number of tweets the user has made, update recency, follower/following ratio and engagement (such as retweet and mention ratio for when a user has interacted/engaged with a tweet).

In [25], an active user is defined as a user who participates in the network consistently over a period of time. We can define the activity of a user as the probability of a user seeing a tweet. However, Twitter users who exclusively read who may be active on the network will not be captured as there is no visible way they interact with the network unless they use an observable metric such as retweeting, if they do we can assume they have read the tweet, meaning that if they have done more active users are likely going to exposed to said new tweets, in turn, have the possibility to interact with them.

Influence can be split into two paradigms: influence is predominately from a small number of users who are very connected or persuasive, or, many users can accidentally become influential depending on unpredictable factors [30]. Twitter influence measures in current research have usually used metrics related to retweets, mentions, and less used, followers. Some researchers have used passive topology of twitter such as a followers/following graph or retweets and mentions graph. Other authors have looked into the problem of influential users given a certain topic. [25]

One method is *TwitterRank*, which is a topic-sensitive measure. It is an addition to the PageRank algorithm (an algorithm that determines how relevant a web page is given a search term [31]). TwitterRank measures the influence taking both the topical similarity between users and the link structure into account [29].

Another two measures are *Retweet Impact* (**RI**) and *Mention Impact* (**MI**) [28]. **RI** estimates the impact of the content created by the user in the aspect of retweets: $RI = RT2 \cdot \log(RT3)$. "The logarithms moderate the impact of overly enthusiastic users who retweet the same content many times" is mentioned in [25], however a user cannot retweet a tweet twice using the current version of Twitter. *RT2* is the number of unique tweets retweeted by other users, *RT3* is the number of unique users who retweeted author's tweets. **MI** estimates the impact of the content created by the user in the aspect of mentions: $M3 \cdot \log(M4) - M1 \cdot \log(M2)$. M1, M2, M3 and M4 can be seen in Table 1.

Next is *Social Networking Potential* (**SNP**) [27]; the equation is: $\frac{Ir(i)+RMr(i)}{2}$, where $Ir(i)$ means Interactor ratio, and is defined as: $Ir(i) = \frac{RT3+M4}{F1}$, meanings of the separate variables are in table 1. $RMr(i)$ is Retweet and Mention Ratio, and is defined as: $RMr(i) = \frac{\#tweets\ of\ i\ retweeted + \#tweets\ of\ i\ replied}{\#tweet\ of\ i}$. **SNP** considers many kinds of actions except from likes [25], this seems like a good measure to use because of accounting for those different actions rather than just retweets or mentions. 25% of the total importance is the number of published tweets and follow-up relationships while the other 75% conciders the numbers of replies, and the number of followers related to the user through retweets

ID	Feature
OT1	Number of original tweets
OT2	Number of shared links
OT3	Self-similarity score of similarity to recent tweet to the users previous tweets
OT4	Number of hashtags used
CT1	Number of conversational tweets
CT2	Number of conversational tweets started by the user
RT1	Number of retweets of other's tweets
RT2	Number if unique tweets (*OT1*) retweeted by other users
RT3	Number of unique users who retweeted author's tweets
M1	Number of mentions of other users by the user
M2	Number of unique users mentioned by the user
M3	Number of mentions by other of the user
M4	Number of unique users mentioning the user
G1	Number of topically active followers
G2	Number of topically active bidirectional following including the user
G3	Number of followers tweeting on topic after the author
G4	Number of friends tweeting on topic before the author

Table 1: List of potential metrics [28][25]

and mentions. The time complexity uses $O(T \cdot k)$ [25] where T is the number of tweets and k is the length of an auxiliary vector.

Two more are *TunkRank* [32] and *UserRank* [33]. They are both adapted from PageRank [31]. TunkRank was the first PageRank translation to be applied to Twitter, and is defined as [32][25]:

$$TunkRank(i) = \sum_{j \epsilon followers(i)} \frac{1 + p \cdot TunkRank(j)}{\#followees\ of\ j}.$$

This method only uses followers/followees only. UserRank was created to measure the influence of a user from their tweets relevance [33][25]:

$$UserRank(i) = \sum_{j \epsilon followers(i)} \frac{1 + \frac{\#followers\ of\ i}{\#tweets\ of\ i} \cdot UserRank(j)}{\#followers\ of\ j}.$$

The benefit of using UserRank versus TunkRank is that they "calculate dynamic coefficient for each user based on a number of his followers and tweets"[33], essentially extending upon Tunkrank to consider other influence metrics like retweets, while TunkRank is an adapted PageRank algorithm without any additional metrics used on-top of that; UserRank requires more data while TunkRank only needs the topological data.

3.6 Sockpuppet Detection

As we mentioned before, sockpuppets reside within online social media platforms; we need a set of methods to be able to detect them to measure their influence.

A formal definition of what a sockpuppet is a fake online identity; a puppetmaster is a person who controls multiple sockpuppet identities and could also be automating posts [34]. They usually are used to try and unfairly support a user's point of view on a topic [1].

Bots are observed to exploit mentions of each other creating a network of these bots to manipulate their influence on the platform [2], and others exploit retweets [5].

There are three main groupings of sockpuppet detection which are: verbal behaviour analysis; non-verbal behaviour analysis and similar-orientation network [1]. Verbal behaviour analysis is based on the textual content a user tweets out, trying to detect a user from a similar writing style, formally known as authorship attribution (**AA**) [35]. Non-verbal behaviour uses extracted features that capture a user's activity or movements such as times they tweet or tweets that are temporally close together [36]. Finally, a similar-orientation network is based on evaluating the similarity of sentiment orientations among user account pairs to construct a similar-orientation network [37][1].

There has been research using **AA** on textual content from Wikipedia users and their edits to detect sockpuppets [38]. They evaluated 239 features that capture grammatical, stylistic, and formatting the writer used; a state vector machine (**SVM**) was created to perform classification. This method has a high time complexity of $O((NR)^2)$; it is

also a verbal behaviour analysis technique [1]. The F-Measure of this particular algorithm is 72%; F-Measure indicates the fraction of valid classifications.

In [34], they identified three features to distinguish legitimate accounts from illegitimate accounts: activity, community, and post features, however, it needs the IPs of the users which we cannot get. The method still could be useful to draw from.

SocksCatch [39] is a non-verbal behaviour analysis technique. It is comprised of three phases: data collection & selection; detection of the sockpuppet accounts using machine learning and finally grouping of sockpuppet accounts using graph theory. This process is complex and lengthy, as seen in figure 3. SocksCatch has a true positive rate of 92.6%; a false positive rate of 7.8%; and an F-Measure of 92.6%. When using an SVM machine learning algorithm, but it can range in between 89% and 95% for correct detection. It is better performing than similar algorithms such as [34] or [40].

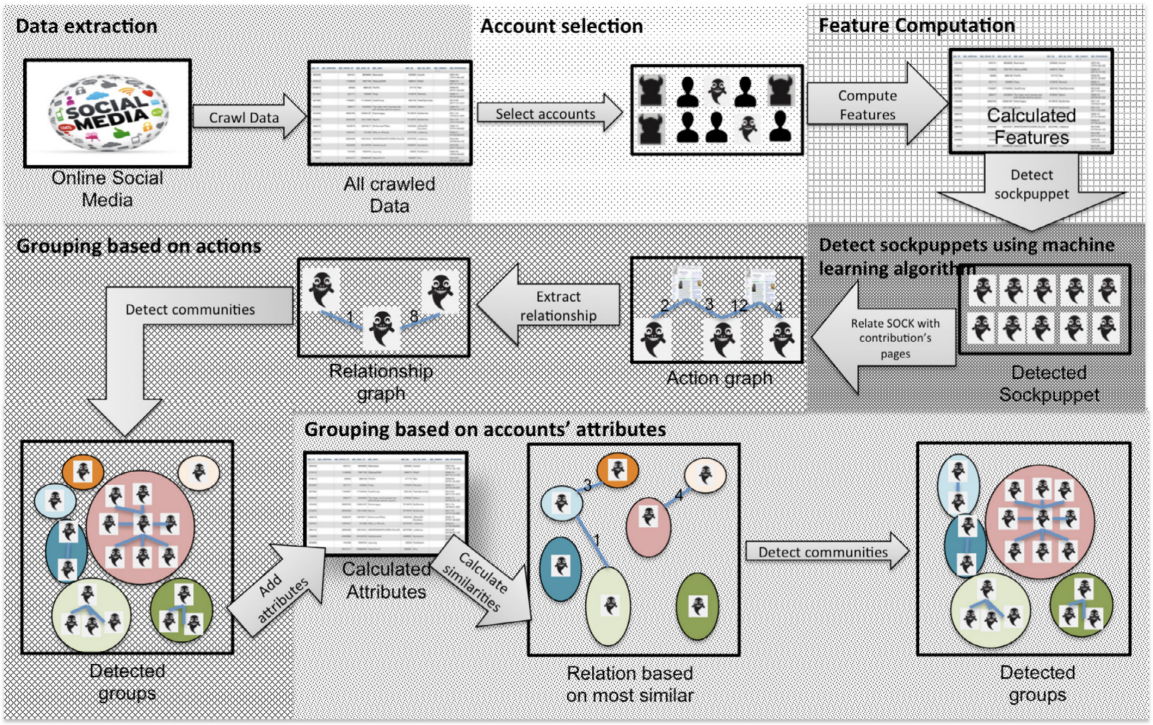


Figure 3: SocksCatch's main steps, image reproduced from [39]

In [37], they use a similar-orientation network method. The paper mentions how verbal and non-verbal analysis can be avoided by trying on purpose to not fall into either category of analysis by changing their language and more, this makes it difficult to gauge how many sockpuppets that are really within a dataset. They observe that sockpuppets want to recover similar social structures as the sockpuppet user can guarantee a similar propagation impact. They turn the problem into a subgraph similarity problem, which slightly outperforms other detection techniques made in the time the paper was published, 2018. They used a dataset from the social media site *Sina Weibo* for their detection algorithm which is a micro-blogging platform so it can still be applied for other micro-blogging platforms such as Twitter. Three users, a, b, c, each user would have the users they follow and interact with such as mentions or retweets. User a and b are a sockpuppet pair, as they share almost all of the same users they follow and interact with. a and c are not a pair of sockpuppets and only have two mutual users they are following, which shows they are not trying to gain the same social network structure to reach the same propagation impact. Collecting a specific discourse community which will likely have a lot of users who follow the same people, but it would seem to be unlikely to have exactly all the same users they are following. This method has the average F-Measure score of 83.5%, which is better than the **AA** method used on Wikipedia [38], but worse than the SocksCatch [39] method.

In another paper [41], they review current sockpuppet detection methods and if they can be used for real-time detection of bots. An example of a bot post points out a tweet that has very high retweets but two likes, which is explicitly a sockpuppet account that is amplified by the 412 other retweeters which show suspicious activity that would be an indicator that they are a bot or sockpuppet. Their results show that detecting bots or a sockpuppet campaign in real-time is impossible with current researched methods. Fortunately, for this research project, we will not need real-time detection.

4 Research Questions

The overarching question is how much influence do sockpuppets have; we are limiting this to the Twitter Transgender discourse that could very likely have sockpuppets residing within; this gives us the opportunity to detect and measure the influence of such detected sockpuppet users to find the influence they have over the given discourse.

Our research questions can be split into two main sections, which are the minimal viable questions to complete the main aim of the research, and the extensions to gain more information from the data if we have enough time. For the main questions, there are four: how do we detect the Twitter Transgender discourse; how do we measure influence; how do we detect sockpuppets and how much influence do the sockpuppets have. We will be implementing methods covered in the background section and try experimental methods.

How do we detect the Twitter Transgender discourse? A community can be found from who follows who and who interacts with who explicitly. From 3.4, we will aim to use the snowball sampling technique with a max of 15 users to be sampled from, we will need to identify large accounts which are part of the trans discourse. We could also try to search for tweets containing words or short phrases associated with the discourse. Experimenting with a combination of both where we sample users who have a keyword or phrase in their bio could be another avenue we could explore to see its effectiveness. The the initial hashtags we will try are: "#trans", "#transgender", "#enby" and "#nonbinary"; we chose this as a starting point as each of the hashtags are relevant to the trans discourse and they would be discussed by users within the discourse mostly.

How do we measure influence? We want to know how much influence a user has over the network to determine who is leading the conversation the most within the network and who users interact with the most. In 3.5 we discuss different methods of measuring influence, we would like to test using basic measures: RI and/or MI [28], as well as more complex methods: **SNP** [27] and/or **UserRank** [33] depending on time constraints. We can apply these measures to users in the collected network to grasp the general influence of users within the network, and then once sockpuppet detection has been complete, we shall select the measures for those users specifically.

How do we detect sockpuppets? To be able to measure the influence of sockpuppets we need to discover them. Sockpuppets try to alter the conversation in the discourse to try to harm them or to push a particular view. From 3.6, we shall implement methods researched; these include subgraph similarity matching technique [37], and depending on time constraints try to use authorship attribution [38][35] and/or a non-verbal analysis such as SocksCatch [39]; however, we do not necessarily need to know the puppetmaster groupings unless we want to measure a particular groups influence as an extension.

How much influence do the sockpuppets have? This is a combination of the influence question and the sockpuppet question, once both are complete we can use the best performing method for them to then collect the statistics of the influence the sockpuppets have over the discourse.

Extension questions are based around natural language processing (**NLP**): What topics are the tweets about? What are the most used words? This would be using a simple approach, using frequency analysis to examine the most used words used additionally to explore a singular keyword topic associated with each tweet to see what topic is used. This is to see if the sockpuppets may only tweet about one topic area on a higher frequency than other members of the twitter transgender discourse which could be a rather simplistic approach to detecting sockpuppets within the network using NLP.

5 Methodology

5.1 Development Environment

5.1.1 Server Setup

To comply with the ethics application, we worked on the university server; we used ssh to connect the Unix server. Applications we needed were outdated, such as git and GPG, so we used conda; conda allows users to manage their python version as well as install new packages for your local user. We installed git, fish shell and poetry via conda. Poetry allows a user to easily install python packages and create python projects.

We connect to the system via ssh, using visual studio code as our editor and using the remote connection extension. We use conda to manage the python version, we set it to 3.9.7. For managing python packages we used poetry which makes it easy to add new packages to the environment. We use a folder for our library of commonly used code in between notebooks and a folder for the notebooks themselves.

We created an environmental variable to store the API key in for the application to take from to use for the requests. The key for encryption will be stored on a local USB stick and will never be stored long term on the server side, only in RAM.

5.1.2 Python and libraries

We decided to use a fast iterative prototype approach while doing this project, we also had a folder which included a self made library for commonly used functions, albeit not all commonly used code was added. We set up a virtual environment with poetry and used conda to install python 3.9.8.

