Ethical Detection of Online Influence Campaigns Using Transformer Language Models

by

Evan Crothers

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
MCS degree in Master of Computer Science,
Concentration in Applied AI

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

The past five years have seen the rapid escalation of online influence campaigns: coordinated attempts to covertly exploit social media platforms to undermine democratic elections and manipulate public opinion. These campaigns threaten the electoral process of democratic countries, erode confidence in integrity of online social spaces, and undermine trust in mainstream news media.

The detection of online influence campaigns (OIC) is a formidable problem, with significant active development within the field of applied artificial intelligence. Models based on the "Transformer" architecture — a specific type of neural network architecture amenable to transferring the capability of large pre-trained language models to novel domains — are a promising instrument for counteracting these campaigns. The focus of this thesis is the intelligent application of such deep learning techniques under real-world conditions for the improved detection of online influence campaigns, while remaining mindful of the ethical implications of automated systems that impact public political expression.

This thesis contributes new methodologies for reducing algorithmic bias in supervised detection of online influence campaigns, as well as a novel unsupervised process for improving OIC detection. In the case of supervised approaches, where labelled text from past influence campaigns is used for detecting new campaigns, we present a method for reducing algorithmic bias through careful additional preprocessing and evaluation procedures. In the case of unsupervised approaches, which operate in the absence of labelled data from prior campaigns, algorithmic bias is mitigated through the incorporation of a human analyst to provide additional oversight.

The supervised detection approach presented in this thesis includes an assessment of the potential for discrimination against non-native English speakers that may result from Transformer-based classifiers when applied to OIC detection in online communities. The findings indicate that while Transformer features derived from the text of user comments can be leveraged to identify suspect activity, this approach can lead to the emergence of algorithmic bias targeting non-native English grammar and keywords over-represented in past influence campaigns. Drawing on research in native language identification (NLI), "named entity masking" (NEM) is demonstrated to create sentence features robust to this shortcoming, while maintaining comparable classification accuracy.

The novel unsupervised process incorporates the creation of a user representation, created through the averaging of multiple Transformer output embeddings for user-provided

submission titles. With dimensionality reduction via Uniform Manifold Approximation and Projection (UMAP), this user representation can be visualized as a projection that a human analyst can use to identify similar posting patterns among active users within a community. By incorporating ethical oversight by trained human operators, this approach results in a practical system that can be used effectively to facilitate analysis of social media communities, while providing a higher ethical standard than a fully automated solution. The usefulness of this solution is demonstrated quantitatively by leveraging past ground-truth data to perform an extrinsic cluster quality analysis on the projection, and a qualitative analysis is performed focused on accounts that have faced disciplinary action from the host social media platform since the analysis took place.

Together, the research and methodologies presented in this work represents substantial improvement to the rigour of contemporary supervised and unsupervised OIC detection systems, and represent a promising future direction for ethical and effective detection techniques.

Acknowledgements

My sincerest gratitude to my thesis supervisors Dr. Herna Viktor and Dr. Nathalie Japkowicz. Working with the two of you has been an immense privilege, and the guidance and insight you've given me along the way have been invaluable. Your knowledge and thoughtfulness are an inspiration to me. Without the two of you, I would not be writing this today. From the bottom of my heart, thank you both.

I have great appreciation and admiration for the talented statisticians, programmers, and analysts I've had the pleasure of collaborating with over the course of this research — particularly John Healy, for your excellent knowledge and guidance on HDBScan and UMAP (and for being a tremendous person with which to talk data science); and Mark Franey, for your keen insights into the principled application of deep learning, and assistance with exploratory data analysis on the 2019 Reddit data.

I would also like to acknowledge the many hard-working people across the Canadian public service who are tasked with protecting democratic institutions, which is tireless work frequently performed under great levels of pressure and scrutiny. Thank you for working hard to look after this wonderful country we live in.

This research owes no small favour to the generosity of the Google TensorFlow Research Cloud (TFRC) team, who provided access to some frankly spectacular high-end hardware to accelerate this research, all while being consistently responsive and receptive. Completing this research required training hundreds of models, and the TPU platform on Google Cloud was invaluable in replicating the results to a high level of confidence. My sincere thanks to those who worked with me personally, and to those behind the scenes making TFRC happen.

And finally, thank you to my wife, Annie, for your depthless patience and encouragement. You are first my heart, always.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Technical Terminology

AI	Artificial Intelligence
BERT	Bi-directional Encoder Representations from Transformers
CH	Calinski-Harabasz
CSS	Cascading Style Sheets
CLS	Classification token (in BERT)
CV	Cross-validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EOM	Excess of Mass
FNE	Frequent named-entity
FN	False Negative
FP	False Positive
GPT2	Generative Pre-trained Transformer 2
GAAC	Group-average Agglomerative Clustering
GLUE	General Language Understanding Evaluation
GMM	Gaussian Mixture of Models
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HTML	Hypertext Markup Language
IP	Internet Protocol
KNN	k-Nearest Neighbours
L1	Language 1 (the first language of an individual)
L2	Language 2 (a second language of an individual)
NLI	Native Language Identification
NLP	Natural Language Processing
NLTK	Natural Language Toolkit

NE	Named-entity
NEM	Named-entity Masking
NEMM	Named-entity Masked Model
NER	Named-entity Recognition
NN	Neural Network
OS	Operating System
PCA	Principal Component Analysis
POS	Parts-of-speech
SEP	Sentence separation token (in BERT)
SVD	Singular Value Decomposition
TF	TensorFlow
TFCO	TensorFlow Constrained Optimization
TPU	Tensor Processing Unit
TFRC	TensorFlow Research Cloud
t-SNE	T-distributed Stochastic Neighbor Embedding
TN	True Negative
TP	True Positive
UMAP	Uniform Manifold Approximation and Projection
URL	Universal Resource Locator
VPN	Virtual Private Network

Other Terminology

AI HLEG	High-Level Expert Group on Artificial Intelligence
BTC	Bitcoin
CBC	Canadian Broadcasting Corporation
CEO	Chief Executive Officer
CIB	Coordinated Inauthentic Behaviour
DFRLab	Digital Forensics Research Lab

EU	European Union
G7	Group of Seven
GAC	Global Affairs Canada
IRA	Internet Research Agency
OIC	Online Influence Campaign
RRM	Rapid Response Mechanism
SITE	Security and Intelligence Threats to Elections
UK	The United Kingdom
US, USA	The United States of America
UTC	Coordinated Universal Time

Chapter 1

Introduction

1.1 Motivation

The recent surge in media coverage of nation-state efforts to influence public perception in online communities [BBC News, 2018][Alba and Satariano, 2019][Coldewey, 2018][Jones, 2018][O'Sullivan, 2019], as well as releases of large datasets by major social media platforms such as Facebook [Schrage and Ginsberg, 2018], Reddit [Reddit, 2018b], and Twitter [Twitter, 2019], have led to a notable increase in public-facing research into detection of online influence campaigns.

Online influence campaigns (OIC) are coordinated operations designed to influence a population by manipulating social media, typically for political or economic advantage. Attempts by nation-states to influence discourse on popular social media platforms in the lead-up to elections have raised concerns about the integrity of democratic institutions, contributing to the G7 launch of an initiative called the "G7 Rapid Response Mechanism" (RRM) in June 2018 to coordinate defenses against threats to democratic institutions [Government of Canada, 2019]. Defending against online influence campaigns is a key element of a broader strategy of safeguarding the democratic process.

For the machine learning community, there are numerous sub-problems within OIC detection that have been the subject of considerable research. Many of them heavily involve the application of advanced natural language processing (NLP) techniques towards the detection of disingenuous online activity, which is the area focused on within this thesis. Research in NLP-based OIC detection has been multi-faceted, including attempts to train high-accuracy text classifiers [Punturo, 2019][Zellers et al., 2019], create effective

user representations [Andrews and Bishop, 2019], and build software tools that can be used to assist a human reviewer [Gehrmann et al., 2019]. Research in all of these areas has leveraged the Transformer architecture: a specific type of neural network architecture amenable to transferring the capability of large pre-trained models to novel domains.

1.2 Thesis Statement

The development of novel neural network architectures that leverage attention mechanisms has had a significant impact on the usage and availability of deep neural network models for natural language processing tasks. These models, which are typically variations on the Transformer architecture introduced in "Attention Is All You Need" [Vaswani et al., 2017] [Devlin et al., 2018], allow widely-available pre-trained models to be easily applied to specialized domains. Transformer models offer promising new capabilities [Punturo, 2019][Weller and Woo, 2019] for counteracting online influence campaigns (OICs) — campaigns which pose a significant threat to the online social landscape and real democratic institutions. The focus of this thesis is on the application of Transformer models for the improved detection of online influence campaigns on active social media communities, while upholding ethical standards in artificial intelligence. In contrast to past work, rather than focusing on obtaining high classification accuracy on small selectively sampled datasets [Punturo, 2019][Weller and Woo, 2019], we focus on quantitatively evaluating algorithmic bias in Transformer classification models, and developing an approach that can be deployed during ongoing social media events — two areas vital to the development and deployment of OIC detection systems in realistic settings, where ethical considerations regarding online expression must be carefully respected, and detection mechanisms cannot enjoy the benefit of pre-labelled data with temporal and topical similarity to the current influence campaign. This work introduces two systems that can improve OIC detection while encouraging ethical application, in both supervised and unsupervised settings. The first introduces techniques for mitigating and evaluating discrimination against minority language communities in OIC text classifiers, while the second introduces a system where machine learning tools provide an analytical interface for a trained human analyst to incorporate human oversight. All research is performed using publicly available corpora from popular social media website Reddit, with the work in Chapter 3 working with comment text and the work in Chapter 4 relying on user-provided submission titles.

1.3 Objectives

The central objective of this thesis is to provide practical guidance on the design and application of systems that leverage state-of-the-art NLP Transformer models for the detection of online influence campaigns. It is the culmination of several research projects exploring this topic, and presents not only promising approaches and techniques, but also highlights pitfalls and ethical considerations to practitioners working in the field of OIC detection. This research should be of use to academia, government, and social media platforms, who are tasked with safeguarding online social spaces and protecting democratic institutions.

This research is narrowly focused on how these models can be applied practically to the domain of OIC detection. This analysis briefly touches a broad spectrum of other research, including native language identification (NLI) [Brooke and Hirst, 2012][Malmasi et al., 2017][Rabinovich et al., 2018], social media user representations [Andrews and Bishop, 2019], constrained optimization [Cotter et al., 2018b], and linguistic forensics [Perkins, 2018]. While informed by work within these fields, this thesis is by no means a complete exploration of these topics, and the reader is encouraged to explore the references for deeper insight into areas of further interest.

The core considerations of this thesis — the evaluation of neural NLP classifiers for discrimination against non-native English speakers, and the unreliable application of past OIC sample data to future detection efforts — will likely remain relevant beyond the current crop of state-of-the-art models. The obstacles in the successful application of deep language models to detection of online influence campaigns are due to characteristics of the available training data and the capabilities of high-capacity neural NLP models, of which Transformer is just the vanguard. The solutions presented herein should similarly serve as a foundation for future efforts to include ethical rigour in the design of classifiers, and creativity in design of tools and methodologies to improve detection of influence operations.

1.4 Thesis Outline

The content of this thesis is organized as follows.

Chapter 2 provides background information required for understanding the contents of this work, beginning with a detailed look at online influence campaigns, framing the problem in more precise terms and explaining in detail the characteristics of a successful

solution. This is followed by a brief overview of Transformer models, how these models can be tailored for new tasks, and what types of semantic and syntactic information are encoded in Transformer representations.

Chapter 3 introduces the first major work of the thesis, the application of a Transformer-based classifier to real-world social media data for the purposes of detecting online influence campaigns. This section includes a discussion of a named-entity masking technique that dramatically reduces algorithmic bias, while minimizing any deleterious effect on classification accuracy.

Chapter 4 introduces an alternative method of addressing the problems encountered in Chapter 3: an unsupervised approach that incorporates a human analyst. The advantages of this approach in applied settings are discussed. Analysis begins with an investigation of the suitability of clustering contextual embeddings of specific keywords, before moving on to clustering sentence representations. A new way to combine sentence representations to create a characteristic embedding of a user is introduced, and combined with dimensionality reduction to create visualizations to augment human analysis. Demonstrations of these visualizations are provided, as well as the early findings obtained from its development and deployment.

Finally, Chapter 5 summarizes the contributions of this thesis and looks ahead at the future of online influence campaigns and corresponding detection methodologies.

Chapter 2

Background

In this chapter we introduce the background knowledge required to understand Transformer-based detection models for online influence campaigns (OICs). This includes a high-level overview of the problem domain: introducing online influence campaigns through the lens of public disclosures from social media companies and research by data journalists. We follow this with a technical explanation of the Transformer model, and the attention mechanism on which it is built. Finally, we outline how deep neural models, such as Transformer, can be applied to NLP tasks on social media (including OIC detection) via transfer learning [Farzindar and Inkpen, 2020].

2.1 Online Influence Campaigns

2.1.1 Overview

The topic of influence campaigns has been featured heavily in news media [BBC News, 2018][Alba and Satariano, 2019][Coldewey, 2018][Jones, 2018][O'Sullivan, 2019], as well as informal conversation on social media platforms. While there is a broad awareness of the overall issue, divergent terminology and abundant misinformation contribute to confusion when discussing OIC detection methodologies. As such, it is important to clarify what we mean by "online influence campaign", as well as address several misconceptions that surround this topic. We define an online influence campaign as a coordinated campaign operated with the intent of influencing Internet users through disingenuous means. The key element of this definition is the notion of the influence being "disingenuous" — a

definition that excludes authentic attempts by an individual to persuade others to their viewpoint, or transparently-run political advertising campaigns. A list of definitions and common nomenclature is discussed in section 2.1.3.

For the purpose of this research, we are furthermore focusing primarily on "political online influence campaigns", which are influence campaigns specifically aimed at achieving specific political outcomes. These are in contrast to "commercial influence campaigns" that focus on producing a material benefit for the operating entity (or the entity sponsoring the operating entity).

When speaking on this topic, in light of the coverage of Russia-backed online influence campaigns operated by the Internet Research Agency (IRA) following the 2016 U.S. election, it is important to understand that influence campaigns are not operated by just a single country or entity, nor are they a partisan issue. Discussion of influence campaigns should not be politicized, otherwise there is a serious risk of research into the issue becoming a political act in itself. This complicates addressing ongoing influence campaigns, to the significant detriment of society as a whole. The central problem of influence campaigns is that a group is able to covertly gain undue influence over public opinion and democratic processes through disingenuous exploitation of the Internet. The extreme result of this would be a "voting bloc" of OIC operators having influence over a country's leadership via channels not subject to domestic electoral regulation — creating conflicting interests for aspiring political leaders and undermining the well-being of the electoral populace. The presence of an online influence campaign in support of a candidate should not be viewed as a defect of that candidate, as long as there is no attempt to coordinate such a campaign themselves and no attempt to prevent watchdogs from discovering and disrupting it.

2.1.2 Common Misconceptions

Political online influence campaigns might not exist

At this point, political online influence campaigns have been heavily reported in the mainstream media [BBC News, 2018][Alba and Satariano, 2019][Coldewey, 2018][Jones, 2018][O'Sullivan, 2019], and numerous publicly-available online datasets from social media websites have been published for review by open data journalists and the academic community [Reddit, 2018a][Twitter, 2019]. Regardless, perhaps due to unnecessary politicization of the topic, some dispute the existence of online influence campaigns altogether. This

presents an obstacle to ongoing research on this phenomenon. Central to any effort to understand the current social media landscape is to accept the reality of political influence campaigns, and focus on understanding the methodology used by different actors and how detection might be improved.

This is not to say that all reporting on influence campaigns is free from bias or inaccuracy, but rather that the phenomenon has been manifested beyond reasonable doubt, and discussion should move on from the understanding that there exists some coordinated groups that seek to manipulate social media for political benefit. Research that challenges the existence of a particular publicly disclosed influence campaign should focus on providing rigorous alternative explanations for the observed data.

All inauthentic behaviour on social media is backed by a government

Influence campaigns can be both political or commercial in nature, and can be backed by either governments or private entities. Inauthentic behaviour is widespread online, and the usage of disingenuous tactics is not sufficient to prove a political agenda. A motivated business owner may operate large numbers of automated accounts to promote their brand, or leave disingenuous reviews for their competitors, in an effort to gain a commercial advantage. Similarly, an individual with a strongly-held political belief (or unethical advertising body) might make multiple accounts to promote their views, or attempt to silence those who oppose them.

There is some blending between different campaign types, as accounts used for commercial influence campaigns may resurface as part of political influence campaigns, such as in the case of compromised accounts being sold, or the same organization being contracted to do multiple types of coordinated inauthentic behaviour on social media. A stark example of this may be found in the available Twitter data for the alleged influence campaign against Hong Kong protesters on Twitter [Twitter, 2019], in which enormous amounts of commercial spam predate the influence campaign by several years.

Influence campaigns and "disinformation" are the same

Research into online influence campaigns is regularly framed as research on "disinformation", but this terminology is inaccurate and limiting. It is very possible — and in some cases, easier — to manipulate a populace without providing information that is explicitly false or misleading. The fundamental problem of these campaigns is not that the

populace being deceived, but of coordinated inauthentic action having undue influence over the minds of the general public through clandestine channels. As such, we will only use the term "disinformation" when we are specifically referring to "false or misleading information, spread with the intention to deceive" [Jack, 2017][Nimmo, 2016].

All influence campaigns are run by "bots"

Informal conversation on influence campaigns often includes mention of "bots" as being the primary driver of influence campaigns, when heavy human involvement has historically been central to their operation. Currently, to the best knowledge of social media researchers working on OIC detection, influence campaigns continue to rely on teams of account operators working in offices with the assistance of anonymization technologies and relatively simple account automation utilities [Volchek and Coalson, 2018][Brooking et al., 2020]. While automation will likely continue to develop in this space, particularly as generative text models improve, it would be myopic to overlook the human element of influence campaigns — particularly when designing detection methodologies.

The human element of modern influence campaigns both creates detection opportunities in the form of human error, but also adds elements of randomness and resourcefulness that can be hard to detect systematically.

Online influence campaigns regularly include "deepfakes"

The term "deepfake" has been extended in common parlance to include a broad variety of generative models that can be applied to create media forgeries. While originally referring specifically to faceswap technology optimized for usage in imagery and video, the term also been used more broadly to refer to GAN-generated imagery [Tolosana et al., 2020], and speech-to-text models [Nguyen et al., 2019] that mimick an individual's speech patterns.

Despite valid concerns regarding deepfakes, it is worth noting that potent online disinformation campaigns using video to date have been accomplished without need for AI-based image forgeries. The ability to easily create image forgeries using AI, while a useful addition to the arsenal available to those seeking to create disinformation, may not be as large a threat as the development of better tools and infrastructure for the creation and operation of large numbers of fake accounts. Even without sophisticated tools and techniques, videos containing false information and spliced together footage of real-world leaders have already proven effective.

GAN models can certainly be used to generate fake profile photos, and this phenomenon has been documented in support of fraudulent online activity [Song, 2019]. These images can be generated and customized in milliseconds by generative models with easily-available open-source models such as StyleGAN [Baylies, 2020][Karras et al., 2019a][Karras et al., 2019b]. It is important to remember, however, that a malicious actor unconcerned with privacy laws or copyright also has an entire Internet full of genuine images of people that they may make basic modifications to and pass off as their own.

With these misconceptions resolved, we can now move on to discussion of the nomenclature of online influence campaigns.

2.1.3 Nomenclature of Online Influence Campaigns

Online influence campaigns are frequently discussed in mainstream political discussion, which has led to broad usage of certain terms in a variety of settings. It is therefore useful to specify exact definitions for many of these terms used in the context of online influence campaigns so that discussion can proceed with precision. The definitions below closely follow those utilized by the Digital Forensics Research Lab (DFRLab) [Brooking et al., 2020], whose lexicon is often useful when discussing online influence campaigns.

Within this thesis, we will consistently use the following definitions:

Online Influence Campaign: A coordinated campaign leveraged using the Internet to influence a group of people through manipulation of social media. This definition is largely synonymous with the associated term "Online Influence Operations".

Information Operations: The broader category of information-based operations, of which OICs are a subset. Officially defined by the U.S.A. Joint Chiefs of Staff in the military context as "the integrated employment of electronic warfare, computer network operations, psychological operations, military deception, and operations security, in concert with specified supporting and related capabilities, to influence, disrupt, corrupt or usurp adversarial human and automated decision making while protecting our own." [U.S. Joint Chiefs of Staff, 2014]

Disinformation Campaign: An online influence campaign specifically with the intent of propagating false or misleading information.

Coordinated Inauthentic Behavior: The Facebook terminology used with reference to online influence campaigns, formally defined as to "the use of multiple Facebook or

Instagram assets, working in concert to engage in inauthentic behaviour [according to the Facebook definition], where the use of fake accounts is central to the operation" [Facebook, 2020].

Bot: A social media account primarily operated automatically according to an algorithm. A bot has minimal direct operation by a human operator, with only basic adjustment to set up automation or test functionality.

Troll: "A person who posts deliberately erroneous or antagonistic messages to a newsgroup or similar forum with the intention of eliciting a hostile or corrective response." [Oxford English Dictionary, 2020]

Commercial Bot: A bot operated to promote the economic interest of a company or individual.

Political Bot: A bot operated with the primary intention of achieving a political goal in the interest of a state, party, candidate, or interest group.

Sockpuppet: "Inauthentic social media accounts used for the purpose of deception which evidence a high likelihood of human operation. This includes catfishing and other highly tailored operations conducted under inauthentic personas." [Brooking et al., 2020]

Disinformation: False or misleading information, specifically intended to deceive or mislead [Jack, 2017] [Nimmo, 2016]

Misinformation: Information that is **unintentionally** false or misleading [Jack, 2017] [Nimmo, 2016]

Foreign interference: Clandestine interference by a foreign state in the internal affairs of a nation (such as an election).

When characterizing different types of accounts, it is important to note that the operator of an account may not be consistent over time. For example, an account may originally be created by a regular user, be compromised via a password leak to become part of a commercial bot network, and then be purchased by a political entity for use as a sockpuppet in an influence campaign. This further complicate detection efforts, as accounts may manifest different behaviours over their lifetime.

An excellent resource for further reading on disinformation and online influence campaigns can be found in the "Disinformation Annotated Bibliography" by Gabrielle Lim [Lim, 2019], which includes a reading list along with descriptions of associated materials,

as well as the "Dichotomies of Disinformation" guide released by DFRLab in collaboration with Google-founded think-tank Jigsaw [Brooking et al., 2020].

2.1.4 Introduction to Reddit

Much of the research in this thesis is performed on the social media website "Reddit", both due to the high availability of diverse data, and the platform's relevance in the "information diet" of many people in North America and across the world. Reddit is a news aggregation and discussion website that has dramatically risen in popularity over the last several years. At the time of writing it is currently the fifth most popular website in Canada [Alexa Internet, 2020a] and sixth most popular website in the United States [Alexa Internet, 2020b] according to Alexa traffic rankings, placing it ahead of Twitter, Instagram, and Wikipedia. Reddit allows for comments on submitted posts, and allows users to reply and vote to the comments of other users. This results in a volume of discussion not found on other popular social media platforms. Comments on Reddit are limited to 10,000 characters, allowing for much more verbose discussion than is easily possible on Twitter, which officially doubled its maximum character length to 280 in November, 2017 [Rosen, 2017].

Reddit enables users to create communities around specific topics with few restrictions. Concern has been expressed by Reddit users and moderators within that the site may be the target of ongoing nation-state efforts to influence popular opinion in order to support political goals. Reddit CEO Steve Huffman addressed the community on April 10, 2018 during the website's 2017 Transparency Report to address these concerns and report the staff's findings [Reddit, 2018a]. This report included a release of 944 accounts "of suspected Russian Internet Research Agency origin". These accounts were preserved for the purposes of transparency, allowing users to scrape their comment histories for further analysis. The full export of all posts and comments made by these suspect accounts was performed by Alberto Coscia and is available on GitHub [Coscia, 2018].

