Detecting Disinformation and Information Warfare on Social Media

Thomas Willingham

A thesis submitted for the degree of
Bachelor of Advanced Computing (Honours)
The Australian National University

November 2021

Except where otherwise indicated, this thesis is my own original work.

Thomas Willingham
18 November 2021

I dedicate this thesis to my Great-Grandfather Bill Haigh, who is sadly no longer with us. His fascinating tales of an honest and exciting life in politics enthralled me as a child, and were a significant cause of the passion I developed for political affairs which lead me to seek out this project. I fear that the political discourse we have today may be unrecognisable to what he knew in his time in office, but I like to think he would be pleased by my seeking out those who corrupt that system.

Acknowledgments

This year has been an extraordinary journey through the world of research, machine learning, and social data science that has only been possible as a result of the support of several people. Thanks to the fellow members of the Behavioural Data Science group, for providing support and guidance at various points in the project, interesting material to discuss in the weekly reading group, and listening to my presentation dry runs.

Thanks also to the people around me personally; my friends, especially Steph and Caitlin, and my family have been immensely supportive and kind throughout the year, even when I've had difficulties adjusting to the intense intellectual work.

Finally, many thanks must go to Andrei, for supervising this work, and teaching me so much of what I now know about machine learning, and the world of research in general. Supervising a third-year student without much advanced training in machine learning must have seemed an enormous risk at the beginning of the year for a data science project, but I am immensely grateful for the opportunity, as well as all of the support I've been provided to do this project well.

Abstract

The spread of disinformation in the 21st century has become of enormous concern for the integrity of democracy, the way we relate to each other online, and in extreme cases, the health and safety of individuals. This project explores how we can utilise information from disinformation campaigns in the past to predict disinformation as it arises into the future. Building on prior work analysing the structure of a limited number of political scandals, we continue to explore three main streams of work in detecting disinformation: content analysis, hashtag analysis, and network analysis. We trial a number of different content analysis methods, finding that a pre-trained BERT is capable of significantly exceeding a random baseline at detecting disinformation in unseen data from a new political context. We then demonstrate that polarising hashtags can be identified by clustering hashtags based on the users who use them. We finally go on to demonstrate an initial approach to combining information from hashtags or content with the interaction networks that have been shown to be effective in past work. All of these techniques combine to provide a platform for a system that could detect disinformation in real time, as it emerges in political contexts that do not yet exist.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context: Mistrust and Fake News, the New Era of Disinformation

With trust in politicians and political systems at an all time low (Cameron and McAllister [2019]), the world has seen the rise of disinformation campaigns on social media. Whether it be to assist in the election of a preferred U.S political candidate, instilling distrust in the Australian Labour Party by claiming they would introduce a "death tax", or producing a rumour that "African gangs" were roaming the streets of St Kilda (Marineau [2020], Henriques-Gomes [2018]), nefarious actors are aware of the power of disinformation. There is also reason to believe that hundreds of people have died as a result of disinformation relating to cures for COVID-19 (Islam et al. [2020]), with social media amplifying messages from former President Trump around injecting hand sanitiser, as well as other harmful or ineffective methods of protecting yourself from the disease.

As a result, there is a significant interest in identifying these campaigns so that they can be stopped. Twitter, the platform which is studied in this project, has begun using "internal systems to proactively monitor content related to COVID-19" (Roth and Pickles [2021]). These systems hide tweets related to COVID-19 that are identified by twitter's machine learning algorithm as being similar to tweets that have been removed in the past by their moderation team (Twitter [2020]). These approaches have significant issues, even excluding the fact that they only consider disinformation relating to COVID-19. Specifically, only considering content that is similar to content that has been removed in the past runs the risk of the actors running the disinformation campaign changing their language slightly so that it sends the same message, but is no longer similar to content that has been previously reported. These actors can see the moderation actions taken, which means that no matter how many of their tweets are reported, they will know what modifications need to be made in order to beat the detection algorithm, leaving the task of removing disinformation to manual reports and fact-checking. This effect is seen in an honours thesis prior to this project, which observes users who are banned by Twitter for posting inappropriate content return with usernames that are identical to their previous identity, but for the

addition of a single emoji (Tripathi [2021]). In summary, approaches which are tested solely on whether the tweets they hide match those which were hidden in the past by the moderation team are doomed to fail at their task of managing disinformation on the platform, because there is an adversary attempting to thwart their detection efforts, who can slightly modify aspects of their online identity, or tweet content, in order to appear like a different individual to the detection algorithms.

Our project seeks to avoid this issue by answering our research question: *How can we detect disinformation campaigns using information that cannot be easily modified in future campaigns*? We search for common linguistic patterns in the content produced by users spreading disinformation in multiple contexts, as well as observing the effects they have on user interaction networks. In doing this, we generate context-independent models for detecting misinformation, which may be used to detect disinformation campaigns as they occur in the political discussions of the future.

1.2 Detecting Disinformation: A three layered approach

In order to answer our research question, we adopt a three-layered approach. Firstly, we examine in detail the content of disinformation tweets, so as to understand the message as it is received by a Twitter user coming across their messaging. This is framed as a classification problem, where we look for a machine learning model the is capable of classifying users as being normal users, or part of a disinformation campaign. We compare a number of Natural Language Processing models by their performance on this task. Ultimately we obtain the best performance by pre-training a BERT classifier for sentence embeddings in an adversarial set-up using synthetic labels and texts produced by GPT2. In doing so we reveal a consistent pattern of similar messages across different political controversies that the model can identify. This indicates the presence of a consistent approach to disinformation across topics and time periods.

Secondly, we extend this analysis by focusing more narrowly on specific parts of the content produced by these users. Namely, we examine the hashtags and entities which are used in their tweets. Hashtags are a unique facet of social media which allows Twitter to display tweets discussing similar topics to users interested in those topics. They therefore provide an indication of the communities which actors seeking to spread disinformation wish to insert themselves into. We see that hashtags are profoundly polarising, with certain hashtags being used near-exclusively by users on the far-right or far-left of Australian politics. This observation is powerful, as it allows us to re-construct the political leanings of users in new political controversies based on their usage of highly polarising hashtags that were identified in previous controversies. From a technical perspective, this provides a mechanism for labelling large novel data sets with minimal expert involvement. The expert labels a small number of hashtags at the beginning of an election, or a new topic being discussed,

and we are able to propagate these labels through the rest of the data set as new tweets emerge. In our work, we apply this technique to produce a set of hashtags which are used near-exclusively by members of one community of users on a previously unlabelled data set.

Finally, we take the knowledge we have gained from analysing the content, hashtag, and entity usage of a user, and combine it with information about the interactions between users on the social media platform to allow us to identify communities of users spreading disinformation. There is discussion in prior work of using the interactions between users to identify if they are a member of a disinformation community. However, we find that for some topics, the interactions between users do a poor job at describing the communities of users in the data. For instance, in the #QandA dataset, there are many non-organic interactions with the official #QandA account due to how the QandA show is structured. To allow for network methods to be useful in these circumstances, we incorporate information about the content and, more specifically, hashtags in use by users alongside their interactions in order to identify their community. This process seeks to identify clusters of users that may be infiltrating or creating communities within which to spread disinformation.

To summarise, our primary contributions are:

- An analysis of multiple Natural Language Processing models for detecting disinformation automatically
- A pre-trained language model capable of detecting disinformation content on an unseen Australian political data set
- A set of polarising hashtags on a previously unlabelled data set based on a small amount of manual labelling
- An approach for improving network methods by way of hashtag analysis in cases where the methods discussed in prior works are not appropriate.

1.3 Thesis Outline

Chapter 2 discusses some prior and related work utilised in this project, as well as a broad statistical analysis of the data utilised in the thesis. Chapter 3 contains the analysis of the common linguistic patterns found in the content of disinformation spreading users. Chapter 4 turns more specifically to looking at the hashtags and entities used by misinformation spreading users on social media platforms. Chapter 5 uses the analysis from previous chapters, as well as observations about the interactions between users in the data, to detect communities of problematic users. Finally, Chapter 6 concludes the work, and discusses further avenues for future work in this area.

Chapter 2

Datasets & Prior Work

In this section, we discuss the key datasets that are referred to throughout the thesis. We examine the statistical features of each dataset, and also discuss our source of truth for whether a user is spreading disinformation. Throughout the thesis, a ground truth label refers to a label for the user that tells us whether or not a particular user is spreading disinformation that we assume to be correct. The opposite of this is a predicted label, which is a label that is produced by one of our computational models, and thus may or may not be accurate. For *#deathtax*, these ground truth labels are taken directly from the prior work, whereas for *#auspol* and thus *#stkilda*, we create our own labels by utilising the procedure proposed in the prior work.
Each of the described datasets are defined by a hashtag. This hashtag appears in every tweet that is contained within that dataset. So every tweet in the *#deathtax* dataset contains the hashtag *#deathtax*. Since tweets can contain multiple hashtags, it is possible for one tweet to be part of more than one dataset. In fact, *#deathtax* and *#stkilda* are both subsets of *#auspol* in this project.