We used several libraries, the important ones include:

1. Requests[42]: Wrapper libraries for the twitter API being limited, `conversation_id` attribute is not commonly implemented in the libraries, for what we wanted to do so it was easier to implement our own wrapper for the API.

2. Cryptology[13]: For encrypting fields in the user database.

3. python-dotenv[43]: For keeping the API key in a file and removed from the GitHub repo of the code through the use of .gitignore

4. numpy[44]: For math which is more complex than python provides without a library

5. nltk[45]: For natural language processing used for authorship attribution and general analysis

6. matplotlib[46]: used for plotting from data

7. networkx[47]: Used for network graphs

8. peewee[48]: The Object-Relational Mapping (ORM) library used for the databases; it makes a database managable via objects in orientated programming

9. seaborn[49]: Also for plotting but more pretty

10. sklearn[50]: For machine learning

5.2 Data Collection

5.2.1 Data Storage

For the database, we have encrypted sensitive fields such as the user id to pseudo-anonymise the data; We are using Fernet encryption provided by the cryptography package; we used Fernet as it was simple to use and guarantees a message cannot be read without the key. To make sure the password never is exposed within the bash logs or in the stored in the environmental variables, we will paste the password into an getpass function input provided by python by default. The database is the file based database sqlite3 which makes it much easier to encrypt and handle.

We stored all the twitter JSON responses. When using .py files and not using juypter, issues had appeared where we didn't program aspects correctly until we collected all the tweets which it then broke so we lost all the data and reduced the amount of data we could process. We had to store the raw JSON while programming but this exposes even more critical information; we created a small python function to use age encryption tool[14] to interface between python and the command line tool to encrypt the whole file as a backup if the code broke. Fernet is not recommended for large files so where we couldn't use it, age was used instead. Fernet was used for strings of text while age was used for a whole file, they both use symmetric encryption. We store the key on my USB so if someone somehow were to decrypt my personal computer SSD they would still not have access to the password.

Experimenting with the raw storage of the requests, once we snowballed and the data also got much larger, storing the raw requests as json meant we had to load into memory iteratively. It took 30 hours to load in all the data to process the tweets after code had broke. We experimented with two other storage types, MessagePack and Pickle. Pickle reduced it to 45 minutes to load in the data without having to iteratively load it in. Loading using MessagePack took greater than the pickle so we stopped loading in the data. When storing large data as backups while working on the data before adding to the database we used pickle files then encrypted them using age.

For the data-set created from all the tweets collected, we encrypted the whole database as the whole thing was sensitive. We used the anonymous IDs for when exporting the feature sets created during the authorship attribution stage which is aggregate data not showing the full contents of the tweet. While working on the data it was decrypted and once finished it was re-encrypted and the decrypted version was deleted.

5.2.2 Database structure

In total there were five databases, three of which were the same but due to file size constraints on the server we could only have files with the size limit of 1GB so it was split into three.

The first was the database named ***data***, this stored the users and the relations between the users. This is primarily to store the users collected through the snowballing, the user ids are encrypted using Fernet and everything else is non-sensitive. The hop is to indicate which hop the user was collected in, and seed is for if the user is a seed they should be used as one when snowballing. This can be seen in figure 4.

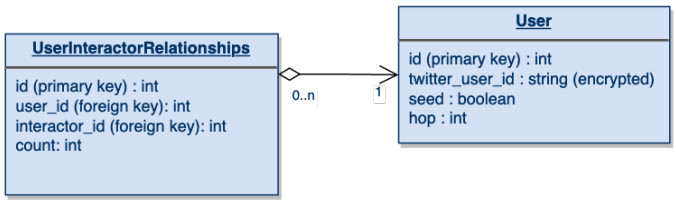


Figure 4: Data database model

The second was the database named ***dataset***, this stores the tweets created by the user. The data split table is for when we train the machine learning algorithm so we split the data 70/30 per user rather than 70/30 over the whole data-set so only some users would be included in the training when we want all the potential sock-puppet users to be within the data. This can be seen in figure 5. The database file was encrypted using age.

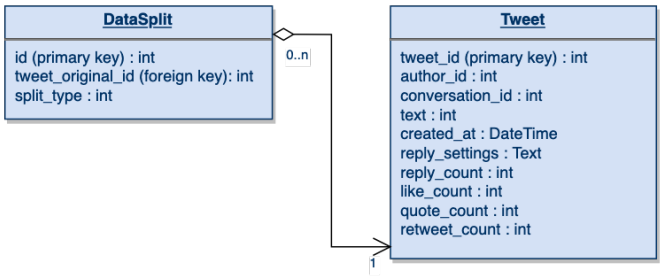


Figure 5: Dataset database model

This last database is comprised of three different databases with the same schema under different names due to the file size constraint of 1GB per file. This stores the pseudo-anonymised user id, the pseudo-anonymised tweet id, the split type (0 for training, 1 for testing) and the feature set derived from the tweet. This can be seen in figure 6.

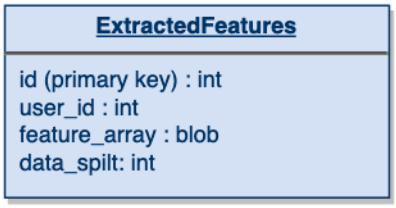


Figure 6: Feature database model

5.2.3 Initial Seed User Collection

We merged community detection and influence measure into this stage due to the initial idea of manually selecting users would be biased due to the users picked may not be as influential as we think or only have a particular view or following. We created a new method of collecting users from a particular discourse using a query and several filtering stages to get the most influential users within a search term given a date. Much of the project time was spent in this section due to having to prototype the method to be fast enough to fit in the limited time-span of the project.

Figure 7: Flow

5.2.3.1 Collection

The first chunk of our method was to collect relevant tweets; calculate a rough influence measure which wasn't dependent on collecting more data (called the naive measure) and get the maximum of 200 top users. We calculated the measure by simply doing a summation for each singular metric (the retweet count, the reply count, the like count and the quote retweet count) and for each tweet from a user within an array. We then do a summation of the values within the array for each user to get their naive metric. We select the maximum of 200 top users.

After, we collect the count of tweets the 200 users make which are relevant to the discourse and select the maximum of 100 top users.

We gather the data necessary for calculating the RI measure, which is the unique retweeters and the count of tweets. Instead of taking the amount of tweets the user has made over their whole lifetime, we will collect tweets from three points in time from the user, collect the unique retweeters and count the amount of tweets we have personally collected. We are doing this because logistically collecting all tweets of the 100 users would take a long time and run down our API tweet count drastically. We select the maximum of 30 top users.

Collecting the initial seed users for snowball sampling followed the method in figure 7, we ran this twice with different queries. The initial query was using neutral terms however this was not giving us a broad view on the discourse, we were only getting pro-trans users post-filtering who were participating in the discourse. In the article [51] there was research showing that there were bots who were participating in anti-trans discourse, meaning that they use different language within the discourse which was hiding it from our results. We used hashtags and phrases which are used within the discourse. We have two queries, the neutral terms query and the non-neutral query which used language which is used within anti-trans discourse, these will be detailed in the next section after this one.

For the time-span collected the tweets in-between (inclusive) the dates `2020/06/10 - 2020/06/20`. We used this date due to JK Rowling having controversy's happening around this time, specifically her essay "TERF Wars". For the JSON request we sent we had to add the bearer token header with the API key generated from the Twitter application. The API endpoint for we used for this stage in the process was ***/2/tweets/search/all***[52]; it is rate-limited to 300 requests per 15 minutes or one request per second. Using the python requests library we sent the JSON request

parameters to ask for 500 tweets per request, get the author id and the metrics, we specified the date-time in the start and end time parameters and attached the query for the trans related content to it. We processed all the tweets between these points of time, totalling to 415200 for neutral query and 199283 for the non-neutral query.

K-Pop accounts consistently appeared above all users if we did not filter them out. Twitter ignores symbols in the query, typically K-Pop accounts use tags such as "[TRANS]" at the beginning of the tweet, but excluding that specific string excludes all content with "trans" (case insensitive) within the result. Instead we had to create a list of regex's to filter the content out post initial tweet collection.

The second step was to calculate the naive measure which we had created, we calculated the measure by simply doing a summation for each singular metric (the retweet count, the reply count, the like count and the quote retweet count) and for each tweet from a user within an array. We then do a summation of the values within the array for each user to get their naive metric. We get the top 200 users and move to the next stage.

To actually know if our initial users are talking about the discourse we get the count of tweets given the discourse query from earlier over the span of the year of 2021 to see if they are still actively talking about it in the future rather than just a set time and that they haven't just made a passing statement just once about the discourse. We used the ***/2/tweets/counts/all***[53] endpoint with the same query used for the initial two points in time data collection, the new start end end time for the whole year of 2021 and the granularity by day so we get the count of the relevant tweets per day rather than per minute or per hour as this means less requests and easier to calculate the counts over the span of the year. The endpoint was rate-limited to also 300 requests per 15-minute window. We pick the top 100 users who discuss the topics and move forward.

To calculate the RI metric, we need to collect the users who have retweeted the users tweets, find the number of unique retweeters and the amount of unique tweets. We collect the user tweets in between three time-spans, we chose: `2021/04/1 - 2021/04/30`, `2021/08/01 - 2021/08/30` and `2021/12/1 - 2021/12/30`. To collect specific user tweets we used the endpoint ***/2/users/:id/tweets*** [54], 1500 requests per 15-minute window but maximum of 100 results per request. Each of the tweets has a unique id, this ID can then be used to get the retweeters of those tweets to calculate the RI measure, we use ***statuses/retweeters/ids***[55], another 300 requests within 15 minutes rate limit. Using is we can create a set per user who are the unique retweeters over all the users tweets.

For the SNP measure, we have most of the data to calculate the influence of these users, however we don't have the unique users who are following the user. Using the ***/2/users/:id/mentions***[56], we can get the mentions between the time-spans and get the max result of 100. 450 requests per 15-minutes with maximum of 100 results per request. We also do not have the follower count, so we can use the ***/2/users***[57] endpoint with a list of users which we want to get the follower counts from. With these results we can calculate the SNP measure using the previous data collected and the new data. We get the top 15 users and these will be the seed users for this query.

We repeat this process twice for the two different queries, resulting in 30 users we can snowball from.

5.2.3.2 Queries

We used two queries, a neutral query and a non-neutral query, using the query construction guide provided by Twitter [58].

The neutral query: ***("trans" OR "enby" OR "transgender" OR "nonbinary") -"eng trans" -"#transporn" -"#porn" -is:nullcast lang:en -is:retweet -is:reply***

The non-neutral query: ***("genderist" OR "genderism" OR "gender cult" OR "adult human female" OR "#SexNotGender" OR "#IStandWithJKRowling" OR "#SexMatters" OR "#BiologyNotBigotry" OR "#WarOnWomen" OR "#IStandWithJKR" OR "Gender Critical" OR "#IStandWithMayaForstater") -is:nullcast lang:en -is:retweet -is:reply***

The neutral terms consist of terms used universally to refer to trans people in a neutral means. A person who may be transphobic still may use a neutral term to refer to the discourse. As they are neutral terms, much more tweets occur with these text as passing statements rather than discourse exclusive; we have to remove certain text from our tweets which are irrelevant.

The non-neutral query consists of hashtags used by typically anti-trans users, we had to include this second search query due to the initial use of neutral only, only got pro-trans discourse users. We want to capture tweets across the discourse, not just neutral users; these hashtags consist of language found used by anti-trans discourse users we have observe used. A pro-trans user can also use these hashtags as a way of trying to combat transphobia which is being circulated within these hashtags.

Within the query we use brackets, quotations and other symbols in our query, a broken down explanation will be detailed. Quotations match the tweets which include the exact text within the tweet. Parentheses can be used to create groupings, so once that group query has been executed we can subtract certain tweets from it; the - operator is used as a prefix to remove results which match to certain criteria. **is:nullcast** captures tweets used for ads, these are negated using the - prefix; we remove retweets through **-is:retweet** otherwise we will get the same tweet many times from the users who have retweeted it; we don't want replies as a tweet which isn't riding on the back of another

popular tweet needs more influence to become popular. We want all our tweets to be English, so we use **lang:en** to limit the captured tweets to English only.

5.2.4 Snowball Sampling

Initially, we thought that simply getting the followers of the people who would follow the given user would be the best approach to the snowball sampling approach. There were two issues with this approach: followers may not indicate that they are active within the discourse but may follow the user for a different reason, and the rate-limit for getting followers of a given user would take far too long (longer than the time we had remaining). Our other approaches were to get the users via likes, retweets or replies. Likes need less user engagement with the topic, and because it's easy to do likes usually are very large volume even compared to followers so the same issue will occur. Retweeting is more impacting, as it re-posts the tweet onto your account showing that you have interest about the topic in the original tweet more than just a like. Replying is more of a higher engagement in the area than a retweet due to the user having to type out their thoughts and tweet rather than just pressing the retweet button. We decided to try and snowball via retweets and replies.

We experimented with snowballing via retweets and replies, we will discuss these results later; in short, we decided to go with replies due to it being less explosive means and taking a shorter amount of time to collect the users. Additionally to also reduce the time, from the collected tweets we took the top ten most influential via the naive measure and collected the retweets/replies to those tweets. Within the results, we have modelled the two graphs of connections to the seed user to see the distribution and separation between the groups within the discourse for the two snowball approaches, figure 9.

Continuing with the snowball approach, using replies only, we use the ***/2/tweets/search/all*** with a query which searches for the tweets given the snowball users from a certain run, the query is a combination of the neutral and non-neutral query with the additional operator of having the ***FROM*** keyword to search for tweets from a certain user. We collect all the users tweets from the year 2021, from there we can get the count of replies per tweet; if the tweet has replies we can get the conversation id. The conversation id is an identifier for a particular thread of tweets [59]. We can use that within the search criteria to then get the users who have interacted with the thread of tweets. For the first hop proceeding the initial user collection we got all users who interacted within the thread, but due to logistical constraints from the tweet count blowing up to thousands of tweets replying to one thread, we also require the user to be replying to the user who we are snowballing from.

Each tweet will be relevant due to each search requires a keyword from the response to be included within it.

The query used: *"trans" OR "enby" OR "transgender" OR "nonbinary" OR "genderist" OR "genderism" OR "gender cult" OR "adult human female" OR "#SexNotGender" OR "#IStandWithJKRowling" OR "#SexMatters" OR "#BiologyNotBigotry" OR "#WarOnWomen" OR "#IStandWithJKR" OR "Gender Critical" OR "#IStandWithMayaForstater") FROM:**UserIdHere** -"eng trans" -"#transporn" -"#porn" -is:nullcast lang:en -is:retweet*

After each snowball, sans the initial user collection, we purge/filter down the database to include users who are connected to two users, meaning: they have replied to two different users within the discourse, that they have two comments to one user and that they have greater than or equal to 2 comments total. This filters out users who have made a passing statement about trans discourse making the data more reliable to have users who actually interact with the discourse on a wider scale.