These suspicious accounts — their account information, posts, and comments — serves as the primary ground-truth for detection efforts on this platform. While other independent researchers have collected other lists of accounts which exhibit some suspicious behavior [Russel, 2018], the official designation of these accounts represents a stronger confidence level not found in other sources.

This thesis is limited to open-source data that can be leveraged by independent re-

searchers without special access to internal company data. Open-source methods are valuable as they increase the number of individuals who can scrutinize activity in online spaces, exposing interference by coordinated groups, without requiring a privileged relationship with the platform holder. The most effective techniques in this problem space likely require access to the platform's internal data, for more detailed user statistics and network information.

In December 2019, following the experiments and analysis within this thesis, an additional official release from Reddit banned and preserved 61 further accounts [Reddit Security Team, 2019] that were suspected of being part of the same influence account network as the 2017 Reddit transparency report. These new accounts had been connected to a leak of official UK government documents on the platform. Given the very limited ground-truth for influence campaigns on this platform, these new accounts represent a valuable and recent addition to the available ground-truth, offering a more recent glimpse into clandestine activity on this platform. In the interests of furthering research in this area, we present a short Python script to collect and structure the contents of these accounts using the Pushshift API, and an archive of both the collection code and output in a GitHub repository [Crothers, 2020].

Combined with the aforementioned collection of accounts hosted by Alberto Coscia, these two samples form the body of publicly available disclosed influence accounts from Reddit administration to date. This data may be of use to future research in this field, as well as for replicating the results of the research presented within this thesis. Information on what fields are available within Reddit OIC datasets can be found in Appendix A.

2.1.5 Fairness and Bias

In order to frame the ethical discussion around OIC detection within this thesis, it is helpful to offer an introduction of the basic concepts and definitions that will be referred to when considering the ethical implications of OIC detection systems. When discussing OIC detection systems, we will focus on two related ethical ideas: that such systems should *promote fairness* and that they should *minimize bias*.

We define fairness in the context of OIC detection as follows: all authentic social media users should have an equal chance to be express themselves online and not face inconvenience, reduced privacy, or censorship based on factors unrelated to participation in an online influence campaign. For example, users of a particular language background should

not be disproportionately identified by such a system. We might utilize "fairness metrics" when attempting to characterize the behaviour of systems to measure quantitatively whether different categories of users are treated differently. In the context of OIC detection, a fairness metric of interes is the likelihood that an innocent individual from a particular population is erroneous flagged as being part of an online influence campaign. While fairness metrics can be useful in understanding the behaviour of systems, it is important to remember that fairness in machine learning requires a sociotechnical approach that incorporates human factors as well [Selbst et al., 2019]. Fairness is fundamentally a human-defined concept, and the human interpretation and operation of these systems is required for a full understanding of the fairness of an OIC detection system.

When referring to bias in the context of ethical system design, we are referring to an inclination of a system towards a particular result, specifically one that has a disproportionate impact on a particular population. As the word "bias" has its own definition in the field of neural networks, when discussing ethical concerns, we will specifically refer to "algorithmic bias". We define algorithmic bias as a systematic tendency that manifests in the execution of computer algorithms, including neural networks. Given that a neural network learns parameters based on training data, it is important to understand that this algorithmic bias may be influenced by other forms of bias during development of the system. This might include *data bias* (bias in the training or evaluation due to its nature, or how the data is collected) and *human bias* (bias in human decisions during system development that stem from a preconceived belief, conscious or unconscious).

Developing effective systems that are fair and minimize harmful biases is often not straightforward. The intent of this work is not to criticize parallel research in Transformer-based OIC detection – on the contrary, this research into OIC detection should be encouraged for seeking potential solutions to a pressing problem. The aim of this work is to promote careful consideration and evaluation before such systems are deployed against real-world individuals for the purpose of law enforcement of social media moderation.

2.2 The Transformer Architecture

The Transformer architecture, first introduced in the seminal paper "Attention is All You Need" [Vaswani et al., 2017], was a key development within the field of deep learning, particularly due to the applicability of the architecture for natural language processing (NLP). In contrast to contemporaneous NLP research, the Transformer architecture eliminated recurrence and convolutional layers from the network design entirely, instead opting to rely entirely on a mechanism called "multi-head attention" (hence the name of the paper). The paper's authors demonstrated that this approach could obtain state-of-the-art performance on a variety of language tasks, sparking a great deal of further innovation in the form of novel models based on the original Transformer architecture (pictured in Figure 2.1). In this section, we provide a brief overview of the Transformer architecture and its attention mechanism. A more detailed explanation is provided in the original paper [Vaswani et al., 2017], from which this discussion borrows heavily.

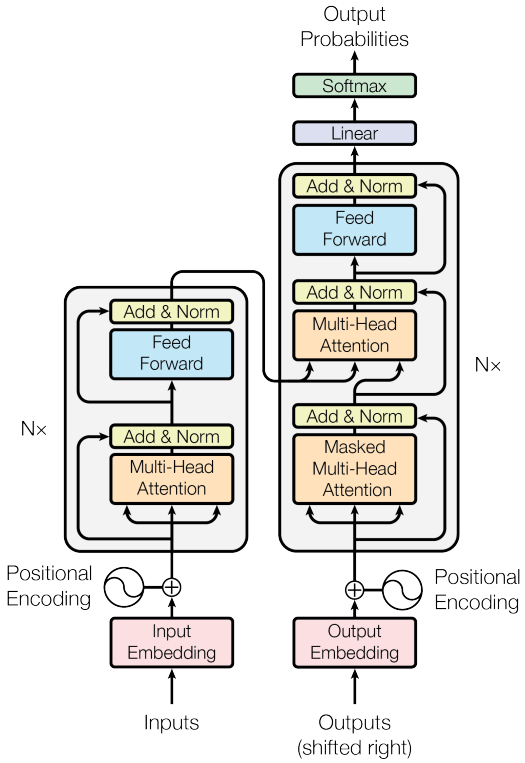

Figure 2.1: The Transformer architecture, as depicted in "Attention is All You Need" [Vaswani et al., 2017]

Transformer has its roots in sequence to sequence (seq2seq) transduction models, which use an encoder-decoder structure [Sutskever et al., 2014][Vaswani et al., 2017] to transform input sequences to output sequences. Transformer follows the same general structure of seq2seq, stacking self-attention and fully-connected layers for both an encoder and decoder [Vaswani et al., 2017]. The encoder is designed to map an input sequence — which in the context of language would be a sequence of input word IDs in some dictionary D — $(x_1, ..., x_n)$ to a sequence of continuous representations $\mathbf{z} = (z_1, ..., z_n)$. Given $\mathbf{z}$, the decoder generates an output sequence $(y_1, ..., y_m)$ — once again, in the context of language models, these would be output word IDs $y_m \in D$ [Vaswani et al., 2017].

In order to gain an intuition regarding Transformer, it is important to understand the self-attention functions within its encoder-decoder architecture. The input to an attention function is a query, q, and a set of key-value pairs k, v. The query and key-value pairs are combined into a single term θ, via some *compatibility function* f_c, computed such that $f_c(q, k) = \theta$.

The Transformer neural network uses "scaled dot-product attention" as its attention function, which defines $f_c(q, k) = \frac{q \cdot k}{\sqrt{d_k}}$, where d_k is the dimension of the keys. This is nearly identical to dot-product attention $f_c(q, k) = q \cdot k$, with the addition of a scaling factor of $\frac{1}{\sqrt{d_k}}$. This scaling factor mitigates vanishing gradients in the softmax function when d_k is sufficiently large. Within the Transformer model, the values of the attention function are computed simultaneously via matrix arithmetic, with q, k, and v placed into matrices Q, K, and V respectively. The resulting attention calculation is thus

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad \text{from [Vaswani et al., 2017]}$$

This attention is performed in parallel by multiple "attention heads", with the results linearly projected together, and the attention process repeated. This allows the Transformer model to simultaneously attend to information in different representation subspaces in a manner that is inhibited by using just a single attention head [Vaswani et al., 2017]. When written with learned parameter matrices, W_Q, W_K, W_V, W_O (query, key, value, and output weights respectively), and output dimension of the model d_{model} this process can be represented as

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W_O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_{Q_i}{}^{d_{model} \times d_q}, KW_{K_i}{}^{d_{model} \times d_k}, VW_{V_i}{}^{d_{model} \times d_v}) \quad (2.1)$$

adapted from [Vaswani et al., 2017]

Transformer models have strong application to the detection of influence campaigns, and much of prior documented influence campaigns to date have a significant textual element [Reddit, 2018b][Twitter, 2019][Hindman and Barash, 2018]. Even when influence campaigns provide images, these images often include text, which can be recognized using optical character recognition (OCR) [Smith, 2007] and processed using text-based models.

Multiple language-focused iterations of the Transformer architecture exist, and can be used to utilized to create token and sentence representations from input text. We will now briefly discuss the function of the specific Transformer model used extensively within this research: "Bi-directional Encoder Representations from Transformers", better known as BERT [Devlin et al., 2018].

2.2.1 Bi-directional Encoder Representations from Transformers (BERT)

The project which currently claims the highest accuracy on OIC comment classification [Weller and Woo, 2019], leverages vector representations of sentences created using deep NLP model BERT [Devlin et al., 2018]. BERT, which stands for "bidirectional encoder representations from transformers", leverages the Transformer encoder architecture and is pre-trained on two unsupervised language modelling tasks. The result is a network that allows for the creation of fixed-length vectors that contain both forward and backward contextual information for every token in the input text, which can be easily fine-tuned for downstream tasks.

BERT is utilized in the research within this thesis to demonstrate the relevance of the findings to current state-of-the-art NLP models, under generic settings. BERT the single most widely utilized and studied Transformer model at this point in time, with over 22,900 stars on GitHub [Google Research, 2019a] and over 5,400 citations recorded by Semantic Scholar at the time of writing [Devlin et al., 2018]. Many novel NLP models that compete for state-of-the-art accuracy under different settings are heavily influenced by the

BERT architecture, such as ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], and XLNet [Yang et al., 2019] — all of which specifically cite and reference BERT by name. Large pre-trained BERT models have been released for several other languages, including Chinese [Google Research, 2019a], French [Martin et al., 2019], and Arabic [Antoun et al., 2020], giving it broad applicability across different language domains. While specialized Transformer models derived from BERT continue to achieve record accuracy on different evaluation tasks, the broad availability, applicability, and foundational role of the BERT model makes it ideal for benchmarks that apply to Transformer models in general.

Whenever possible, we use the most basic and heavily-downloaded variant of BERT, the uncased base model (12-layer, 768-hidden, 12-heads, 110M parameters) [Google Research, 2019a]. The basic architecture is selected, as its smaller size encourages reproducibility, and it is typically the first to be available for other written languages [Antoun et al., 2020][Martin et al., 2019] or application libraries [Explosion, 2019]. While larger BERT architectures models may obtain higher accuracy on tasks, the relative trends in behaviour are expected to be consistent with the base model [Devlin et al., 2018], which is has led to the model being selected as the prime candidate when attempting to characterize BERT behaviour [Clark et al., 2019]. When assessing BERT-base alongside other BERT variants, such as BERT-large, research has found that behaviours such as the ordering in which information and coreferences are learned are equally observed in results from both models [van Aken et al., 2019].

2.2.2 BERT Architecture and Training

The BERT architecture implementation is a multi-layer bidirectional Transformer encoder architecture that is nearly identical to the original Transformer implementation [Vaswani et al., 2017]. The distinction that differentiates BERT is the selection of input/output representations, and a distinct unsupervised (i.e., without labelled training data) training process applied across a large NLP corpus. This allows pre-trained BERT models to be easily specialized to specific language tasks, which on release, allowed it to substantially outperform state of the art models on every General Language Understanding Evaluation (GLUE) evaluation task.

There are two main steps in the training of a BERT model for a particular task, *pre-training* and *fine-tuning*. Pre-training is done on a very large corpus, where the model is trained on two unsupervised language modellings tasks. These tasks are:

1. Predicting a masked word in a sentence

2. Predicting whether one sentence follows another

Training a bi-directional Transformer encoder architecture on these tasks creates a model that internally represents considerable information about the relationships between words and sentences in natural language. As a result, these pre-trained models can then be fine-tuned on downstream language tasks through transfer learning (which will be discussed further in §2.2.4).

Several variations of the BERT architecture exist, varying based on the number of layers, the size of the hidden dimension, and the number of parallel attention heads. For example, the uncased BERT base model uses a 12-layer, 768-hidden, 12-head, architecture (for a total of 110M parameters). Smaller, distilled, BERT models have been released as small as 2-layer, 128-hidden, 2-head, architecture (4M parameters) [Turc et al., 2019]. NVIDIA has trained their own BERT models as large as 48-layer, 2560-hidden, 40-head, architecture (3.9B parameters) [Shoeybi et al., 2019]. Each of these models still use the same basic Transformer encoder architecture as the original BERT implementation, albeit at very different scales.

2.2.3 BERT Sentence Embeddings

Sentence embedding refers to the family of techniques whereby sentences are mapped to vector representations within a continuous vector space, as in the case of word and phrase embedding [Bengio et al., 2003][Mikolov et al., 2013]. These "sentence embeddings" are useful for classification and clustering of sentences. Fixed-length sentence embeddings are particularly valuable as they convert variable length sentences into fixed-length feature vectors.

When using BERT, an additional token, referred to as `[CLS]`, is inserted at the start of the input sentence sequence, alongside WordPiece tokenized input [Google Research, 2019b]. The neural network's output representation of this token represents the full forward context of the sentence, and can be used as a fixed-length vector representation of variable-length sentences. This is the procedure used to obtain sentence representations in the original BERT paper [Devlin et al., 2018].

Several papers have focused heavily on understanding the representations created by BERT [Clark et al., 2019][Rogers et al., 2020][van Aken et al., 2019]. Analysis of BERT's

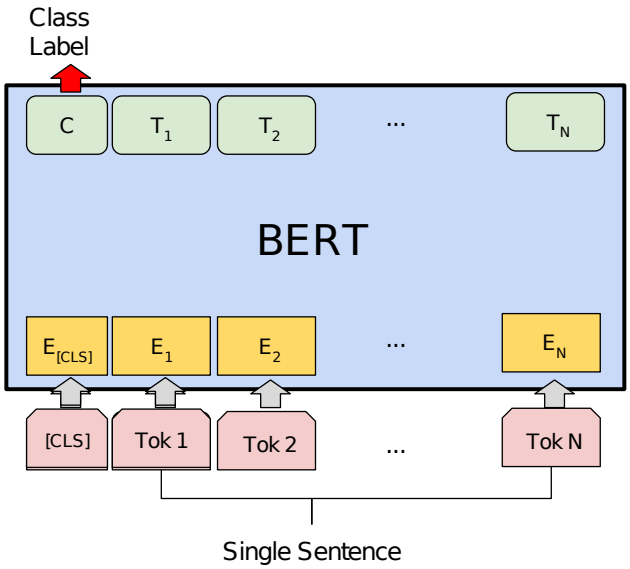

Figure 2.2: BERT single-sentence classification architecture, as depicted in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [Devlin et al., 2018]. An input sequence of token IDs *Tok* of length N is prepended with a special `[CLS]` token and mapped to a set of corresponding embeddings, E. These are fed through the Transformer encoder architecture of BERT, creating a corresponding vector for each input token. The output vector C for the preprended `[CLS]` token can be used as a sentence representation for downstream tasks, such as classifying the input sentence.

attention has found that BERT seems to possess a property described as "syntax-aware attention" [Clark et al., 2019]. Different attention heads can be identified as attending to different aspects of the sentence based on the syntax. For example, noun modifiers attend to their linked nouns, and direct objects attend to their linked verbs. This property is distributed across the attention heads in BERT, and seems to responsible for the impressive performance of models derived from this architecture. This phenomenon can be visualized using the BertViz tool [Vig, 2019], and several examples of specialized attention heads in $\text{BERT}_{\text{BASE}}$ are demonstrated in Figure 2.3.

The end-to-end process of using pre-trained BERT model for sentence embedding is as follows:

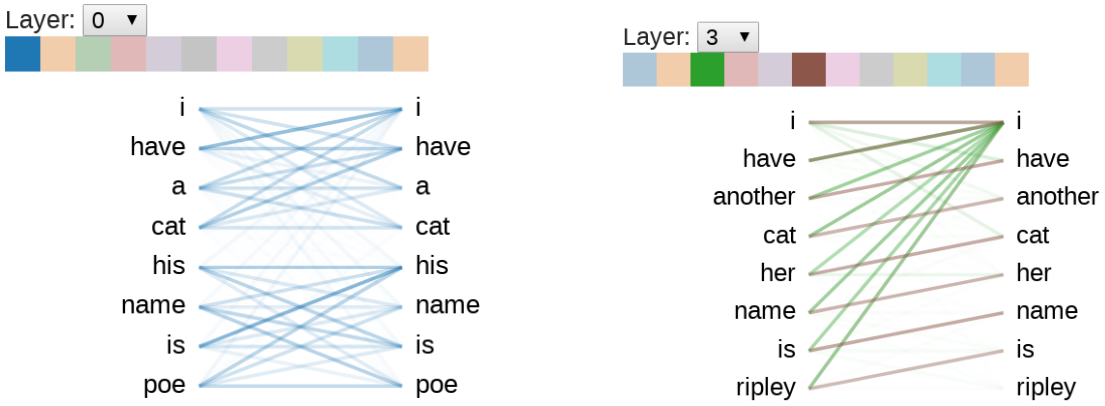

Figure 2.3: Attention mechanism specialization in BERT. Left: Attention head 0 of layer 0 primarily attends to other words in the same sentence. Right: Attention head 2 of layer 3 primarily attends to the first word of the sentence, while head 5 attends to the previous word.

Algorithm 2.2.1 BERT unpaired sequence embedding with prepended classification token

Require: A pre-trained BERT model (Transformer encoder) B with hidden dimension H

Require: Array A containing k input sequences to be embedded

Require: WordPiece dictionary D

$R \leftarrow [\,]$

for all $S \in A$ **do**

  $S_T \leftarrow$ WordPiece tokenization of S according to dictionary D

  $S_{ID} \leftarrow$ Word token IDs corresponding to S_T according to dictionary D

  $S_{input} \leftarrow$ [`[CLS]`, $s_0, s_1, .., s_n$, `[SEP]`], where s_i is the i^{th} token of S_{ID}, and `[CLS]`, `[SEP]` $\in D$ are special IDs for classification and separator tokens

  $E_{in} \leftarrow$ Input embeddings for S_{input}

  $W \leftarrow B(E_{in})$, where W is a $k \times H$ vector containing final layer neuron activations

  Append element $w_0 \in W$ to R, where w_0 the encoded representation for `[CLS]`

return R

2.2.4 Transfer Learning

An advantage of using pre-trained language models such as BERT for NLP tasks is the ease with which they can be tailored for downstream tasks through transfer learning.

Transfer learning refers to the practice of leveraging a model trained on one domain to improve performance in another domain. Common examples can be found in convolutional

neural networks in computer vision, where many of the neurons in a trained network are performing the tasks of edge-detection, understanding textures, or determining what elements are in the foreground — while the latter layers are performing higher-level tasks, such as object identification. As a result, a large trained classifier on a particular problem domain can be a very useful starting point for other models, simply by retraining the top few layers of a model. This is the principle that underpins the TensorFlow Object Detection API [Huang et al., 2016].

The BERT model was an important milestone in the application of transfer learning to the field of natural language processing. BERT was pre-trained on two unsupervised language modeling tasks: predicting a masked word in a sentence and predicting which sentences follow others. These two tasks, when applied over a very large corpus, result in a model that has a very good "understanding" of language. This is the reason that, at the time of its release, BERT was able to easily obtain state-of-the-art results on a broad range of language tasks [Devlin et al., 2018].

Formally, we define transfer learning as follows (this definition itself is abridged from an in-depth review of transfer learning [Pan and Yang, 2010]). To define transfer learning formally, we must also provide a formal definition of a "domain" and "task" in the context of machine learning systems.

*A **domain** $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, ..., x_n\} \in \mathcal{X}$* [Pan and Yang, 2010].

*A **task** $\mathcal{T}$, consists of two components: a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$, given a specific domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$. A task is denoted as $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$* [Pan and Yang, 2010].

Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$ [Pan and Yang, 2010].

With respect to BERT in this thesis, the source domain $\mathcal{D}_S$ is the large corpus of predominantly English text consisting of BooksCorpus [Zhu et al., 2015] and English Wikipedia that was used to train BERT, with the source task $\mathcal{T}_S$ being the unsupervised language modelling tasks of masked word prediction and next-sentence prediction [Devlin et al., 2018]. The knowledge obtained from $\mathcal{D}_S$ and $\mathcal{T}_S$, which is represented by the parameters of the BERT neural network, is applied to improve performance on target

tasks $\mathcal{T}_T$ of sentence classification (Chapter 3) and sentence embedding (Chapter 4), in the target domain $\mathcal{D}_T$ of online influence campaign detection on Reddit.

BERT is one variation of Transformer-based models, which can be applied to a diverse range of tasks. In the basic *fine-tuning* approach, an additional classification layer of neurons is added on top of the BERT network, and gradient updates are performed to specialize the entire network for the new domain. This has large memory requirements as pre-trained Transformer models may have hundreds of millions [Devlin et al., 2018] or billions [Radford et al., 2018][Shoeybi et al., 2019] of trainable parameters. As an alternative, sentences can be fed into the unaltered pre-trained BERT model, and the resulting vector representations of the sentence can be used to train an auxiliary model. This approach avoids having to load the entire BERT model into memory during training. It does, however, have the trade-off of requiring the intermediate sentence representations to be stored on disk, which can be quite large — recording the bottom four layers of neuron activations for even the basic uncased BERT model creates approximately 15KB of data per input token [Google Research, 2019a].

Within this thesis, we leverage transfer learning by feeding the output activations from a pre-trained BERT network as training input to a neural net classifier. This is equivalent to adding an additional classification layer to the top of the network and freezing the weights in all pre-trained layers [Devlin et al., 2018]. Breaking this into two steps has the result of substantially improving memory efficiency during training, and the two trained networks may be "re-assembled" to form a single end-to-end network that can transform input sequences into classification results for inference. The outputs of a pre-trained BERT network are a form of deep sentence representation, and may even be visualized to demonstrate what similarities in input sequences result in similar representations in the BERT embedding space (as we will demonstrate later in Chapter 4).

The approach of using output activations from a pre-trained BERT model is referred to as the *feature-based approach*. Ablation experiments on the feature-based approach present a slightly lower performance than fine-tuning the entire model, while providing other advantages, including memory efficiency and the creation of sentence vectors for downstream tasks [Devlin et al., 2018].

The capacity of pre-trained Transformer models to represent both syntactic and semantic detail, as well as amenability to transfer learning to new domains, makes them a promising technology for OIC detection.

2.3 Concluding Remarks

We have now introduced online influence campaigns (OICs), outlined what is known about their operation to date, and discussed the (often NLP-based) challenges specific to the problem of OIC detection. We have also discussed Transformer models, starting with an overview of the attention mechanism that underpins the Transformer architecture, and then focused specifically on BERT: a particular Transformer-based model with great applications for transfer learning in NLP, which makes it an excellent candidate for use in OIC detection systems. With this knowledge in mind, we now move on to the results of a series of experiments that characterize the behaviour of Transformer models within the OIC detection domain, revealing potential ethical pitfalls in supervised OIC detection strategies, and develop an effective unsupervised process for OIC detection — demonstrating its efficacy in the high-risk environment of a federal election.

Chapter 3

Reduction of Algorithmic Bias in Supervised OIC Detection

3.1 Introduction

The supervised approach, in which labelled training data are used to create a model, is a common formulation of the online influence campaign detection problem, typically evaluated as a user or comment classification task [Punturo, 2019][Weller and Woo, 2019]. In haste to develop countermeasures, there has been little research into characterizing the behaviour of OIC classification models. An understanding of the shortcomings of existing state-of-the-art methods is important to demonstrate the fairness and legitimacy of such models as a means of arbitrating online communities. These concerns are particularly important as the automated suppression of speech presents significant ethically considerations. As such, the development of robust techniques for evaluation and reduction of algorithmic bias must develop alongside new detection methods.