2.1 Ground Truth Labels

By extending the prior work done in understanding the structure of the social network (Tripathi [2021]), we generate a network graph of all users in the *#auspol* data set. We run the same clustering algorithm (Blondel et al. [2008]) on the entire *#auspol* data set to determine communities of users.

This shows that the conclusions from the prior work hold, with a clearly identifiable misinformation cluster visible in purple in Figure 2.1. We can identify this as the misinformation cluster in two main ways. Firstly, we note that this cluster contains all of the misinformation opinion leaders identified in (Tripathi [2021]), indicating that this is the same community as was identified there. Secondly, we can look at some example tweets from the cluster, which lend credibility to the idea that this is the misinformation cluster. These samples include:

- *sigh* Mark, it's called Summer. And guess what? There's gonna be more climate change in a few months.. it's gonna get cold
- @SkyNewsAust Still pumping his tyres up. Spouting renewables guff. Snowy

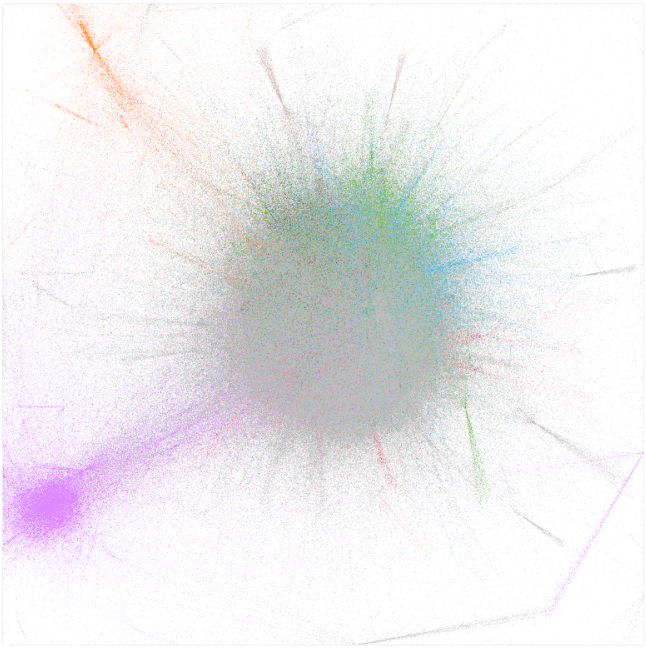

Figure 2.1: Gephi Visualisation of *#auspol* Network

2 is a joke.

- Man ambushed in a #Sunshine #Melbourne park by a gang. #3aw livable city police @3AW693 @3AWNeilMitchell #9news

These tweets are examples of fake news connected to the African Gangs controversy described below under *#stkilda*, and fake news suggesting that climate change is a hoax.
Through the remainder of the thesis, where we require ground truth labels for whether a user is a disinformation spreader in *#deathtax*, we use those produced in (Tripathi [2021]). Where we require labels for *#stkilda* or *#auspol*, we consider those users in this purple cluster in Figure 2.1 to be disinformation spreaders. As will be discussed in Section 2.5, *#qanda* cannot be meaningfully labelled in this way. Labelling this dataset is discussed in Chapter 4 using a more sophisticated methodology.

2.2 *#auspol*

The *#auspol* hashtag encompasses a wide variety of Australian political tweets. This tag is typically used simply to indicate that a tweet relates in some way to Australian Politics. The *#deathtax* and *#stkilda* datasets are subsets of *#auspol*. It contains 17

million tweets and it was collected in 2019/2020, during the 2019 Australian Federal Election. This dataset can be labelled using the above procedure, and the network graph may be seen in Figure 2.1.
Since this dataset contains so many tweets on so many different topics, it is not analysed in detail in the subsequent chapters of the thesis.

2.3 *#deathtax*

During the 2019 election campaign, there was a rumour that the Australian Labour Party (ALP) were seeking to secretly introduce a "death tax" if elected. This was to be a significant tax on the estate of those who passed away, and was described by some twitter commenters as "stealing grandma's gold teeth". This rumour was unable to be validated, with ALP members taking to social media to indicate that there was no death tax (Murphy et al. [2019]).
The dataset in use in this project consists of tweets that utilised the hashtag *#deathtax* or *#inheritancetax* during 2019 alongside *#auspol*. There are 30,234 such tweets, with 7,983 users involved in the discussion. Prior work suggests the dataset contains 4981 normal users, and 2907 users involved in perpetuating the death tax rumour. This amounts to 36.4% of users in the dataset being disinformation spreaders. As such, we would expect that randomly guessing whether a user is spreading disinformation or not to result in an F1 score of approximately 0.364. Approaches with a higher F1 score than this can therefore be considered better than a random baseline.

2.4 *#stkilda*

Throughout 2018 and 2019, there was also a rumour perpetuated that there were significant numbers of gangs of African youths in the Melbourne suburb of St. Kilda who were committing violent crimes. Whilst it is more difficult to emphatically prove that there were no such gangs, no strong evidence has been found to support their existence. As such, we consider that the claim that such gangs definitely existed to be an instance of disinformation. This disinformation was deployed by various right-wing politicians, such as Fraser Anning, who attempted to use it to paint the Victorian ALP government as incapable of preventing crime (Henriques-Gomes [2018]).
Our dataset consists of tweets that utilised the hashtag *#stkilda* during 2019 alongside *#auspol*. There are 16,570 such tweets. Ground truth labels are taken using the disinformation cluster identified in Figure 2.1, since all users in *#stkilda* are neccesarily also in *#auspol*. We find that approximately 6.0% of users in this dataset are disinformation spreaders. Since this is a very imbalanced class problem, we define a random dataset to determine whether classification methods are doing better than a random baseline on this dataset. This is discussed in Section 3.1.

2.5 #qanda

The #*qanda* dataset contains approximately 700,000 tweets, that were collected throughout 2020. Each of these tweets contains the hashtag #*qanda*. This dataset is unusual, since it is explicitly tied to the QandA talk show, a weekly political television program run by the ABC. In the show, prominent figures in Australia, including politicians, activists, and professionals, are invited to participate in a panel discussion centred around a theme. Questions to the panel are sourced from viewers of the show, who sit in the studio-audience and provide questions which are discussed by that week's panel. Tweets from users using the #*qanda* hashtag may be featured on-screen during the airing of the show. This contributes to the first observation about this dataset, that can be seen in Figure 2.2. This figure graphs the number of tweets made to #*qanda* on each day throughout 2020. We see regular peaks, which correspond to the airing of episodes. Throughout this thesis, we discuss tweets taken from an episode. This refers to tweets posted on the date of the episode, or during the week after the episode. This is based on the observation that the majority of discussion in the dataset occurs during the airing of episodes, and the assumption that most discussion directly after an episode likely relates to the topics discussed in that episode.
The interactions network graph for #*qanda* as a whole can be seen in Figure 2.3. This graph does not have a clear disinformation cluster. While there are three coloured clusters that separate from the main bulk of tweets, two of these (in black and red) are composed exclusively of discussions in languages other than English, and the final one (in purple) is centred around the twitter account of the host of the show. This becomes even clearer observing the interactions graph for a single episode, as can be seen in Figure 2.4. Here, we see a number of distinct clusters. Each of these clusters is centred around the account of one of the panellists on that week's episode of the show. As such, the methods described in the prior work cannot be used to create ground truth labels for this dataset. Instead, we analyse the hashtags in use in this dataset in Chapter 4, and discuss a compound similarity metric that may perform better than interactions alone in Chapter 5.