We snowballed twice, the first was the users who interacted with the initial collected users, and the second was the users who interacted with the aforementioned users to tie up the loose ends to close the graph. We collected the tweets of the last users but didn't snowball with them and purge using the filter.

In the last snowball, there was an issue with our method where we originally was collecting all users who replied to a given thread where a user replied to it with the relevant text from our query however it was acting too well, where we were collecting thousands of tweets from just one thread, so we has to restrict the collection to just a direct reply to that user within that thread which drastically reduced the data collected.

5.3 Sockpuppet Detection

5.3.1 Finding Suspicious Users

The first group of suspicious users were the intersection of three sets, the suspicious retweeters, the suspicious times and the suspicious text. We classified a tweet as suspicious based on retweet if they had more retweets than likes. Sock-puppets may potentially all tweet on the same minute, we collect the top 250 times where the tweets have hit a peak within a ±minute time span. The last set was the users who participated within the most used strings of text within the data pulled from the discourse; we normalised the text through removing mentions, links and tokenizing the text; and finally hashing it using FarmHash64[60] as a low collision hashing algorithm rather than cryptographically difficult algorithm.

The second set was using a partially implemented version of the sockpuppet detection method in [37] using graph theory. We collected users who interacted with one other user more than 150 times; we used only the graph similarity measure for interactions and selected the users who have a score > 0.5.

The graph is constructed where a user is a node, and an edge is directed from one node to another if one node replied to another. The edges are weighted with the log of the total count of replies from one node to another node.

Symbol	Description
V_u	The set of nodes which have replied to centre node u
$X_{u,v}$	The set $V_u \bigcap V_v$
x_i	The ith element of X
$w_{u_i}^{u_j}$	The weight of the shortest path between two nodes u_i and u_j
$p_{u_i}^{u_j}$	The path of the shortest path between two nodes u_i and u_j

Table 2: Symbol definitions [37]

The graph similarity measure equation for the graph follows, using table 2:

$$\Phi(u_i, u_j) = \frac{w_{u_i}^{u_j}}{|p_{u_i}^{u_j}|^2}$$

$$H(u, v) = \frac{\sum_{i=1}^{|X|}(\Phi(u, x_i) + \Phi(v, x_i))}{|V_u \bigcup V_v|}$$

The third set was similar to the previous, just using users who had interacted more than 50 times with one user to get more users within the intersection measures.

5.3.2 Authorship attribution: Feature Set

We extract the features from the text after normalising it through removing identifying information and links it using an extremely similar extraction technique to [38]. Each tweet is it's own document and each user has their own set of documents.

There are 240 features, one more than the aforementioned, reinterpreted from [38]:

Total number of characters: the amount of characters used within the document, this is used to get the users general length of their document as they may type at similar lengths per document.

Sentence count: This tries to capture the general amount of sentences a user writes per document, as there is a limit on word count on twitter, users may use longer sentences to fill that space or several smaller sentences to fill the character count, or maybe small one sentence statements.

Total number of tokens: A token is a sequence of characters which do not include white-space or punctuation. This can help find how many words the user usually uses per document.

Total alphabet count: The count of all alphabetic characters within the document. The user without knowing may create a pattern within the characters they use consistently across documents.

Total punctuation count: The count of punctuation symbols within the text. A user may favour using many punctuation marks or a sparse amount.

Two/three continuous punctuation count: Where a user has typed three consecutive punctuation marks, two example being "!!!" or "!!.". A user may do this stylistically and use this across documents.

Total contraction count: Contractions are words which are combined such as "won't" and "shouldn't". Users may favour contractions over the long version of such text.

Parenthesis count: The amount of parenthesis' used, there are three types we counted: brackets, comma and dash parenthesis.

All caps letter word count: The count of words which are all uppercase, usually showing emphasis.

Emoticons count: The count of emoticons used in the text. We have included both ASCII and Unicode emoticons, such as ":)" and "😄" compared to the original paper which only included the ASCII variant.

Happy emoticons count: Same as the aforementioned but for positive emojis, however we included a wide range of larger range ASCII happy emoticons using [61] and positive emojis.

Sentence count without capital letter at the beginning: If a user doesn't include a capital at the beginning of the sentence, such as using a number or lowercase alpha character, it will be counted.

Quotation count: Counts the amount of quotations used, only counts double quotes and has to be in a pair.

Parts of speech tags frequency: This is split into 36 separate features, one for each POS tag we will use, we ignore the punctuation tags however.

Frequency of letters: This is the frequency of letters used within the text, normalised by the amount of alpha characters in the text, contributes 26 features.

Function words frequency: Takes up 150 features, using the list of function words in [62] and counts the frequency of them appearing in their text.

Small "i" frequency: Some users may consistently make the mistake of using a lower case i, or for stylistic choices.

Full stop without white space frequency: Similar to the previous point, where users accidental or on purpose consistently put a period and a word after with no white-space in-between.

Question frequency: Some users may use question marks more than others, the question marks don't have to be used in a grammar correct way.

Sentence with small letter frequency: Versus **Sentence count without capital letter at the beginning**, this is where a user explicitly only starts a sentence with a alpha character, and that character is lowercase.

Alpha, digit, uppercase, white space, and tab frequency: The original paper found out that the distribution of these characters varies from author to author making it helpful with attributing text to an author.

'A', and 'an' error frequency: Authors may consistently make errors when using the incorrect usage of "a" in place of "an" and vice versa.

"he", "she" and "they" frequency: Users may use a particular pronoun more than others or a particular distribution of using such pronouns across documents.

5.3.3 Authorship attribution: Machine Learning

We assume the suspicious users collected are sockpuppets, training the machine learning algorithm and running it against the non-suspicious users to check if they have the same author.

We split the data into a 70:30 training:testing split for each user in the database and created a new pseudo-anonymised database with the pseudo-anonymised user id and the feature set for each document so it cannot be reversed back to the original text given just the feature set.

Using a pipeline with sci-kit learn, we created a process which tool the data, over-sampled it using SMOTE[63], used a simple scaler and fed that into a linear SVM. In [35], they also used an SVM, this is why we used one too.

For training we got the training data for each user in a given group of suspicious users, using this we trained the SVM. To check if the users are being identified correctly to the tweet text we check it against the testing data for the user.

For each user within the suspicious group we used the classifier to get the potential original tweet author pseudo-anonymised. A voting method was created from the briefly mentioned method in [35] to create a confidence; for each user we classified their tweets, e.g., User A could have 50 tweets and each tweet has an suspicious user attributed to it from the SVM. Going through each tweet we collect the occurrences of the detected accounts, using this we can create a confidence where we get the highest counted suspicious account and divide it by the amount of tweets they have. An example could be that a user has 50 tweets but 49 of them were associated to one user, that would be high confidence, while another could have 50 tweets but the highest occurrence could of just been once for a particular user which would be a very low confidence.

After extracting the feature sets and creating the anonymised databases, we trained the linear SVM on Google Colab [64]. Example row of anonymised data: `ID: 6116, USER_ID: 106039, SPLIT_TYPE: 1, FEATURE_SET: [190, 4, 40, 11, 144, 7, 0, 0, 1, 3, 0, 0, 2, 0, 2, 1, 2, 0, 0, 3, 4, 0, 0, 0, 0, 5, 3, 0, 1, 0, 0, 2, 0, 3, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 3, 0, 1, 0, 0, 0.04861111111111111, 0.0, 0.020833333333333332, 0.027777777777777776, 0.1597222222222222, 0.013888888888888888, 0.020833333333333332, 0.034722222222222224, 0.0763888888888889, 0.013888888888888888, 0.006944444444444444, 0.027777777777777776, 0.013888888888888888, 0.08333333333333333, 0.06944444444444445, 0.034722222222222224, 0.0, 0.08333333333333333, 0.09027777777777778, 0.0625, 0.027777777777777776, 0.034722222222222224, 0.013888888888888888, 0.013888888888888888, 0.020833333333333332, 0.0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 144, 4, 5, 30, 0, 0, 0, 0, 1, 0]` we believe this doesn't break any ethics as it's only aggregate and anonymised data so no sensitive data was leaked when using it on google servers. The ID is the pseudo-anonymised tweet id, the user id is the pseudo-anonymised id of the person who posted it; no other data touched the google servers which could connect the pseudo-anonymised ids to the actual ids.

5.3.4 Measuring Sockpuppet Influence

As data collection has concluded, we cannot find the influence via twitter, however we do have the graph created from our user database with the interactions and the count of interactions between each user. The graph is the same as the

graph similarity measure however instead of the log of the count between users for the weight, we just use the raw count value. The edges are weighted with the total count of replies from one node to another node.

Using this network, we originally thought using TunkRank or UserRank could be appropriate, but an easier method was just to use PageRank[31] as the way our network is now structured and that we cannot collect more data as for UserRank[33] and TunkRank[32] relies on the follower count while we only have the interaction counts. PageRank will use the weight of the replies for measuring the influence for how connected they are between them and the amount of connections they have inbound to the user. We use the networkx implementation of PageRank to calculate the values for each user within the network regardless of if their tweet data is stored in the dataset database. Using these values we can check the average score for the sockpuppets and compare them to the non-sockpuppets with their rankings.

5.4 Category Ratios

This section was originally an extension however it became intertwined with finding suspicious users. We used the query keywords as a list to be used to search if one of the terms appears within the tweet text. We categorise them into a count of each type if a term appears in the text; because twitter uses a different search algorithm than ours there is an error margin where twitter switched around words within a query to match a tweet. We will have four categories: both, includes neutral, includes non-neutral and neither. Using the data returned we can compare if the suspicious tweets has a different distribution of tweet category versus the non-suspicious tweets. The categories doesn't strictly mean that the people in one category is exclusively made by pro or anti trans users as mentioned previously.

6 Results

6.1 Data Collection

6.1.1 Initial Seed User Collection

Interestingly, when using neutral terms only trans positive users appeared after filtering; this shows that even though there is a lot of negative discourse about trans people, the people who hold the most influence over terms which are used to refer to them neutrally is dominated by users who support trans people which is a great thing.

We were able to collect users influence while narrowing down the data. The means in 3 show that the non-neutral have higher SNP score as a mean but it's primarily held by just two users seen on the right in 8, while neutral users had a lower mean score but each users had a value visible. This could mean that neutral term users are more intertwined creating a more level influence between them while non-neutral is help by few users but they are very influential over their query.

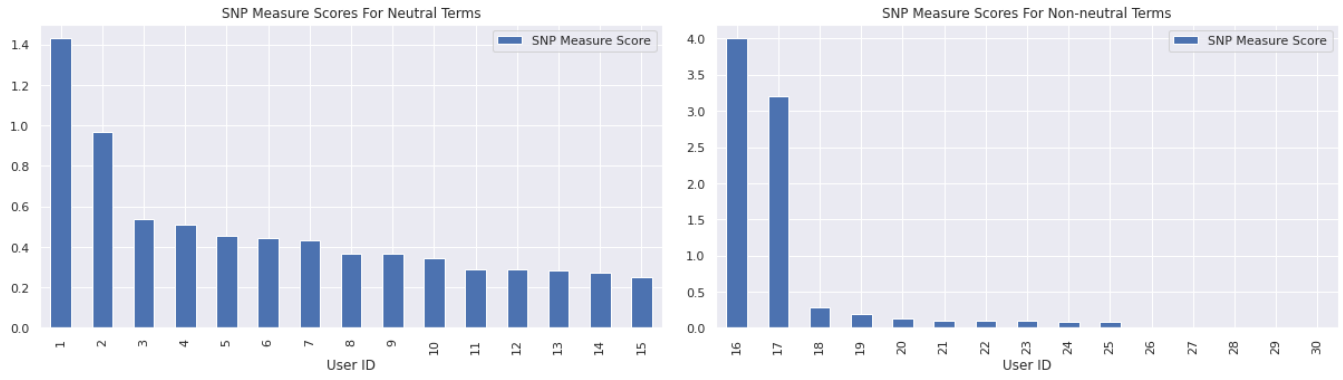


Figure 8: SNP Measure Distribution

Measure	Neutral Members	Neutral Mean	Non-neutral Members	Non-neutral Mean
Naive	200	84153.21	200	8841.89
Discourse Count	43	290.67	16	324.8
RI	30	581.70	16	497.04
SNP	15	0.48	15	0.55

Table 3: Mean influence results from each stage

6.1.2 Snowballing

In figure 9 on the left, there are 135185 users and 173290 connections between them. An average of 1.28 connections per node, however all users except the first thirty initial users are interacting with someone rather than someone interacting with them. We used Gephi [65] using ForceAtlas 2 [66] to distribute the nodes.

Additionally in figure 9 on the right, there are 44413 users and 58739 connections. An average of 1.32 connections per node, with the same caveat as the previous average result.

Comparing the two figures in 9 show that there are two communities on the left graph, the large one and smaller one, these could be the two different search queries communities. While on the right is the replies graph, less users because users have to be more engaged in the discourse to do that, and much more centralised meaning that they reply to the different query discourse types regardless of their position.

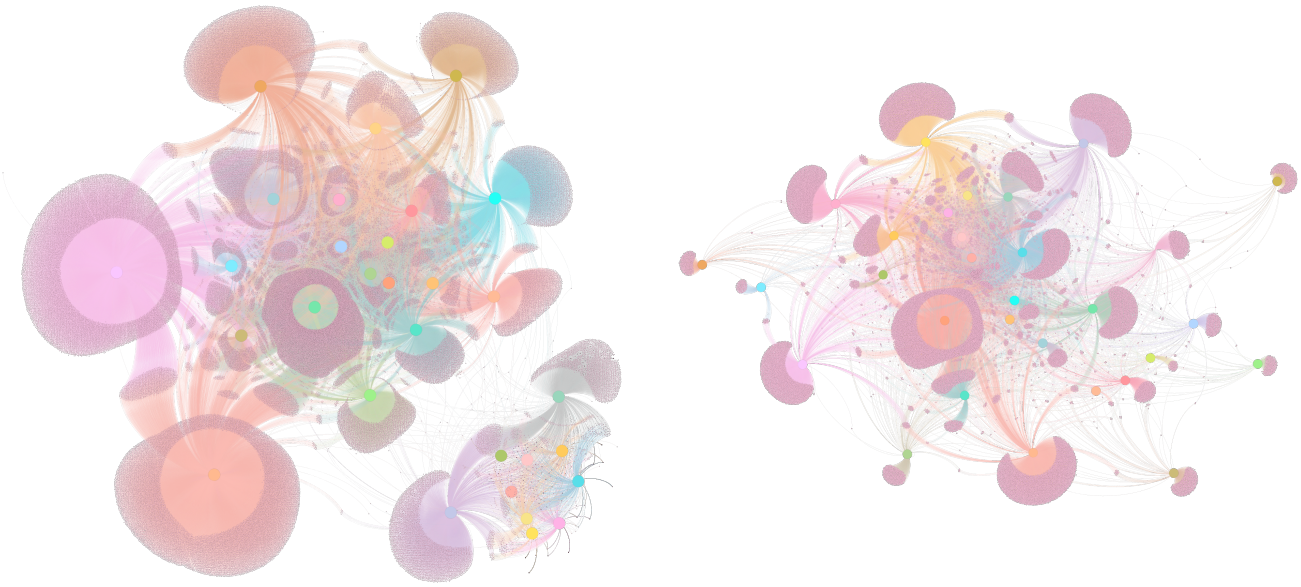


Figure 9: Left: Retweet Graph, Right: Reply Graph

Analysing the hubs only within the network for the first hop in figure 10, there were thirty nodes with fifty-five edges between them, with three of the nodes not being connected to the other nodes in any way. This shows that although the discourse was split into two separate queries the main thirty are interconnected using the combined query to connect them together and that the influential figures are actively engaged with it through communicating with the other members.