By analyzing classification results from Transformer-based OIC classifiers trained on corpora used in OIC detection projects, we demonstrate that features derived from the text of user comments are useful for identifying suspect activity, but lead to increased erroneous identifications ("false positive" classifications) when keywords over-represented in past influence campaigns are present. Drawing on research in native language identification (NLI), we use "named entity masking" (NEM) to create sentence features robust to this shortcoming, while maintaining comparable classification accuracy. We demonstrate that while NEM consistently reduces false positives (i.e., erroneous identifications of online influence

accounts) when key named entities are mentioned, both masked and unmasked models exhibit increased false positive rates on English sentences by Russian native speakers, raising ethical considerations that should be addressed in research. Much of the experiments and analysis within this section were detailed within the IEEE MLSP 2019 paper "Towards the Ethical Detection of Online Influence Campaigns" [Crothers et al., 2019], from which significant portions of this text are reproduced with permission.

In the context of this chapter, "L1" refers to an individual's first (native) language, while "L2" refers to an individual's second (non-native) language. Historically, the largest reported online influence campaigns targeting English users have been operated by countries with populations that do not typically speak L1 English [Twitter, 2019]. As a result of this, content-based natural language processing (NLP) models trained on text from past influence campaigns may inadvertently develop a significant bias towards detection of writing by L2 English speakers — particularly those who share an L1 language with the country to which the influence campaign has been attributed.

It is important to note that the classification problem studied in this chapter uses data that has been class balanced between influence accounts and regular users, while real influence campaigns typically operate in a setting of extreme class imbalance, where the number of influence accounts is heavily outnumbered by non-influence accounts. While studying classifiers trained on balanced data is sufficient for characterizing Transformer representations, detection methods "in the wild" will likely require further efforts to obtain sub-samples with less class imbalance, or the deployment of a detection approach better suited to predictive modeling in imbalanced domains [Branco et al., 2016].

For the purpose of this research, we classify comments individually rather than grouping comments by user and performing a user-level classification task. The intent of this is to demonstrate that Transformer feature representations, without any subsequent processing or manipulation, behave in a manner that raise ethical questions for their application to the domain of OIC detection. While user-level tasks are an important part of building practical detection systems (and as such, user representations will be discussed in Chapter 4), an analysis that also includes an ethical investigation into the impact of different user representation techniques is left to future work.

3.1.1 Ethical Focus

There is a distinction between authentic speech from those representing themselves online, and speech written under the direction of a government with the intent of manipulating a populace — particularly when it is designed to misrepresent the author or spread misinformation. The goal of detection systems should be to differentiate genuine expression from deliberate manipulation, focusing on signs that may indicate that online activity is disingenuous and directed.

It is not unexpected that a model trained to positively classify sentences written by Russia-operated influence accounts would demonstrate increased false positive rates (proportion of negative examples erroneously classified as positive) on English comments by L1 Russian speakers. Similarly, it is reasonable that the false positive rate increases if the sentence contains named entities frequently found within the training data, such as those referring to American politics or cryptocurrency. The combination of these tendencies, however, sets the groundwork for the automated suppression of speech — and in particular, political speech — by native Russian speakers. This represents a serious ethical consideration that should influence the decision to deploy any content-based influence campaign detection model. Influence campaigns operated by other countries with a high proportion of non-native English speakers will likely cause other language populations to face a similar risk.

User-submitted comments provide a significant variety of features that are useful for classification problems. Features derived from the textual content of the comment itself — or content-based features — have been shown to hold predictive power on a number of classification problems related to the writer of the text [Brocardo et al., 2014], including work specifically on detection of online influence campaigns [Weller and Woo, 2019]. It is unrealistic to expect that development in influence campaign detection should ignore a rich set of feature data. Development of content-based models is encouraged to continue with careful consideration to potential algorithmic bias against language communities at high-risk of false positive classification. Such mitigations may include negative training examples from L2 English language communities, or further attention to differentiating between genuine accounts and influence accounts with similar linguistic features.

3.2 Prior Art

Past work during the 2017 NLI Shared Task [Malmasi et al., 2017] has explored the current state-of-the-art in NLI, demonstrating successful combinations of semantic and syntactic features for differentiating language learners from native English speakers. However, this task did not reflect highly fluent advanced non-native speakers, which represent a much more challenging classification task. Further research identified that the language level of L2 English Reddit users posting in European communities was much more sophisticated than most English learners and approached the level of the majority-English Reddit community as a whole [Kyle and Crossley, 2014][Rabinovich et al., 2018]. Classification of sophisticated non-native English speakers on Reddit was the subject of a comprehensive analysis that included both comment content and metadata [Goldin et al., 2018].

NLI research has also contributed the concept of "topic bias" [Brooke and Hirst, 2012][Malmasi et al., 2017] as an undesirable property in NLI datasets. Topic bias may occur when the key themes and topics of texts are not evenly distributed across classes. Within online influence campaigns, there is a significant skew towards political topics within comments by influence accounts. As a result, positive detection may be heavily influenced by the presence of these topics. This creates weaknesses in the classification model as influence campaigns may not refer to the same topics as past influence campaigns, and presence of discussion of specific topics may cause a classifier to perform well on uniformly sampled sentence data, but poorly on data with similar topic content. Named entity masking (NEM) has been used as an effective means of reducing topic bias in past NLI work [Malmasi, 2016][Malmasi and Dras, 2014][Rabinovich et al., 2018], and should be applied to diminish topic bias within content-based influence campaign detection as well.

Prior work on detection of influence campaigns has mentioned L2 language features, such as differing stopword frequencies [Im et al., 2019]. Much of the more formal research on influence campaigns focuses around Twitter due to the substantial quantity of available OIC data [Twitter, 2019] and high media profile. Past work has demonstrated a holistic approach to troll detection designed to incorporate features intended to match propaganda agents as well [Fornacciari et al., 2018]. While Reddit has seen an enormous surge in popularity, little formal research has been performed so far on online influence campaigns on Reddit, with a handful of graduate research projects forming the current state-of-the-art for classification [Punturo, 2019][Weller and Woo, 2019].

There is some overlap between the work in this section and the field of forensic lin-

guistics, which has seen useful applications for NLI in cybercrime investigations [Perkins, 2018]. A forensic attribution of an influence campaign to a particular nation based on linguistic data is beyond the scope of this research, and would necessitate evaluating linguistic similarities between not just English and Russian, but other commonly used languages as well.

3.3 Methodology

This chapter focuses on analyzing the language characteristics of comments posted by accounts "of suspected Russian Internet Research Agency origin" released by link-aggregation and discussion website Reddit within their 2017 transparency report on April 10, 2018 [Reddit, 2018a]. Using natural language processing (NLP) model BERT [Devlin et al., 2018], contextual embeddings for sentences within these comments are generated, and a classifier trained to distinguish between sentences from Reddit accounts randomly sampled with equal probability from all users and those from suspected influence accounts. This classification methodology and training dataset is designed to be comparable to that used by the project that currently claims state-of-the-art classification performance on Reddit [Weller and Woo, 2019], with the distinction of being a sentence-level, rather than comment-level, task. This process is repeated with the same data after performing "named entity masking" (NEM) to replace named entities with their corresponding parts-of-speech (POS) tag.

The performance of this model is evaluated not only against a holdout test set of suspect sentences and random sentences, but also against two separate evaluation datasets based on the L1 language of the user – each described in §3.3.2. We create a dataset of sentences from comments by users who self-identify as being from L1 English countries, as well as a set of comments by users who self-identify as being from Russia. These datasets are constructed using similar methodology to recent work in native language identification [Goldin et al., 2018], and rely on the assumption that these self-identified affiliations are typically indicative of the native language of the author, which has been experimentally demonstrated to be correlated [Rabinovich et al., 2018]. This test is used to demonstrate the tendency of each model to generate more false positives when considering English comments written by users who speak Russian as a first language, as opposed to English native speakers. Also evaluated is a more demanding test set which filters these sentences to those that contain "frequent named entities" (FNE): the top ten named entities most frequently mentioned in suspect comments within the ground-truth data.

The purpose of the research in this section is to form a compelling case for the development of safeguards in the deployment of content-based moderation methods, particularly those that may target distinctive linguistic characteristics (e.g., L2 English) shared by users outside of the target group. Online influence campaign detection offers a useful example where this problem is evident. The results of this work offer some direction for the leveraging of content-based features for influence account detection, which may be integrated into downstream model for influence campaign detection on Reddit, similar to recent work in building holistic troll detection models on Twitter [Fornacciari et al., 2018]. Synthesizing these content features with past work on metadata-based approaches to influence campaign detection on Reddit [Punturo, 2019] may contribute to an ethically sound and effective approach.

3.3.1 Classification Architecture

Fixed-length sentence embeddings can be obtained from BERT by feeding WordPiece tokenized input sequences into the BERT model and reading the final layer activations for the prepended special classification token (`CLS`), as per the design of the model [Devlin et al., 2018] and as outlined in §2.2.3. This allows us to perform transfer learning, using the mechanism discussed in §2.2.4. A single layer classifier is trained on sentence embeddings from the BERT model, as is standard for fine-tuning BERT for sentence classification tasks. This straightforward configuration is used to improve repeatability, and avoid undue emphasis on the specifics of the neural network's construction. More complex multi-layer classifier architectures that directly leverage the BERT word embeddings have been shown to improve overall classification accuracy on this task [Weller and Woo, 2019], at the cost of increased storage overhead and computational complexity over the highly performant base model.

A maximum sequence length of 128 is chosen for the model. While BERT supports sequence lengths up to 512, a shorter sequence length is recommended by Google Research [Google Research, 2019a] as the relationship between Transformer attention and sequence length is quadratic [Devlin et al., 2018], leading to dramatic increases in computation time. A training batch size of 64 is used to maximize the efficiency of the TPU v3-8. We use the uncased variant of the BERT base model (12-layer, 768-hidden, 12-heads, 110M parameters) for our experiment. The selection of this variant — the most basic model with the most general application — reflects the ethos of using standard models

with high availability and ease of study discussed in §2.2.1. The hyperparameter settings of learning rate and training epochs ($2e-5$ and 3, respectively) reflecting the effective configuration recommended by the official BERT GitHub repository at the time of writing [Google Research, 2019a].

3.3.2 Experimental Setup and Corpora

All experiments were conducted on a n1-standard-2 (2 vCPUs, 7.5 GB memory) Google Cloud instance, with a Tensor Processing Unit (TPU) v3-8. The code for the experiment is available on GitHub [Crothers, 2019]. These experiments required the following three corpora:

Corpus I: Submissions and comments from 2017 Reddit transparency report

This corpus is comprised of Reddit submissions and comments made by accounts on a list released by Reddit staff on April 10, 2018 as "of suspected Russian Internet Research Agency origin" [Reddit, 2018a][Reddit, 2018b]. These accounts were preserved for the purposes of transparency, allowing users to scrape their comment histories for further analysis. A full export of all comments made by these suspect accounts was performed by Alberto Coscia and is available on GitHub [Coscia, 2018].

This corpus represents the largest official collection of Reddit accounts released to date as "suspicious" in the context of coordinated influence campaigns. While other independent researchers have collected other lists of accounts which exhibit some suspicious behavior [Russel, 2018], the official designation of these accounts represents a stronger confidence level not found in other sources. As such, for this analysis, we will only use the officially designated suspect accounts as ground-truth training examples.

As a platform holder, Reddit has access to additional data not publicly available. This includes access to IP logs, private actions (such as upvotes/downvotes), and more granular user activity tracking. These features, as well as Reddit's access and subject-matter expertise in their data, allows for this attribution to be considered accurate with a high degree of confidence.

Submission titles from this dataset will be used later in the research performed in Chapter 4.

Corpus II: Randomly sampled Reddit comments

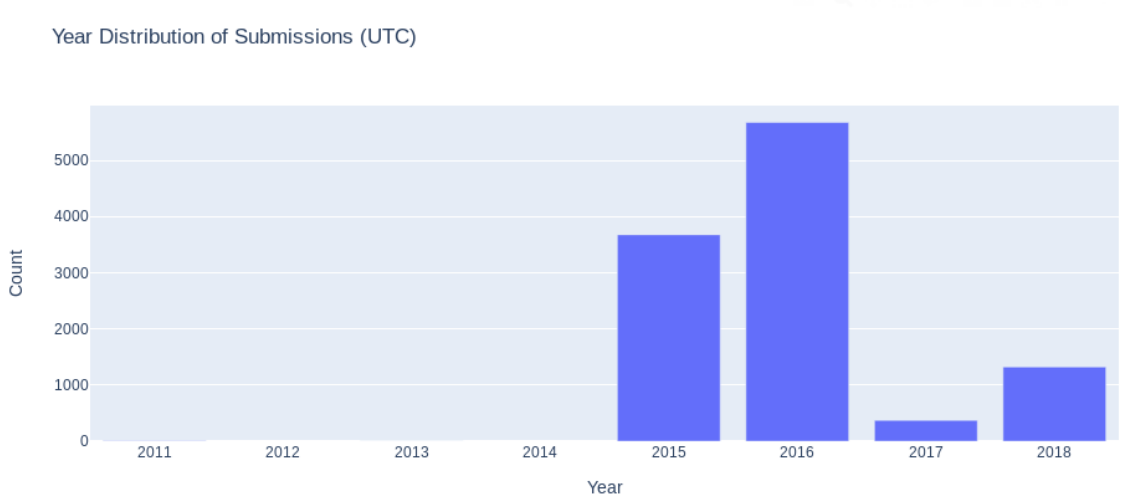

Figure 3.1: Breakdown of number of submissions by year in Corpus I. Activity in these accounts peaked in 2016, the year of the U.S. federal election

This corpus of random sampled comments was created by Brandon Punturo [Punturo, 2018], and has been used in two past online influence detection projects on Reddit [Punturo, 2019][Weller and Woo, 2019]. The corpus consists of Reddit accounts chosen with equal probability from all current Reddit users, and represents a typical random sampling approach for acquiring a negative class for influence campaign detection. We use this dataset to compare our results to past work in the field. It is important to note that more sophisticated sampling techniques may better address false positive similarities between influence accounts and genuine accounts.

For the purpose of this study, we assume that none of these randomly-sampled comments are attached to an influence campaign, based on the current understanding of the scale of past influence campaigns [Reddit, 2018b] and the volume of daily comments on Reddit [Baumgartner, 2019].

Corpus III: Augmented L2 Reddit dataset

Reddit has been the data source for past work on Native-Language Identification (NLI) on sophisticated second-language speakers [Goldin et al., 2018][Rabinovich et al., 2018]. This work entailed the creation of datasets of Reddit comments from users of a variety of different languages by looking for self-identified "flair" in European subreddits. This corpus includes a sizeable number of comments from self-identified Russian users: 31,167 from European subreddits and 586,398 comments from other subreddits [Goldin et al.,

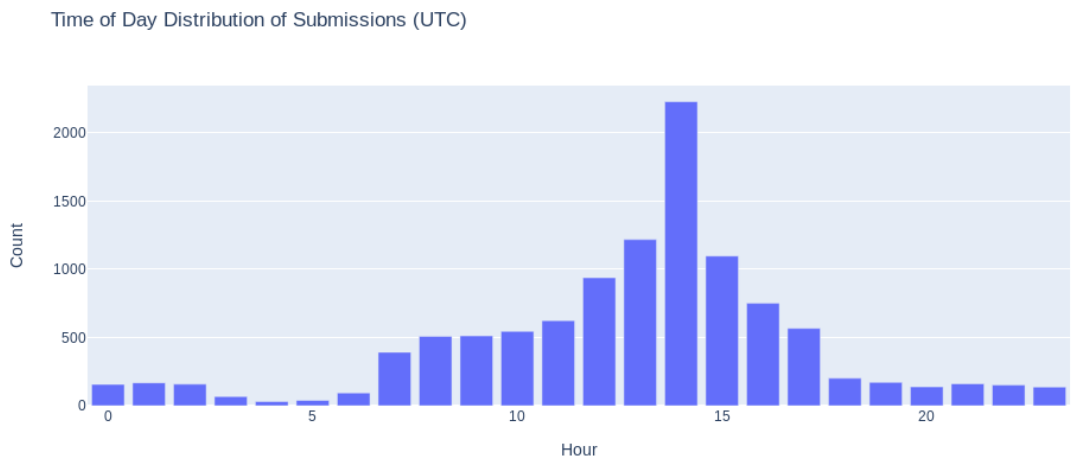

Figure 3.2: Breakdown of suspicious submissions by hour of day in Corpus I. This timing of submissions may indicate waking or working hours among groups that are predominantly co-located geographically.

2019].

We augment the Russian content of this corpus by leveraging the user-specified flair in the subreddit "AskARussian", which has 4,072 users at the time of writing. We collect the comments of users with self-identified flairs that indicate Russia or a specific Russian region. Deduplication is performed, and users who are already present in the original L2-Reddit dataset are discarded, as are self-declared bot accounts. The results in a total of 774,702 comments. When tokenized into sentences greater than 10 characters long, the result is 1.9 million sentences. We believe this is the most comprehensive corpus of online comments made by highly-fluent L1 Russian / L2 English language speakers.

The L2-Reddit dataset also includes comments by users who self-identify as being from countries that typically speak L1 English. We assess Australia, Ireland, New Zealand, the United Kingdom, and the United States of America as fulfilling this criterion. These L1 English comments are compared to the L1 Russian comments to determine the difference in classification accuracy.

Similar to Corpus II, we assume that none of these comments are affiliated with online influence campaigns, based on the current understanding of the scale of past influence campaigns [Reddit, 2018b] and the volume of daily comments on Reddit [Baumgartner, 2019].

3.4 Data Preprocessing and Masking Procedure

The first two corpora are considered the two distinct classes for the classification task, and are used together to train a classifier. Corpus III is used to evaluate false positive rates of this classifier against native English and Russian speakers.

As these corpora contain real comments in original formatting, data preprocessing is an important consideration. The comment data is cleaned according to a multi-step process.

1. Normalize text into standard format using regular expressions to remove extraneous escape characters.
2. Perform sentence tokenization using Python NLTK [Loper and Bird, 2002] to extract sentences from comments.
3. Remove newline characters, Reddit quote markdown characters, and horizontal tab characters (`&#009;`).
4. Remove all URLs and replace with [URL] token.
5. Discard sentences shorter than 10 characters. Very small sentences are poorly-suited to the classifier and may introduce noise.
6. Run full BERT tokenization pipeline [Google Research, 2019a], which includes converting to lowercase, WordPiece [Google Research, 2019b] tokenization, punctuation splitting, and invalid character removal.

The available data on influence campaigns contains frequent mentions of specific named entities. As a result, past classification work using BERT embeddings for classification on Reddit [Weller and Woo, 2019] has highlighted posts containing these words as a prominent failure case. In order to demonstrate the impact that these keywords have on BERT feature classification, we perform named-entity recognition (NER) to extract named entities (NEs) from sentences in the three corpora described in §3.3.2. These are used to form three additional datasets, with each named entity replaced by the corresponding tag. This approach, similar to that taken in native language identification research on Reddit [Rabinovich et al., 2018], emphasizes the other content features in the text, such as grammatical structure and word choice. To perform named entity masking, we use the largest (and most accurate) implementation available in the "spaCy" Python package, which supports

Entity	Entity Type	Count
US	GPE	79
TIE	ORG	79
Trump	ORG	67
Bitcoin	ORG	52
Hillary	PERSON	39
America	GPE	38
Russia	GPE	37
Russian	NORP	31
ISIS	ORG	29
BTC	ORG	28

Table 3.1: Most common named entities within Corpus I comments, used to generate additional "frequent named entity" (FNE) evaluation dataset. © 2019 IEEE

detection of a broad range of entities at accuracy comparable to state-of-the-art [Explosion, 2019][Kiperwasser and Goldberg, 2016].

Table 3.1 shows the ten most frequent named entities in Corpus I comments, omitting less-distinctive results for "DATE", "CARDINAL", and "PERCENT" entities. We have also omitted one named "PERSON" entity: "`:D`". While the frequent presence of this emoji within the suspicious comments may be a distinguishing feature, it does not meet the criteria of a valid named entity for this analysis. Each unique named entity is only counted once per comment that it occurs in, to prevent highly repetitive comments in which a named entity is mentioned multiple times from dominating the results.

The entities in Table 3.1 will be considered "frequent named entities" (FNEs) and are used to filter the prepared datasets to create a final evaluation dataset that emphasizes known failure modes in content-based models.

3.5 Results and Analysis

Table 3.2 illustrates that the performance of the NE masked model (NEMM) is comparable to that of the unmasked model when distinguishing between sentences written by randomly sampled Reddit users and sentences written by suspected influence accounts.

	Unmasked Model	Masked Model
Accuracy	**0.7409**	0.7266
AUC	**0.7409**	0.7266
F1 Score	**0.7433**	0.7302
Precision	**0.7359**	0.7202
Recall	**0.7512**	0.7409

Table 3.2: Mean evaluation results on masked and unmasked models trained to differentiate between suspect sentences (positive class derived from Corpus I comments) and random comments (negative class derived from Corpus II) on an evaluation set containing an equal number of comments from each class.

Corpus	Unmasked Model	Masked Model	t-statistic
L1Ru	63.82%	**43.72%**	9.92
L1En	38.82%	**36.10%**	3.67
L1Ru-FNE	70.09%	**56.55%**	20.46
L1En-FNE	54.46%	**51.97%**	3.80

Table 3.3: Type I error rates on Corpus III sentences written by L1 Russian and L1 English users, as well as t-statistic of difference between unmasked and masked Type I error rates.

The unmasked model does however retain a slight advantage on the trained classification task.

Table 3.3 shows the performance of both models on an evaluation dataset of randomly sampled English-language comments by L1 Russian (L1Ru) and L1 English users (L1En). Both models demonstrate a significant increase in false positives when applied to the L1Ru dataset compared to the L1En dataset. The false positive rate is highest for the unmasked model when classifying comments by L1 Russian speakers that contain a named entity frequently mentioned within the training data (L1Ru-FNE), followed by the false positive rate on arbitrary L1 Russian comments (L1Ru). The increased error caused by the presence of frequent named entities is substantially improved by the NEMM.

The results of each test in this experiment were rigorously validated by repeating 10 runs of 10-fold cross-validation. As the corpora are not of equivalent size, undersampling is used

on Corpus II, with each run performed using a new random sample, to reduce the impact of the randomly selected sample. We then perform significance testing by using the corrected repeated k-fold CV test to calculate the t-statistic [Bouckaert and Frank, 2004][Nadeau and Bengio, 2003]. This approach relies on fewer assumptions regarding independence of variation between folds, and is appropriate for comparing two distinct approaches. For the accuracy of differentiating random sentences from suspect sentences as displayed in Table 3.2, we attain a score of $t = 4.9394$ (two-tailed $p < 0.00001$). Significance testing results for the difference in false positive rates are displayed alongside results in Table 3.3, computed using a paired-sample t-test. All of the readings fall within a two-tailed significance level of $p < 0.001$.

The results described in §3.5 indicate that models trained exclusively on content features of existing influence campaigns disproportionately misclassify speakers of that language, as well as users who refer to specific named entities common to past influence accounts. When both of these conditions coincide, the effect is magnified substantially, giving the highest percentage of false positives in the evaluation set.

Simply put: users with Russian as a first language, particularly those who are discussing the United States, politics, or cryptocurrency, are at increased risk of false positive classification when writing in English.

When the classification model and test data are masked, the model becomes more resistant to the presence of FNEs and topic bias in L1 Russian comments, but a pronounced gap between the performance on L1 English and L1 Russian sentences remains. This gap demonstrates the interplay between powerful language models that naturally learn features indicative of a user's native language, and sampling approaches that draw positive and negative examples from users with largely distinct native language backgrounds.