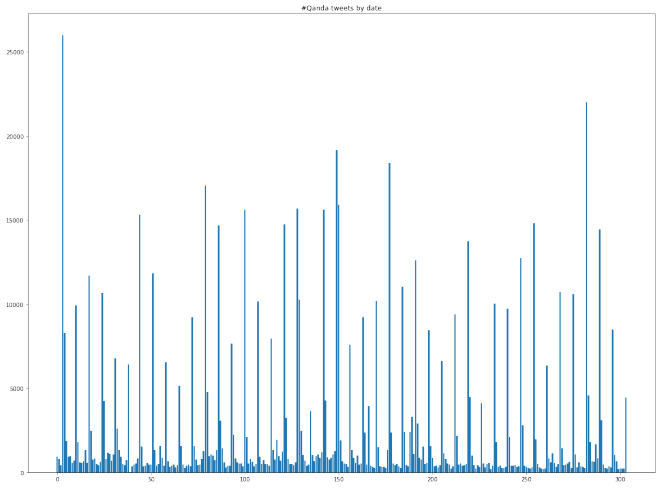

Figure 2.2: Tweet Counts per Day for *#qanda*

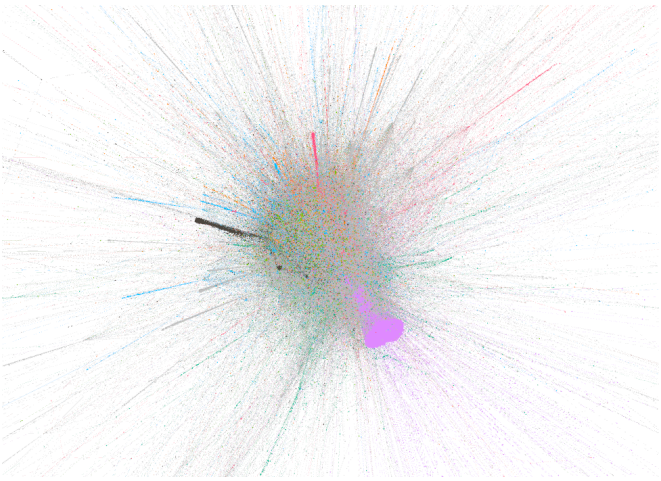

Figure 2.3: Gephi Visualisation of Entire *#qanda* Network

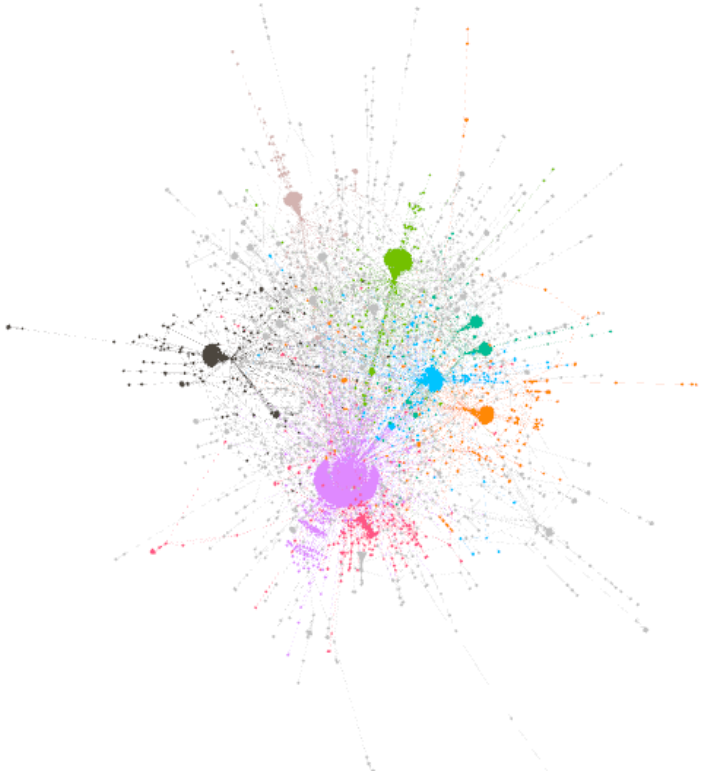

Figure 2.4: Gephi Visualisation of *#qanda* Network for Episode 1

Chapter 3

Finding the Disinformation Message: Content Analysis

The content produced by a user on social media has been the main way that disinformation and other problematic content has been identified in the past, and is the key part of the message that is portrayed to other users. We seek to identify users spreading disinformation messages through their content, since any significant change to the content being produced by these users would mean fundamentally changing the message that they are sending. In this way, our goal is to identify the arguments and structures in use by disinformation spreaders as they spread their problematic message across Australian political twitter. We focus especially on the ways that disinformation content in different political contexts is similar, explicitly training and testing our models on data drawn from different datasets.
This chapter begins with a discussion of how do this, and how to evaluate performance of such models that are designed to work on unseen datasets. Then, each of the techniques we use to analyse user content are introduced, before we examine the performance of each of these classifiers. In doing so, we show that it is possible to obtain classification results that perform significantly better than the control tests, indicating that we can identify unchanging linguistic patterns across disinformation campaigns. In other words, while the context of a tweet might change, the techniques used to spread disinformation do not.

3.1 Validation & the Random Dataset

Since our research question explicitly considers applying our models to future disinformation campaigns, we approach the classification problem in this work in an unusual way. Typical machine learning approaches take a dataset, split it into a training set and a test set, train the model on the training set, and examine the generalisation error on the test set.
While we do use this approach for the #deathtax dataset, we also consider applying the model that is built from training on #deathtax to campaigns with content that was not a part of the #deathtax dataset at all. Specifically, we take the model trained on the #deathtax dataset, and test it on the entire contents of the #stkilda dataset.

This provides information around how well our models are performing at the task of real-time disinformation detection.
However, this introduces a new problem. Since we do not split the #stkilda dataset into a test and train section, we need a baseline to compare our results to, in order to verify that we are accurately measuring how well the model is picking up disinformation. To do this, we employ a random dataset. This dataset is constructed by taking the #stkilda dataset, including it's class labels, and replacing the text of each tweet with the text of a randomly sampled tweet from #auspol. The effect of doing this is to create a control dataset, controlling for the size and class balance of #stkilda, but removing the disinformation content. As such, the main metric we use to measure the performance of our model is how well it performs at identifying disinformation spreaders on #stkilda, as compared to the result received by running the same model on the random dataset.

3.2 Approaches

3.2.1 `TFIDF` Modelling Function Words

Our first approach is to attempt to directly input all words from all tweets into a classifier, to observe whether there are words that are much more likely to be used by disinformation spreaders. To do this, we first use a bag-of-words model, which produces vectors with a length equal to the number of unique words in the entire dataset. Then, if a token t is used n times in a given tweet, the tth entry in the vector is equal to n. In order to avoid vectors being disproportionately affected by stopwords, such as "the", "a", or "and", we apply a Term Frequency Inverse Document Frequency (`TFIDF`) model to the vectors obtained from the bag-of-words model. In the `TFIDF` model, rather than the tth entry in the vector being equal to n, it is equal to

$$\frac{n_w}{\sum_{k=0}^{l} n_k}$$

Where n_w is the number of occurances of this token in tweet w, and l is the total number of tweets in the dataset. Intuitively, this means that tokens which are used infrequently will be given larger weightings in the vectors provided to the classifier.
For classification, we use a Random Forest classifer (Breiman [2001]). This classifier is provided with `TFIDF` vectors as training inputs, and classifies them as either being produced by a disinformation spreader, or a normal user. The results of this classification are discussed in Section 3.3.

3.2.2 Word Embedding Based Classification

Moving beyond an analysis of the tokens used by disinformation spreaders, we also conduct an analysis of how users talk on the platform. In order to begin doing this, we start by looking at semantic embeddings of individual tokens using the `Word2Vec` library [Mikolov et al., 2013]. The produced embeddings are vectors with arbitrary

Table 3.1: Relative Disinformation Usage of Identified Semantic Clusters

Cluster ID	% Usage in Disinformation Cluster	Total Number of Usages
25	70.6%	5160
1	63.4%	8478
16	63.3%	6327
55	29.5%	11835
29	19.0%	36791
28	17.2%	4515

Table 3.2: Tokens from Identified Semantic Clusters

Cluster ID	Tokens				
25	Planned	Implement	Draconian	Carbon	Scrapped
1	Parasitic	Purse	Rob	Hardworking	Rorted
16	Unloseable	Green	Rigging	Electoral	Illegitimate
55	Accused	Claim	Accuse	Warning	Urging
29	Liar	Lied	Lying		
28	Fearful	Fear	Scared		

dimensions, but where two vectors having a low distance in the vector space indicates that the two tokens are semantically similar. Accordingly, two vectors which are very far from each other in the vector space are considered to be semantically dissimilar.