We repeated this process again, resulting in the total users in the **final** user database was 14745, 121028 interaction relationships, while we only stored tweets from 7281 of those users (49.38%) after the snowball purges and reducing the data to users who have minimum of 50 tweets. There was an average 8.2 connections inbound for each node. In total we stored 1873093 tweets after collecting the relevant tweets of each user from all the hops within the snowball sampling. Both of these database data was collected over the span of a year. In 17 it shows the final network visualisation, with the size of the node being how many inbound connections there are for the node.

Finally in figure 11 analysing the tweets collected, we can see that the data collected is relevant as the most used word is trans and the other words are related to the discourse only meaning the collection method was successful and with the data collected we can use it to find sockpuppets. Surprisingly when programming this there was issues with filtering out usernames mentioned to pseudo-anonymise the output, there was one user who was mentioned so much that they were put into the top 100 words which would mean they are very influential.

6.2 Sockpuppet Detection

6.2.1 Finding Suspicious Users

There were 6783 suspicious tweets with more likes than retweets which was made up from 2473 users with an average of 2.74 suspicious tweet per user; which is a surprisingly small amount of tweets, only 0.004% of the total set, however this might of been lowered through the purges.

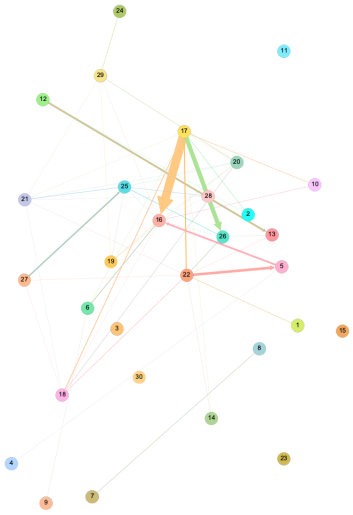


Figure 10: Hub Only

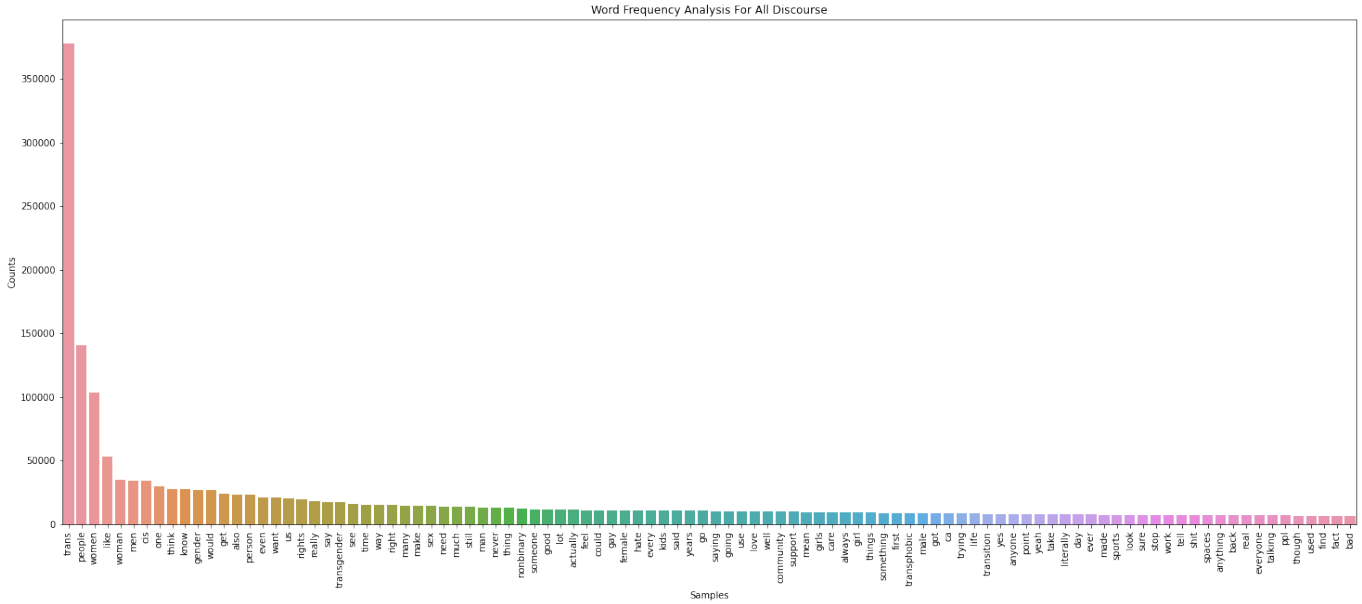


Figure 11: Word distribution within tweet text

For suspicious times, we plotted several time series graphs to analyse the tweet volume made. There was a mean of 3.83 tweets per minute from the users within the tweet data-set, and a mean of 5131.76 tweets per day, showing that it very active with the seven thousand users. In 12 we can see that the tweet volume trend is increasing but at a low speed although the topic is more talked about within recent years which is a juxtaposition; at the beginning there is a sharp increase which could mean that 2020 data trend could be very different.

For suspicious text, in figure 13 we see the distribution for the top 100 hashes while in table 5 we see the top 50 anonymised text results. We can see that one particular term is much higher than others, this being ”trans women are women”; this is a popular mantra by activists which makes it hard to classify the users participating within this group of tweets as it’s not clear cut. These results are very mixed with the categories, it’s almost a 50:50 split. The non-neutral anti-trans comments seem much more conspiratorial verses the other text within the table which means that sockpuppets could be contributing to these fringe ideas found within the discourse bumping up the numbers.

For tweets, there is a mean of 257 tweets per user over the span of the year, 0.7 tweets a day per user in the discourse. There are some users who have a substantial amount of tweets over the year, seen in 7, one user tweeted twelve thousand times within the year around the discourse; that is 32 tweets a day over the span of the year for the discourse only, this could be indicative of them being a sockpuppet. We also have the statistics for the amount of times

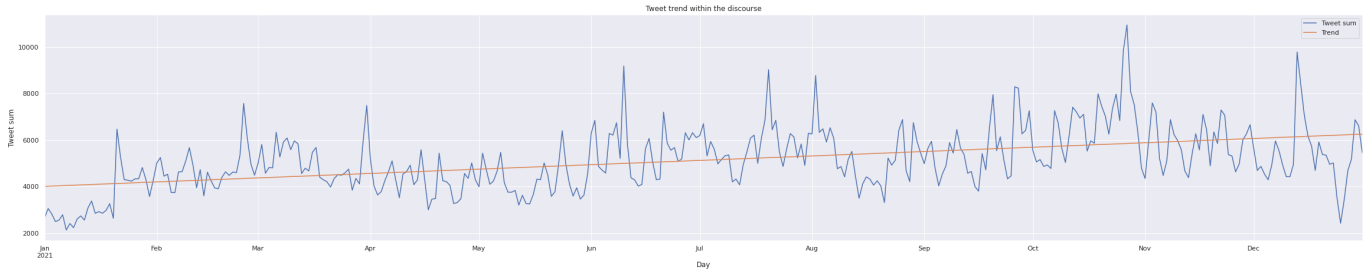


Figure 12: Time series plot of tweet volume per day

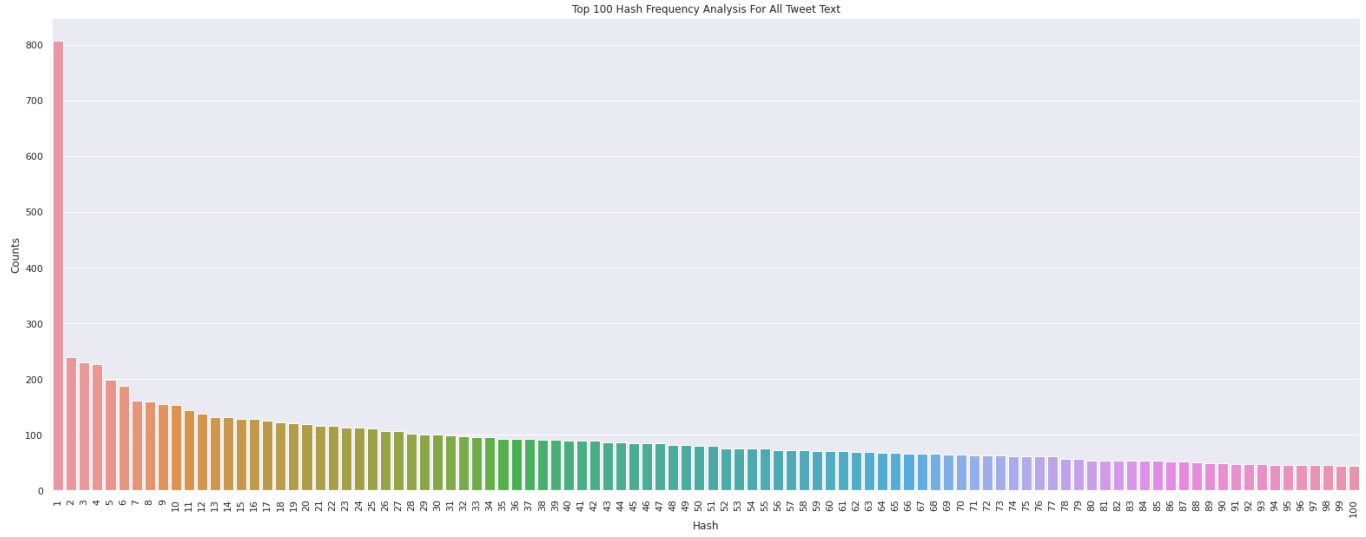


Figure 13: Distribution of top 100 hashes

a user has replied to another within the discourse in 8, showing that there are many users with these huge numbers. Looking into some of these accounts across top reply and tweeter further they seem human raising concerns about potential addictions on twitter as well as making it hard to differentiate between high volume creating sockpuppets and users who are potentially addicted.

For subgraph similarity, we chose users who interacted more than or 150 times with one other user directly, there were 126 users with 7875 combinations between them. In 9 there is one user who has extremely similar graph to another user but analysing these two users manually they had only one similar user which inflated this with a small union set length between the users; due to only getting the users within the discourse it seems that these numbers are not truly accurate to their definitive score, however it's still useful for a general pointer of suspiciousness which could be combined with another method.

6.2.2 Machine Learning

There were 42 members in the intersection of the top 250 suspicious hash users and top 250 suspicious times; 163 users within the intersection of suspicious top 250 hash users and suspicious retweeters; 163 users in the intersection of top 250 suspicious times and the suspicious retweeters; 33 users in intersection of top 250 suspicious hash users, top 250 suspicious times users and suspicious retweeters; 43 users who replied to a specific user $>$ 365 times directly; 26 users with $>$ 0.5 similarity score.

We trained three SVMs using different suspicious users groups captured from the previous section. Specifically, the 33 users from the intersection from all three suspicious tweet groups, and the two suspicious network groups: 43 users who replied to a specific user $>$ 365 times directly; 26 users with $>$ 0.5 similarity score. The confidences were derived from the voting method described in 5.3.3.

The first group we trained the SVM from was the group of 33 users. The average F1 score was 0.48, with the average accuracy of 0.97; we can see the diagonal line in the confusion matrix 14 Running this classifier against the other feature sets with the voting method resulted with all except one having a confidence lower than 3%, while the top was the confidence of 3.92%.

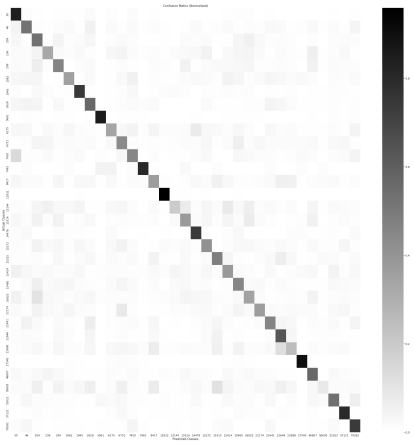


Figure 14: SVM First Group Confusion Matrix

The second group average F1 score was 0.29 and an average accuracy of 0.96. The highest confidence was 5.8% which was the highest confidence value we got from the three classifiers created. The confusion matrix 15 has a harsh line on the diagonal.

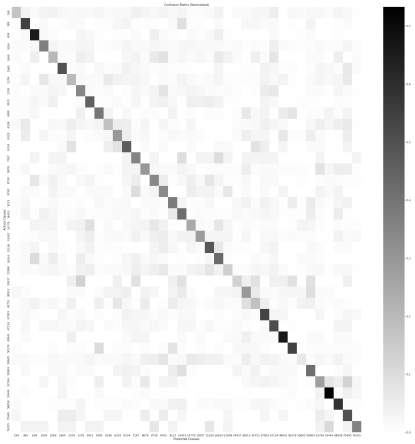


Figure 15: SVM Second Group Confusion Matrix

For the third group, the results were the worst performing from the three classifiers. With an average F1 score of 0.27 and average accuracy of 0.95. We can see in the confusion matrix 16 that it is more dithered versus the previous two. The highest confidence was 2.3%.

Each confidence was low which makes us believe that the suspicious users are not sockpuppets if our machine learning algorithm works as intended, but this does not mean that there isn't sockpuppets it's just that we may of not used an appropriate method for finding the actual suspicious sockpuppet users for our dataset. It means that these large volume of tweet creation for finding sockpuppets isn't the most efficient method of finding sockpuppets within a given dataset.

6.2.3 Category Ratios

Using our original query keywords we analysed the category of the text of the top occurring repeated tweets. In table 4 we have the ratios for each group of tweets and the different tweet categories. Non-suspicious ratio versus the suspicious ratio has a difference ratio between them of `0.055, -0.031, 0.008, -0.032`, showing that there is a higher polarisation between the two different categories as their is less of both terms co-accruing and that there is a higher ratio of non-neutral category type posts. Even higher is the high hash count ratio; the normal suspicious ratio was hashes which had more than or at 15 counts of it occurring while the high hash count was above or at 75; this could potentially mean there is more sockpuppet activity within the non-neutral terms.