3.6 Concluding Remarks

We conclude that the use of content-based features without safeguards creates the potential for discrimination against users of specific language backgrounds, especially when they are engaged in speech that contains common named entities that often reflect political topics. As protection of genuine free expression of political opinions on the Internet is a value of many organizations and governments, designers of online influence detection models should consider constructing test datasets of L2 English speakers using contextual data clues, such

as flair or IP address, for the purpose of identifying avenues of discrimination. While some measurable bias towards detection of users who speak the same L1 language as the target distribution may be inevitable, this behaviour should be tracked and mitigated whenever possible. development of improved contextual and content-based influence campaign detection methods should be done with minority language communities in mind to prevent large-scale discrimination.

Supervised approaches to the OIC detection problem face a number of challenges. First and foremost, the availability of high-quality training data is a significant challenge, a problem that is further exacerbated by the changing topics and tactics utilized by influence campaign originators as time goes on (as well as the increasing number of documented "bad actors" in the space). A classification approach must also be able to deal with significant class imbalance, little access to valuable back-end data, and be robust to changes in tactics over time — while being mindful of the potential for automated political censorship at scale. In online communities, the number of influence accounts is typically heavily outnumbered by non-influence accounts, thus requiring significant effort for detection systems to first address the problem of the imbalanced domain. Research into more sophisticated approaches that extend beyond simple binary text classifiers, therefore, is a key area of development.

Chapter 4

Novel Unsupervised Process for OIC Detection

4.1 Introduction

In the previous chapter we encountered several obstacles to the successful application of OIC detection techniques. We might reduce these challenges to two main problems:

- Automated detection or flagging systems based on content features may systematically discriminate against minority language communities
- Labelled OIC text is limited in both quantity and variety, with text characteristics highly specific to each campaign and its operator

We have demonstrated that we can partially mitigate the first problem using named-entity masking (NEM), but the second problem remains a significant obstacle. When relying on training data from labelled influence campaigns for detection of other campaigns, supervised approaches are likely to be most effective at discovering new influence campaigns similar to labelled campaigns, or finding "the rest" of an online influence campaign after obtaining an initial labelled sample to direct future detection efforts.

In this chapter, we introduce a process based on an unsupervised approach that incorporates a human analyst. This process addresses the limitations of labelled OIC data through a detection methodology that can be performed without labels, though may still leverage labelled data when searching for similar content to publicly-disclosed OICs. To

reduce algorithmic bias, the presence of a trained human analyst assists in addressing the possibility of automated discrimination, by meeting the criterion for "human agency and oversight" cited by the High-Level Expert Group on Artificial Intelligence (AI HLEG) as the first of seven key requirements for trustworthy artificial intelligence [HLEG, 2019]. We acknowledge, however, that human oversight is just one component of an ethical system, and algorithmic bias may persist in spite of human involvement [Angwin et al., 2016]. This chapter is dedicated to the design of this detection process, which leverages an unsupervised methodology to improve human analysis of social media communities for OIC detection.

While such a detection process could be designed to have the underlying model dynamically updated based on user feedback (as in the "active learning" approach), the role of the human in the system described here is purely as an end-user, not as a feedback mechanism to the model.

In the design of this process, we hypothesize that user-generated text from online influence campaigns contains common characteristics that may manifest in Transformer sentence embeddings. By reducing the text embeddings of Reddit submission titles into a visualizable number of dimensions using a structure-preserving projection method, it is possible for a human analyst to quickly explore similar Reddit submissions. Within this analysis, we consider submission titles as "sentences" for two reasons: first of all, submission titles on Reddit regularly take the form of complete sentences or questions; and secondly, BERT is trained in an unsupervised setting on a large corpus that includes an enormous variety of text — BookCorpus [Zhu et al., 2015] (800M words) and English Wikipedia (2,500M words) — and is expected to form meaningful representations for a large variety of English input sequences, including unusual or partial ones. Naturally, further research may perform additional BERT pre-training using Reddit titles as an additional corpus, according to the pre-training procedure discussed in §2.2.2. While submission titles are often (but may not always be) grammatically complete and correct "sentences", we will use the term "sentence embedding" to refer to the process of embedding sequences of words within the title field. Some submissions to Reddit may include body text along with a title, but this is only for text posts (also known as "self posts") or polls. This field is not present when the submission is an external link or multimedia element. We will be focusing on the title field in this chapter, as it is present in all cases for all submissions to the website.

While online influence campaigns on Reddit have historically contained a diverse variety of titles [Reddit, 2018b], we theorize that submission titles from online influence

campaigns will, overall, have BERT embeddings with a comparatively lower Euclidean distance to other OIC title embeddings than the embeddings of regular submission titles, even when embedding distributions are multi-modal. This is a reasonable expectation due to similarities in the linguistic background of OIC operators, and the external coordination of subject matter that an orchestrated political OIC necessitates — both factors that impact the resulting BERT embedding, as observed in §3.5. Under this expectation, we anticipate that by taking the mean embedding of titles submitted by each user, clear concentrations of similar OIC accounts may emerge in the resulting meta-embedding space. In low-dimensional projections of this space, the human ability to quickly identify patterns and reason about the world will allow them to triage what information is relevant to detection of an online influence campaign, and improve detection ability.

4.1.1 Purpose of the Study

The central purpose of this study is the formal introduction of an unsupervised approach to detecting online influence campaigns that allows a human analyst to easily leverage state-of-the-art NLP research through a visual interface. This includes significant new development in the area of OIC detection, including the application of a simple meta-embedding technique for Reddit user modelling, the use of BCubed on HDBSCAN clusters within 2D and 3D projections as a performance metric for visualization quality, and a case study based on the 2019 Canadian Federal Election in which the process is used to identify suspicious accounts — several of which were later verified to have been removed independently by the social media platform.

4.1.2 Ethical Focus

The key ethical advantage of the approach this process encourages, is the incorporation of a human analyst into the analytical process. This does not mean that the methodology used is incapable of manifesting algorithmic bias in the data presented to the analyst, but that the integration of a human analyst into the detection methodology is widely regarded as an improvement in the standard of rigour for the deployment of an automated system, as noted by the ethics guidelines for trustworthy AI developed by the High-Level Expert Group on AI (AI HLEG) [HLEG, 2019]. The first of seven key principles of trustworthy AI systems is "human agency and oversight", which is supported by the introduction of a

human arbiter. Further, by relying less on labelled influence campaigns (which was demonstrated quantitatively in §3.5 to have potential for discrimination), it is also an arguable improvement for the fifth key principle: "diversity, non-discrimination and fairness".

One area that creates potential for bias within machine learning systems is that the selection of training data and detection objectives can itself represent manifest underlying data and human biases [Siapka, 2018]. These problems, along with the bias towards false positive detection of text with superficial topical similarity to issues in previous campaigns, might be reduced through development of unsupervised detection methods. An unsupervised method still requires the formulation of some objective, and the system must still process a collection of input data, both of which may still internalize some human bias and/or data bias. Furthermore, the data from past influence campaigns may still be used alongside an unsupervised method, which would reintroduce some of these issues as well. These points aside, the absence of training data and a domain-specific training objective offers a reduction in avenues for other forms of bias to influence the model. This may be further improved through minimizing the arbitrary selection of hyperparameters, and instead relying on algorithms that either do not require a particular hyperparameter (such as those that do not require the number of clusters to be specified *a priori*), or determining hyperparameters via a rigid experimental methodology or according to a cited defensible baseline recommendation.

In light of the sociotechnical nature of this problem, the human element of the detection system should not be ignored in assessment of fairness and bias [Selbst et al., 2019]. A full ethical assessment should ideally include a holistic trial of how the OIC system operates when applied to a real detection task, (including the human oversight), compared against other methods. This process would be similar to the proposed EU process for algorithmic impact assessments for automated decision making in policing [Kaminski and Malgieri, 2019]. This type of trial is beyond the scope of this research, but is a valuable area of future work in ethical OIC detection.

4.2 Prior Art

Social Media Analysis

Extensive work has been done in the space of social spam detection that overlaps with the area of political OIC detection. In either case, the goal is typically to disseminate

a message to a broad audience that contains vulnerable targets. Many approaches for spam detection exist, such as regex matching on shared spam URLs [Gao et al., 2010], or modelling friend/follower relationships between users [Yang et al., 2012]. In political online influence campaigns, the presence of a malicious website is not necessarily required to influence a reader, and filtering methods that detect unusual domains or URLs can be bypassed by linking to existing published articles, or popular social media and blogging platforms, instead of bespoke websites. Modelling friend/follower relationships is also less valuable on a platform such as Reddit, which, at the time of writing, has very low emphasis on relationships between user accounts. Common participation in the same subreddit communities or comment threads, however, may provide some opportunity to create links between accounts.

Transformer-based solutions for detecting malicious behaviour in social media are still in the very nascent stages. Several projects [Punturo, 2019][Weller and Woo, 2019] and papers [Crothers et al., 2019][Kennedy et al., 2019] have demonstrated the performance of fine-tuned BERT models on text classification of malicious comments or falsified reviews. While this research forms useful groundwork for further investigation, the use of balanced training and evaluation datasets means that the resulting systems may not translate well into real-world detection methodologies, where such systems must contend with significant class imbalance.

Previous work has been performed on clustering social media users based on their written text [Gencoglu, 2019][Singh et al., 2016], and specifically using BERT embeddings to cluster written arguments between humans [Reimers et al., 2019]. Attempts to better understand BERT embeddings have also leveraged visualizations obtained through dimensionality reduction techniques, resulting in insights into the combination of syntax and semantics represented within BERT embeddings [Coenen et al., 2019]. These works do not consider the combination of multiple BERT embeddings to create user representations based on textual content, and UMAP visualization of mean BERT user embeddings has never yet been used within a published social media analysis process, for OIC detection or otherwise.

Combining Transformer embeddings to create a representation of a user based on their written text requires aggregating multiple sentence embeddings into a single meta-embedding for each user. The approach of combining multiple embeddings into a meta-embedding through averaging has been demonstrated as effective in a theoretical setting, without regard for applications towards user representation [Coates and Bollegala,

2018]. Parallel work in the field of user representation has taken the approach of sampling "episodes" of written text, and combining episode representations using an auxiliary neural network trained on an author identification task [Andrews and Bishop, 2019]. This approach incorporates not only textual features but temporal features as well, and produces very strong results for author identification on Reddit. The comparatively simple method presented within this research relies purely on text representations from pre-trained BERT models, and is nevertheless able to effectively characterize groups of users through average title embeddings.

Cluster Evaluation

Extrinsic cluster evaluation is used within this chapter as a measure of visualization quality that approximates human perception. The rationale for this approach is that what humans perceive as a "structure" within point visualizations is best described in machine learning as a "cluster". To select the most appropriate clustering method for usage on Reddit data, a preliminary intrinsic cluster evaluation process is used to assess the suitability of several clustering methods.

The intrinsic performance of several clustering algorithms is assessed using the Calinski-Harabasz (CH) criterion [Caliński and JA, 1974], which is well-regarded as an metric for comparing disparate clustering techniques. This criterion has been used for analyzing clustering performance on social media text representations in prior art, including evaluation of clustering methods for deep vector embeddings derived from Twitter posts [Gencoglu, 2019]. Where appropriate, additional pre-processing and feature-space reduction will be performed to improve the performance of techniques that are not suited for being applied directly to high-dimensional vectors.

Formally the Calinski-Harabasz criterion is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. Mathematically this is represented as:

$$CH = \frac{BGSS}{k-1} / \frac{WGSS}{n-k}$$

where k is the number of clusters, $BGSS$ is the between-cluster sum of square distances, $WGSS$ is the within-cluster sum of square distances, and n is the number of points. This

criterion is analogous to the F-statistic in univariate analysis [Caliński and JA, 1974]. This is an intrinsic cluster quality metric that reaches a higher value when clusters are more distinct from one another, and their constituent points closer together.

After selection of an appropriate clustering technique for Reddit data based on intrinsic performance, the results of this technique will be evaluated with an extrinsic cluster evaluation metric as a proxy for human perception. BCubed is an extrinsic cluster evaluation metric that can be used to compare the performance of a clustering against a set of labels indicating the ideal cluster assignment for each point [van Rijsbergen, 1974]. An extension of BCubed, known as Extended BCubed [Rosales-Méndez and Ramírez-Cruz, 2013], is a superset of BCubed, which also supports overlapping clusterings. BCubed provides precision, recall, and F-score metrics for clustering results. The BCubed F-score is a particularly desirable extrinsic evaluation metric as it upholds conditions surveyed to be important for an effective cluster evaluation metric. These conditions were summarized in a comprehensive review of extrinsic cluster methods [Amigó et al., 2009] as: *"homogeneity", "completeness", "rag bag",* and *"clusters size versus quantity"* [Amigó et al., 2009].

The following equations (4.1, 4.2, 4.3) come from the formal definition of Extended BCubed [Rosales-Méndez and Ramírez-Cruz, 2013], precisely as they are given by Amigó et alia. Extended BCubed precision is defined as:

$$P = \frac{1}{|U|} \sum_{o \in U} \frac{1}{|\bigcup_{g \in G(o)} g|} \sum_{o' \in E(o,G)} \frac{min(|G(o) \cap G(o')|, |C(o) \cap C(o')|)}{|G(o) \cap G(o')|} \tag{4.1}$$

where U represents the collection, G stands for the candidate clustering, C for the gold standard (i.e., perfect cluster assignment based on labelled data), $G(o)$ represents the set of candidate clusters containing object o, $C(o)$ is the set of classes of the gold standard containing o, $E(o, G)$ is the set of objects co-occurring with o in at least one candidate cluster, and $E(o, C)$ is the set of objects co-occurring with o in at least one class of the gold standard.

Using the same definitions, Extended BCubed recall is defined as:

$$R = \frac{1}{|U|} \sum_{o \in U} \frac{1}{|\bigcup_{g \in C(o)} g|} \sum_{o' \in E(o,C)} \frac{min(|G(o) \cap G(o')|, |C(o) \cap C(o')|)}{|C(o) \cap C(o')|} \tag{4.2}$$

and the Extended BCubed F-measure is defined in terms of precision and recall as

$$F_\alpha(P, R) = \frac{1}{\alpha(\frac{1}{P}) + (1 - \alpha)(\frac{1}{R})} \tag{4.3}$$

When $\alpha = 0.5$, the F-measure is the harmonic mean between precision and recall, which is the setting used in this analysis.

A Python implementation of extended BCubed is available on GitHub [Hromic, 2019], and was used for the analysis in this chapter.

4.3 Methodology

We break down the methodology in the creation of the analytical process into two "steps", each of which result in an associated visualization process that can assist a human analyst with the detection of an online influence campaign. The first step visualizes individual BERT title embeddings, while the second combines title embeddings to create visualizable user representations. The second step, which provides a more powerful process by leveraging many BERT embeddings in aggregate, is the primary focus of this chapter, and the subject of our perception quality measures through the proxy of BCubed extrinsic cluster evaluation.

The first step focuses on detection of user-generated submission titles that have been written as part of online influence campaigns. We consider sentences written as part of the campaign as the "influence title class", while sentences not part of the influence campaign are considered "background title class". As the presence of the influence title class is

uncertain within the visualization, we rely on prior influence campaign data as a means of evaluating whether the model colocates influence accounts — however, in application, the process can be operated without labelled OIC data.

The second step groups submission titles by the user that authored them, conceptualizing a "user" as the collection of all posts authored by a particular person. Users which are operated as part of an online influence campaign are considered part of the "influence user class", while users not operated by an online influence campaign are considered the "background user class". Once again, as the presence of the influence user class in the collected sample is uncertain, we rely on prior influence campaign data to evaluate whether the model is performing effectively. As analysis of this visualization provides users of interest, we can perform a query several months after the analysis to determine which accounts in the analysis have been subsequently banned or suspended. This serves as a coarse heuristic for evaluating whether the process can be used effectively to detect undesirable accounts.

We leverage the power of large Transformer models for both of these tasks, by averaging Transformer embeddings derived from user-generated text into features that represent individual users, and projecting them into 2 or 3 dimensions using Uniform Manifold Approximation and Projection (UMAP) [McInnes and Healy, 2018][McInnes et al., 2018]. In contrast to other dimensionality reduction methods, such as t-distributed stochastic neighbour embedding (t-SNE) [van der Maaten and Hinton, 2008], UMAP arguably better preserves the global structure of the underlying data [McInnes and Healy, 2018], providing a projection where not only are the distances between nearby points meaningful (local structure), but their relative position with respect to other groups of points within the projection space is also meaningful (global structure). Additionally, UMAP can be easily performed directly on high-dimensional data, whereas t-SNE typically requires first reducing the data to a smaller space via a less computationally expensive method such as principal component analysis (PCA) [McInnes and Healy, 2018]. A preliminary comparison to verify these characteristics was performed, the results of which can be found in Appendix C. Similar to Chapter 3, we will once again perform these experiments using the BERT base uncased model [Google Research, 2019a].

A projection is useful for facilitating human detection if it places posts that are part of an influence campaign into visually identifiable structures within the projection, either in a single part of the embedding space, or in several distinct clusters. To evaluate whether the resulting projections are likely useful for aiding human detection of online influence campaigns, we will use "Hierarchical Density-Based Spatial Clustering of Applications

with Noise" (HDBSCAN) [McInnes et al., 2017] to create a density-based clustering of embeddings from labelled influence campaigns. We consider HDBSCAN a useful proxy for human perception of similarity in the projected data points, based on internal quality metrics and understanding of Gestalt perception principles, as discussed in §4.4.1. By using labelled influence campaigns as a form of ground-truth, we can leverage extrinsic cluster evaluation methods, which require some amount of labelled data. We find that BCubed, an extrinsic cluster evaluation metric which captures cluster quality very well under diverse scenarios [Amigó et al., 2009], scores these clusters as substantially outperforming a random baseline, indicating that exploring user representations through UMAP visualization is beneficial to human perception.

As in the experiments in the previous chapter (§3.3.1), we use the transfer learning strategy described in §2.2.4 to create sentence embeddings using the prepended `[CLS]` tokens from the BERT model (a process also described in §2.2.3). We can then use these high-dimensional sentence representations as input to dimensionality reduction algorithms for the purpose of visualization. These representations were then combined to form user representations that are better suited for detection of suspicious accounts (rather than considering each individual text submission in isolation).

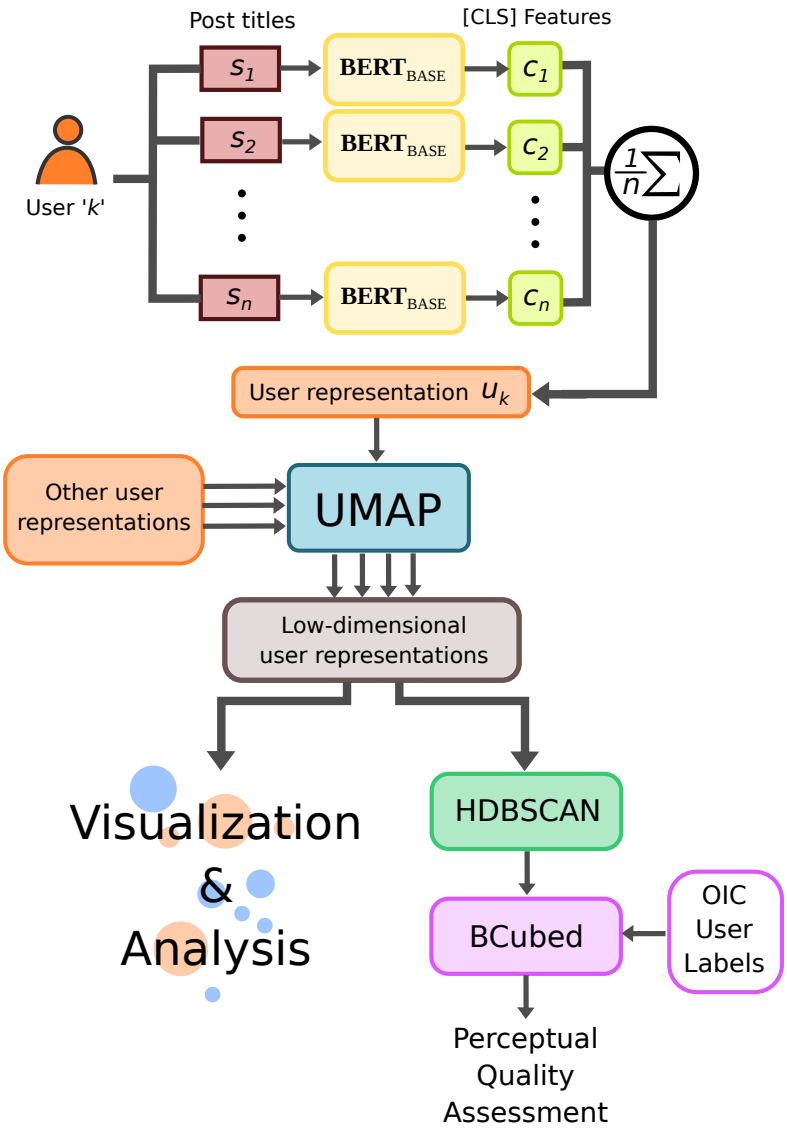

Figure 4.1: Flow diagram of user embedding, visualization, and evaluation methodology

Community of Focus

In this section, we focus on a Reddit dataset containing popular communities for Canada-focused discussion. As a very popular social network in Canada [Alexa Internet, 2020a], Reddit is an important platform for online foreign influence research. Canada-oriented subreddits were selected based on the reasoning that an influence campaign targeting Canadians would likely target subreddits that are specifically geared towards Canadian cities or Canada at large, as communication here would have the greatest chance of influencing Canadian voters.

All posts within Canada-focused subreddits were scraped using the Pushshift API [Baumgartner et al., 2020] and organized into a dataset. A full list of the included subreddits can be found in Table D.1. This is by no means an exhaustive list of all Canada-affiliated subreddits, but represents a diverse sample of many of the most popular communities.

To better understand the dataset, the volume of activity in each subreddit was measured by month (Figure 4.2), and the hour of day was calculated for each comment and post to determine at what time of day users were most active (Figure 4.3). From Figure 4.2, we determine that the dataset contains activity from a time range that is relevant to the election (and in fact increases with proximity to the election). From Figure 4.3, the timestamps of activity within these subreddits indicate that the majority of their users are likely located within a range of longitudes that includes Canada, supporting the relevance of these subreddits towards influencing Canadian voters.

In contrast to the previous chapter, we will focusing on submission titles rather than comments in our methodology, though we will once again be relying on submission titles from past online influence campaigns (Corpus I in the previous chapter). The technique utilized and the rationale for the selected process are described in the next section.

A more detailed breakdown of the total number of posts and comments from each subreddit can be found in Table D.1.

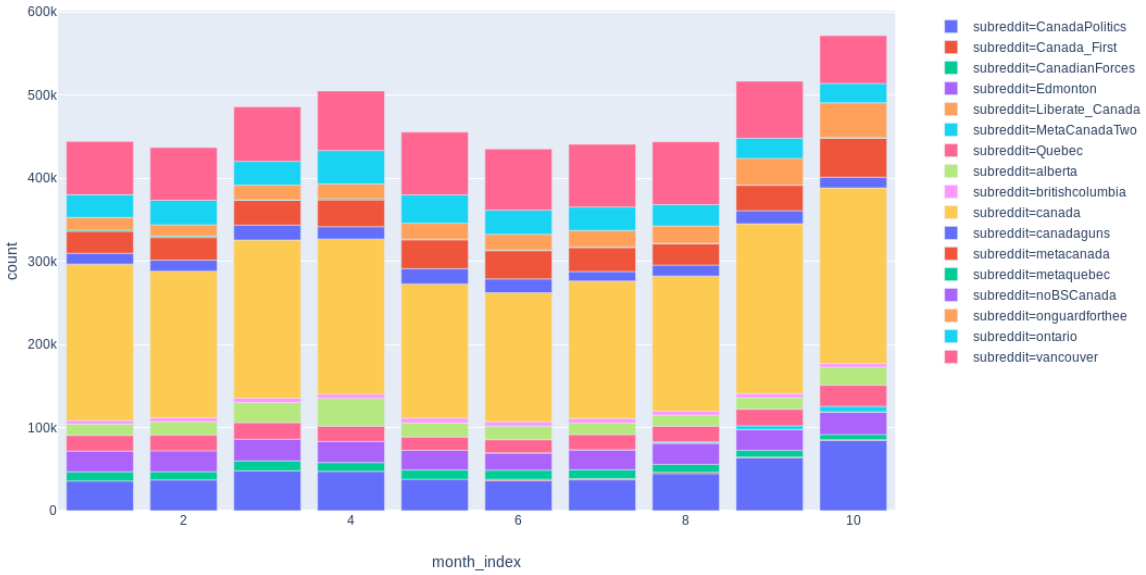

Figure 4.2: Number of posts and comments to each subreddit over months of 2019. The activity levels remains relatively stable, with overall volume increasing as the election approaches.