3.2.2.1 Observing Semantic Clusters Directly

In order to determine whether there are distinct semantic clusters associated with disinformation or debunking, we use a clustering algorithm on the text of the *#deathtax* dataset. To do so, we take the `Word2Vec` encoding of all the words of the tweets in the dataset. Then, we find clusters through an implementation of the K-Means algorithm provided by the `scikit-learn` library [Pedregosa et al., 2011], looking for 60 clusters. Then, for each token in each cluster, the number of times that the token appears in the misinformation corpus and debunking corpus is counted, yielding the percentage of times that the token is used by each group. Sorting the clusters by the proportion of times their tokens are used in the disinformation cluster, then taking the top 3 clusters from each side that contain at least 1,000 references across clusters yields Table 3.1.

In order to determine which semantics are being used by each group, we then extract tokens from these clusters in the table. This is done by taking the top 5 highest used tokens in a cluster, except in the case where there are fewer than 5 tokens in the cluster, where the whole cluster is reported. These tokens are reported in Table 3.2 for the identified clusters.

From this, we see that there are semantic clusters for verbs, adjectives, and adverbs that are more likely to appear in a tweet spreading disinformation than a nor-

mal tweet. This is notionally independent of the nouns/people/objects involved in a particular scandal. That's because many of the clusters contain semantics for abstract concepts, such as fear, or lying. This also provides an initial insight into a key fact we discover about disinformation spreaders; they are very unlikely to talk about disinformation in the abstract. Correspondingly, we see that users attempting to debunk this disinformation appear to call out the other side for spreading disinformation.

3.2.2.2 Classification

Having determined that there are distinct semantic clusters that can be attached to disinformation spreaders, we attempt to determine which types of tokens are most useful in discriminating between disinformation spreaders and normal users. This is done for all tweets in the *#deathtax* dataset by collecting the text of each tweet and obtaining semantic embedding vectors for each token that is being included in a given analysis. Then, these vectors are added together and normalised to yield a single vector representing the semantic content of a tweet. These vectors are used as training data for a Random Forest classifier (Breiman [2001]).

We observe the effect of allowing this classifier access to some tokens and not others by training classifiers that were only exposed to a subset of the total number of tokens. Specifically, the texts are also run through a part of speech tagger, associating each token in a text with its part of speech. We utilise these tags to perform an ablation study over parts of speech, removing particular parts of speech from the information available to a classifier. This is done to examine whether there are some particular words in tweets that hinder generalisation performance. For instance, we might expect that removing all nouns in the dataset would allow better generalisation, since nouns are a large part of what uniquely identifies a particular political context.
In order to yield a score for the classifier, we withhold some data from the *#deathtax* dataset to use as a test set, train the classifier on the appropriate vectors from the training set, and then score it on the test set.

Finally, in order to assess generalisation performance, we also assess the classifier on the *#stkilda* and random dataset described in Section 3.1.

Reporting the F1 score for each of the three datasets on each trained classifier yields Table 3.3.

This ablation study does not appear to reveal any positive effects on generalisation performance obtained by removing parts of speech. The case where no tokens are removed performs the best of all on both *#deathtax* and *#stkilda*, with the random dataset scores remaining roughly consistent. As such, we report the scores obtained from using the Random Forest classifier on `Word2Vec` embeddings with no tokens removed for the remainder of this thesis.

Table 3.3: Results of Ablation Study of Parts of Speech using `Word2Vec`

Tags Excluded	*#deathtax* F1	*#stkilda* F1	Random F1
None	0.873	0.250	0.151
Adjectives	0.884	0.229	0.147
Verbs	0.867	0.207	0.155
Adverbs	0.868	0.215	0.153
Nouns	0.824	0.160	0.155
Determiners	0.827	0.241	0.151
Nouns, Adjectives	0.866	0.169	0.152
Nouns, Adjectives, Verbs, Adverbs	0.855	0.156	0.167
All Except Punctuation	0.767	0.201	0.166
All Except Verbs	0.844	0.150	0.155

3.2.3 Sentence Embedding Based Classification

We additionally utilise BERT [Devlin et al., 2018], a deep neural network based tool pre-trained on a very large dataset to generate sentence embeddings. These embeddings are similar to the word embeddings, in that sentences with similar semantics (i.e that could be substituted for each other in a text) are given similar vectors.

We expect better results from using BERT to create a single embedding for the textual content of a tweet than normalising sequences of word embeddings, as when the BERT model is trained on the initial dataset, it is able to pick up details of sentence structure which cannot be encoded in the individual word embeddings. These details may include the order of words in a sentence, or where punctuation usage significantly changes the meaning of the sentence. Examples of such cases abound, but one example was when The Associated Press tweeted "BREAKING: Dutch military plane carrying bodies from Malaysia Airlines Flight 17 crash lands in Eindhoven". This sentence was read by many to mean that the Dutch military plane had crashed, but what they intended to write was "Dutch military plane, carrying bodies from Malaysia Airlines Flight 17 crash, lands in Eindhoven". Such nuances are unlikely to be picked up on using an approach that only focusses on the individual words in a sentence, but may be picked up by a more sophisticated method that takes the whole sentence into account.

In order to allow BERT the opportunity to identify larger-scale details of a user's total textual output, we initially provide the BERT embedding generator with the total textual output of a user in the *#deathtax* dataset, as opposed to the previous methods, where the classifier is only provided with a single tweet at a time. Then, once the embeddings are generated for each user, we acquire the label for that user as being part of an identified disinformation spreading community or a normal user. Finally, we feed this information into a classification head that sits on top of the BERT architecture, which uses a neural network to classify users based on whether

they belong to an identified disinformation-spreading cluster.

We opt not to repeat the ablation study approach of removing tokens tagged with particular parts of speech from the data available to the BERT model, as this would have the effect of negating the larger-scale linguistic patterns that we wish for BERT to pick up on.

Results from this classification are discussed in Section 3.3.

3.2.3.1 Pre-training `BERT` using `DINO`

Finally, we fine-tune the underlying language model in our `BERT` architecture using a large scale, self-supervising pre-training process, initially proposed as a tool called `DINO` in a 2021 paper (Schick and Schütze [2021]). This tool allows for adversarial training of the model, by having GPT2 (Radford et al. [2019]), a generative language model capable of generating texts based on a prompt, generate training data.
In our set-up, `DINO` takes in a set of tweets from #auspol, and for each tweet, generates texts by providing GPT2 with the original tweet, and a prompt. Specifically, there are two prompts which can be provided: "write a sentence which means the same thing", and "write a sentence that is not at all similar in meaning". In the pre-training process, the task of the `BERT` model is to discriminate which prompt was used to generate a particular sentence.
Some example texts which were generated by the GPT2 model, and were provided to the `BERT` as training examples, include:

- `The Death Tax is an idea that has been out there for over 100 years. The idea was first proposed in a book by Thomas Sowell. He was not a conservative.`
- `I have a very good friend who is very, very good in maths. I'm very afraid of the Australian government's maths skills, which have been on the wane for the past 20 years.`
- `The Trump Administration is not a government of, by, or for the rich.`

Of particular interest is the first of these examples. Thomas Sowell is a real person, a prominent economist who has written about inheritance taxes before (Sowell [2003]). Of course, he was not the first person to think of taxing estates, but the model has produced a convincing tweet using a figure who is involved in discussions around estate taxation. Many of the tweets generated by the model are convincing enough that they could plausibly exist.
The adversarial task is not the final task which we wish for our model to perform, but it is a task which requires the model to develop an understanding of the language in use in Australian political twitter. This pre-training is especially useful for our research, since it allows for easy generation of over a million classification problems to train on, with ground truth labels that do not require an expert in Australian

politics to validate. This is in contrast to the disinformation detection problem we seek to solve in this project, where reliable ground truth labels are only available for a small subset of the data, and those labels require an expert in Australian politics to validate.

3.3 Results

In order to asses the performance of each technique, we assess each technique on its performance at predicting whether users are spreading disinformation on *#deathtax*, *#stkilda*, and the random dataset, as described in Section 3.1. In all experimental runs, the relevant model is trained on *#deathtax*, and tested on all three of these datasets. The results of these tests can be seen in Figure 3.1. This figure reports the performance of each model on each dataset, utilising the F1 score to do so. This score combines the precision and recall of the model, and is utilised because it is highly accurate in imbalanced class problems, such as this one.
We observe that in all cases, the models perform best on the test data from the scandal they were trained on, consistently reporting an F1 score of approximately 0.8 for *#deathtax*. The `Word2Vec` and standard `BERT` models perform very similarly, with scores on *#stkilda* being around 0.25 and score on the random dataset being roughly 0.14 for each. We also see that pre-training the `BERT` model using the `DINO` process, as described in Section 3.2.3.1, gives the best relative performance when we compare the result on *#stkilda* (0.412) to the baseline score on the random dataset (0.14).
We also observe a significant change in the score assigned to the random baseline between `TFIDF` and the remainder of the techniques, as well as some small fluctuations in the score for the other three techniques. It is likely that the score is significantly lower when using `TFIDF` because this technique produces data with significantly more noise than the other methods, which we would expect to cause a naieve classifier (as any classifier working on a random dataset will neccesarily be) to perform worse than the other methods, which all use 100-dimensional vectors. Similarly, the small fluctutations in the scores associated with the random baseline in the other three techniques (`Word2Vec`, `BERT`, and `BERT` with `DINO`) are likely due to small changes in the specific random tweets that are selected to be the random dataset across many experimental runs.