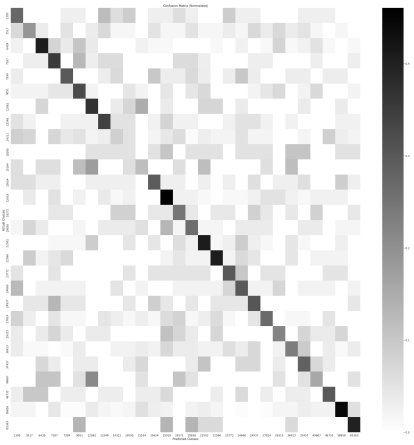


Figure 16: SVM Third Group Confusion Matrix

Type	Includes Only Neutral	Includes Only Non-neutral	Includes Both	Neither
Non-Suspicious	0.919	0.041	0.015	0.023
Suspicious	0.863	0.073	0.007	0.055
High Hash Count Suspicious	0.868	0.096	0.0	0.034

Table 4: Category ratios for each tweet group type

6.3 Measuring Influence

The top 25 influential users within the network are listed within 10. The average PageRank score for the suspicious users is 0.00062, which would position them 216 in a list of 14745 which is very high; the highest influential user within the suspicious accounts was 0.0108 which was position 9 in the top PageRank scores, analysing this account seems to be a real user. These users who were suspicious were collected via looking at high volumes, these users may be influential within the community because of the high amount of tweets they produce versus the standard user within the discourse; we learned that high volume doesn't mean that these users are sockpuppets just because they produce a lot of content and are concerningly extremely engaged with the discourse. We couldn't get measures for the detected sockpuppets because of the methods we implimented we did not detect any.

7 Discussion and further work

7.1 Main Findings

The results in summary tell us that the data we collected is relevant; but that the methods for detecting the sockpuppets are not working as intended.

The answer to our research title is inconclusive as we do not know if the users we collected were sockpuppets and that looking at some of the accounts manually they seem very human but are heavily invested within the discourse.

Our data collection method to get discourse related tweets worked very well; however the sockpuppet detection results was ambiguous; it needs more data and research into trying to find the suspicious users to use to find sockpuppets with.

The data collection method we created is reliant on the first choice, the queries, within the process so there is a difficult choice at the start of the research project which should be analysed more thoroughly, but if correctly formatted it yields data which is very relevant to the discourse with the benefit of being able to map the connections between users on a network easily via an interaction type, makes the collection process less explosive and that the users participating are not just passers by.

We implemented the methods for the sockpuppet detection however issues appeared when trying to find the users to train our methods with. There is a merging of users who are very active and bots within high volumes, and that because we have many tweets it is difficult to find the collections of suspicious users as they may have been clever and tried to hide their activity through tweeting moderately or over many accounts a smaller amount of times rather than the thought of that the sockpuppets will be overt when the data is collected and analysed looking at traditionally suspicious indicators such as high direct amount of tweets to a specific user.

The mean results in 3 show that the non-neutral have higher SNP score but it's primarily held by just two users within 8, while neutral users had a lower mean score but each users had a value visible. This could mean that the non-neutral group has large accounts supporting with influence not being well distributed across users; while neutral terms are more evenly distributed with lower influence potentially meaning it is more interconnected with their group gaining more level influence between them.

In 9 for the retweet graph we can see that there are two groups, one huge group and one smaller group in the bottom right; this could be the split between the different users captured via the two queries. For the reply graph as the count of users is lower versus the retweet graph and still relevant to the discourse. Due to the lower count it makes it more manageable to snowball with due to the lesser number of people to explore versus retweeters; additionally it also means that we get users who are more engaged within the discourse than a passive indicator which is retweeting; this is because as a user you only need to press one button, while replying means you have to be more engaged within the discourse itself.

7.2 Limitations

The collection method used through snowballing with a specific query worked well for getting the relevant discourse community data, seen within the network diagram 17 and the word distribution in 11, it helped with reducing the amount of tweets collected but have highly relevant tweets. If we were able to collect infinite amount of tweets theoretically, just searching twitter and taking all the tweets using the query within the snowball stage but not to a specific user would been a better method to gather all the relevant data as all our method does it help with the explosion of tweets being returned. The approach we took took a lot of time to create which effected the remaining sections of the project, if someone else were to implement this they will have to take in account a potential large time spent developing it.

Snowballing via replies was very fast for the return we got from using it, however the bot detection now was restricted to trying to detect the tweets via replies explicitly, which removes other vectors of attack such as likes and retweets; although collecting these would of made the project infeasible with the time it would of taken to collect that information too. It possibly may of filtered out some of the bots with the purging section in combination with only replies. Although replies mean they are definitely involved within the discourse, it has the downside of users who are active within the discourse but do not interact via replies are cut out.

In the time series analysis 12 we can see that the tweet volume trend is increasing but at a low speed although the topic is more talked about within recent years; this could be due to us purging the data to just the users who are most involved in the discourse online so users who are posting due to online events around the discourse just once within the discourse would be removed from the purge. At the beginning of 2021 it was low compared to the rest of the year; if we had data on 2020 we may of seen the amount of tweets per day increase drastically between the two years from this big jump at the beginning which would be suspicious, but we don't have this data.

In 8 there are some users which have direct replied to another user more than two thousand times, although we haven't stored all of these replies due to the purges. These are substantial amount of replies just to one user, typically indicative of sockpuppet activity, but looking closer at these users they seem to be real people who are just heavily invested within the discourse and/or addicted to using Twitter. Looking over all the other tables related to the suspicious users we have been looking at large numbers to aim to find users but this could of been a fault of ours as the bots may of attempted to camouflage through tweeting less frequently or obscuring in some way. There is also the possibility that the bots were throw away accounts who just posted once, although a waste of accounts as a resource. Manually looking through the users who post thousands just to one user or just tweets in general, from a non-computational view, it seems like these users are real humans; the issue is, how would we be able to differentiate between the sockpuppets and addicted users? This could be a potential research topic in the future especially as more people are becoming addicted to these social media platforms so this may be an bigger issue for future research in highly volume of posts topics.

The feature set looking at 14 showing it is working relatively well, and that the feature set which is extracted is working well with identifying a specific user. This method ignores the fact that some of these users in a group could be the same user, which is a downside; this could potentially explain why 16 had worse results if some users had very similar typing patterns which may of affected the accuracy of the classifier. The F1 average scores were not amazing across the board so improvement needs to be made to the feature set, maybe splitting the emojis and emoticons into separate features could help with identifying users too but this would not actually fully fix the issue, just could help raise the score. The difference of the feature set **sentence with small letter frequency** and **sentence count without capital letter at the beginning** are very similar, how much information is actually being held to help with these small changes?

In 10 we can see these users are mostly the initial users collected which could either mean that due to using the snowball method that the users who were first collected will have the highest influence due to the users all being derived from them, however we are also seeing users who are not within the first thirty which shows that the formation through the method allowed to find more influential users who were not in the initial week within the initial snowball user

collection stage. Maybe including some random users from the discourse stage who have reached a certain threshold could be a way of adding randomness to seed from finding more users who may be interacting with a cluster of bots that would be caught while snowballing.

Everything is reliant on the original search queries, so the results can only be as good as this query; the project expands like a tree from that single point. We did not do any query optimisation apart from using our own knowledge of the discourse on Twitter which helped with the starting point. The original method was to select specific users manually which would of lead to bias, but there could also be bias using these keywords; however it would be a lesser bias than manually selecting specific users.

We only snowballed twice, not including the initial users collected, this may of left out potential users who are deeper within the discourse network which we never got to. Possibly the further into a discourse network we may find more bots.

Some methods implemented were designed for a fully fleshed out network, the SNP measure uses the amount of followers to scale the result, however these followers would of been accumulated over time so we would of had to collect all tweets for the SNP measure to be truly accurate to what it is supposed to be. Another is the sub-graph similarity, it assumes that all of the users interactions are taken rather than just specific discourse related interactions which may of affected the end results.

Within certain discourses, especially ones which are around peoples rights, people use short phrases / mantras; this can become an issue when trying to find bot activity via text and looking at high volumes as for us in 5 the mantra "trans women are women" appeared first which is good but did become an issue with trying to find sockpuppets as this phrase appeared at the top and may of contributed to accidentally assigning a user as suspicious for participating within the discourse via one of these phrases but is not a sockpuppet.

There is a higher ratio of non-neutral terms versus the suspicious ratio, 4, showing that potentially there is more bot activity within the non-neutral terms if volume is a way of measuring the bot activity precisely. In 5 there are 25 definitive anti-trans tweets within the top 50 duplicate tweets even though we see in the ratios that the neutral terms are used most but compared to non-neutral it's much larger; in the future we should try and determine if a tweet is supportive or not to see if particular groups have more suspicious activity happening.

7.3 Future Research

For future research, it would be beneficial to try and differentiate between sockpuppets who use high volume tweets and twitter users who are highly active but are humans. This could be done by getting the full interaction graph than just the specific discourse interaction graph. One method extending what we done is where you could collect the discourse users then after you have collected all the members you want, then get the interactions each user has made regardless of if it's outside of the discourse. This would take a very long time to collect depending on the timespan the tweets will be collecting from, especially as particular endpoints such as getting the retweeters of a particular tweet has frequent ratelimits happening. With this and the follower relations per user we could use the full potential of [37] to help find more subtle sockpuppets which interact and follow in a way a sockpuppet would. The one issue while developing the interaction sub-graph similarity method is that it takes a long time to compute the values. You would have to either try every combination of users or just use the top users and strike out the high volume users who are below a certain similarity score as they would likely not be sockpuppets.

It would also be beneficial to research more into subtle ways sockpuppets can appear within data. The users controlled by a specific puppet master may rotate the phrases they use so the copied text may not get picked up using high volume analysis because the counts for a particular copied text is lower than most of the other copied text. Possibly fully implementing [39] for the discourse may be a potential fix, however this requires a large amount of data again so a short project time may not be appropriate to use this method. As discussed in [51] that "name+number" usernames were another indicator of suspiciousness within the network, maybe using other indicators such as the bio of a user and comparing that to other bios could be another potential route; but this also means collecting more data on the individual accounts which increases overhead and project time.

The feature set used was based from a wikipedia sockpuppet authorship attribution method [35] which was aimed to attribute a user to text with short text, however Twitter communication is different such as emojis being used more frequently. Potentially emojis could be used as a seperate feature for the identification rather than concatenated with ASCII emoticons as a user may favour using ASCII over emoji which may help identify a user. Some features may not need to be used and may make it more difficult, the only one thought of was the sentence starting features mentioned in the previous subsection; how much do these features actually help?

While using the wikipedia authorship attribution method we had the problem of not actually having the users to use the detection on, this is why we tried to find suspicious users. The paper had confirmed sockpuppets it could use the data with to test of the method was working. Maybe creating an generalised dataset which contains labelled sockpuppets that is ethically produced could be beneficial to test if the feature set used is actually working rather than hoping that it has or that you use a method already created and test how well the method actually works.

8 References

[1] A. Alharbi, H. Dong, X. Yi, Z. Tari, and I. Khalil, "Social media identity deception detection: A survey," *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021, ISSN: 0360-0300. DOI: 10.1145/3446372.

[2] M. C. Benigni, K. Joseph, and K. M. Carley, "Bot-ivistm: Assessing information manipulation in social media using network analytics," in *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, N. Agarwal, N. Dokoohaki, and S. Tokdemir, Eds. Cham: Springer International Publishing, 2019, pp. 19–42, ISBN: 978-3-319-94105-9. DOI: 10.1007/978-3-319-94105-9_2. [Online]. Available: https://doi.org/10.1007/978-3-319-94105-9_2.

[3] S. Bradshaw and P. N. Howard, "Troops, trolls and troublemakers: A global inventory of organized social media manipulation," p. 37,

[4] Twitter. "Twitter API for academic research | products." (), [Online]. Available: https://developer.twitter.com/en/products/twitter-api/academic-research (visited on 11/08/2021).

[5] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 280–289, May 3, 2017, Section: Full Papers. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14871.

[6] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, Jun. 24, 2016, ISSN: 0001-0782. DOI: 10.1145/2818717.

[7] S. Wojcik, S. Messing, A. W. Smith, L. Rainie, and P. Hitlin, "Bots in the twittersphere," Pew Research Center, Report, Apr. 9, 2018, Journal Abbreviation: An estimated two-thirds of tweeted links to popular websites are posted by automated accounts – not human beings. [Online]. Available: https://apo.org.au/node/141291.

[8] B. Heredia, J. Prusa, and T. Khoshgoftaar, "Exploring the effectiveness of twitter at polling the united states 2016 presidential election," in *2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*, Oct. 2017, pp. 283–290. DOI: 10.1109/CIC.2017.00045.

[9] E. Ferrara, "What types of COVID-19 conspiracies are populated by twitter bots?" *First Monday*, May 19, 2020, ISSN: 1396-0466. DOI: 10.5210/fm.v25i6.10633. arXiv: 2004.09531.

[10] T. U. G. H. Office. "Hate crime, england and wales, 2020 to 2021," GOV.UK. (), [Online]. Available: https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2020-to-2021/hate-crime-england-and-wales-2020-to-2021 (visited on 10/27/2021).

[11] Brandwatch. "The scale of transphobia online," Brandwatch. (), [Online]. Available: https://www.brandwatch.com/reports/transphobia/ (visited on 10/27/2021).

[12] "BCS code of conduct | BCS." (), [Online]. Available: https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/ (visited on 11/14/2021).

[13] *Cryptology: Decrypt/encrypt text using various cipher techniques*, version 0.0.14.

[14] F. Valsorda, *FiloSottile/age*, original-date: 2019-05-18T20:44:54Z, May 12, 2022. [Online]. Available: https://github.com/FiloSottile/age (visited on 05/12/2022).

[15] I. C. Education. "What is an application programming interface (API)." (Oct. 15, 2021), [Online]. Available: https://www.ibm.com/cloud/learn/api (visited on 11/08/2021).

[16] "List of LGBTQ+ terms," Stonewall. (May 28, 2020), [Online]. Available: https://www.stonewall.org.uk/help-advice/faqs-and-glossary/list-lgbtq-terms (visited on 05/20/2022).

[17] W. O. Bockting, M. H. Miner, R. E. Swinburne Romine, A. Hamilton, and E. Coleman, "Stigma, mental health, and resilience in an online sample of the US transgender population," *American Journal of Public Health*, vol. 103, no. 5, pp. 943–951, May 2013, ISSN: 0090-0036. DOI: 10.2105/AJPH.2013.301241.

[18] O. Jenzen, "Trans youth and social media: Moving between counterpublics and the wider web," *Gender, Place & Culture*, vol. 24, no. 11, pp. 1626–1641, Nov. 2, 2017, Publisher: Routledge, ISSN: 0966-369X. DOI: 10.1080/0966369X.2017.1396204.

[19] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, p. 2658, Mar. 2, 2004. DOI: 10.1073/pnas.0400054101.