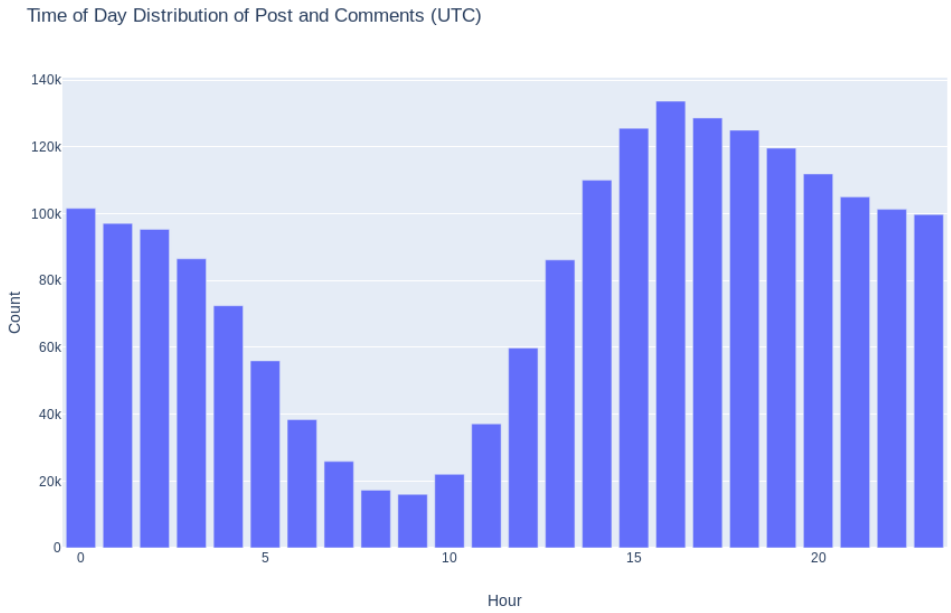

Figure 4.3: Time-of-day analysis for posts and comments in the 2019 Canadian Reddit dataset. Activity levels exhibit a strong correlation with daytime hours in Canada (which covers time zones from UTC-8h during Pacific Standard Time to UTC-2h30 during Newfoundland Daylight Time).

4.3.1 Title Visualization via BERT Embedding and UMAP

We begin by generating BERT [Devlin et al., 2018] embeddings for every submission title within the entire dataset, one again leveraging the uncased variant of the BERT uncased base model (12-layer, 768-hidden, 12-heads, 110M parameters) [Google Research, 2019a]. Once again, the selection of this variant reflects the ethos of using standard models with high availability and ease of study discussed in §2.2.1. From these embeddings, we select only the representation of the `[CLS]` token prepended to each input sequence.

Uniform Manifold Approximation and Projection (UMAP) [McInnes and Healy, 2018][McInnes et al., 2018] is then used to reduce these vectors into a visualizable number dimensions for review by trained analysts specializing in OIC analysis. This dimensionality reduction was performed on a single high-capacity processing node with 61GB of RAM and a powerful central processing unit (Intel® Xeon® Platinum 8175M CPU @ 2.50GHz). We perform UMAP dimensionality reduction twice, both times setting the number of neighbours to 15 and using cosine distance as the similarity metric, but varying the number of output components to create 2 and 3-dimensional representations. The parameter of 15 was found to provide a high-quality projection with an appropriate balance of local and global structure. Similarity in the UMAP embedding is represented in Euclidean space, so Euclidean distance is used as the similiarity metric.

These representations are then plotted and visualized using a number of annotation and colouring schemes including by subreddit or by the results of a clustering algorithm. The UMAP configuration used within these experiments is non-deterministic, resulting in embeddings that — while very similar in structure — may have differences in orientation.

Finally, we repeat the above processes after overlaying suspicious submission titles from documented influence campaigns (Corpus I in the previous chapter) to visualize a sample of labelled data alongside unlabelled data. This provides an idea of where an influence campaign might appear within the space, and allows us to obtain a measure of visualization quality.

4.3.2 User Visualization via Meta-Embedding and UMAP

To create user representations, submission titles in the 2019 Canadian Reddit dataset were grouped by author, and the dataset was filtered to only those users who posted 10 submissions or more over the course of the year. The BERT embeddings of the `[CLS]`

tokens for these submissions were then averaged together to create a meta embedding of each user's 2019 submission titles.

These representations are once again fed into UMAP dimensionality reduction and HDBSCAN density-based clustering algorithms, using the same hyperparameter settings for the meta-embedding as the basic `[CLS]` embeddings.

Interactive Visualizations and Embedding Projector

Two utilities are used to visualize title and user embeddings for analysis. The first leverages the Bokeh visualization library [Bokeh Development Team, 2019], which creates interactive HTML documents. By plotting the embeddings in two-dimensional space, specifying the colour based on cluster or subreddit, and setting hover tooltips that map to the original textual content, this provides an easily-shareable self-contained analytical visualization for understanding this data. An example can be found in Figure 4.4.

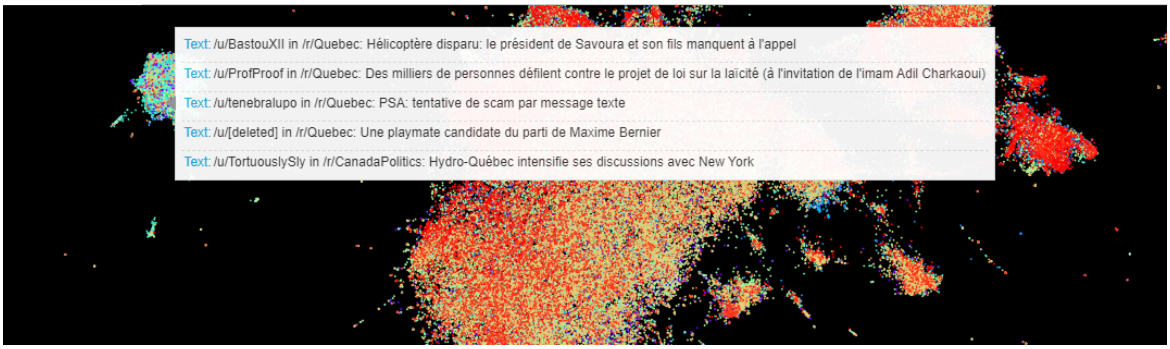

Figure 4.4: Screenshot of interactive Bokeh visualization of submission title embeddings, coloured by subreddit. Hovering over a point provides the source data corresponding to that embedding.

For cases where three-dimensional embeddings need to be visualized, or when more advanced capabilities are required — such as browsing nearby points, or searching for specific users within the data — the TensorBoard embedding projector [Google, 2016] is used. TensorBoard Embedding Projector is a software developed by Google for visualizing embeddings. A version of this software is freely available online [Google, 2016], as is the source code for self-hosting the application [Google, 2019]. A preview of TensorBoard applied to the user representations can be found in Figure D.4.

Using TensorFlow embedding projector, a user can load a custom embedding and labels, and then search these embeddings. Since the data contains both usernames as well as a

series of post titles, this provides an easy opportunity for the user to rapidly determine the nature of different clusters, and characterize similar posting activity with relative ease.

4.4 Results and Analysis

4.4.1 Intrinsic Clustering Comparison for Evaluation Criterion

To motivate the selection of which clustering algorithm to use as an extrinsic evaluation criterion, we compare intrinsic clustering performance on a sample of BERT sentence embeddings using several clustering techniques available in the scikit-learn library [Pedregosa et al., 2011]. Each technique is introduced here, alongside the resulting Calinski-Harabasz scores for varying numbers of clusters k (in cases where the numbers of clusters is set *a priori*). The results of quantitative analyses for k-means, GAAC, and GMM, can be found in Table 4.1, while results for DBSCAN can be found in Tables 4.2 and 4.3.

K-means

As k-means requires an *a priori* specification of the number of clusters, we will run a CH-score evaluation of the cluster quality at a number of different values for number of clusters, k, and highlight the value with the best performance. Due to the scale of the data involved, we randomly sample 1000 sentences from a dataset of BERT sentence embeddings from Reddit, which we assess to be sufficient for determining the approximate quality of the clustering criterion. This same sample will be used for all quantitative comparisons to ensure a fair comparison. A chart depicting the change in CH score as k is manipulated can be found in Figure 4.5.

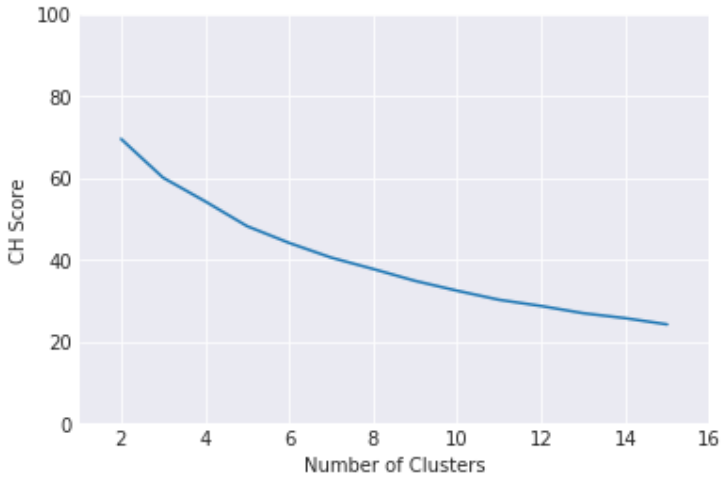

Figure 4.5: Graph of CH scores for differing cluster counts, k, for k-means algorithm. Optimal value is determined to be found at $k = 2$,

Group-average Agglomerative Clustering

Similar to k-means, group-average agglomerative clustering (GAAC) requires an a priori specification of the number of clusters. We evaluate the performance of GAAC at varying numbers of clusters k, and calculate the CH score at each step. A chart depicting the change in CH score as k is manipulated can be found in Figure 4.6.

The result of GAAC clustering on the feature vectors was among the poorest of the clustering algorithms. To determine whether this was due to the dimensionality of the input data or due to another characteristic of the agglomerative algorithm, the built-in singular value decomposition (SVD) was used to further reduce the dimensionality of the data. However, this did not result in any additional improvement in the results.

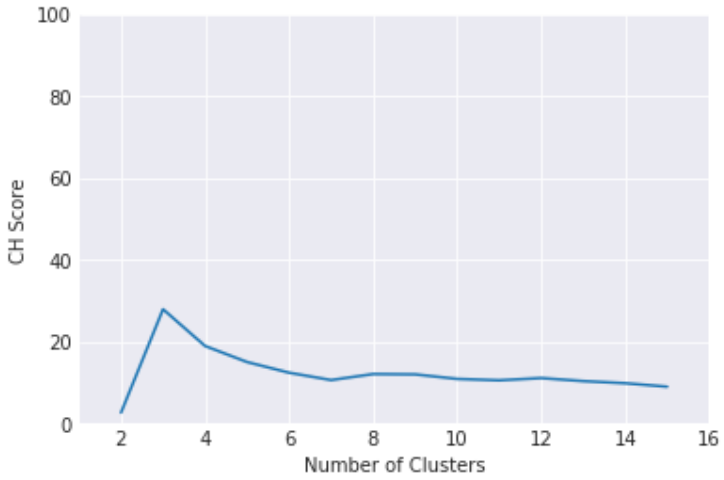

Figure 4.6: Graph of CH scores for differing cluster counts, k, for GAAC algorithm. Optimal value is determined to be found at $k = 3$,

Gaussian Mixture Model

We evaluate the gaussian mixture model (GMM) algorithm using the same CH internal clustering criterion at different values of k, with the added manipulation of the covariance type. There are three types of covariance that we will evaluate for the purpose of this analysis: spherical, diagonal, and full. Spherical covariance limits the shape of the cluster such that all dimensions are equal, which gives it similar properties to the k-means algorithm. Spherical covariance limits the expressive capability, but has a corresponding increase in performance. Diagonal covariance is very commonly used (and is the default in scikit-learn), and allows for the formation of elliptical clusters that cannot be expressed with k-means. Full covariance allows for oblong clusters that may be stretched at an angle, but requires the most computational power.

A visual plot of the three covariances as the number of clusters k is manipulated can be found in Figure 4.7. The numerical results for each of the three GMM covariances, as well as the results for the previous two methods, can be found in Table 4.1.

k	k-means	GAAC	GMM		
			Spherical	Diagonal	Full
2	**69.81**	2.70	**68.47**	**67.38**	**69.15**
3	60.10	**27.97**	49.49	59.34	55.26
4	54.32	18.99	54.24	46.15	46.12
5	48.32	15.06	47.71	47.69	46.30
6	44.20	12.46	43.66	42.66	43.56
7	40.60	10.66	40.01	40.13	39.90
8	37.82	12.13	36.37	37.45	37.73
9	34.99	12.05	34.45	34.81	34.66
10	32.44	10.94	31.02	32.00	32.25
11	30.48	10.59	29.87	29.96	30.28
12	28.63	11.15	27.92	28.40	28.31
13	27.15	10.40	26.90	26.67	26.96
14	25.60	9.88	25.25	25.29	25.67
15	24.58	9.04	24.21	24.03	23.66

Table 4.1: Calinski-Harabasz scores per number of clusters, k, under different clustering approaches.

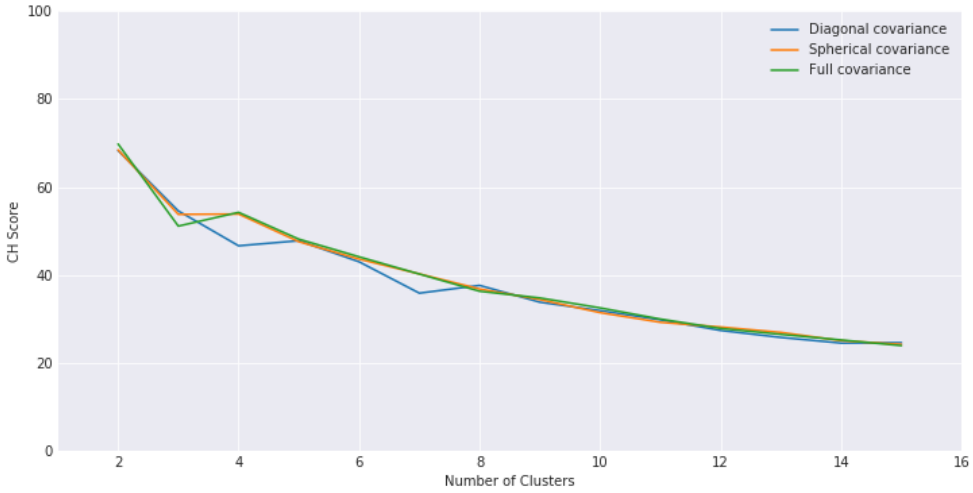

Figure 4.7: Graph of CH scores for differing cluster counts, k, for different covariances of Gaussian mixtures. Optimal value is determined to be found at $k = 2$ when using a full covariance.

DBSCAN

Unlike the previous methods, DBSCAN does not require the number of clusters to be specified a priori. However, as discussed earlier, it does possess some weaknesses that are likely to be exasperated by the sparseness and high-dimensionality of the BERT activations.

We test a number of different values for *min_samples*, m, the parameter which determines the number of samples required to consider a region "dense", as well as the value ϵ which specifies the distance between points below which they are considered neighbours. The results on the raw feature activations can be found in Table 4.2. For the top-performing combination of hyperparameters, DBSCAN identified two separate clusters (in addition to the 'noise' class).

The observed performance of DBSCAN on the raw network activations was quite low relative to k-means and GMM. To improve this, we use principal component analysis (PCA) to reduce the feature space of the random sample from 768 to 50, and then repeat testing values for m and ϵ on this data. This had the intended effect of greatly improving the quality of the resulting clusters. However, it is important to note that some information is lost in this process, and this should be taken into account when comparing this result to the results for the previous clustering methods. Once again, for the top-performing combination of hyperparameters in this method, DBSCAN identified two separate clusters (in addition to the 'noise' class).

m	10	11	12	13	14	15
ϵ						
9.0	26.60	26.63	24.10	23.42	23.53	22.13
9.2	28.15	28.09	28.13	28.07	27.13	27.06
9.4	28.10	28.23	28.32	28.34	28.34	28.46
9.6	28.83	28.81	28.87	**28.88**	28.03	28.02
9.8	6.91	7.06	7.06	7.06	6.97	6.97
10.0	6.55	6.55	6.55	6.55	6.55	6.94

Table 4.2: Calinski-Harabasz scores for different values of ϵ and min_samples, m, for DBSCAN on unprocessed feature sample.

Selection of Cluster Algorithm based on Intrinsic Criteria

Based on the results of intrinsic cluster evaluation, we observe DBSCAN had the highest performance on the data after dimensionality reduction, and has the added benefit of not requiring the number of clusters to be set *a priori*. Rather than setting the hyperparameters via grid search, we leverage HDBSCAN [McInnes et al., 2017] – a variant of DBSCAN – that sets epsilon and distance parameters based on stability, and (unlike DBSCAN) can find clusters of variable density. This allows us to leverage the observed performance of the density-based clustering method on this type of data, while streamlining selection of hyperparameters. Furthermore, the ability to find variable density clusters better matches Gestalt psychology for identifying visual patterns [Zahn, 1971], which is an important characteristic for an evaluation metric that is meant to be an analog for human perception. Human perception of "structures" or "clusters" is strongly related to boundary assignment (or segmentation) in neuroscience [Zhou et al., 2000]. When humans perceive a data projection, groups of points that are close to one another and disjoint from other groups of points are interpreted by the brain as belonging to one "object" (i.e., cluster) or another. We suggest, in the absence of a rigorous neurological survey that compares human clustering point assignment against computational clustering algorithms, that a density-based clustering algorithm with strong intrinsic performance that dynamically adapts to variations of density and number of clusters, is an adequate proxy for human point assignment, and therefore a suitable measure for the usefulness of a projection.

m / ϵ	15	16	17	18	19	20	21	22	23	24
6.3	67.18	65.45	62.93	54.96	50.37	50.43	33.41	32.94	32.65	32.46
6.4	67.62	68.20	68.67	66.00	66.12	51.94	51.48	33.34	32.62	32.97
6.5	69.83	69.89	69.46	69.56	**70.07**	63.66	55.43	55.53	55.27	55.38
6.6	68.57	68.32	69.22	67.75	67.83	68.74	66.45	56.07	56.13	56.10
6.7	65.32	67.16	67.70	68.57	68.83	68.53	68.43	64.93	58.31	54.96
6.8	66.01	65.10	65.48	65.34	65.15	67.89	68.34	67.40	67.34	66.07
6.9	63.81	64.03	62.98	63.09	63.63	64.47	64.04	65.32	65.70	65.76
7.0	62.77	62.88	62.43	62.43	62.09	62.12	63.23	63.86	63.26	63.47

Table 4.3: Calinski-Harabasz scores for different values of ϵ and min_samples, m, for DBSCAN on PCA50 processed feature sample.

4.4.2 Assessment of Visualization Quality

We finally consider the results of overlaying the submission titles from the suspicious accounts (Corpus I) over top of the Canadian subreddit data. This allows us to use HDBSCAN clustering to perform a quantitative evaluation using BCubed [Amigó et al., 2009] as an extrinsic cluster quality metric. This is useful to measure the quality of the resulting projections. A projection that places many influence accounts into a similar local density-based cluster is assessed to have greater value for a human analyst. We provide a baseline for each result by comparing it to a random baseline with the same number of clusters. This is accomplished by assigning labels to each sample from a uniform distribution over $1..k$ where k is the number of clusters obtained by the density-based algorithm — an approach that has been used for quite some time for obtaining a baseline of extrinsic cluster quality [Strehl and Ghosh, 2003]. This process is performed 10 times, and the result is averaged.

We perform clustering under two different settings of HDBSCAN, excess of mass (EOM) clustering and leaf clustering. This setting determines how HDBSCAN selects flat clusters from the the resulting cluster tree hierarchy. Put simply, EOM tends to select a small number of large clusters, with excluded points forming several smaller clusters. In contrast, leaf clustering selects the leaf nodes from the HDBSCAN tree, thus producing a larger number of smaller homogeneous clusters. Leaf clustering is performed to assess visualization

Embedding Type	Cluster Selection (#)	BCubed Precision	BCubed Recall	F-Score
2D Users	HDBSCAN-EOM (3)	0.8857	**0.8407**	**0.8626**
	HDBSCAN-Leaf (20)	**0.9116**	0.3187	0.4723
	Random (3)	0.8739	0.3340	0.4833
	Random (20)	0.8750	0.0509	0.0962
3D Users	HDBSCAN-EOM (3)	0.8850	**0.8398**	**0.8618**
	HDBSCAN-Leaf (21)	**0.9149**	0.3746	0.5315
	Random (3)	0.8738	0.3339	0.4831
	Random (21)	0.8749	0.0486	0.0920

Table 4.4: Extrinsic BCubed scores of projection space density-based clusters when ground-truth provided. Cluster counts provided in parentheses.

quality within areas considered a single cluster under EOM.

The results of this can be found in Table 4.4. The strong performance of each clustering method over their corresponding random baseline demonstrates that clustering with HDBSCAN in either EOM or leaf settings results in substantially higher F-Scores than a similar number of randomly selected clusters. This indicates that UMAP projection of the user representations places OIC users in closer and denser regions, quantitatively improving utility to a human analyst over random navigation of the data.

4.4.3 Qualitative Analysis and Application

Samples of the visualization lenses obtained by running the procedure described in the methodology are displayed within this chapter for human visual assessment. The embedding of `[CLS]` tokens can be found in Figures 4.9 and 4.10, coloured by the results of HDBSCAN cluster assignment and source subreddit respectively. Note that this HDBSCAN clustering is not part of the earlier extrinsic cluster evaluation process, but merely an analytical approach for better visualizing the data. The visualization of the meta-embedding user representations can be found in Figure 4.8.

Finally, the setting in which labelled Reddit OIC data is visualized alongside the 2019 Canadian Reddit data, can be found in Figure 4.13.

CLS Embeddings

The BERT embeddings of the prepended `[CLS]` tokens form structures in the UMAP visualization based on both the sentence structure and subject matter of the words within the submission title. Questions appear in a contiguous cloud, with questions about specific subject matter closer to one another. Likewise, news article headlines can be found in close proximity to one another due to similar writing style.

We can see from colouring the `CLS` embeddings by source subreddit (Figure 4.10) that there is a large variety in titles across subreddits. A notable exception is the collections of cyan points in several groupings which represent French-language posts, primarily concentrated in a single subreddit in the dataset (r/Quebec). The sentences in each small cyan cluster displays commonalities, with distinct groupings for French-language headlines, French-language questions, etc.

Meta-Embeddings

User embeddings based on BERT representations of submission titles reflect both the common topics in a user's submission titles, as well as the typical grammatical structure of their sentences. For example, users who tend to share news articles are in a similar location, and users who post polling results for political parties are in another. Performing an HDBSCAN excess-of-mass (EOM) clustering on the projection resulted in two main types of accounts: primarily English users and primarily French users, as seen in Figure 4.8. Note that once again this HDBSCAN clustering is not part of the earlier extrinsic cluster evaluation process, but merely an analytical approach for better visualizing the data. To improve this visualization, and address the differences caused by the English-French language divide, we might subsample users to isolate only those users who primarily post in English. Alternatively, we can once again leverage HDBSCAN leaf clustering as an alternative to subsampling (Figure D.6).