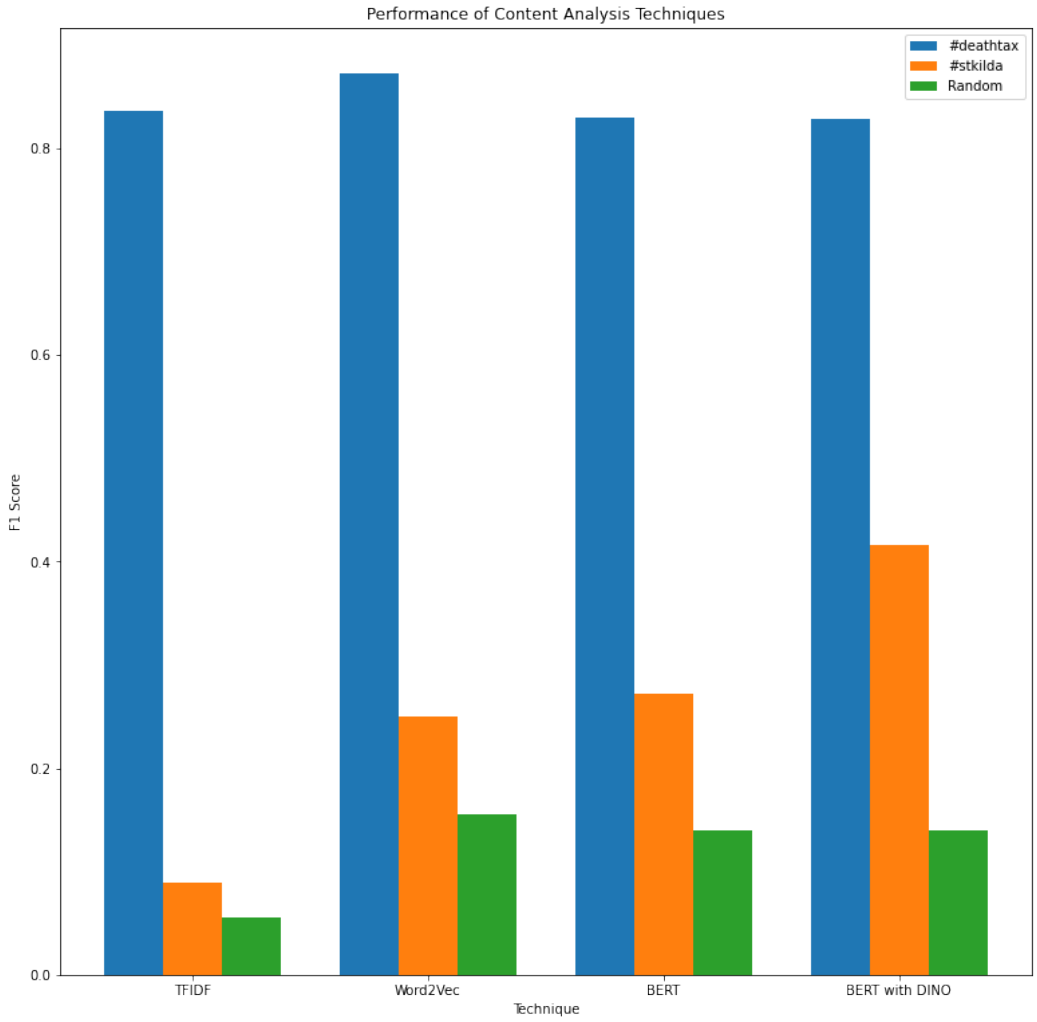

Figure 3.1: F1 Scores Obtained From Various Content Analyses of Australian Political Data

Chapter 4

From #COALition to #Eachwayalbo: Tracing Inflammatory Hashtags

Hashtags are a phenomenon unique to social media which serve to identify the topic of discussions on social media. These can give us an insight into the intended target audience of a user's message, as well as what communities they wished to post it in. This chapter treats the hashtags from a user independently of the content of their tweets, allowing us to follow disinformation spreaders as they seek to enter the discussion at different times.
Specifically, we are interested in creating groups of hashtags, attempting to automatically create clusters of hashtags that are similar to each other. We seek to group hashtags together that are used by similar users. Essentially, we're looking for small collections of hashtags that are near-exclusively used by a small group of users. Since this means that only people of one political viewpoint are willing to use these hashtags, this would indicate that these hashtags are highly polarising.
We begin by investigating hashtag and entity clusters produced for *#deathtax*, comparing the clusters produced to the ground truth labels obtained in Section 2.1. We then proceed to consider clustering hashtags on the *#qanda* dataset, using labels on a small number of initial episodes to identify similarly polarising hashtags in later episodes. In doing so, we provide a promising avenue for creating labels to evaluate content-based disinformation detection on the unlabelled *#qanda* dataset.

4.1 Hashtag and Entity Clustering on #deathtax

4.1.1 Methodology

Before we can use a clustering algorithm on the hashtags in use in *#deathtax*, we need to define what it means for two hashtags to be similar to each other. We say that two hashtags are similar if they are used by similar users. In order to capture this mathematically, we create a user space for each dataset by computing, for each hashtag, a vector with length equal to the number of users in the dataset. Each entry into this vector will be equal to the number of times that the corresponding user uses that hashtag. Then, the hashtag space matrix is created with each of these vectors

as a row of the matrix, meaning that the columns of the matrix correspond to users, and the rows of the matrix correspond to hashtags. As such, if two rows are equal in this matrix, that would imply that those two hashtags were used exactly the same number of times by exactly the same users. We are then able to define the distance between two hashtags as the cosine distance between their corresponding vectors in the hashtag space matrix. Having done this, we have a distance metric, and may utilise standard clustering algorithms.
We also compute a distance metric based on the transpose of this matrix for clustering users based on the hashtags they use. When clustering this way, the distance between two users is defined as the cosine distance between their hashtag vectors. We would expect clustering on this metric to provide clusters of users who are utilising similar hashtags, so who might share similar ideologies or be members of similar communities. Similarly to the user space, we refer to the matrix containing all of these vectors as the hashtag space.
We use the KMedoids clustering algorithm provided by the SciKit Learn library (Pedregosa et al. [2011]) to compute the clusters based on each of these metrics. Since we are seeking to match the resulting clusters against the ground truth labels, which can either be a 0 or a 1, we produce 2 clusters for each metric.

4.1.2 Results

The results of this clustering may be seen in Figure 4.1 and Figure 4.2. These figures represent the results of using the t-SNE dimension reduction tool (Maaten and Hinton [2008]) on hashtag vectors in Figure 4.1 and user vectors in Figure 4.2. In each figure, the dots (representing a single hashtag or user respectively) are coloured according to the cluster that is assigned to them by the KMedoids algorithm. This allows us to see from the figures the separation of clusters. Ideally, all of the orange dots would be in one part of the figure, and all of the blue dots in another part, as this would indicate that there are two distinct communities being identified. We observe that there is some separation between the orange and blue clusters in both figures, but in order to determine the quality of the clusters, we need a quantitative metric. For this, we compute a pairwise confusion matrix for the clustering shown in Figure 4.2; validation of hashtag clustering in the user space is discussed in Section 4.3.
The pairwise confusion matrix takes all pair combinations of users in the dataset, and for each pair, compares the clusters assigned by the KMedoids algorithm to the ground truth labels. A pair is considered to be correct in two cases: where both users have the same label in both the KMedoids cluster output and the ground truth, or where both users have different labels in both the KMedoids cluster output and the ground truth. For example, if user A has KMedoids cluster 0 and is labelled a disinformation spreader, and there is another user B, then there are two ways this pair can be considered correct. Either user B also has KMedoids cluster 0 and is labelled a disinformation spreader, or user B has KMedoids cluster 1 and is labelled as

a normal user. Intuitively, this method describes how similar the KMedoids clusters are to the ground truth in a way which doesn't make any assumptions about which KMedoids cluster corresponds to which ground truth label.
Similarly to Chapter 3, we compute the F1 score for this confusion matrix, which yields a result of *0.579*. As discussed in Section 2.3, with 36.3% of users being disinformation spreaders, we would expect a random baseline to produce an F1 score of 0.363. As such, this clustering method significantly outperforms a random baseline.