[20] T. D. Jayawickrama. "Community detection algorithms," Medium. (Feb. 1, 2021), [Online]. Available: https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae (visited on 11/06/2021).

[21] M. C. Benigni, K. Joseph, and K. M. Carley, "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on twitter," *PLOS ONE*, vol. 12, no. 12, e0181405, Dec. 1, 2017, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0181405.

[22] L. A. Goodman, "Snowball sampling," *The Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 148–170, Mar. 1961, Publisher: Institute of Mathematical Statistics, ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177705148.

[23] A. Campan, T. Atnafu, T. M. Truta, and J. Nolan, "Is data collection through twitter streaming API useful for academic research?" In *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 3638–3643. DOI: 10.1109/BigData.2018.8621898.

[24] L. Ben Jabeur, L. Tamine, and M. Boughanem, "Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks," in *String Processing and Information Retrieval*, L. Calderón-Benavides, C. González-Caro, E. Chávez, and N. Ziviani, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2012, pp. 111–117. DOI: 10.1007/978-3-642-34109-0_12.

[25] F. Riquelme and P. González-Cantergiani, "Measuring user influence on twitter: A survey," *Information Processing & Management*, vol. 52, no. 5, pp. 949–975, Sep. 1, 2016, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2016.04.003.

[26] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, pp. 10–17, May 16, 2010, Number: 1, ISSN: 2334-0770. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14033.

[27] I. Anger and C. Kittl, "Measuring influence on twitter," in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, ser. i-KNOW '11, New York, NY, USA: Association for Computing Machinery, Sep. 7, 2011, pp. 1–4. DOI: 10.1145/2024288.2024326.

[28] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, Hong Kong, China: Association for Computing Machinery, 2011, pp. 45–54. DOI: 10.1145/1935826.1935843.

[29] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10, New York, New York, USA: Association for Computing Machinery, 2010, pp. 261–270. DOI: 10.1145/1718487.1718520.

[30] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "In the mood for being influential on twitter," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Oct. 2011, pp. 307–314. DOI: 10.1109/PASSAT/SocialCom.2011.27.

[31] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 107–117, Apr. 1998, ISSN: 01697552. DOI: 10.1016/S0169-7552(98)00110-X.

[32] "A twitter analog to PageRank," The Noisy Channel. (Jan. 13, 2009), [Online]. Available: https://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/ (visited on 11/12/2021).

[33] T. Majer and M. Šimko, "Leveraging microblogs for resource ranking," in *SOFSEM 2012: Theory and Practice of Computer Science*, M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser, and G. Turán, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2012, pp. 518–529. DOI: 10.1007/978-3-642-27660-6_42.

[34] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian, "An army of me: Sockpuppets in online discussion communities," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 3, 2017, pp. 857–866. DOI: 10.1145/3038912.3052677.

[35] R. M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y-Gómez, and P. Rosso, "Authorship attribution using word sequences," in *Progress in Pattern Recognition, Image Analysis and Applications*, J. F. Martínez-Trinidad, J. A. Carrasco Ochoa, and J. Kittler, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2006, pp. 844–853. DOI: 10.1007/11892755_87.

[36] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, 2014. DOI: 10.1109/TIFS.2014.2332820.

[37] J. Wang, W. Zhou, J. Li, Z. Yan, J. Han, and S. Hu, "An online sockpuppet detection method based on subgraph similarity matching," in *2018 IEEE Intl Conf on Parallel Distributed Processing with Applications, Ubiquitous Computing Communications, Big Data Cloud Computing, Social Computing Networking, Sustainable Computing Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, Dec. 2018, pp. 391–398. DOI: 10.1109/BDCloud.2018.00067.

[38] T. Solorio, R. Hasan, and M. Mizan, "A case study of sockpuppet detection in Wikipedia," in *Proceedings of the Workshop on Language Analysis in Social Media*, Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 59–68. [Online]. Available: https://aclanthology.org/W13-1107.

[39] Z. Yamak, J. Saunier, and L. Vercouter, "SocksCatch: Automatic detection and grouping of sockpuppets in social media," *Knowledge-Based Systems*, vol. 149, pp. 124–142, Jun. 1, 2018, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2018.03.002.

[40] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *Trans. Info. For. Sec.*, vol. 9, no. 8, pp. 1311–1321, Aug. 2014, ISSN: 1556-6013. DOI: 10.1109/TIFS.2014.2332820.

[41] O. Beatson, R. Gibson, M. C. Cunill, and M. Elliot, "Automation on twitter: Measuring the effectiveness of approaches to bot detection," *Social Science Computer Review*, p. 08 944 393 211 034 991, Aug. 6, 2021, Publisher: SAGE Publications Inc, ISSN: 0894-4393. DOI: 10.1177/08944393211034991.

[42] K. Reitz, *Requests: Python HTTP for humans.* Version 2.27.1. [Online]. Available: https://requests.readthedocs.io (visited on 05/12/2022).

[43] S. Kumar, *Python-dotenv: Read key-value pairs from a .env file and set them as environment variables*, version 0.20.0. [Online]. Available: https://github.com/theskumar/python-dotenv (visited on 05/12/2022).

[44] *Numpy: NumPy is the fundamental package for array computing with python.* Version 1.22.3. [Online]. Available: https://www.numpy.org (visited on 05/12/2022).

[45] N. Team, *Nltk: Natural language toolkit*, version 3.7. [Online]. Available: https://www.nltk.org/ (visited on 05/12/2022).

[46] J. D. H. Droettboom Michael, *Matplotlib: Python plotting package*, version 3.5.2. [Online]. Available: https://matplotlib.org (visited on 05/12/2022).

[47] A. Hagberg, *Networkx: Python package for creating and manipulating graphs and networks*, version 2.8. [Online]. Available: https://networkx.org/ (visited on 05/12/2022).

[48] C. Leifer, *Peewee: A little orm*, version 3.14.10. [Online]. Available: https://github.com/coleifer/peewee/ (visited on 05/12/2022).

[49] M. Waskom, *Seaborn: Seaborn: Statistical data visualization*, version 0.11.2. [Online]. Available: https://seaborn.pydata.org (visited on 05/12/2022).

[50] *Scikit-learn: A set of python modules for machine learning and data mining*, version 1.1.0. [Online]. Available: http://scikit-learn.org (visited on 05/12/2022).

[51] "#KeepPrisonsSingleSex: How botnets pushed a hashtag to westminster." (), [Online]. Available: https://www.logically.ai/articles/keepprisonssinglesex-pushed-by-botnet (visited on 05/12/2022).

[52] "GET /2/tweets/search/all." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all (visited on 05/12/2022).

[53] "GET /2/tweets/counts/all." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tweets/counts/api-reference/get-tweets-counts-all (visited on 05/12/2022).

[54] "GET /2/users/:id/tweets." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-tweets (visited on 05/12/2022).

[55] "GET statuses/retweeters/ids." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-retweeters-ids (visited on 05/12/2022).

[56] "GET /2/users/:id/mentions." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-mentions (visited on 05/12/2022).

[57] "GET /2/users." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/users/lookup/api-reference/get-users (visited on 05/12/2022).

[58] "Search tweets - how to build a query." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query (visited on 05/12/2022).

[59] "Conversation ID." (), [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/conversation-id (visited on 05/12/2022).

[60] *Google/farmhash*, original-date: 2015-08-14T21:01:50Z, Apr. 23, 2022. [Online]. Available: https://github.com/google/farmhash (visited on 05/12/2022).

[61] R74n. "List of text faces," Copy Paste Dump. (), [Online]. Available: https://c.R74n.com (visited on 05/12/2022).

[62] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006. DOI: 10.1002/asi.20316.

[63] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 1, 2002. DOI: 10.1613/jair.953. arXiv: 1106.1813.

[64] "Google colaboratory." (), [Online]. Available: https://colab.research.google.com/ (visited on 05/12/2022).

[65] "Gephi - the open graph viz platform." (), [Online]. Available: https://gephi.org/ (visited on 05/12/2022).

[66] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PLOS ONE*, vol. 9, no. 6, e98679, Jun. 10, 2014, Publisher: Public Library of Science. DOI: 10.1371/journal.pone.0098679.

9 Appendix

9.1 Ethics Application and Data Management Plan

Ethical Review Application (ER/EP396/1) Ethan Pannell	
Project Title	Assessing the influence of sock puppets on the transgender Twitter discourse
Status	Approved
Email	ep396@sussex.ac.uk
Phone No.	
Applicant Status	UG
Department	Informatics
Supervisor	Berthouze, Luc
Project Start Date	26-Nov-2021
Project End Date	17-May-2022
External Funding in place	No
External Collaborators	No
Funder/Project Title	
Name of Funder	

Ethical Review Application ER/EP396/1 (continued)

Project Description

Sockpuppets are malicious users of online social media platforms such as Twitter who try to skew opinions online to an extreme, such as the far right. These users are typically bots but occasionally humans. Users within a social media platforms have influence, whereby their posts are more likely to be interacted with by other users. It would be beneficial to characterise how much influence these sockpuppets have over the network so we can control their impact over the long-term. The discourse on Twitter about the Transgender topic has a high likelihood of being targeted by sockpuppets because they concern a minority group who are misunderstood and considered a "hot topic"; for this reason, it is likely that the discourse features sockpuppets.

There are three separate stages:
1. Identification of the group of users participating in the target discourse
2. Measuring Influence
3. Sockpuppet Detection

There are separate methods for each point above, explained more within the protocol document.

All the data displayed within the paper will be aggregate data, using visualisations or tables. No tweets will be explicitly displayed.

I will pseudoanonymise the data using my own generated id's to remove the dependency on the twitter identifiers which could identify a user; they will be collected initially but then replaced with our own IDs as we need to store the relation between users, tweets and other users. I will be collecting data on both user accounts and tweets on the Twitter platform, as follows.
For users, I shall collect:
- User ID
- Following
- Followers
- Creation Date
- Bio / Description

For tweets:
- Tweet ID
- User ID
- Creation date
- In reply status id
- In reply user id
- Text content
- Number of likes
- Number of retweets
- Number of quote retweets
- Number of replies
- Is a quote retweet
- Quoted status id
- Entities (specifically hashtags and user mentions)

Importantly, this research is not to identify the users behind the sockpuppet accounts nor the actual users on the platform. More specific details such as the Twitter API ToS are described within the data management plan.

Ethical Review Form Section A (ER/EP396/1)	
Question	**Response**
>> Checklist	
A1. Will your study involve participants who are currently or potentially vulnerable or unable to give informed consent or in a dependent position (e.g. people under 18, people with learning difficulties, over-researched groups or people in care facilities)?	Yes
A2. Will participants be required to take part in the study without their consent or knowledge at the time (e.g. covert observation of people in non-public places), and / or will deception of any sort be used? Please refer to the British Psychological Society Code of Ethics and Conduct (or similar guidelines) for further information.	Yes
A3. Unless specifically and clearly consented (e.g. a media release form), will it be possible, through a research output, to identify participants in any way? (This does not include taking email details for participant prize draws or identifying participants from signed consent forms or holding identity encryption spreadsheets that are stored securely separate from the research data).	No
A4. Might the study induce psychological stress or anxiety, or produce humiliation or cause harm or negative consequences beyond the risks likely to be encountered in the everyday life of the participants?	No
A5. Is there a risk that the research topic might lead to disclosures from the participant concerning their beliefs, involvement in illegal actions or any other activities that may represent a threat to themselves or others?	No
A6. Will the study involve collecting any personal special category information* in a form that could allow the participant/ participants to be identified? [* identifiers relating to race, ethnic origin, politics, religion, trade union membership, philosophical beliefs, genetics, biometrics, health, sex life or sexual orientation]	No
A7. Will any drugs, placebos or other substances (such as food substances or vitamins) be administered as part of this study and will any invasive or potentially harmful procedures of any kind will be used?	No
A8. Will your project involve working with any substances and / or equipment which may be considered hazardous?	No
A9. Will your study involve the taking and/or storage of human tissue that falls under the Human Tissue Act (HTA)? http://www.sussex.ac.uk/staff/research/governance/erp_overview/humantissue	No
>> Risk Assessment	
A10. If you have answered Yes to ANY of the above questions, your application may be considered as HIGH risk. If, however you wish to make a case that your application should be considered as LOW risk please enter the reasons here. Researchers should note that SREOs or C-RECs may decide NOT to agree with the case that you have made.	

Ethical Review Form Section C (ER/EP396/1)	
Question	**Response**
>> Risk Checklist - Participants	
C1. Is DBS (Disclosure and Barring Service) clearance necessary for this project? If yes, please ensure you complete Section C24a below	No
C2. Are alcoholic drinks, drugs, placebos or other substances (such as food substances or vitamins) to be administered to the study participants?	No
C3. Can you think of anything else that might be potentially harmful to participants in this research?	No
C4. Does the project involve working with any substances and/or equipment which may be considered hazardous? (Please refer to the University's Control of Hazardous Substances Policy http://www.sussex.ac.uk/hso/policies).	No
C5. Could the nature or subject of the research potentially have an emotionally disturbing impact on the researcher(s)?	No
C5a. If yes, briefly describe what measures will be taken to help the researcher(s) to manage this.	
C6. Could the nature or subject of the research potentially expose the researcher(s) to threats of physical violence and / or verbal abuse?	Yes
C6a. If yes, briefly describe what measures will be taken to mitigate this.	I could ask the University not to make the dissertation avalible.
C7. Does the research involve any fieldwork - Overseas or in the UK?	No
C7a. If yes, where will the fieldwork take place? (All research requiring overseas travel will require the submission of a fully completed OTTSRA form). In the event that the Foreign and Commonwealth Office has travel warnings in place for the country (ies) to be visited you will also need to provide a detailed risk assessment.	
C8. Will any researchers be in a lone working situation?	No
C8a. If yes, briefly describe the location, time of day and duration of lone working. What precautionary measures will be taken to ensure safety of the researcher(s)?	
C9. Can you think of anything else that might be potentially harmful to the researcher(s) in this research?	No
>> Data Collection and Analysis (Please provide full details)	

C10. PARTICIPANTS: How many people do you envisage will participate, who are they, and how will they be selected?	The data is collected via the Twitter API. I am unsure the total amount of users where data will be collected from, but the maximum amount could be in the hundreds of thousands. Matthew C. Benigni et al (2017) used snowball sampling with a depth of two (2-hop) collected 119,156 users, and processed all their user tweets which resulted in 862 million tweets, we will only be using a span of a month which will be very likely to drastically reduce this down. Brandwatch reported in 2019, there were 5.5 million tweets from 2016 to 2019 classified as transgender discourse; using this, there would be an average of 115 thousand tweets per month overall and not all of these tweets would be discovered to be processed. The tweets of interest are those of users active within the Twitter trans discourse topic, this will likely include transgender individuals, allies or people who are generally interested

in the "topic".