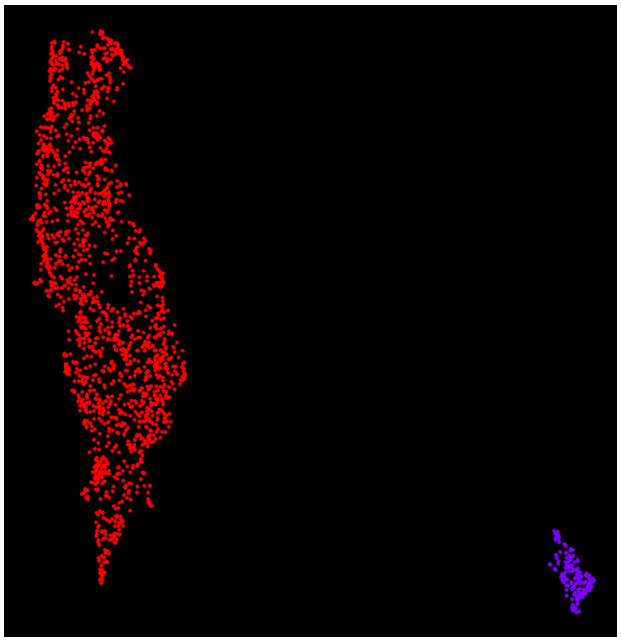

Figure 4.8: 2D UMAP meta-embedding of users based on averaged `[CLS]` token embeddings, coloured by HDBSCAN cluster. The greater distance between predominantly English and predominantly French user accounts in the user representation projection causes a marked decrease in cluster count compared to Figure 4.9

We find that the meta-embedding of `[CLS]` tokens qualitatively separates similar user accounts, where users who tend to post similar topics or structure their submission titles in

similar ways, tend to appear closer to one another. Figure D.2 displays the 3D visualization of the concentrated tip of the English-language cluster, within which users who make submissions with news headlines are densely concentrated, allowing these users .

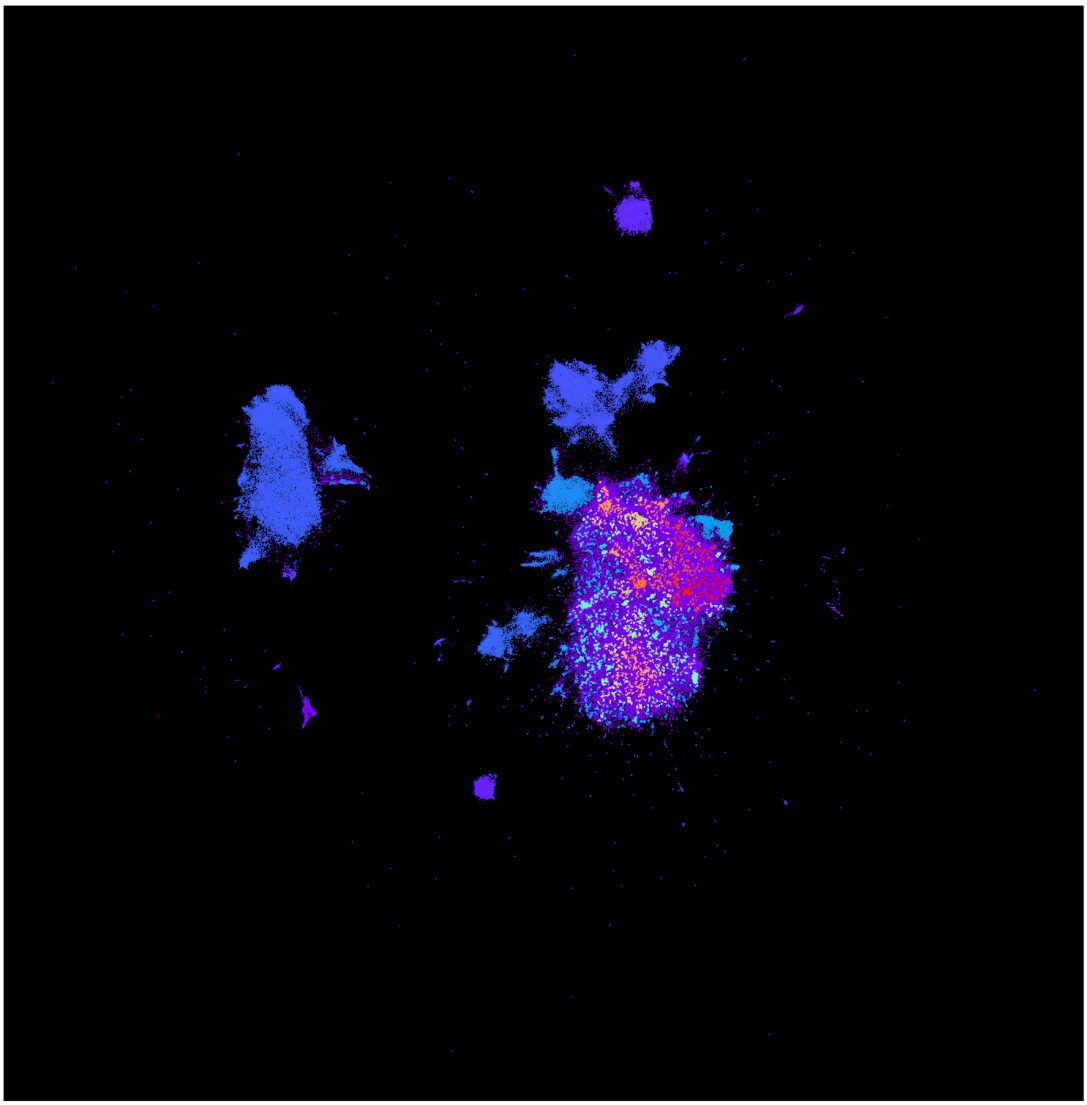

Figure 4.9: UMAP embeddings of submission title CLS token embeddings, coloured by HDBSCAN cluster. Many small, disparate clusters are created, which can be interpreted as separate "topics" within the data (e.g. "exclamations that involve American politics")

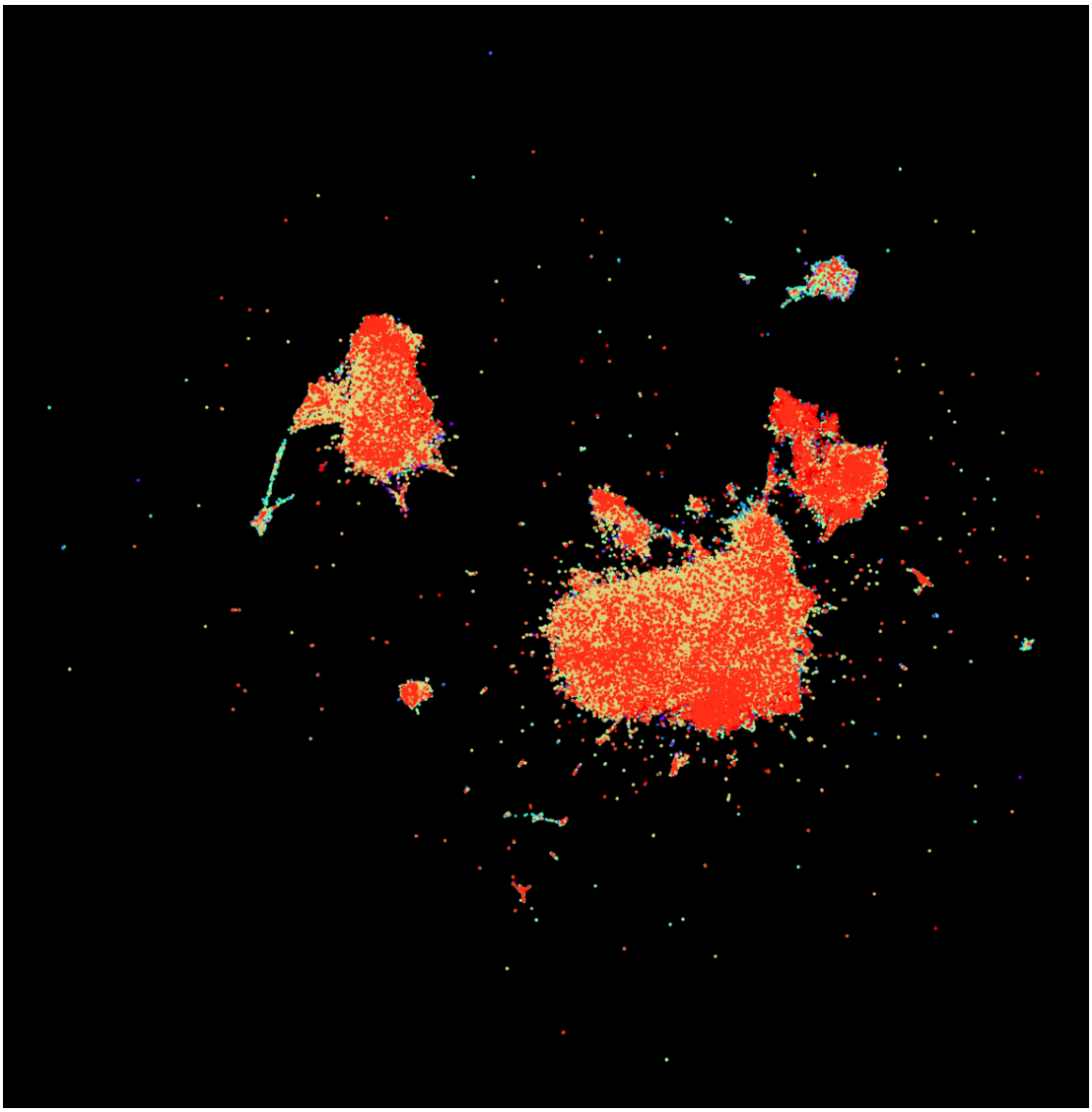

Figure 4.10: UMAP embeddings of submission title CLS token embeddings, coloured by source subreddit. BERT embeddings vary across source subreddits, with the notable exception of posts from r/Quebec, which colocate more strongly due to linguistic variation.

It was observed that users who are heavily active in alt-right Canadian subreddits, such as r/Canada_First and r/metacanada [Balgord and Zhou, 2017][Milton, 2018], tend to be close to one another, shown in Figure D.3. Inspecting groups of user accounts that were nearly overlapping revealed one instance where three user accounts were created within a short time of one another, posted similar material, and were active in similar communities. This was interpreted as these accounts being linked, potentially through a common operator. Concrete attribution, however, is difficult to attain without external validation, such as via internal user IP logs from Reddit servers (data which is not available for research). There is no specific restriction on having multiple Reddit accounts, so the similarity in users accounts does not indicate any suspicion of wrongdoing. This finding does however, demonstrate how these user representations might be used to form investigative leads for coordinated inauthentic behaviour [Facebook, 2020], where multiple accounts are controlled by a single operator.

Exploring the HDBSCAN EOM clustering of 3-dimensional meta-embeddings of users alongside labelled OIC Reddit Data (Corpus I from the previous chapter) resulted in 3 clusters, displayed in Figure 4.13. The largest cluster is composed of a diverse range of English-language titles. The second largest cluster contains a broad assortment of French-language titles. The last — and smallest — cluster, is observed to be a "spam cluster". The spam cluster is the smallest cluster, and contains the highest relative proportion of OIC accounts. Content within this cluster is entirely English-language, but is focused on tech news articles and promotion of specific websites. Of the 20 accounts within this cluster, 16 of them were among those banned and preserved as part of the Reddit transparency report account disclosure. The remaining four are users from the 2019 Canadian subreddit dataset. Upon revisiting these accounts several months later, it was discovered that all four of these users had also been banned, raising the proportion of banned users in this cluster to 100%. This indicates that the embedding and dimensionality reduction technique in this chapter, chained with assessment of small outlying user clusters, is a promising method for detection of accounts that exhibit spam behaviour deemed unsuitable by social media platforms.

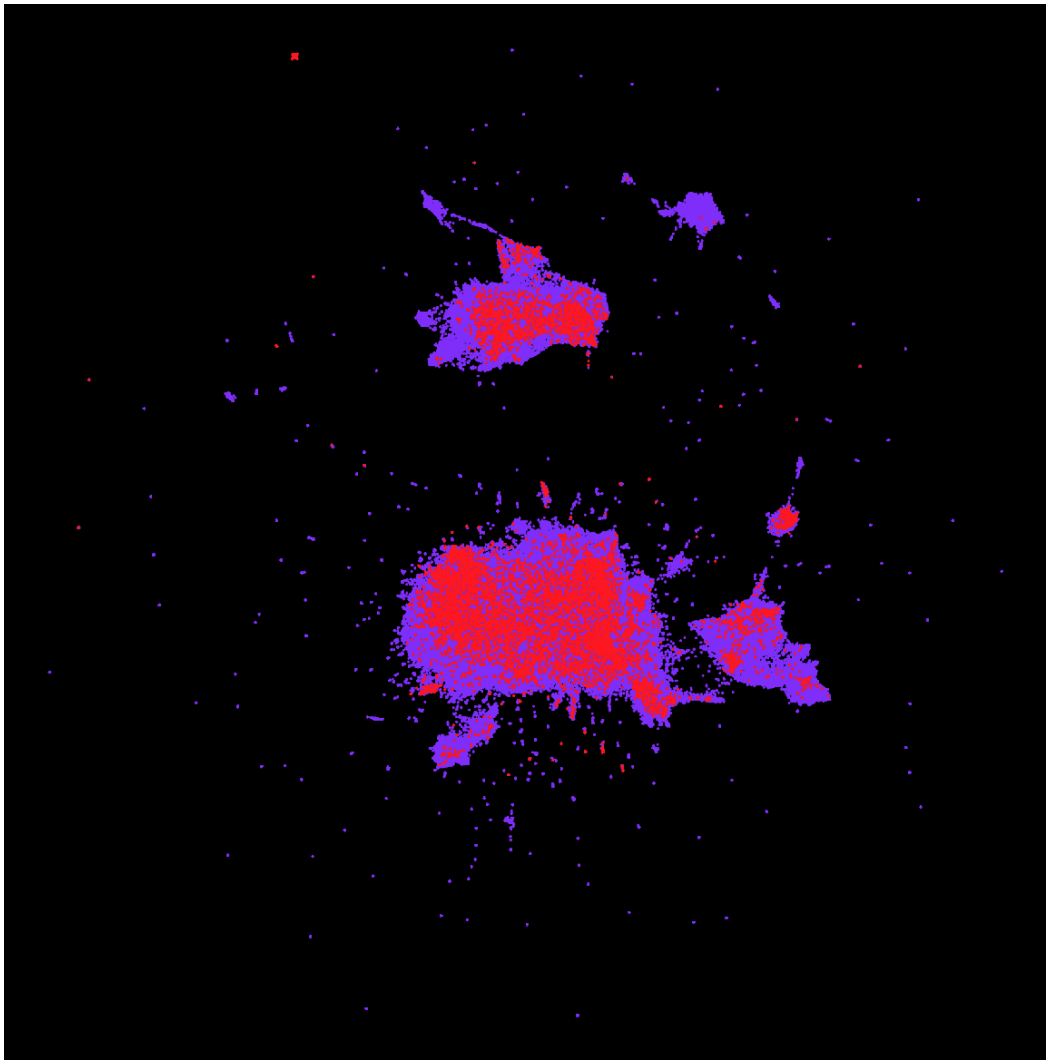

Figure 4.11: Overlay of labelled Reddit OIC title embeddings (red) over titles in Canadian subreddit (purple). OIC titles are scattered throughout the space, though there exist some types of titles that are very infrequent, such as those in French.

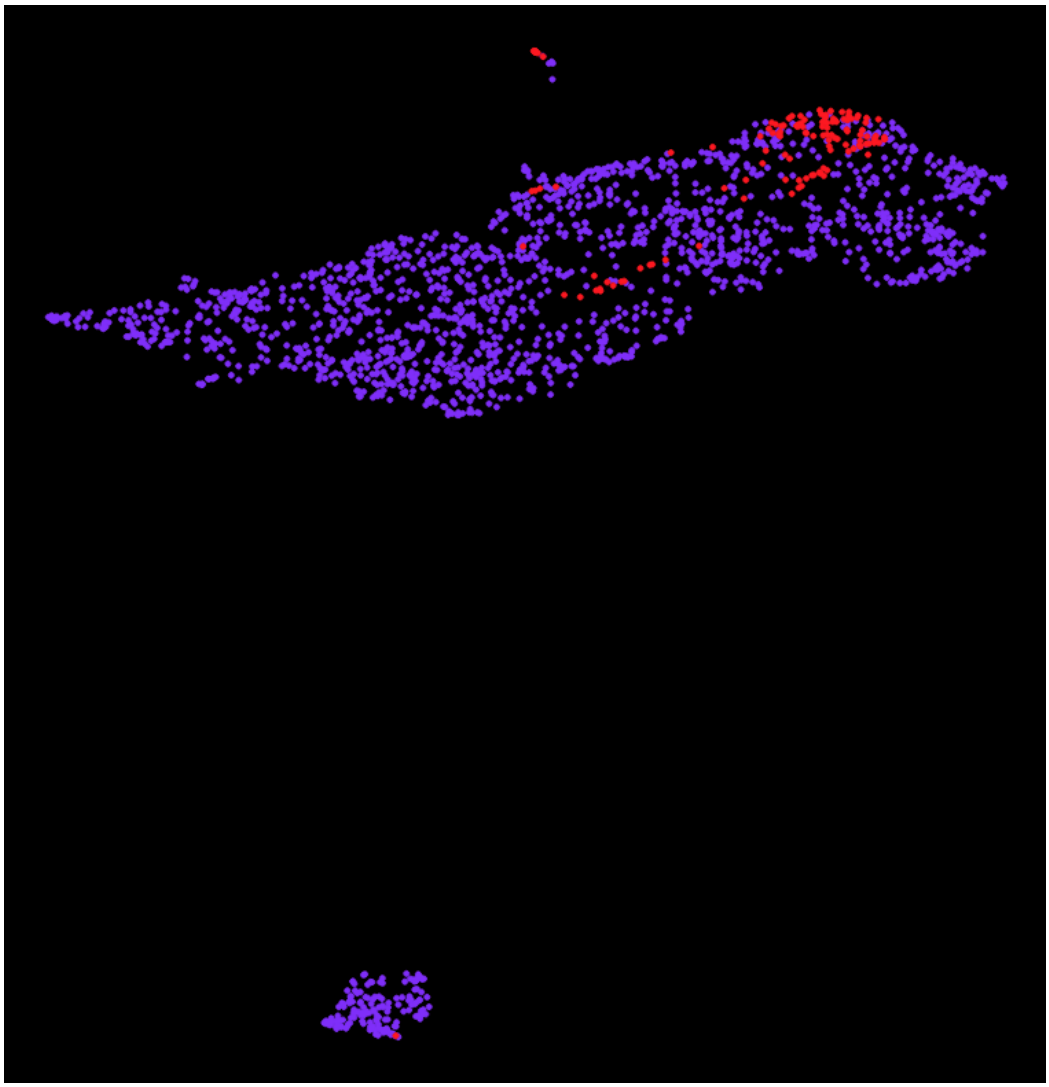

Figure 4.12: Overlay of labelled Reddit OIC user embeddings (red) over user embeddings from Canadian subreddits (purple). As hypothesized, we find OIC users are better colocated in the user embedding space than the title embedding space (which can be seen in Figure 4.11)

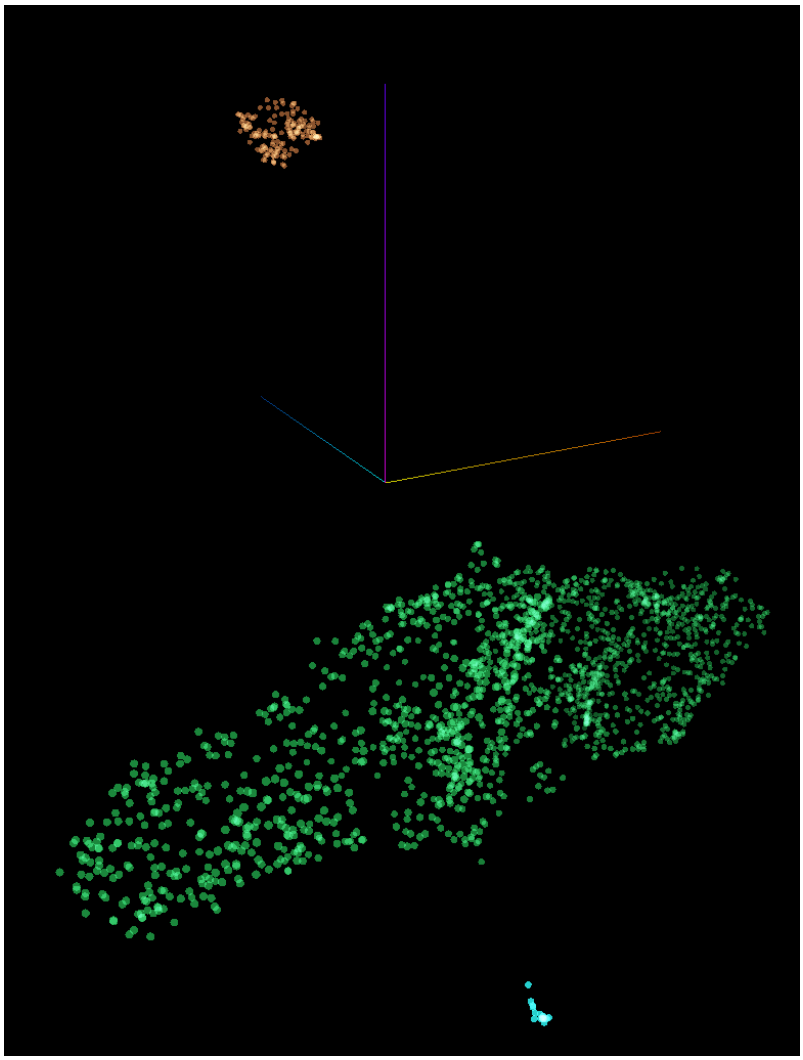

Figure 4.13: UMAP embedding of BERT CLS representations of submission titles in Canada-oriented subreddits along with ground-truth, coloured by HDBSCAN EOM cluster. Orange represents the French-language cluster, green represents the English-language cluster, and blue represents the "Spam" cluster

Figure 4.14: Left: Cluster of spam accounts with names (coloured by proximity to "AngelaSmith345") Right: Newly banned accounts highlighted in red; blue accounts are linked to labelled IRA influence campaigns

Users within the spam cluster exhibit very similar posting behaviour, and share significant resemblance in naming convention. We theorize that the usernames of these accounts were likely generated automatically by selecting two names from a list of common names, and optionally appending a number. The four recently-banned accounts are named "AngelaSmith345", "VickyXue", "angelashoe", and "kalvinjon8". A close-up view of these user embeddings can be found in Figure 4.14.

The commonality in shared material and proximity to IRA accounts, indicates that these accounts have common features to similar accounts used during the 2016 IRA influence campaign. This does not necessarily mean that these accounts are owned and operated by the IRA. They may, for example, belong to a spam network that sold accounts to the IRA at some point in time. Alternatively, it may be that these accounts are operated using the same software utilities, or simply that spam tends to strongly resemble other spam — much more so than regular discourse.

Whether or not there is a direct link between the IRA accounts and these new spam accounts, the separation of a cluster that exhibits suspicious characteristics is of benefit to OIC detection efforts, and will likely be useful in analyzing future campaigns.

4.5 Concluding Remarks

In this chapter we demonstrated a novel approach for using BERT embeddings and dimensionality reduction to create processes that assist OIC detection. We have shown that even

rudimentary meta-embedding approaches can be effective at combining Transformer sentence embeddings into user embeddings that are assessed quantitatively and qualitatively to have desirable properties under an applied OIC detection setting for the 2019 Canadian Federal Election.

Future work in this area should incorporate additional pre-processing steps to separate English and French language content for different processing pipelines, as the presence of French text created large variations that impacted the granularity of the HDBSCAN clustering process in the excess-of-mass (EOM) setting. French pre-trained Transformer models, such as CamemBERT [Martin et al., 2019] could be used to better investigate these French-language submission titles separate from the English-language analysis.