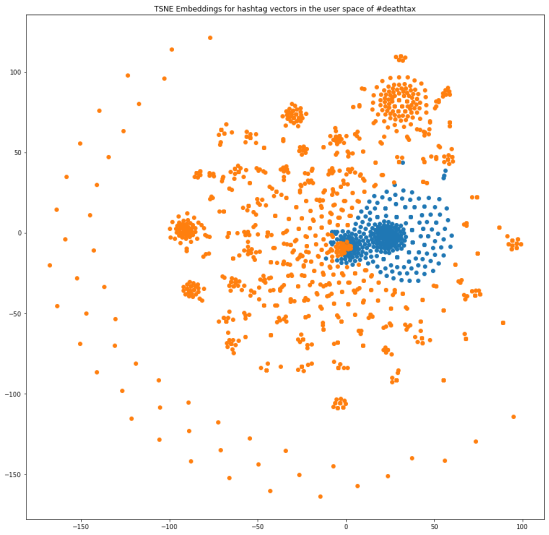

Figure 4.1: TSNE visualisation of Hashtags in the User space of *#deathtax*

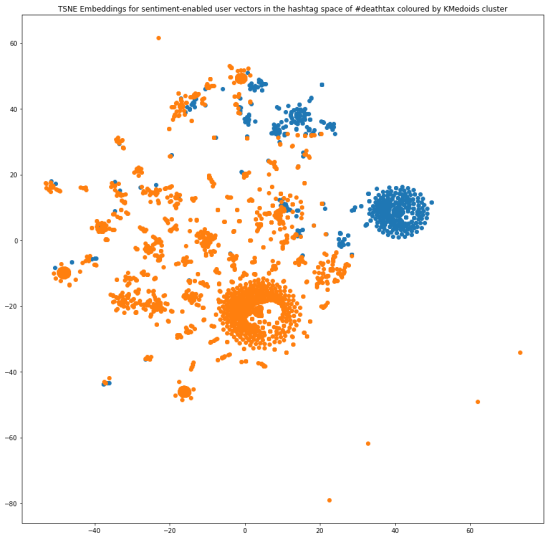

Figure 4.2: TSNE visualisation of Users in the Hashtag space of *#deathtax*

4.2 Hashtag and Entity Clustering on #Qanda

4.2.1 Methodology

As discussed in Section 2.5, the network based methods included in prior work are not appropriate to the #qanda dataset because of the large number of interactions with the host and panellists of the show, which make it difficult to detect communities of users. As such, we require an alternative method of generating labels for this dataset. Having observed that hashtag analysis can produce sensible clusters for differentiating between regular users and disinformation spreaders, we seek to generate these clusters for *#qanda*.
In pursuit of this goal, we extend the work on hashtags in the user space. We choose to analyse hashtags in the user space rather than users in the hashtag space since it is easier for the research team to assess whether a hashtag is likely to be polarising quickly. Users may tweet many times, or in ways that seem contradictory, but since hashtags are designed to categorise tweets into particular categories, they are far less likely to be as noisy. Furthermore, users who are attempting to spread disinformation seek to mask their intentions using more benign hashtags. This means that clustering users directly is unlikely to capture the problematic hashtags which are being used, since they may be masking those with many more benign ones. As such, we modify our approach to focus primarily on identifying the problematic hashtags, such that this masking will not cause problems in our analysis.
Additionally, unlike *#deathtax*, *#qanda* discusses a large variety of political topics. Seeking to simply cluster hashtags into two clusters will not produce sensible results, since we would expect to see a disinformation cluster and a normal user cluster associated with each independent topic discussed on the show. As such, we move away from the KMedoids clustering algorithm, and instead utilise the DBScan algorithm (Pedregosa et al. [2011]). This algorithm utilises the same distance metric described above, but rather than computing a pre-defined number of clusters, the algorithm determines an appropriate number of clusters. This is done by setting a parameter for how close together data points need to be in order to be part of the same cluster. This means that points which are far from all other points in the graph are not assigned to any cluster, and that many clusters are generated in cases where there are lots of small groupings of data points, as is the case in this application.

4.2.2 Results

Table 4.1 gives some examples of hashtags that are clustered together. Each row in the table represents a cluster of hashtags that are in the same cluster. We can see that these hashtags are quite polarising, with each row being able to be assigned as supporting the LNP or ALP. This is significant, because it demonstrates that a number of the hashtag clusters in the dataset can be manually identified as being polarising.

Table 4.1: Polarising Hashtags Clusters in #qanda

Supports	Hashtags				
ALP	ClimateEmergency	StopAdani	ExtinctionRebellion	MurdochMedia	ClimateSolutions
LNP	Pmlive	Eachwayalbo	CoalForBrains	Albo	LaborHypocrites
LNP	THEIRABC	Climatecult	ClimateHoax	Gofundyourselves	notmyABC
ALP	COALition	Fossilfools	LakeMacquarie	LakeMac	RupertMurdoch
ALP	LNPCorruption	IPAcoalition	FarRightCoalition	FeatheringTheirCaps	ImpeachScomo

4.3 Hashtag Label Propogation on #Qanda Episodes

We are now able to identify clusters that relate to polarised left or right wing activity on social media, but the DBScan algorithm on the entire #qanda dataset produces hundreds of clusters. Manually labelling each of these would be prohibitively time-consuming, and not a scalable solution for future disinformation campaigns, which does not fit with our research question. As such, we explore one final aspect of detecting clusters of hashtags that are highly polarising, which is how the hashtags in the #qanda dataset interact with each other over time. Specifically, we seek to manually label the tweets attached to the first episode of the show in 2020, and identify cases where the hashtags which are labelled as being polarising to the right or left appear again in discussion around later episodes. Since these hashtags are highly polarising, we assume that if the hashtags appear again in clusters formed from data from later episodes, then the other hashtags in those clusters must also be polarising.

4.3.1 Methodology

To achieve this, we again utilise the DBScan clustering algorithm, and manually annotate each cluster for the first episode in the #qanda dataset. We then create a pool of hashtags that are considered left-wing polarised or right-wing polarised based on our annotations.
Then, we utilise the DBScan clustering algorithm to create clusters for subsequent episodes. For each of these clusters on a subsequent episode, if the cluster contains a hashtag that was identified as polarised on the initial episode, then that cluster is also identified as polarised in the same direction as the initial hashtag.
While we cannot empirically validate the correctness of the clustering obtained due to the lack of a ground truth, we can examine examples of hashtags that are automatically identified as being polarised in order to determine if this identification is sensible. Some examples which were identified as being polarised left wing from episode 2 which did not appear in episode 1 are as follows:

- ActNow
- ClinareCriminals

- GobalHeating
- GrantGate
- mateGate
- BeingLiberalMeans
- Skyfuqwits
- MAFS
- Poppy
- FIVERR

This largely appears to have worked, with hashtags very similar in meaning but not included in the initial labelling of polarised left-wing hashtags due to spelling errors (e.g "ClinareCriminals" and "GobalHeating") being included alongside additional anti-LNP slogans not present in episode 1 (e.g "BeingLiberalMeans"). However, we also see some irrelevant hashtags being caught up, such as "MAFS", "Poppy", and "FIVERR". For more discussion on the likely cause of this issue, see Section 4.3.2
On a larger scale, we also analyse all of the clusters produced by using the polarising hashtags from episode 1 to label the clusters produced on two other randomly selected episodes, specifically episodes 3 and 18. Doing this resulted in 21 clusters receiving a label as either polarised left wing or polarised right wing. 7 of these were classified as polarised right-wing, of which 4 appeared to be correctly labelled according to the experimenter. This was determined based on whether the identified hashtags corresponded to known right-wing political campaigns, and whether the clusters included hashtags that were not politically charged. In cases where there was ambiguity as to the meaning of a hashtag, we also examined the users who utilised the hashtags for further information. For instance, in one case we utilised the following user description to identify that the hashtags they were using were likely to be polarised right-wing:

```
WARNING !
THIS IS GOING TO HURT A LOT.
AUSTRALIAN CONSERVATIVE.
IF I'M NOT REPLYING TO YOUR TWEETS IT'S BECAUSE I DON'T COMMUNICATE WITH
PEDOPHILES LIKE YOU.
```

Of the remaining 14 clusters classified as polarised left-wing, 13 appeared to be correctly labelled.