C11. RECRUITMENT: How will participants be approached and recruited?	Through the Twitter API using tweets with specific search queries and the relation to members within the discourse, e.g., followers and following.
C12. METHOD: What research method(s) do you plan to use; e.g. interview, questionnaire/self-completion questionnaire, field observation, audio/audio-visual recording etc.?	Collecting data through the Twitter API and stored in a relational database.
C13. LOCATION: Where will the project be carried out e.g. public place, in researcher's office, in private office at organisation?	On the university servers and accessible to the researchers involved with the project
>> Ethical Considerations (Please provide full details)	
C14. PARTICIPANT WELLBEING: Will the study involve engaging participants in the discussion of distressing or sensitive topics? (e.g. sexual activity, drug use, ethnicity, political behaviour, potentially illegal activities). If so, please set out how you will manage the well-being of participants.	Data is taken from existing public tweets, so no communication is initiated. There are therefore no concerns for the well-being of the participants.

C15. INFORMED CONSENT: Please describe the process you will use to ensure your participants are freely giving fully informed consent to participate. This will usually include the provision of an Information Sheet and will normally require a Consent Form unless there is justification for not doing so. (Please state this clearly).	No tweets will be explicitly displayed within the results, and all data will be deleted at the end of the project. It would be difficult to gain consent from every user to participate as this is taken from a large number of users and additionally because we are searching for sockpuppets, accounts get deleted or permanently banned so keeping the data up to date may be difficult. I will make sure that the research output will be unidentifiable. When users accept the Twitter T&Cs they consent for their data to be used for research.
C16. RIGHT OF WITHDRAWAL: Participants should be able to withdraw from the research at any time. Participants should also be able to withdraw their data if it is linked to them and should be told when this will no longer be possible (e.g. once it has been included in the final report). Please describe the exact arrangements for withdrawal from participation and withdrawal of data for your study.	No data will be linked to a user via the data displayed within the dissertation, all outputs are aggregate, and the data will all be dropped at the end of the project.

C17. OTHER ETHICAL ISSUES: If you answered YES to anything in A.1 above (participants who are potentially vulnerable or unable to give informed consent or in a dependent position) you must specifically address this here. Please also consider whether there are other ethical issues you should be covering here. Please also refer to the professional code of conduct you intend to follow in your research.	This study differs from most psychology or medical studies with participants, because the data to be used is readily available on Twitter. The users (our version of participants) have accepted the T&Cs for their data to be used within research projects and their data are publicly available. A1: Some of these users will be vulnerable due to being from a minority group, however they are not vulnerable in the described way. A2: Users have accepted the T&Cs of Twitter which include the acceptance of their data being used within research. The users do not know what studies their data is being used for, such as this case, so there is no informed consent. There will be no deception in this study e.g., posting something as a bait for users to interact with. A5: Tweets and account data may display their beliefs however none of this information would be displayed in the dissertation. In addition, these users

have already posted publicly rather than via a private one-to-one interview. The sockpuppet detection method seeks to establish whether a user is a sockpuppet not to identify the user.

A6: Yes, however data will be pseudo-anonymised, and further only aggregate data will be used in the output of the analysis.

>> Data Protection, Confidentiality, and Records Management	
C18. DPA: Will you ensure that the processing of personal information related to the study will be in full compliance with the Data Protection Act 2018 (GDPR)? Please give full details under C19a. (http://www.sussex.ac.uk/ogs/policies/information/dpa)	Yes
C18a. C18a. If you are transferring any personal data outside of the United Kingdom, you must explain how you will comply with data protection legislation. Further guidance is available at: https://www.sussex.ac.uk/ogs/policies/information/dpa/transfersoutsideeea	No
C19. CONFIDENTIALITY: Will you take steps to ensure the confidentiality of personal information? Please explain how any identifiable personal records and research data will be managed and stored whilst ensuring that participants have given appropriate informed consent for this in C19a and or C20 below.	Yes
C19a. DATA MANAGEMENT: Please provide details of any anonymisation procedures and of physical and technical security measures (including secure storage) of identifiable personal and research data here. Indicate how these will be employed in the collection, analysis and research output and dissemination stages.	Data will be pseudo-anonymised. The project does not seek to collect personal data about the users. Rather, it is about characterising whether they are a sockpuppet and whether they have strong influence on the discourse. The data will be stored on the university servers.
C20. CONSENT: If personal information related to this study will be retained and shared (i.e. in outputs) in a form that is not fully anonymised (separated from information that can identify the participant) please outline how participants will be made aware of this (including any limitations) and indicate their consent.	
C21. Will the Principal Investigator take full responsibility during the study, for ensuring appropriate storage and security of information (including research data, consent forms and administrative records) and, where appropriate, will the necessary arrangements be made in order to process copyright material lawfully?	Yes
C21a. If you answered "No" to the above question, please give further details of how data and records will be managed:	
C22. Who will have access to personal information and data relating to this study?	Me and my supervisor.
C23. DATA MANAGEMENT PLAN: AFTER the study: State how long study information including research data, consent forms and personal identification will be retained, in what format(s) and where the information will be kept. http://www.sussex.ac.uk/ogs/policies/information/recordsmanagementguidance	Data will be deleted once analysis is complete, which will be 17 May 2022 at the latest.
>> Other Ethical Clearances and Permissions	
C24. Are any other ethical clearances or permissions (internal or external) required? Please see the help text (i) for further details.	No
C24a. If yes, please give further details including the name and address of the organisation. If other ethical approval has already been received please attach evidence of approval, otherwise you will need to supply it when ready. (You do not need to provide evidence of a current DBS check at this point).	

Protocol Document

Influence Measure

- RI[1] means Retweet Impact, where a user can repost content from another user which means they have influence over them. It uses the amount of tweets which have been retweeted and multiplied by the logarithm of the number of unique users who have retweeted the author's tweets to calculate the influence of the user.
- MI[2] means Mention Impact, measuring the amount of impact the user has surrounding mentions of a user on twitter; it uses several metrics surrounding mentions, such as the number of mentions a user writes in total.
- SNP[3] means Social Networking Potential, this combines many metrics together, these generally surround the number of retweets, number of unique mentions, reply count and the amount of tweets a user has created.
- TunkRank[4] is an adaptation of the PageRank algorithm to be applied to twitter, this is where it uses exclusively the count of followees to determine the influence through recursively finding the influence of the followees (an recursive algorithm) with a probability of retweet times one divided by the amount of followees the user has.
- UserRank[5] is an extension of TunkRank, where it uses the amount of followers divided by the tweets of a certain user for the probability for influence instead of the probability of retweet and divides by the count of followers of the user rather than retweet.

Sockpuppet Detection

- Subgraph similarity[6] is where two users could follow and interact with the same users a lot, showing high similarity, or could not share any following users or interactions between them showing low similarity. Users who could be sockpuppets want to try and recreate the influence networks they had using a different account, this is a potential method.
- Authorship Attribution (AA)[7] is where a particular writing style can be attributed to a specific user. We would compare the writing styles of each user with each other to try to detect if a user is a sockpuppet or not.
- SocksCatch[8] is a particular method which has several stages, the method itself uses non-verbal behaviour analysis where you have to extract features to capture a users activity or movements such as how close temporally they post on their twitter account. The method uses a machine learning algorithm such as a state vector machine (svm) to detect sockpuppets from such features.

Natural Language Processing

- Word frequency is where we would attempt to find the most commonly used words from the set for users and truncate the results to show the top results within the community.
- Topic detection is where we will try to find a single word to attribute the tweet topic to, we would then try and see if sockpuppets mention more about the community they are in over the other users within the community to see a potential method to detect sockpuppets within a community.

---

[1] Pal and Counts, 'Identifying Topical Authorities in Microblogs'.
[2] Pal and Counts. 'Identifying Topical Authorities in Microblogs'.
[3] Anger and Kittl, 'Measuring Influence on Twitter'.
[4] 'A Twitter Analog to PageRank'.
https://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/
[5] Majer and Šimko, 'Leveraging Microblogs for Resource Ranking'.
[6] Wang et al., 'An Online Sockpuppet Detection Method Based on Subgraph Similarity Matching'.
[7] Coyotl-Morales et al., 'Authorship Attribution Using Word Sequences'; Solorio, Hasan, and Mizan, 'A Case Study of Sockpuppet Detection in Wikipedia'.
[8] Yamak, Saunier, and Vercouter, 'SocksCatch'.

DATA MANAGEMENT PLAN

Project title:	Assessing the influence of sock puppets on the transgender Twitter discourse
Researcher name:	Ethan Pannell
Supervisor name:	Professor Luc Berthouze
Date:	09/12/2021

1. Overview

The overall aim of this project is to try and assess the extent to which sockpuppets influence specific online communities on social media networks. Sockpuppets are malicious bots or users that try to harm or invite discourse to advocate their view, such as using misinformation to create discourse within a community. Usually, users who are banned from their main accounts do this to circumvent the ban or to control a group of sockpuppets (in which case they are referred to as puppetmaster). Their aim is to try and share their ideas on a larger scale. This project will specifically focus on detecting the presence, and characterising their influence, of such malicious operators in the Transgender Twitter discourse in order to help inform of the community about those who try and poison the discourse from within.

Concretely, and in short, this project will make use of Twitter data over a limited period of time (2 time points across a month). The data will be collected via the Twitter's Academic Research API (applied for through my supervisor). Given that the data collected will include data from a protected minority group, only aggregated data will be used to ensure that no single individual can be identified from it.

The purpose of this Data Management Plan is to outline how the project will collect, manage and secure data – both personal data[1] and other data – during the research project.

2. Data summary

- ***Data Sources:*** *Data will be collected from Twitter using the Twitter API*
- ***Data Collected:*** *Account and tweet data will be collected. For accounts, we will collect their account id, bio, account creation date, who they follow, and who follows them. For tweet data, we will collect data on the tweet textual content, the like count, retweet count, reply count and if a tweet is a reply or retweet (either a default retweet or a quote retweet).*
- ***Deriving Personal Identity through Triangulation / Reverse Search:*** *We will be only displaying aggregate data of users involved within the discourse, statistics or particular topics discussed, none of which can be used to identify a particular user. We will not collect data on users' GPS coordinates or general location, so there will be no way to locate them.*

[1] Personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

- ***Data Collection Methods:*** *The data collection methods are to search using particular hashtags[2] using the Twitter API, or to manually find large accounts which are active in the transgender Twitter discourse and use snowball sampling[3] to map the set of users topologically from who follows who. Snowball sampling denotes a technique whereby a set of users are selected, and then their immediate followers are collected (this is one-hop). Repeating the process on each of those followers identifies the two-hop group of users, and so-on. In this project, we will cap the depth to three hops.*
- *We will collect tweets from two points in time, over a month, if the data is too much (tens of millions) we will reduce the time span iteratively until it's below that. We will start off with the following strings for the hashtag search method for identifying actors of the target discourse: #TransRightsAreHumanRights", "#TransLiberationNow", "#trans", "#transgender", "#enby" and "#nonbinary".* These hashtags were chosen by myself based on the fact that the hashtags are centered around the trans topic. For snowball sampling we will collect the top 15 of the latest top tweets and use those users as the seed users, the search terms will be the same as the hashtags before.
- ***Data Volume:*** *We will collect all tweet data between two points in time over a span of a month. As we cannot calculate this in advance, we cannot give a definitive answer, however an estimate is in the hundreds of thousands.*
- ***Data Analysis:*** *Once we have collected our dataset, we will be undertaking a number of different analysis techniques:*
    - *Influence measuring - whereby an algorithm attempts to calculate a measure for how much influence a user has in a network. This takes account of different metrics from a tweet, such as retweet count, mention count, likes and replies to create a measure to determine the influence of a user.*
    - *Sockpuppet detection - whereby an algorithm attempts to detect sockpuppets within a network through methods such as authorship attribution and subgraph similarity techniques.*
    - *NLP tweet topic and word frequency analysis - whereby an algorithm attempts to identify a topic a piece of textual information is about, and frequency analysis analyses the most frequently used words used within a dataset.*
- ***Data Usage:*** *The results of the analysis will be displayed within a paper. The raw data will not be shared outside of the people involved with the project within the university.*
- ***Data Security:*** *All of the data will be held on secure University servers and only will be available to the project researchers.*
- ***Data Deletion:*** All data will be deleted on completion of the project.

3. Social media platform requirements

The project will use the Twitter Application Programming Interface ('API') to download and conduct analysis of relevant Twitter content, including Tweets and Tweet IDs.

Any use of Twitter content will be in accordance with the requirements of the Developer Agreement and Policy, and the 'Incorporated Developer Terms' as defined in the Agreement. In

[2] Benigni, Joseph, and Carley, 'Online Extremism and the Communities That Sustain It', DOI: 10.1371/journal.pone.0181405
[3] Goodman, 'Snowball Sampling', DOI: 10.1214/aoms/1177705148.

particular, the project will not use the Twitter content in a way prohibited by the 'Restrictions on use of Licensed Materials' in Section II of the Developer Agreement and the project will comply with the API Restricted Use Rules.

4. Fair and lawful processing of personal data

The project will involve the processing[4] of personal data and any such processing will be in compliance with UK data protection legislation, namely the Data Protection Act 2018 and the UK General Data Protection Regulation ('UK GDPR').

The lawful basis for processing personal data in the project is the University's 'public task' under Article 6(1)(e) of the UK GDPR. Under the University's Royal Charter, our purpose is "*to advance learning and knowledge by teaching and research to the benefit of the wider community*". Therefore, the processing of personal data for the purpose of our research activities is necessary for the performance of a task carried out in the public interest or in the exercise of the University's official authority. Consent, therefore, is not required to process the personal data.

It is not expected that the project will involve the processing of special category data[5], unless such data forms part of Tweet content, for example, an individual disclosing data concerning their health in a Tweet.

Should special category data be collected or otherwise processed, then such processing is permitted under Article (9)(2)(j) of the UK GDPR on the basis that it is necessary for '*scientific or historical research purposes*' and / or under Article (9)(2)(e) because it has manifestly been made public by the data subject.

5. Pseudonymisation, anonymisation and deletion

Pseudonymisation means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and securely. Anonymous data means that an individual cannot be identified, either directly or indirectly.

Once Twitter content is downloaded, personal data will initially be pseudonymised and then, where possible, anonymised.

Firstly, the Twitter handles will be replaced with arbitrary labels (a pseudonymised identifier). The mapping – to ensure that the label replacements are consistently applied to individual users / Twitter handles – will then be deleted. Prior to deletion, the mapping information which cross references Twitter handles and pseudonymised labels, will be stored securely and held separately to the API downloaded data. Access to the mapping information will be limited to the researcher and their supervisor.

---

[4] Processing means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

[5] Special category data : http://www.sussex.ac.uk/ogs/policies/information/dpa/sc-crim-convictions

The Tweet IDs, which are the unique identification numbers generated for each Tweet and could enable individuals to be identified, will be removed. However, even with the handle and Tweet ID removed, individuals may still be identifiable by searching against the Tweet's textual content. This will be obscured for Tweets that are included in any publication. Further, if the content of the Tweet is sensitive, contains special category information, or if the Tweeter is potentially vulnerable, the text should be paraphrased for inclusion in final reports or in any publication.