The text produced as part of online influence campaigns can vary significantly over the course of a campaign's operation [Reddit, 2018b], as well as between disparate campaigns [Twitter, 2019], impacting text representations accordingly. This type of variation in a target variable is often referred to in machine learning as "concept drift" [Sammut and Harries, 2017]. OIC campaigns are expected to undergo gradual concept drift as topics on social media change and OIC goals are adjusted, while also undergoing the abrupt shifts resulting from new influence campaigns by new operators, or sudden changes in tactics. Analysis of concept drift within an online influence campaign, and comparison of content in disparate OICs, may improve the applicability of past OIC data to future campaigns, and assist in the development of more robust detection methodologies.

While there is grammatical and semantic information on the sentence embedded in every token representation generated by BERT[Clark et al., 2019][Rogers et al., 2020][van Aken et al., 2019], the `[CLS]` tokens are of particular use as they represent the complete context of the sentence [Devlin et al., 2018]. For future development, however, it is likely unnecessary to choose a single token representation, as the Transformer architecture might rather be integrated into a larger model that integrates additional temporal and metadata features, as in parallel work on user representations [Andrews and Bishop, 2019].

Chapter 5

Conclusion

We conclude that the use of content-based features for OIC detection without safeguards creates the potential for discrimination against users of specific language backgrounds, especially when they are engaged in speech that contains common (often political) named entities from past influence campaigns. As protection of genuine free expression of political opinions on the Internet is a value of many organizations and governments, designers of online influence detection models should consider constructing test datasets of L2 English speakers using contextual data clues, such as flair or IP address, for the purpose of identifying avenues of discrimination. While some measurable bias towards detection of users who speak the same L1 language as the target distribution may be inevitable, this behaviour should be tracked and mitigated whenever possible. Future development of improved contextual and content-based influence campaign detection methods should be done with minority language communities in mind to prevent large-scale discrimination.

We further conclude that Transformer sentence embeddings of user-generated text can be combined to create effective user representations that can be used to immediately surface suspicious behaviour in real online communities, and demonstrate the deployment of such a system. We find that even simple methods of combining multiple representations to create a single user representation dramatically improves utility for human analysts in an applied setting. Leveraging more specialized content-based models, such as entity sentiment detection, or separating the input based on source language, would likely improve the content-based elements of this approach. Furthermore, combining content representations with additional features based on the available metadata, such as time of day or distribution of activity by community, would likely dramatically improve the quality of user representations for OIC detection.

5.1 Contributions

In this paper we provided the following contributions:

- **An evaluation process for assessing OIC detection models for discrimination against L2-English language communities.** The ethical implications of automated suppression of political speech are important and consideration of this should be integrated into the development of future solutions. By considering a dataset segmented by the native language of the author, we have a new way of evaluating the fairness of models, and a possible metric to be used in new fairness constraint systems such as those underlying TensorFlow Constrained Optimization (TFCO) [Cotter et al., 2018a][Cotter et al., 2018b][Narasimhan et al., 2019].
- **Novel unsupervised OIC detection methodology that integrates Transformer embeddings and meta-embeddings.** We discover that [CLS] tokens hold more utility than mid-sequence tokens for clustering qualitatively similar users together due to syntactic information embedded within BERT activations, and that BERT representations can be effectively combined to form simple and effective user representations. This investigation culminated in the creation of a process to assist with characterization of social media users in the context of searching for election interference.

5.2 Future Work

Detection of online influence campaigns is a field that is likely to change rapidly over the coming years as new advances in detection prompt corresponding advances in evasion. Successful classifiers will likely require diverse content-based and metadata features to attain a solution that is both effective and ethical. Machine learning provides a number of improved offensive and defensive capabilities and is likely to dramatically change both how these operations are conducted and how they are detected. To that end, it is useful to consider both the likely direction of online influence campaigns offensively, as well as future work on defenses as they apply to current campaigns — and whether they might also apply to future campaigns as well.

5.2.1 The Role of AI in Online Influence Campaigns

At the time of writing, the threat of coordinated disinformation is most heavily posed by accounts run by a mixture of humans and relatively simple automated systems. Neural fake news, such as that anticipated by detection models like Grover [Zellers et al., 2019], has yet to emerge as a significant problem, and likely will remain a limited threat for at least some time. The infrastructure and training overhead of using a computer to write a fake news article and then review it (as opposed to simply having a farm of human employees to do it), particularly when combined with the tedium of automating domain purchases and website configuration, is likely not a scalable disinformation approach. More concerning is the ease with which language models could be used to generate large numbers of fake comments for influence accounts, automating social media feeds that combine normal human behaviour with political or commercial messaging. Such a phenomenon would be concerning, particularly as short text sequence lengths have been found to be more challenging to detect [Radford et al., 2018][Zellers et al., 2019].

5.2.2 Advancements in OIC Detection Methodology

The evaluation of language classification models with regard to ethical considerations is a challenging and worthwhile area of research. Broader analysis that evaluates a greater number of classifiers, and deeper research on mitigating discrimination in the domain of influence campaign detection, would be a promising direction for future research.

The study of linguistic features that accurately target deceptive or manipulative behaviours may assist in addressing the ethical concerns highlighted in this paper. A related field, the detection of hate speech inciting violence, is another worthwhile area of future development both in improving detection methods and in answering ethical questions around the classification boundary between personal opinion and hate speech.

Due to the observed linguistic similarities between English sentences by self-identified Russian users and sentences from influence accounts, our results may be interpreted in the context of attribution of influence campaigns to their originators. Such an analysis is beyond the scope of this paper. As the existing online influence campaign datasets released by Twitter [Twitter, 2019] and Reddit [Coscia, 2018][Reddit, 2018b] do not contain any IP information that may be approached using geolocation techniques, it is difficult to independently determine the origin of suspect accounts. A more thorough linguistic

analysis similar to that performed by Goldin et al. [Goldin et al., 2018] may provide more insight into online influence campaigns on this platform. Furthermore, this could allow for detection of future trends, such as recruitment of native language speakers, or the use of generative text models, such as GPT-2 [Radford et al., 2018].

A clear and substantial area of improvement lies in fine-tuning the BERT encoder on the body of Reddit comment data specifically, or training a custom architecture from scratch for this purpose. This was not done during the course of this analysis due to processing constraints and a desire to demonstrate the findings on the most basic and accessible version of the Transformer architecture (as discussed in §2.2.1), but this will likely be a component of future research. Further computational power would also allow a deeper investigation of clustering hyperparameters and pre-processing steps for dimensionality reduction, which were beyond the scope of these studies.

While this research was focused primarily on Transformer embeddings for content-based analysis, there are still numerous other possible approaches that leverage content. One idea within this area is searching for the presence of grammatical and spelling errors within content produced by individuals operating out of countries that do not speak English natively. It may be possible to improve the accuracy of detection methods by extracting features for detecting L2 English. However, this kind of approach should be considered with the highest level of caution, as it would likely be at high risk of unfairly targeting users of specific combinations of viewpoints and first languages.

Aside from the content, a large number of metadata features may also be useful in more accurate detection of online influence accounts. While no single feature alone is likely to provide a comprehensive model for detecting and eliminating this form of behaviour, an ensemble model is likely to be highly effective. Masking or removing any 'tells' across some or all of these areas is likely to involve considerable effort and cost, which reduces the prevalence and effectiveness of inauthentic behaviour.

- Time of day of most frequent activity. For influence accounts run by human operators, they are most likely vulnerable to some form of fingerprinting based on daily activity, particularly those that have employees concentrated in a specific geographic area.

- Profiling of active communities. The provided Reddit data indicates that many of the suspect accounts are disproportionately active in specific communities, some of which appear to be unrelated to their online messaging goals. This potentially provides some insight into the interests and mentalities of the operators running the

account. The ground-truth data in the Reddit analysis showed a strong skew towards cryptocurrency, online gambling, and the video game "Counter-Strike: Global Offensive".

- Grouping users based on account creation timestamps and performing username similarity analysis. Many of the provided ground-truth accounts, particularly the group which was active within the r/worldnews community, showed very strong username similarity with tightly grouped account creation timestamps. These kinds of accounts could likely may have been purchased from an individual who had created them according to a particular name generation scheme. Detecting these blocks of procedural registered accounts could assist in reducing coordinated influence activity.

- Deeper analysis of connections to confirmed malicious activity on other platforms or websites. Twitter, has for example, released an enormous dataset of accounts they have assessed to be linked to nation-state influence campaigns. Individuals who frequently link to these accounts may be heavily influenced by, or a part of, these campaigns. Linking to websites known to be operated or endorsed by nation-state influence campaigns may be a potentially useful feature.

- Outlier detection in user self-identification groups. For example, the r/AskEurope subreddit allows a user to personally affiliate themselves with a country by choosing a "flair" to appear next to their name. The selection of affiliation made by an influence account may create a mismatch between their social media activity and social media activity typical of users from that country, causing them to appear as an outlier from the population.

Careful selection of features to determine those that change the least between influence campaigns is therefore an interesting area of future development, though will likely require cooperation of social media platforms to release more detailed internal data, presenting significant challenges to open research.

In addition to the above listed features, an individual with access to internal Reddit data would, for example, be able to leverage the following features:

- IP addresses. IP addresses contain useful geolocation information. Even if disguised through the use of proxies or VPN, this may create commonalities with other

suspected malicious accounts. The practical near intractability of faking the natural usage of thousands of residential IP addresses without aberrations or cross-contamination is likely a source of significant information gain.

- Vote manipulation. On sites such as Reddit where positive voting actions influence visibility, it is possible that bad actors may attempt to game the system by operating multiple accounts which can promote one another's posts. This creates opportunities to link disparate accounts together, as well as detect unnatural rates in the increase of votes given a particular piece of content.

By leveraging many features together in an intelligent fashion, strong user representations might be created from the available social media data, substantially accelerating defensive efforts against online influence campaigns. High-capacity NLP representation models, such as Transformers, will likely be instrumental to this development. Through thoughtful consideration and rigorous evaluation, development of these models can be directed in a manner that is both effective towards mitigating a serious threat to democratic societies, while protecting the well-being and rights of Internet users.

APPENDICES

Appendix A

Reddit Data Schema

The Pushshift dataset contains all Reddit data submissions and comments from June 2005 onward, and is regularly updated. According to the published paper on this dataset, as of April 2019 it contained 651,778,198 submissions and 5,601,331,385 comments posted across 2,888,885 subreddits [Baumgartner et al., 2020].

The significant fields within the submissions and comments from the Pushshift dataset are presented in the following tables for reference as to what features are available to researchers using Reddit data from Pushshift.

Field	Description
id	The submission's identifier, e.g., "5lcgjh" (String).
url	The URL that the submission is posting. This is the same with the permalink in cases where the submission is a self post. E.g., "https://www.reddit.com/r/AskReddit/
permalink	Relative URL of the permanent link that points to this specific submission, e.g., "/r/AskReddit/comments/5lcgj9/what_did_you_think_of_the_ending_of_rogue_one/" (String).
author	The account name of the poster, e.g., "example_username" (String).
title	The title that is associated with the submission, e.g., "What did you think of the ending of Rogue One?" (String).
created_utc	UNIX timestamp referring to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the submission is posted. Note that it excludes the prefix /r/. E.g., 'AskReddit' (String).
subreddit_id	The identifier of the subreddit, e.g., "t5_2qh1i" (String).
selftext	The text that is associated with the submission (String).
num_comments	The number of comments associated with this submission, e.g., 7 (Integer).
score	The score that the submission has accumulated. The score is the number of upvotes minus the number of downvotes. E.g., 5 (Integer).
is_self	Flag that indicates whether the submission is a self post, e.g., true (Boolean).
over_18	Flag that indicates whether the submission is Not-Safe-For-Work, e.g., false (Boolean).
distinguished	Flag to determine whether the submission is distinguished by moderators. "null" means not distinguished (String).
edited	Indicates whether the submission has been edited. Either a number indicating the UNIX timestamp that the submission was edited at, "false" otherwise.
domain	The domain of the submission, e.g., self.AskReddit (String).
stickied	Flag indicating whether the submission is set as sticky in the subreddit, e.g., false (Boolean).
locked	Flag indicating whether the submission is currently closed to new comments, e.g., false (Boolean).
quarantine	Flag indicating whether the community is quarantine, e.g., false (Boolean).
hidden_score	Flag indicating if the submission's score is hidden, e.g., false (Boolean).
retrieved_on	UNIX timestamp referring to the time we crawled the submission, e.g., 1483228803 (Integer).
author_flair_css_class	The CSS class of the author's flair. This field is specific to subreddit (String).
author_flair_text	The text of the author's flair. This field is specific to subreddit (String).

Table A.1: Submissions data description, adapated from Pushshift documentation [Baumgartner et al., 2020]

Field	Description
id	The comment's identifier, e.g., "dbumnq8" (String).
author	The account name of the poster, e.g., "example_username" (String).
body	The comment's text, e.g., "This is an example comment" (String).
link_id	Identifier of the submission that this comment is in, e.g., "t3_5l954r" (String).
parent_id	Identifier of the parent of this comment, might be the identifier of the submission if it is top-level comment or the identifier of another comment, e.g., "t1_dbu5bpp" (String).
created_utc	UNIX timestamp that refers to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the comment is posted. Note that it excludes the prefix /r/. E.g., 'AskReddit' (String).
subreddit_id	The identifier of the subreddit where the comment is posted, e.g., "t5_2qh1i" (String).
score	The score of the comment. The score is the number of upvotes minus the number of downvotes. E.g., 5 (Integer).
distinguished	Flag to determine whether the comment is distinguished by the moderators. "null" means not distinguished (String).
edited	Flag indicating if the comment has been edited. Either the UNIX timestamp that the comment was edited at, or "false".
stickied	Flag indicating whether the submission is set as sticky in the subreddit, e.g., false (Boolean).
retrieved_on	UNIX timestamp that refers to the time that we crawled the comment, e.g., 1483228803 (Integer).
gilded	The number of times this comment received Reddit gold, e.g., 0 (Integer).
controversiality	Number that indicates whether the comment is controversial, e.g., 0 (Integer).
author_flair_css_class	The CSS class of the author's flair. This field is specific to subreddit (String).
author_flair_text	The text of the author's flair. This field is specific to subreddit (String).

Table A.2: Comments data description, adapted from Pushshift documentation [Baumgartner et al., 2020]

Appendix B

SpaCy Annotation Specification

During named entity recognition processes, we use the spaCy model trained on the OntoNotes 5 dataset. Here is provided the description of the fields returned from this library.

Type	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Table B.1: Named-entity recognition specification for spaCy OntoNotes 5 Model [Explosion, 2020]

Appendix C

Visualization Comparison

The work in this thesis uses universal manifold approximation and projection (UMAP) [McInnes and Healy, 2018][McInnes et al., 2018] to perform dimensionality reduction for visualization. This algorithm is selected because of the balance between global and local structure in the resulting embedding, which we consider advantageous for human visual analysis.

In order to illustrate the advantages of this particular dimensionality reduction method, this appendix contains results from preliminary visualization experiments comparing projection techniques for BERT sentence representations using sentences from the Reddit r/syriancivilwar community, with overlaid OIC ground-truth data. We compare results from UMAP to t-distributed stochastic neighbour embedding (t-SNE) [van der Maaten and Hinton, 2008] as well as 2-dimensional principal component analysis (PCA).

The /r/syriancivilwar subreddit [Reddit, 2020], is dedicated to discussion of ongoing civil conflict in Syria. Concern has been expressed by Reddit users and moderators within this community that they may be the target of ongoing nation-state efforts to influence popular opinion in order to support political goals. Reddit CEO Steve Huffman replied to community concerns on April 10, 2018 with the comment "[/r/syriancivilwar] is on our radar for a variety of reasons and we're investigating" [Reddit, 2018a]. The subreddit has a very specific area of discussion and is heavily moderated, reducing variance in topic and user behaviour.

In the following projections, OIC sentences are represented in yellow, while sentences from other users are in blue. As OIC users post a variety of sentences on a variety of topics, the BERT embeddings of their sentences are expected to vary substantially.

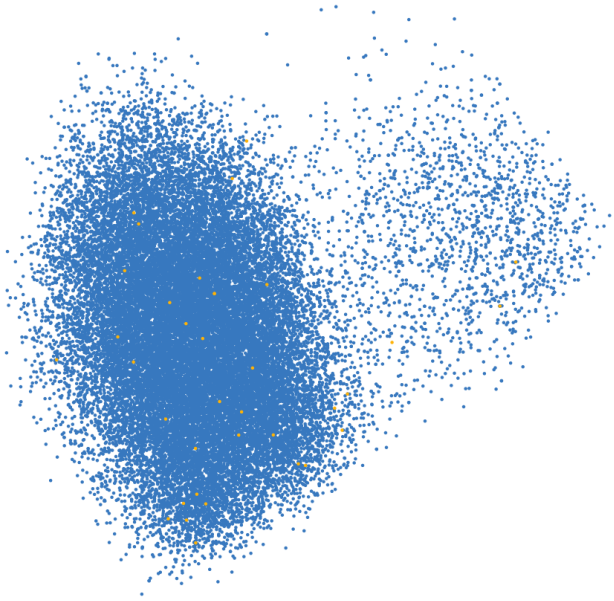

Figure C.1: PCA plot of the top 2 principal components in BERT features, responsible for ≈18.2% of the variance. Some separation of the data is observed, but comparatively little interpretable information is present in point coordinates.

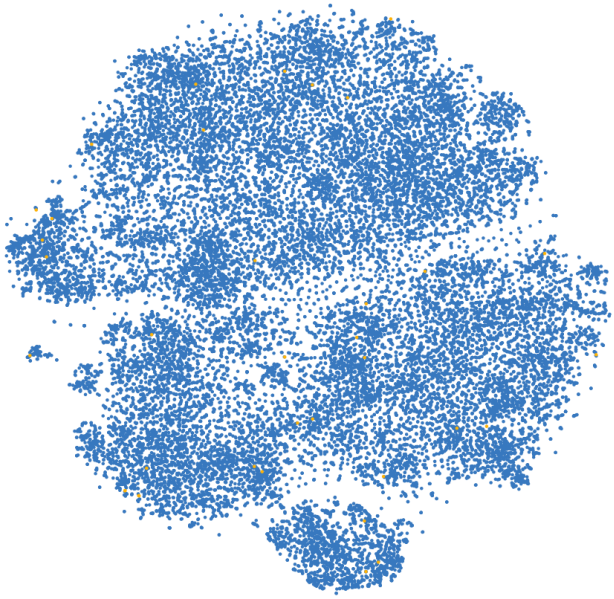

Figure C.2: T-SNE clustering of top 50 principal components of BERT features. This projection is a dramatic improvement over the 2-dimensional PCA projection. Local structure is observed, as similar vectors draw closer together and higher-density areas emerge.

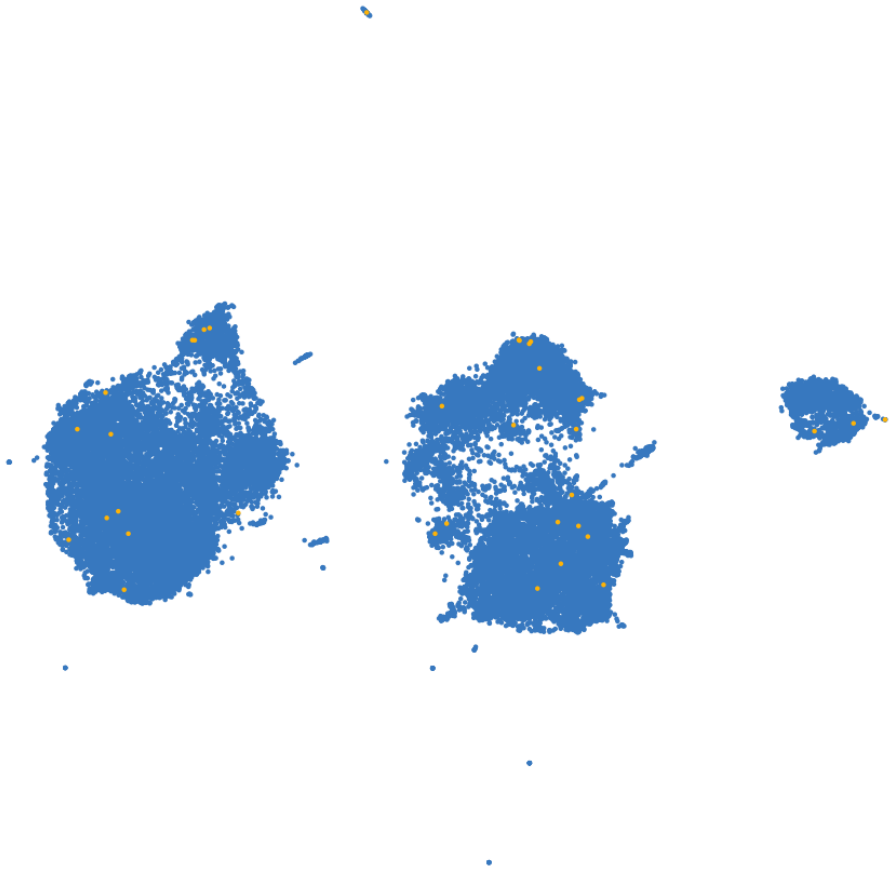

Figure C.3: UMAP visualization of BERT features. In addition to local structure, we observe the emergence of global structure, where distances between groupings of points becomes semantically meaningful. Related groups appear nearby one another, while more distinct groups appear further away. Protrusions of points are interpretable, and follow a "proximity is similarity" relationship.

From review of the available dimensionality reduction techniques and in light of these comparisons, it was assessed that UMAP was the projection method most suited for human analysis of Transformer representations for the purpose of facilitating OIC detection.

Appendix D

Additional Figures from Chapter 4

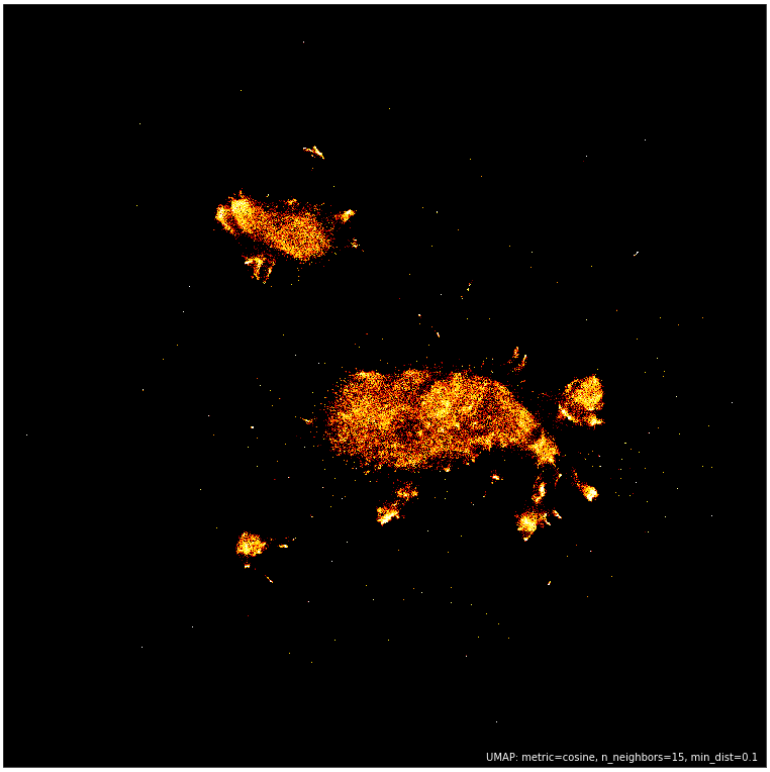

Figure D.1: UMAP embeddings of submission title `[CLS]` tokens, according to the process in §4.3.1, coloured by density. "Hotter" areas contain larger numbers of posts. Several distinct structures have formed, several of which are highly dense. OIC submission titles are expected to be dispersed throughout these structures, as they vary in syntactic and semantic content.