4.3.2 Methodological Issues: Hashtag Stuffing

The work undertaken in this section assumes that where users include a hashtag in their tweet, that is because they identify with the message embedded in that hashtag. However, this is not always the case. On twitter, including popular or trending

hashtags in your tweets makes it more likely that tweet will be shown to other users. This creates an incentive for people who wish to receive more publicity for their tweets to simply include as many popular hashtags as they can find in their tweets, irrespective of whether they agree with those hashtags. Take this tweet for example, which makes use of the hashtag #DefundtheBBC, which is typically considered a right-leaning hashtag:

```
"#DefundtheBBC #StormDennisUK #PresidentsDay #qanda #4corners #StKevins
#SurvivorAU #MAFS #DisChem3SIXTY5
Do you want to create a professional Clickable email signature in your
Outlook, Gmail?
contact me or order me
email signature"
```

It is quite clear that this user, and similar users who are advertising products, or attempting to spread other messages, may not neccesarily believe in the hashtags they are using. This could be a source of experimental error in the hashtag propogation in use, and goes some way to explaining the cases where hashtags are included in the list of polarising ones when they are clearly apolitical (for example #MAFS). Since these tweets are characterised by using a large number of unrelated hashtags, solutions to this issue could include removing tweets with large numbers of hashtags or hashtags that are infrequently used together from the analysis.

4.4 Conclusion

We find that it is possible to generate clusters of hashtags in the user space for unlabelled datasets which correspond to highly polarised communities. We also find that we can utilise a small amount of labelling of these hashtags to automatically obtain labels for the remainder of the hashtags in the dataset. These automatically assigned labels appear to be accurate, especially for polarising left-wing hashtags, but are currently hampered by users including hashtags in their tweets where they do not neccesarily agree with the message of the hashtag.

Chapter 5

Identifying the Social Graphs of Disinformation Spreaders: Network Features

This section builds heavily on prior work around interaction networks, seeking to supplement this work with further information gained from our study of content and hashtags. We introduce the idea of a weighted social graph, before examining how the weights on an edge between two users are calculated. Then, we utilise this method to attempt to reconstruct our ground truth labels on *#deathtax*, and consider how it might be improved to make it more suitable for use on *#qanda*.

5.1 Compound Interaction-Hashtag Similarity Metrics

Having discovered that hashtags are often polarising, and can be used to assist in identifying polarised communities of users, we seek to incorporate this information into our interaction metrics. Network methods for detecting communities of users treat the social network as a graph, with each node on the graph representing a user. In prior work focussed on the interactions between users, if two users have interacted, then there is an edge between them, and if not, then there is no edge (Tripathi [2021]). We seek to add to this by conceiving of the social network as a weighted graph. Then, we incorporate two pieces of information into determining the weighting of an edge between two users; their interactions, and how similar their hashtag usage is.

5.1.1 Defining the Weights

We first seek to define a metric to weight graph edges based on the interactions between users. This is done by first looking at every pair of users, and observing the maximum number of interactions between any two users. Then, an interaction score out of 1 is computed as follows, where $s_{i_{(a,b)}}$ is the interaction score between users a and b, $i_{(a,b)}$ is the number of interactions between users a and b, and $i_{\max}$ is the

maximum number of interactions between any two users:

$$s_{i_{(a,b)}} = \frac{i_{(a,b)}}{i_{\max}}$$

For defining the component of the edge weight between two users that is derived from the similarity of their hashtags, we simply take the cosine similarity between the hashtag vectors for each of the users. Calling this $s_{h_{(a,b)}}$, we can then define a weight for each edge between all users in the graph. Calling that weight $w_{(a,b)}$, it is defined as follows:

$$w_{(a,b)} = s_{i_{(a,b)}} + s_{h_{(a,b)}}$$

This produces a similarity score out of 2 for each pair of users in the dataset.

5.1.2 Reconstructing Clusters on *#deathtax*

In order to verify whether the weighted method is suitable for use on *#qanda*, we first attempt to reconstruct the network clusters observed in the prior work. To do this, we follow the method used in (Tripathi [2021]). We import the weighted edges into Gephi, and then use the provided network modularity algorithm (Blondel et al. [2008]) to obtain modularity labels for each user.
Having obtained labels for each user, we use the pairwise confusion matrix method described in Section 4.1.2 to assess how well this clustering works. We find that doing so, we obtain an F1 score of 0.535.

5.2 Evaluation of Performance

This score is, similarly to the case in Section 4.1.2, significantly better than a random baseline, so it performs adequately in determining whether a user is a member of a disinformation spreading community. However, given that this clustering tool has the same information available to it that was used to obtain the ground truth labels, it is strange that it would perform so much more poorly than, for example, the content analysis. This could be for a number of reasons.
Firstly, it is possible that there is a meaningful difference created by utilising the number of interactions that two users have, rather than just the fact that they had an interaction at all. There remains an open question as to which of these approaches would be more likely to give accurate results, which could be further analysed in future work.
Secondly, the weightings may be poorly balanced, or thrown off by outliers in the dataset. The interaction score is taken relative to the pair of users in the dataset that have interacted the most times, meaning that if there is one pair of users that has interacted far more than anyone else, this would effectively mean that interactions are no longer considered in the edge weighting. This is because all of the more normal interaction pairs would have comparatively extremely small interaction scores.

Chapter 6

Conclusion

6.1 Summary

In our project, we have constructed a number of models and approaches for detecting disinformation. We have successfully implemented a variety of content-based models for tracking the content produced by disinformation spreaders, with all of these performing better than random on completely unseen data. We also show that Australian political data contains a number of polarising hashtags, and that these hashtags propogate - highly polarised users tend to use similar hashtags when discussing many topics. This insight allows us to automatically label new hashtags, providing an initial avenue for labelling the *#qanda* dataset. Finally, we propose a method of addressing the issues with interaction-based networks, although more work is required to achieve the full potential of this method as an avenue for detecting disinformation.

6.2 Future Work

This project has been largely exploratory, and there is scope for future work in turning the insights we have obtained about disinformation in the Australian political context into functioning algorithms that can be used in realistic settings. We may seek to turn the highest performing content analysis model into a real-time learning system. Having shown that the DINO pre-training procedure works well at training the BERT model to detect disinformation, it stands to reason that providing more real-world data for pre-training, and allowing the BERT to see more and more data in a real-time set-up would further improve our results.
Additionally, we may seek to further incorporate combinations of techniques. We show that hashtag analysis and content analysis are both capable of generalising to new political contexts; we believe that a combination of these methods would likely yield even better identifications of users spreading disinformation into the future.
Finally, we can improve the compound similarity described in Chapter 5. We may be able to incorporate information about when users' tweets are semantically similar to each other, as defined by the pre-trained BERT model, to do a better job of identifying users that are forming communities of disinformation.
In these ways, we will be able to further improve on this work, and bring it closer to

a system that can be feasibly implemented in the real world, in real-time, to monitor disinformation as it occurs during future Australian elections.

Bibliography

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 10 (Oct 2008), P10008. doi:10.1088/1742-5468/2008/10/p10008. http://dx.doi.org/10.1088/1742-5468/2008/10/P10008. (cited on pages 5 and 28)

Breiman, L., 2001. Random Forests. *Machine Learning*, 45 (2001), 5–32. (cited on pages 12 and 14)

Cameron, S. and McAllister, I., 2019. Trends in australian political opinion. results from the australian electoral study 1987-2019. Technical report, ANU School of Politics & International Relations. (cited on page 1)

Dawson, N.; Rizoiu, M.-A.; Johnston, B.; and Williams, M. A., 2019. Adaptively selecting occupations to detect skill shortages from online job ads. In *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 1637–1643. IEEE, Los Angeles, CA, USA. doi:10.1109/BigData47090.2019.9005967. http://arxiv.org/abs/1911.02302https://ieeexplore.ieee.org/document/9005967/.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805. (cited on page 15)

Henriques-Gomes, L., 2018. Channel seven accused of fear-mongering over 'african gangs' story. Available at https://www.theguardian.com/media/2018/jul/09/channel-sevens-african-gangs-beat-up-prompts-fear-among-african-australians. (cited on pages 1 and 7)

Islam, M. S.; Sarkar, T.; Khan, S. H.; Kamal, A.-H. M.; Hasan, S. M.; Kabir, A.; Yeasmin, D.; Islam, M. A.; Chowdhury, K. I. A.; Anwar, K. S.; Chughtai, A. A.; and Seale, H., 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103 (2020), 1621–1629. (cited on page 1)

Issa, F.; Monticolo, D.; Gabriel, A.; and Mihăiţă, A., 2014. An intelligent system based on natural language processing to support the brain purge in the creativity process. *IAENG International Conference on Artificial Intelligence and Applications (ICAIA'14) Hong Kong*, (Mar. 2014).

