Deploying Artificial Intelligence to Combat Covid-19 Misinformation on Social Media: Technological and Ethical Considerations

Barry Cartwright
International CyberCrime Research Center, Simon Fraser University
barry_cartwright@sfu.ca

Richard Frank
International CyberCrime Research Center, Simon Fraser University
rfrank@sfu.ca

George Weir
University of Strathclyde
george.weir@strath.ac.uk

Karmvir Padda
International CyberCrime Research Center
Simon Fraser University
karmvir_padda@sfu.ca

Sarah-May Strange
International CyberCrime Research Center
Simon Fraser University
sarahmaystrange@gmail.com

Abstract

This paper reports on AI research into online misinformation pertaining to the COVID-19 pandemic within the Canadian context. This is part of our longer-term goal, i.e., development of a machine-learning tool to assist social media platforms, online service providers and government agencies in identifying and responding to misinformation on social media. We report on predictive accuracies accomplished by applying a combination of technologies, including a custom-designed web-crawler, The Dark Crawler, the Posit toolkit, and four different machine-learning models based on Naïve Bayes, Support Vector Machines, LibLinear and LibShortText. Overall, we found that Posit and LibShortText models showed higher levels of correlation to the pre-determined (manual and machine-driven) data classifications than the other machine-learning algorithms tested. We further argue that the harms associated with COVID-19 misinformation — e.g., the social and economic damage, and the deaths and severe illnesses — outweigh the right to personal privacy and freedom of speech considerations.

Keywords: COVID-19; misinformation; social media; machine-learning.

1. Introduction

This paper reports on the findings of research into online misinformation pertaining to the COVID-19 pandemic. The project was part of our longer-term goal, i.e., the development of an artificial intelligence (AI) tool to assist social media platforms, online service providers and government agencies in identifying and responding to misinformation on social media (Cartwright et al., 2019). The COVID-19 research — sponsored by the Canadian government's Digital Citizenship Cooperation Program — was conducted by the International CyberCrime Research Centre (ICCRC) at Simon Fraser University in Canada, in cooperation with the Department of Information and Computer Sciences at the University of Strathclyde in Scotland. As this was a Canadian-based project, funded by the Canadian government, the data that we analyzed was derived from Canadian social media sources.

This paper will focus on predictive accuracies attained by the Posit toolkit, a text-reading software solution designed by George Weir of Strathclyde University (Owoeye & Weir, 2018; Weir et al., 2016; Weir et al., 2018), in combination with J48 and Random Forest, when it comes to automated classification of real information and misinformation about COVID-19 on Canadian social media. We compare the Posit results to those attained by a variety of other machine-learning models, including Naïve Bayes (NB), Support Vector Machines (SVM), LibLinear (LBL) and LibShortText (LST). We also hope that our review of some of the legal precedents and ethical considerations will help to shed light on the complexities and pitfalls that legislators and regulators can expect to encounter when seeking to remediate the threat posed by COVID-19 misinformation on social media.

2. Framing the Problem

Over the past 20 years, people have come to rely increasingly on social media for their personal health and medical information (Suarez-Lledo & Alvarez-Garcia, 2021). Every time there is a new epidemic, social media becomes saturated with false claims about the suspicious origins of the disease, unproven (and

URI: https://hdl.handle.net/10125/102897
978-0-9981331-6-4

even harmful) cures and remedies, and conspiracy theories about biowarfare (Bernard et al., 2020).

Misinformation is simply incorrect or false information (Berghel, 2017; Jankwoski, 2018) and may not involve malicious intent. Individuals and groups that circulate misinformation may be misinformed themselves. Indeed, the anti-vaccination movement associated with COVID-19 is by no means "new" (Bester, 2016) — it has been alive and well on social media for decades (Kata, 2010), and can be expected to persist well beyond the present COVID-19 crisis (Wilson & Wiysonge, 2020).

COVID-19 misinformation has been described variously as an "infodemic," "misinfodemic" or "disinfodemic" — fast-spreading yet false information disseminated primarily on social media, which can cause serious harm by persuading people to act contrary to public health policies and regulations and scientific guidance in general (Krause et al., 2020; Posetti & Bontcheva, 2020; Yang et al., 2021). Despite growing awareness of the risks that COVID-19 poses to the public, online misinformation regarding the virus often drowns out more credible sources of information. Research indicates that one-out-of-three people reported having encountered false or misleading information about COVID-19 on social media during the early stages of the pandemic (OECD, 2020). Another study found that approximately 30% of people continued to believe false COVID-19 information that they found on social media, and more importantly, over 40% of them shared that information with others without considering whether or not the content was accurate (Pennycook et al., 2020). More importantly, this misinfodemic has led to unnecessary transmission and death rates, through the propagation of anti-science theories that portray COVID-19 as a "hoax," or that claim that social-distancing and mask-wearing are ineffective or assert that younger people are not at risk of infection.