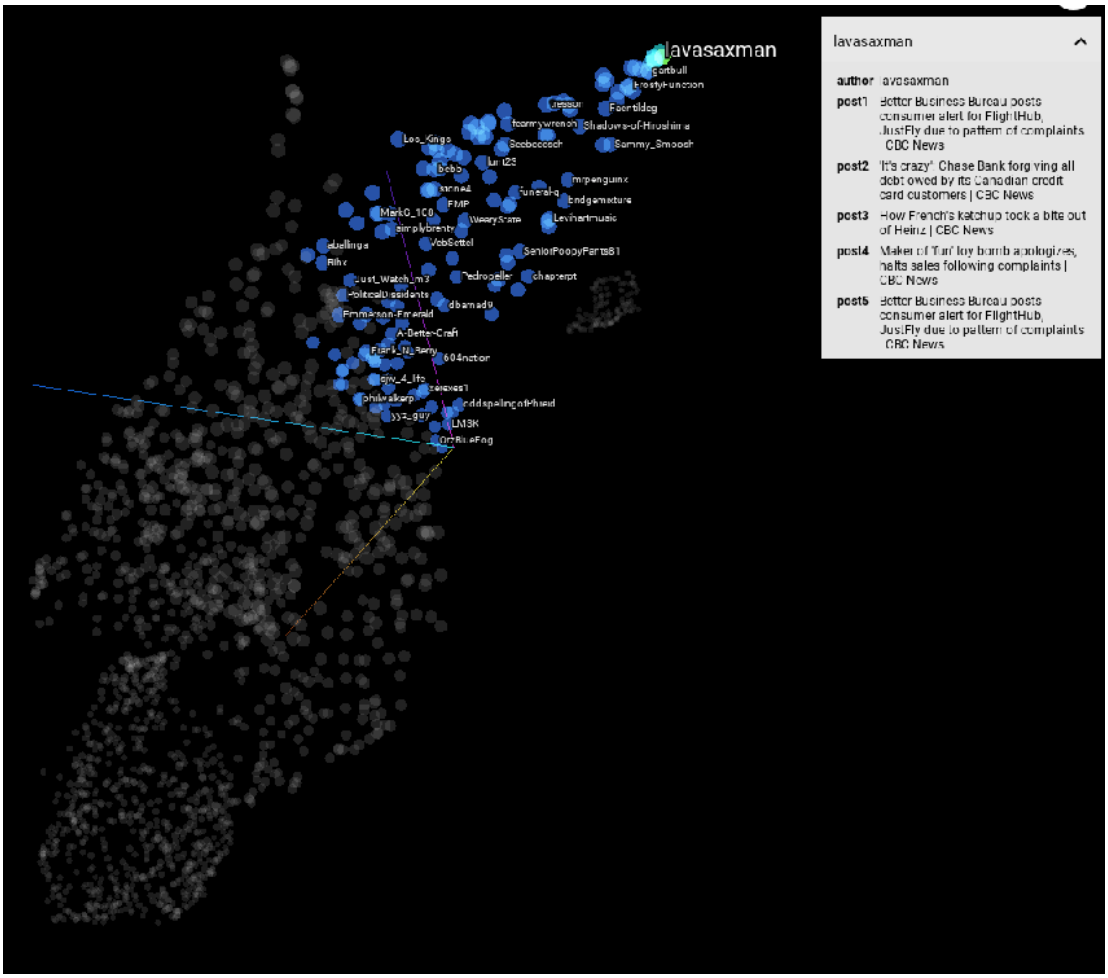

Figure D.2: Grouping of users who post primarily news articles. The "tip" of the area is occupied by nearly overlapping users who specifically share CBC news headlines.

Figure D.3: Grouping of users heavily active within alt-right Canadian subreddits.

Figure D.4: Full view of TensorBoard interface for the projector application.

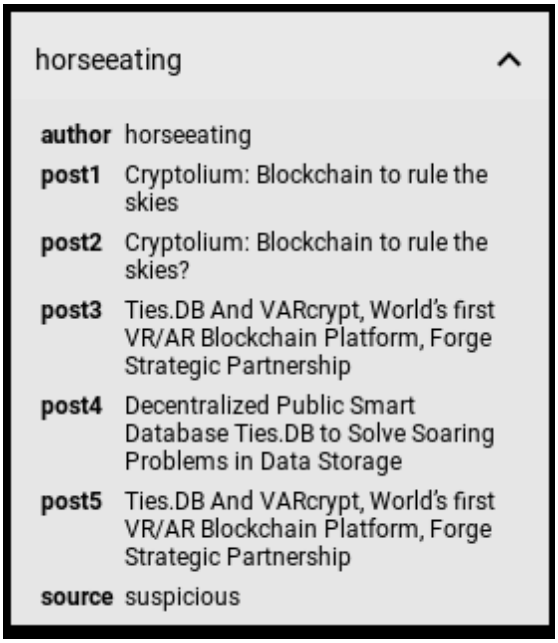

Figure D.5: Left: Screenshot of submission titles by recently banned spam account active in Canadian subreddits Right: Screenshot of submission titles by IRA influence account nearby modern spam accounts

Subreddit	Submission and Comment Count
canada	1 801 026
vancouver	690 945
CanadaPolitics	471 875
metacanada	316 732
ontario	290 296
Edmonton	247 054
onguardforthee	216 723
Quebec	190 293
alberta	183 231
canadaguns	147 384
CanadianForces	99 986
britishcolumbia	50 155
MetaCanadaTwo	13 013
metaquebec	8 198
Canada_First	6 225
Liberate_Canada	707
noBSCanada	318

Table D.1: Combined submission and comment counts across Canadian Reddit dataset

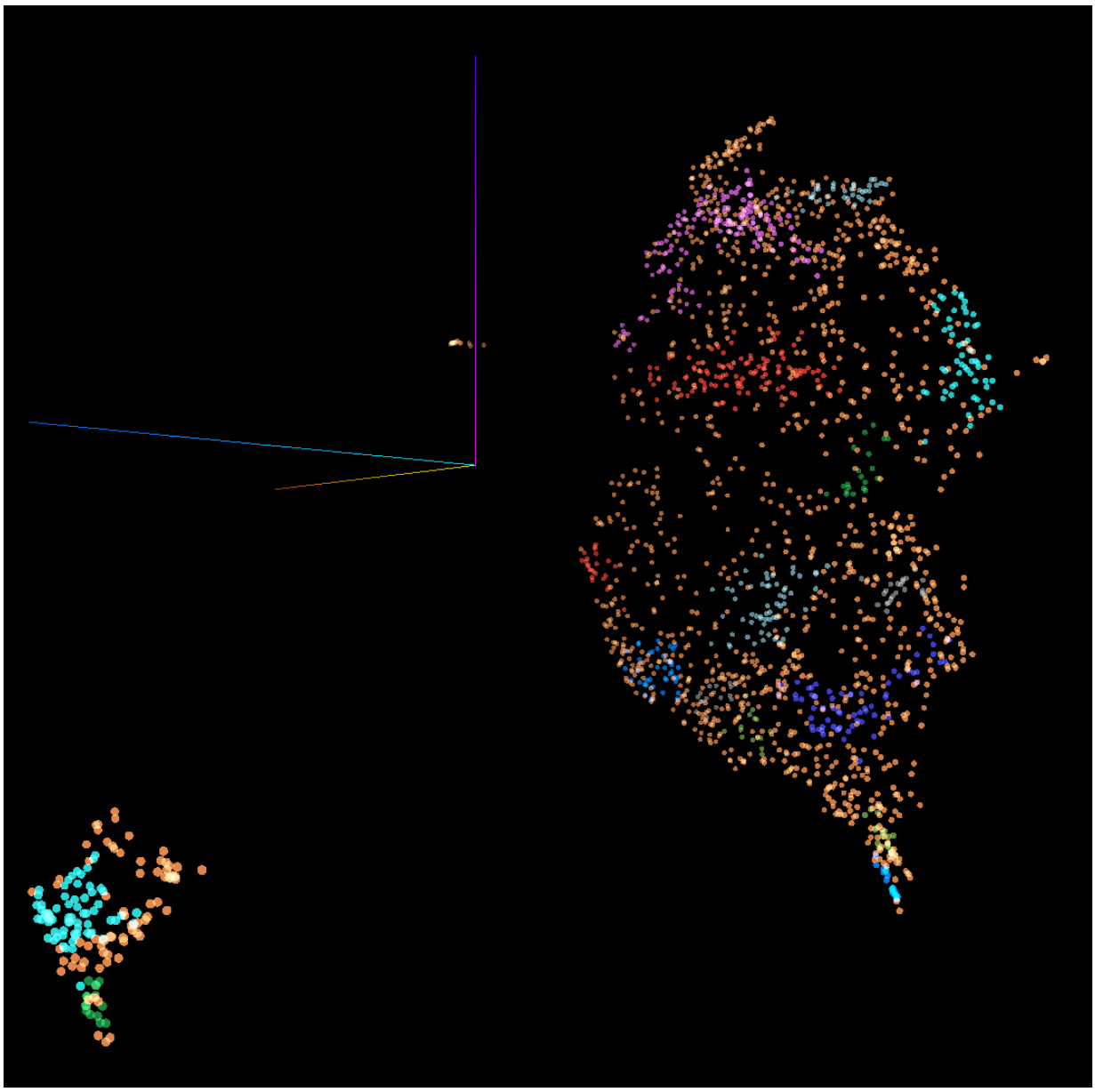

Figure D.6: Visualization of HDBSCAN leaf clustering in 3D projector view. Colours represent different cluster assignment. Orange represents the "background" cluster.

References

[Alba and Satariano, 2019] Alba, D. and Satariano, A. (2019). At least 70 countries have had disinformation campaigns, study finds.

[Alexa Internet, 2020a] Alexa Internet, I. (2020a). Canada alexa rankings. https://www.alexa.com/topsites/countries/CA. Accessed: 2020-03-19.

[Alexa Internet, 2020b] Alexa Internet, I. (2020b). United states alexa rankings. https://www.alexa.com/topsites/countries/US. Accessed: 2020-03-19.

[Amigó et al., 2009] Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

[Amigó et al., 2009] Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

[Andrews and Bishop, 2019] Andrews, N. and Bishop, M. (2019). Learning invariant representations of social media users. In *EMNLP/IJCNLP*.

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias.

[Antoun et al., 2020] Antoun, W., Baly, F., and Hajj, H. M. (2020). Arabert: Transformer-based model for arabic language understanding. *ArXiv*.

[Balgord and Zhou, 2017] Balgord, E. and Zhou, S. (2017). Conservative party leadership advisor helped create anti-islam organization.

[Baumgartner, 2019] Baumgartner, J. (2019). Reddit comment archive datasets. https://files.pushshift.io/reddit/comments/. Accessed: 2019-04-20.

[Baumgartner et al., 2020] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. *ArXiv*, abs/2001.08435.

[Baylies, 2020] Baylies, P. (2020). Pbaylies stylegan-encoder fork. https://github.com/pbaylies/stylegan-encoder. Accessed: 2019-10-03.

[BBC News, 2018] BBC News (2018). Russia 'meddled in all big social media' around us election.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

[Bokeh Development Team, 2019] Bokeh Development Team (2019). *Bokeh: Python library for interactive visualization*.

[Bouckaert and Frank, 2004] Bouckaert, R. R. and Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In Dai, H., Srikant, R., and Zhang, C., editors, *Advances in Knowledge Discovery and Data Mining*, pages 3–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Branco et al., 2016] Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2).

[Brocardo et al., 2014] Brocardo, M., Traore, I., Saad, S., and Woungang, I. (2014). Verifying online user identity using stylometric analysis for short messages. *Journal of Networks*, 9.

[Brooke and Hirst, 2012] Brooke, J. and Hirst, G. (2012). Measuring interlanguage: Native language identification with l1-influence metrics. In *LREC*.

[Brooking et al., 2020] Brooking, E. T., Kann, A., Rizzuto, M., Cole, R. T., and Gully, A. (2020). Dfrlab dichotomies of disinformation.

[Caliński and JA, 1974] Caliński, T. and JA, H. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.

[Clark et al., 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341.

[Coates and Bollegala, 2018] Coates, J. and Bollegala, D. (2018). Frustratingly easy meta-embedding - computing meta-embeddings by averaging source word embeddings. In *NAACL-HLT*.

[Coenen et al., 2019] Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. B., and Wattenberg, M. (2019). Visualizing and measuring the geometry of bert. In *NeurIPS*.

[Coldewey, 2018] Coldewey, D. (2018). How russia's online influence campaign engaged with millions for years.

[Coscia, 2018] Coscia, A. (2018). Reddit suspicious accounts dataset. https://github.com/ALCC01/reddit-suspicious-accounts. Accessed: 2019-04-20.

[Cotter et al., 2018a] Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. (2018a). Training well-generalizing classifiers for fairness metrics and other data-dependent constraints.

[Cotter et al., 2018b] Cotter, A., Jiang, H., Wang, S., Narayan, T., Gupta, M., You, S., and Sridharan, K. (2018b). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals.

[Crothers, 2019] Crothers, E. (2019). Code for classification experimental evaluation. https://github.com/ecrows/l2-reddit-experiment. Accessed: 2020-03-12.

[Crothers, 2020] Crothers, E. (2020). Reddit uk leak account data. https://github.com/ecrows/reddit-uk-leak-accounts. Accessed: 2020-03-12.

[Crothers et al., 2019] Crothers, E., Japkowicz, N., and Viktor, H. L. (2019). Towards ethical content-based detection of online influence campaigns. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Explosion, 2019] Explosion (2019). Spacy python library. https://github.com/explosion/spaCy. version 2.0.16.

[Explosion, 2020] Explosion (2020). Spacy annotation specification. https://spacy.io/api/annotation.

[Facebook, 2020] Facebook (2020). Community standards.

[Farzindar and Inkpen, 2020] Farzindar, A. A. and Inkpen, D. (2020). *Natural Language Processing for Social Media: Third Edition*. Morgan & Claypool.

[Fornacciari et al., 2018] Fornacciari, P., Mordonini, M., Poggi, A., Sani, L., and Tomaiuolo, M. (2018). A holistic system for troll detection on twitter. *Computers in Human Behavior*, 89:258–268.

[Gao et al., 2010] Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, page 35–47, New York, NY, USA. Association for Computing Machinery.

[Gehrmann et al., 2019] Gehrmann, S., Strobelt, H., and Rush, A. M. (2019). GLTR: statistical detection and visualization of generated text. *CoRR*, abs/1906.04043.

[Gencoglu, 2019] Gencoglu, O. (2019). Deep representation learning for clustering of health tweets. *CoRR*, abs/1901.00439.

[Goldin et al., 2018] Goldin, G., Rabinovich, E., and Wintner, S. (2018). Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3591–3601.

[Goldin et al., 2019] Goldin, G., Rabinovich, E., and Wintner, S. (2019). The reddit l2 corpus. http://cl.haifa.ac.il/projects/L2/. Accessed: 2019-04-20.

[Google, 2016] Google (2016). Embedding projector.

[Google, 2019] Google (2019). Embedding projector standalone code. GitHub repository.

[Google Research, 2019a] Google Research (2019a). Google research bert git repository. https://github.com/google-research/bert. Accessed: 2019-05-28.

[Google Research, 2019b] Google Research (2019b). Google sentencepiece git repository. https://github.com/google/sentencepiece. Accessed: 2019-05-28.

[Government of Canada, 2019] Government of Canada (2019). G7 rapid response mechanism.

[Hindman and Barash, 2018] Hindman, M. and Barash, V. (2018). Disinformation, 'fake news' and influence campaigns on twitter.

[HLEG, 2019] HLEG, A. (2019). Ethics guidelines for trustworthy ai.

[Hromic, 2019] Hromic, H. (2019). Extended bcubed python implementation. https://github.com/hhromic/python-bcubed. Accessed: 2020-03-28.

[Huang et al., 2016] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K. (2016). Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012.

[Im et al., 2019] Im, J., Chandrasekharan, E., Sargent, J., Lighthammer, P., Denby, T., Bhargava, A., Hemphill, L., Jurgens, D., and Gilbert, E. (2019). Still out there: Modeling and identifying russian troll accounts on twitter. *CoRR*, abs/1901.11162.

[Jack, 2017] Jack, C. (2017). Lexicon of lies: Terms for problematic information. *Data & Society*.

[Jones, 2018] Jones, R. (2018). Secret online influence campaigns are the future of american politics.

[Kaminski and Malgieri, 2019] Kaminski, M. and Malgieri, G. (2019). Algorithmic impact assessments under the gdpr: Producing multi-layered explanations. *SSRN Electronic Journal*.

[Karras et al., 2019a] Karras, T., Laine, S., and Aila, T. (2019a). A style-based generator architecture for generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Karras et al., 2019b] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2019b). Analyzing and improving the image quality of stylegan.

[Kennedy et al., 2019] Kennedy, S., Walsh, N., Sloka, K., McCarren, A., and Foster, J. (2019). Fact or factitious? contextualized opinion spam detection. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 344–350, Florence, Italy. Association for Computational Linguistics.

[Kiperwasser and Goldberg, 2016] Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

[Kyle and Crossley, 2014] Kyle, K. and Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

[Lan et al., 2019] Lan, Z.-Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

[Lim, 2019] Lim, G. (2019). *Disinformation Annotated Bibliography.* Munk School of Global Affairs.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *ACL Workshop on Effective Tools and Methodologies in NLP and Computational Linguistics*.

[Malmasi, 2016] Malmasi, S. (2016). Subdialectal differences in Sorani Kurdish. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 89–96, Osaka, Japan. The COLING 2016 Organizing Committee.

[Malmasi and Dras, 2014] Malmasi, S. and Dras, M. (2014). Arabic native language identification. In *EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 180–186, Doha, Qatar. Association for Computational Linguistics.

[Malmasi et al., 2017] Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., and Qian, Y. (2017). A report on the 2017 native language identification shared task. In *12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

[Martin et al., 2019] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, E. V., Seddah, D., and Sagot, B. (2019). Camembert: a tasty french language model. *ArXiv*, abs/1911.03894.

[McInnes and Healy, 2018] McInnes, L. and Healy, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426.

[McInnes et al., 2017] McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

[McInnes et al., 2018] McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[Milton, 2018] Milton, J. (2018). Canada's largest subreddit accused of harbouring white nationalists.

[Nadeau and Bengio, 2003] Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3):239–281.

[Narasimhan et al., 2019] Narasimhan, H., Cotter, A., and Gupta, M. (2019). Optimizing generalized rate metrics with three players. In Wallach, H., Larochelle, H., Beygelzimer, A., dÁlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 10747–10758. Curran Associates, Inc.

[Nguyen et al., 2019] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., and Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*.

[Nimmo, 2016] Nimmo, B. (2016). Identifying disinformation: an abc approach.

[O'Sullivan, 2019] O'Sullivan, D. (2019). Facebook announces first takedown of influence campaign with ties to saudi government.

[Oxford English Dictionary, 2020] Oxford English Dictionary (2020). Oxford english dictionary (online).

[Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

[Pedregosa et al., 2011] 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. *J. Mach. Learn. Res.*, 12:2825–2830.

[Perkins, 2018] Perkins, R. C. (2018). The Application of Forensic Linguistics in Cybercrime Investigations. *Policing: A Journal of Policy and Practice*.

[Punturo, 2018] Punturo, B. (2018). Reddit-trolls repository of brandon punturo on github. https://github.com/brandonjpunturo/Reddit-Trolls. Accessed: 2019-05-28.

[Punturo, 2019] Punturo, B. (2019). Predicting russian trolls using reddit comments. https://towardsdatascience.com/predicting-russian-trolls-using-reddit-comments-57a707653184. Accessed: 2019-04-20.

[Rabinovich et al., 2018] Rabinovich, E., Tsvetkov, Y., and Wintner, S. (2018). Native language cognate effects on second language lexical choice. *TACL*, 6:329–342.

[Radford et al., 2018] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018). Language models are unsupervised multitask learners.

[Reddit, 2018a] Reddit (2018a). Reddit 2017 transparency report and suspect account findings. https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/. Accessed: 2019-05-23.

[Reddit, 2018b] Reddit (2018b). Reddit transparency report: Suspicious accounts. https://www.reddit.com/wiki/suspiciousaccounts. Accessed: 2019-04-20.

[Reddit, 2020] Reddit (2020). Syrian civil war community on reddit. https://www.reddit.com/r/syriancivilwar/. Accessed: 2019-04-12.

[Reddit Security Team, 2019] Reddit Security Team (2019). Suspected campaign from russia on reddit. https://www.reddit.com/r/redditsecurity/comments/e74nml/suspected_campaign_from_russia_on_reddit/. Accessed: 2020-03-12.

[Reimers et al., 2019] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (2019). Classification and clustering of arguments with contextualized word

embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

[Rogers et al., 2020] Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *arXiv e-prints*, page arXiv:2002.12327.

[Rosales-Méndez and Ramírez-Cruz, 2013] Rosales-Méndez, H. and Ramírez-Cruz, Y. (2013). Cice-bcubed: A new evaluation measure for overlapping clustering algorithms. *Lecture Notes in Computer Science*, 8258:pp 157–164.

[Rosen, 2017] Rosen, A. (2017). Tweeting made easier: official blog post. https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html. Accessed: 2019-06-05.

[Russel, 2018] Russel, J. (2018). Josh russel reddit investigation. https://www.reddit.com/user/eye_josh/comments/843beq/russian_reddit_accounts_and_links/. Accessed: 2019-04-20.

[Sammut and Harries, 2017] Sammut, C. and Harries, M. (2017). *Concept Drift*, pages 253–256. Springer US, Boston, MA.

[Schrage and Ginsberg, 2018] Schrage, E. and Ginsberg, D. (2018). Facebook launches new initiative to help scholars assess social media's impact on elections.

[Selbst et al., 2019] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

[Shoeybi et al., 2019] Shoeybi, M., Patwary, M. A., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

[Siapka, 2018] Siapka, A. (2018). The ethical and legal challenges of artificial intelligence: The eu response to biased and discriminatory ai. SSRN.

[Singh et al., 2016] Singh, K., Shakya, H., and Biswas, B. (2016). Clustering of people in social network based on textual similarity. *Perspectives in Science*, 8.

[Smith, 2007] Smith, R. (2007). An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, page 629–633, USA. IEEE Computer Society.

[Song, 2019] Song, V. (2019). Spy used ai to create fake linkedin photo to fool targets, report finds.

[Strehl and Ghosh, 2003] Strehl, A. and Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

[Tolosana et al., 2020] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*.

[Turc et al., 2019] Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Well-read students learn better: The impact of student initialization on knowledge distillation. *ArXiv*, abs/1908.08962.

[Twitter, 2019] Twitter (2019). Twitter elections integrity dataset. https://about.twitter.com/en_us/values/elections-integrity.html. Accessed: 2019-04-20.

[U.S. Joint Chiefs of Staff, 2014] U.S. Joint Chiefs of Staff (2014). *Joint Publication 3-13: Information Operations*. United States. Joint Chiefs of Staff.

[van Aken et al., 2019] van Aken, B., Winter, B., Löser, A., and Gers, F. A. (2019). How does bert answer questions?: A layer-wise analysis of transformer representations. In *CIKM '19*.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

[van Rijsbergen, 1974] van Rijsbergen (1974). Foundation of evaluation. *Journal of Documentation*, 30:365–373.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Vig, 2019] Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

[Volchek and Coalson, 2018] Volchek, D. and Coalson, R. (2018). Kremlin 'trolls' meddling in russia's own elections? 'of course we are!'.

[Weller and Woo, 2019] Weller, H. and Woo, J. (2019). Identifying russian trolls on reddit with deep learning and bert word embeddings. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15739845.pdf. Accessed: 2019-05-22.

[Yang et al., 2012] Yang, C., Harkreader, R., Zhang, J., Shin, S., and Gu, G. (2012). Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 71–80, New York, NY, USA. Association for Computing Machinery.

[Yang et al., 2019] Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

[Zahn, 1971] Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20(1):68–86.

[Zellers et al., 2019] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

[Zhou et al., 2000] Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20 17:6594–611.

[Zhu et al., 2015] Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.