Kong, Q.; Rizoiu, M.-A.; Wu, S.; and Xie, L., 2018. Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos. In *The Web Conference 2018*

- *Companion of the World Wide Web Conference, WWW 2018*, 175–178. ACM Press, Lyon, France. doi:10.1145/3184558.3186972. https://arxiv.org/abs/1801.04117http://dl.acm.org/citation.cfm?doid=3184558.3186972.

Kong, Q.; Rizoiu, M. A.; and Xie, L., 2020. Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *International Conference on Information and Knowledge Management, Proceedings*, 645–654. ACM, New York, NY, USA. doi:10.1145/3340531.3411861. https://arxiv.org/pdf/2001.11132.pdfhttps://dl.acm.org/doi/10.1145/3340531.3411861.

Maaten, L. v. d. and Hinton, G., 2008. Visualizing data using t-sne. *JMLR*, (2008). (cited on page 20)

Mao, T.; Mihaita, A.; and Cai, C., 2019. Traffic signal control optimisation under severe incident conditions using genetic algorithm. *Proc. of ITS World Congress (ITSWC 2019), Singapore*, (Oct. 2019).

Marineau, S., 2020. Fact check US: What is the impact of russian interference in the US presidential election? Available at https://theconversation.com/fact-check-us-what-is-the-impact-of-russian-interference-in-the-us-presidential-election-146711. (cited on page 1)

Mihaita, A.; LI, H.; and Rizoiu, M., 2020. Traffic congestion anomaly detection and prediction using deep learning. doi:arXiv:2006.13215.

Mihaita, A. S.; Benavides, M.; Camargo, C.; and Cai, C., 2019a. Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models. *International Journal of Intelligent Transportation System Research (IJITSR)*, 17, 2 (2019), 125–141. doi:DOI:10.1007/s13177-018-0160-z. https://link.springer.com/article/10.1007/s13177-018-0160-z.

Mihaita, A. S.; Dupont, L.; Cherry, O.; Camargo, M.; and Cai, C., 2018. Air quality monitoring using stationary versus mobile sensing units: a case study from lorraine, france. *Proc. of ITS World Congress (ITSWC 2018), Copenhagen, Denmark*, (Sep. 2018).

Mihaita, A.-S.; Li, H.; He, Z.; and Rizoiu, M.-A., 2019b. Motorway Traffic Flow Prediction using Advanced Deep Learning. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 1683–1690. IEEE, Auckland, New Zealand. doi:10.1109/ITSC.2019.8916852. https://ieeexplore.ieee.org/document/8916852/.

Mihaita, A.-S.; Liu, Z.; Cai, C.; and Rizoiu, M.-A., 2019c. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. In *Proceedings of the 26th ITS World Congress*, 1–12. Singapore. http://arxiv.org/abs/1905.12254.

Mihăiţă, A.; Camargo, M.; and Lhoste, P., 2014. Evaluating the impact of the traffic reconfiguration of a complex urban intersection. *10th International Conference on*

Modelling, Optimization and Simulation (MOSIM 2014), Nancy, France, 5-7 November 2014, (Nov. 2014).

Mihăiţă, A. S.; Tyler, P.; Menon, A.; Wen, T.; Ou, Y.; Cai, C.; and Chen, F., 2017. An investigation of positioning accuracy transmitted by connected heavy vehicles using dsrc. *Transportation Research Board - 96th Annual Meeting, Washington, D.C.*, (Jan. 2017).

Mihăită, S. and Mocanu, S., 2011. An energy model for event-based control of a switched integrator. *IFAC Proceedings Volumes*, 44, 1 (2011), 2413–2418. doi:https://doi.org/10.3182/20110828-6-IT-1002.02082. https://www.sciencedirect.com/science/article/pii/S1474667016439741. 18th IFAC World Congress.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv*, (2013). (cited on page 12)

Mishra, S.; Rizoiu, M.-A.; and Xie, L., 2018. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, 1–10. Stanford, CA, USA. https://arxiv.org/pdf/1804.02101.pdf.

Monticolo, D. and Mihăiţă, A., 2014. A multi agent system to manage ideas during collaborative creativity workshops. *International Journal of Future Computer and Communication (IJFCC)*, 3, 1 (Feb. 2014), 66–70. doi:10.7763/IJFCC.2014.V3.269.

Murphy, K.; Knaus, C.; and Evershed, N., 2019. 'it felt like a big tide': how the death tax lie infected australia's election campaign. Available at https://www.theguardian.com/australia-news/2019/jun/08/it-felt-like-a-big-tide-how-the-death-tax-lie-infected-australias-election-campaign. (cited on page 7)

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12 (2011), 2825–2830. (cited on pages 13, 20, and 22)

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I., 2019. Language models are unsupervised multitask learners. (2019). (cited on page 16)

Rizoiu, M. A. and Velcin, J., 2011. Topic extraction for ontology learning. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* (Eds. W. Wong; W. Liu; and M. Bennamoun), 38–60. IGI Global. ISBN 9781609606251. doi:10.4018/978-1-60960-625-1.ch003. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-625-1.ch003.

Rizoiu, M.-A. and Xie, L., 2017. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *International*

AAAI Conference on Web and Social Media (ICWSM '17), 182–191. Montréal, Québec, Canada. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553https://arxiv.org/pdf/1703.01012.pdf.

Rizoiu, M. A.; Xie, L.; Caetano, T.; and Cebrian, M., 2016. Evolution of privacy loss in wikipedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, 215–224. ACM, ACM Press, New York, New York, USA. doi:10.1145/2835776.2835798. http://dl.acm.org/citation.cfm?doid=2835776.2835798http://arxiv.org/abs/1512.03523http://dx.doi.org/10.1145/2835776.2835798.

Roth, Y. and Pickles, N., 2021. Updating our approach to misleading information. Available at https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information. (cited on page 1)

Schick, T. and Schütze, H., 2021. Generating datasets with pretrained language models. *CoRR*, abs/2104.07540 (2021). https://arxiv.org/abs/2104.07540. (cited on page 16)

Shafiei, S.; Mihaita, A.; Nguyen, H.; Bentley, C. D. B.; and Cai, C., 2020. Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation. In *Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.* doi:http://hdl.handle.net/10453/138721.

Shafiei, S.; Mihăiţă, A.-S.; Nguyen, H.; and Cai, C., 2022. Integrating data-driven and simulation models to predict traffic state affected by road incidents. *Transportation Letters*, 14, 6 (2022), 629–639. doi:10.1080/19427867.2021.1916284. https://doi.org/10.1080/19427867.2021.1916284.

Sowell, T., 2003. On estate-tax issues of merit and money. *Sun*, (2003). (cited on page 16)

Tripathi, K., 2021. Profiling information warfare on social media: a forensic analysis of the 2019 australian elections. Honours Thesis. (cited on pages 2, 5, 6, 27, and 28)

Twitter, 2020. Coronavirus: Staying safe and informed on twitter. Available at https://blog.twitter.com/en_us/topics/company/2020/covid-19#misleadinginformationupdate. (cited on page 1)

Unwin, J. T.; Routledge, I.; Flaxman, S.; Rizoiu, M. A.; Lai, S.; Cohen, J.; Weiss, D. J.; Mishra, S.; and Bhatt, S., 2021. Using hawkes processes to model imported and local malaria cases in near-elimination settings. *PLoS Computational Biology*, 17, 4 (apr 2021), e1008830. doi:10.1371/JOURNAL.PCBI.1008830. http://medrxiv.org/content/early/2020/07/17/2020.07.17.20156174.abstracthttps://dx.plos.org/10.1371/journal.pcbi.1008830.

Wen, T.; Mihăiţă, A.-S.; Nguyen, H.; Cai, C.; and Chen, F., 2018. Integrated incident decision-support using traffic simulation and data-driven models. *Transportation*

Research Record, 2672, 42 (2018), 247–256. doi:10.1177/0361198118782270. https://doi.org/10.1177/0361198118782270.

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2019. Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW (nov 2019), 1–25. doi:10.1145/3359285. http://dl.acm.org/citation.cfm?doid=3371885.3359285.

Wu, S.; Rizoiu, M. A.; and Xie, L., 2020. Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, 715–725. https://arxiv.org/abs/2003.09557.

Zhang, R.; Walder, C.; and Rizoiu, M.-A., 2020. Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 04 (apr 2020), 6803–6810. doi:10.1609/aaai.v34i04.6160. http://arxiv.org/abs/1905.10496https://aaai.org/ojs/index.php/AAAI/article/view/6160.