Any personal data will only be retained for as long as is necessary for the purpose of the research project. The data will be deleted once all analysis has been completed as it is required for each stage of the project and analysis, this will be the max of May 17th 2022. We both need sockpuppet detection and influence measures which rely on twitter data until the full analysis is complete due to the project being based exclusively around twitter itself, this isn't a sub-component of the project.

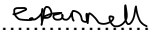
6. Data storage and security.

All project data will be stored securely on University servers / storage[6], and in accordance with the University's Information Security Policy. Appropriate access controls will be in place to ensure that access is limited to the researcher and supervisor, or other individuals authorised by the researcher and supervisor.

All project data will be classified as either 'Sensitive' or 'Protected' in accordance with the University's Information Classification and Handling Policy and password protection, encryption and other information security arrangements will be in place, as detailed in the associated Matrix.

7. Third parties

No third parties will be used during this project.

Signed:  ……

Researcher name: Ethan Pannell

Dated: 09/12/2021

Signed:  ……

Supervisor name: Luc Berthouze

Dated: 09/12/2021……

[6] https://www.sussex.ac.uk/its/services/research/researchdata/datastorage

1. Question A6, asks "Will the study involve collecting any personal special category information* in a form that could allow the participant/participants to be identified? [* identifiers relating to race, ethnic origin, politics, religion, trade union membership, philosophical beliefs, genetics, biometrics, health, sex life or sexual orientation]. The answer you have given here is 'No' but a check of the # search terms indicates that the data collection will include some of these identifiers above the level of part of a Tweet content, so please review this.

It was agreed (see email from Lauren Shukru dated 6 December 2021) that, we quote, “since the project is not targeting any particular special categories of personal data, the incidence of special category data collected is likely to be similar to that such as incidental health data, and is therefore in keeping with the data processing set out in the Outline Data Management Plan Template for student projects”. Please note that in the interest of avoiding any doubt/confusion, we have reworded the application in various places (including the title) to clarify that we are not seeking to identify members of a particular community but rather studying Tweets dealing with a particular discourse. This discourse will likely involve members of a special category community but will also involve non-members as well as artificial entities – the sockpuppets that we are interested in.

In addition, all efforts will be made to process (but not store) tweets on the fly.

2. The same applies to the statement in the Outline Data Management Plan template: "It is not expected that the project will involve the processing of special category data."

As above.

3. Please note that the collection of special category data on a mass scale requires a Data Protection Impact Assessment (DPIA) (https://www.sussex.ac.uk/ogs/policies/information/dpa/dpia for the DPIA screening tool.)

As above. Email from Alexandra Elliott (Head of Information Management and Compliance) confirms that, we quote, she “[I] don’t think a DPIA is required on this occasion”.

4. The Outline Data Management Plan also states that "the research project will only disclose the project data to third parties – for example, through the use of third-party tools for machine learning and data analysis, or other third party software or platforms – where such third parties are compliant with the UK data protection legislation requirements. Advice will be sought from the University's Data Protection Officer as necessary." In these circumstances, please contact the DPO now in order to establish whether a DPIA is required for the project, as reviewers will need to know this.

This was an error. No third-party tool will be used.

Finally, in relation to C13, we originally indicated that the project would be carried out on a “personal computer either at university or home”. This has been corrected. The data will only exist on university servers.

E Pannell and L Berthouze
20 December 2021

UNIVERSITY OF SUSSEX

Sciences & Technology C-REC
crecscitec@admin.susx.ac.uk

Certificate of Approval

Reference Number	ER/EP396/1
Title Of Project	Assessing the influence of sock puppets on the transgender Twitter discourse
Principal Investigator (PI):	Luc Berthouze
Student	Ethan Pannell
Collaborators	
Duration Of Approval	5 months
Expected Start Date	13-Jan-2022
Date Of Approval	13-Jan-2022
Approval Expiry Date	17-May-2022
Approved By	Karen Long
Name of Authorised Signatory	Lauren Shukru
Date	13-Jan-2022

*NB. If the actual project start date is delayed beyond 12 months of the expected start date, this Certificate of Approval will lapse and the project will need to be reviewed again to take account of changed circumstances such as legislation, sponsor requirements and University procedures.

Please note and follow the requirements for approved submissions:

Amendments to protocol

* Any changes or amendments to approved protocols must be submitted to the C-REC for authorisation prior to implementation.

Feedback regarding the status and conduct of approved projects

* Any incidents with ethical implications that occur during the implementation of the project must be reported immediately to the Chair of the C-REC.

Feedback regarding any adverse(1) and unexpected events(2)

* Any adverse (undesirable and unintended) and unexpected events that occur during the implementation of the project must be reported to the Chair of the Science and Technology C-REC. In the event of a serious adverse event, research must be stopped immediately and the Chair alerted within 24 hours of the occurrence.

Monitoring of Approved studies

The University may undertake periodic monitoring of approved studies. Researchers will be requested to report on the outcomes of research activity in relation to approvals that were granted (full applications and amendments).

Research Standards

Failure to conduct University research in alignment with the Code of Practice for Research may be investigated under the Procedure for the Investigation of Allegations of Misconduct in Research or other appropriate internal mechanisms (3). Any queries can be addressed to the Research Governance Office: rgoffice@sussex.ac.uk

(1) An "adverse event" is one that occurs during the course of a research protocol that either causes physical or psychological harm, or increases the risk of physical or psychological harm, or results in a loss of privacy and/or confidentiality to research participant or others.

(2) An "unexpected event" is an occurrence or situation during the course of a research project that was a) harmful to a participant taking part in the research, or b) increased the probability of harm to participants taking part in the research.

(3) http://www.sussex.ac.uk/staff/research/rqi/policy/research-policy

9.2 Network Images

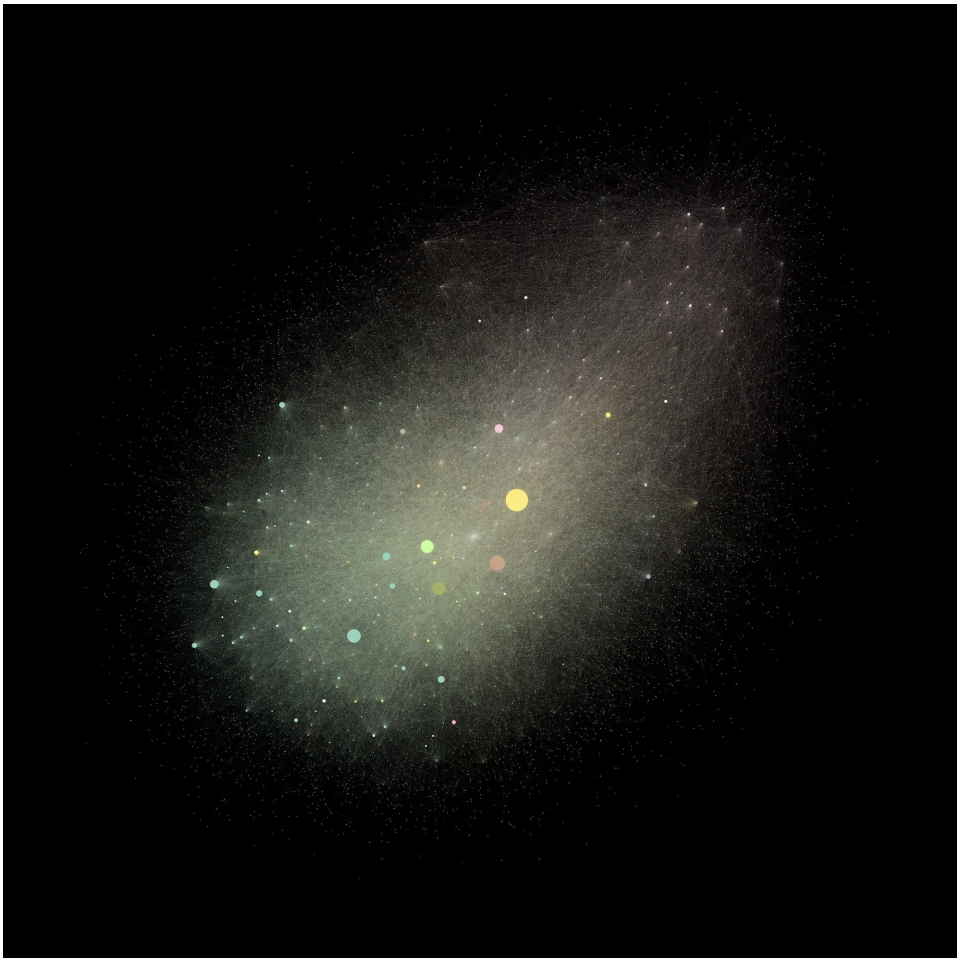


Figure 17: Final Reply Graph

9.3 Tables

Position	Count	Anonymised Text
1	804	Trans women are women
2	235	Trans
3	226	Adult human female
4	222	Petition to suspend olympic transgender policy
5	194	Describing men's rights activists as hating trans people
6	184	Asking if trans hatecrime should be questioned within legislation
7	158	Are you trans
8	156	Trans rights
9	151	Trans women aren't men
10	150	Reworded version of position 5
11	140	Gender critical
12	134	Woman adult human female
13	128	Endorsement of charity for guidance on trans inclusion
14	127	Describing that there has not been any trans olypians and that they do not hold a record in sports
15	125	Reworded version of position 13
16	125	Stats about people within the olymics, that a total of two have been transgender women and that they haven't won anything, citing it's due to androgen deprivation
17	121	Thank you
18	118	Abusive tweet against a TERF in response to their transphobia
19	116	Trans women
20	115	Information about a trans support group
21	112	A person calling left wing people who are accepting of trans people that they help companies implying it's bad
22	112	Describing the discourse as a gender debate and proposing there is medical deceit against trans children in anti-trans tone
23	109	American black trans athletes did not get scholarships, neither have an record and will not continue sport after graduating
24	109	Athlete discusses the impact that trans women have in women's sports, and that they have to save women's sports from trans women
25	107	An endocrinologist criticises medical help for trans people because they think they are just doing it to fit into a stereotype of a particular sex
26	103	Describes WHRC discussing that trans people are a threat for women's rights in the US
27	102	Trans women are not a threat to women's rights
28	98	Describes that sex matters in law and signing a petition about how trans people are politically erasing sex
29	97	An interview between two women's rights activists is happening, saying that feminism doesn't exist within leftism
30	97	Asking Why are people are interested in trans-exclusionary feminist ideas and asking what they gain from harassing trans lesbians
31	95	Trans women are men
32	94	Discuss rise in referrals to the GIC in London, using negative hashtags against trans people
33	92	Describing that trans women have an advantage in sport and that fear of being bigotted is a higher priority than science
34	91	Describing a type of person who uses privilege around being transgender further influence others to become trans, that the type of person call everyone else transphobic, and that cis is a slur
35	89	Reworded position 34
36	88	General petition supporting tweet
37	88	Calling GLAAD is a lobby group and that it has created lists of people to stay from who argued against helping trans children medically
38	87	Trans people
39	87	Endorsing a book which says it's impossible to be trans and that they are just becoming stereotypes of the gender they identify as

Table 5: Text content of top 39 repeated text with names removed, 40 to 50 in 6

Position	Count	Anonymised Text
40	85	Reworded position 20
41	85	Trans rights are human rights
42	85	Describing using the trans flag as a form of racism and saying that trans people are creating a false sense of support
43	82	Define trans
44	82	General affirming comment for trans and nonbinary people
45	81	Criticising activism for trans people as maintaining social structure and describing trans people as re-hashing homophobia and misogyny as progressive
46	81	Describing a that they foresee that the affirmation approach to helping trans people will stop and describing a charity for trans children to be taken apart
47	80	A petition to not allow trans topic being discussed in schools in Scotland
48	77	Trans women are trans women
49	77	I'm trans
50	76	An author discusses book about gender dysphoria and looks at the political discourse around trans identity in a negative light

Table 6: Text content of text which repeated text with names removed who were inbetween the top 40 and 50, 0 to 39 in 5

Position	User	Tweet Count
1	46	12064
2	30	8316
3	236	6958
4	1106	6079
5	5061	5760
6	5908	5705
7	1004	5237
8	204	5229
9	46075	4524
10	248	4488
11	5147	4483
12	222	4440
13	6751	4387
14	15272	4192
15	46658	3871
16	79282	3849
17	209	3842
18	12211	3821
19	299	3767
20	22646	3674
11	48972	3639
22	13314	3626
23	26	3498
24	83	3432
25	348	3390

Table 7: Top users who tweeted the most

Position	User	Interactor	Count
1	9213	8765	2540
2	7187	75445	2232
3	22698	3915	2016
4	22698	6329	1810
5	7187	7187	1791
6	9213	389	1495
7	15315	49641	1414
8	7187	76355	1323
9	22698	16914	1305
10	9213	694	1170
11	15087	15087	1159
12	1004	1004	1154
13	7187	76192	1152
14	14365	1960	1045
15	14365	50276	966
16	24937	2204	900
17	2051	53444	792
18	28021	28021	773
19	14365	199	718
20	25	36913	700
21	9213	1060	700
22	14365	50605	668
23	15315	8678	651
24	7187	58970	612
25	24937	24937	600

Table 8: Top people who directly replied to another person

Position	User One	User Two	Similarity Score
1	36913	12062	0.963
2	14616	24937	0.274
3	25	36913	0.274
4	47003	8678	0.148
5	75445	76355	0.134
6	2544	47003	0.129
7	47003	14365	0.129
8	36913	14776	0.129
9	36913	656	0.127
10	24937	4967	0.122
11	36913	4380	0.117
12	46992	15480	0.112
13	14365	8678	0.112
14	2544	8678	0.109
15	50205	14776	0.103
16	2544	14365	0.101
17	47003	15511	0.101
18	58686	14776	0.099
19	22698	14776	0.094
20	8678	15511	0.090
21	14365	15511	0.088
22	25	12062	0.086
23	36913	16914	0.083
24	2544	15511	0.083
25	47003	15835	0.082

Table 9: Similarity of users who replied to one other user > 150 times

Position	User	PageRank Score
1	22	0.052
2	25	0.033
3	17	0.031
4	16	0.026
5	20	0.019
6	27	0.018
7	13	0.013
8	12	0.012
9	22698	0.010
10	5	0.010
11	21	0.008
12	15602	0.007
13	6	0.007
14	26	0.007
15	13821	0.007
16	10	0.007
17	9	0.006
18	14499	0.005
19	23197	0.005
20	28	0.005
21	3915	0.005
22	28514	0.005
23	22944	0.005
24	29315	0.005
25	2	0.005

Table 10: Users with the highest PageRank Scores