According to the World Health Organization, as of June 2022 there had been over a half-billion confirmed cases of COVID-19 around the world, and well over six million COVID-related deaths (WHO, 2022). According to the Centers for Disease Control and Prevention (2022a), as of June 2022, there had been almost 85 million cases and over one million COVID-related deaths in the United States. In Canada, there had been over 3.5 million cases and over 41,000 COVID-related deaths as of June 2022 (Government of Canada, 2022). Thus, the magnitude of this problem should not be downplayed.

While effective vaccines for COVID-19 exist, there nevertheless remains a considerable amount of vaccine hesitancy (Centers for Disease Control and Prevention, 2022b; Lavoie et. al, 2022; StatCan COVID-19, 2021), not to mention resistance to preventative measures such as stay-at-home orders and mask-wearing mandates. Much of this reluctance or resistance can be traced back to the COVID-19 misinformation found on social media. It is essential that this problem be addressed, not only when it comes to online misinformation regarding the COVID-19 pandemic (Krishnan et al. 2021), but with future diseases and pandemics of this nature as well.

3. Methodology

3.1 The Research Sample

At the outset, the research team manually downloaded textual content from known social media sources of real information and misinformation about COVID-19, attempting to focus on Canadian content, in keeping with our research mandate. In the first instance, this included a dataset of 800 messages taken from four social media platforms: Facebook, Instagram, Twitter and Reddit. This was a purposive sample, intended to provide us with "ideal type" examples of real information and misinformation. From qualitative analysis of these messages, we identified key words and key phrases associated with real information and misinformation about COVID-19 and pinpointed other Canadian social media sites where greater amounts of such real information and misinformation could be found.

When focusing on real, or genuine, information about COVID-19 we identified specific hashtags, keywords and key phrases (Figure 1) for qualitative analysis. Similarly, we identified hashtags, key words and key phrases (Figure 2) that contained large amounts of "misinformation" about COVID-19.

We employed The Dark Crawler (TDC), a web-crawler designed by Richard Frank of SFU, to harvest 131,337 tweets, using the "real information" and "misinformation" hashtags, key words and key phrases described above, harvesting 86,586 "real information" tweets and 44,751 "misinformation" tweets, with a focus on Canadian social media content. These were organized in a preconfigured Excel spreadsheet that included the date harvested, the date and time stamp on the tweet, whether or not it was from a verified account, whether it was posted by a member or guest, the URL subforum, the hashtag, the user ID of the person posting the message, the original source of the post (if it came from another source), the title of the tweet, the textual content, and links to images, videos and memes and to external websites (if any existed).

While gathering misinformation from Reddit, the research team collected comparator datasets of similar size, drawn from what were determined through careful inspection to be sources of accurate information about

COVID-19. The purpose of this was twofold: (1) to train machine-reading algorithms to discern between real information and misinformation about COVID-19 as well as to evaluate the accuracy of their automated classification; and (2) to train members of the research team to discern between real information and misinformation about COVID-19, as well as evaluate the accuracy of their manual classification.

For Reddit, the team collected examples of misinformation from r/LockdownSkepticism, a sub-Reddit whose name speaks for itself, and real information from r/Coronavirus_BC, a sub-Reddit moderated by a doctor in British Columbia. Other misinformation Reddit sites that were sampled using ParseHub included r/antimaskers, r/CovidSkepticsCanada, r/NoNewNormal and Coronavirus (FOS) (with FOS being an acronym for freedom of speech), while other real information Reddit sites included r/COVID-19, r/COVID_CANADA, r/CanadaPolitics and r/CanadaCoronavirus, again to name a few.

The research team harvested, organized and saved 22,559 "real information" and 21,972 "misinformation" messages from Reddit using ParseHub. It cannot be said that these were random samples, as the searches were targeted, with the specific objective of generating large datasets of real information and misinformation about COVID-19, within the Canadian context.

Figure 1

Hashtags, Keywords and Key Phrases Associated with COVID-19 "Real Information"

a) Hashtags	b) Keywords		c) Key phrases
#BCCDC@CDCofBC	aerosol	protect	air travelers
#CORONAVIRUS	arrest	quarantine	care home(s)
#coronaviruscanada	CDC	restrictions	coronavirus pandemic
#CoronavirusVaccine	charges	rollout	COVID cases
#covax	COVID-19	transmission	COVID-19 cases
#covidcanada	curfew	vaccinated	COVID patients
#COVID19	deaths	vaccines	COVID-19 vaccine
#covid19	dying	variant	flatten the curve
#Covid-19	enforce	ventilator	government officials
#COVID19AB	exposure	violations	health authorities
#covid19bc	fined	WHO	ICU cases
#Covid19canada	fines		infection rate
#COVID19Ontario	ICU(s)		long-term care
#covid19vaccine	immunization		negative COVID test
#getvaccinated	infections		on a ventilator
#slowthespread	mandatory		public health crisis
#stayhome	measures		public health order
#stopcovid19	non-essential		safety regulations
#vaccine	pandemic		severe illness
	prevent		social distancing
	prevention		vaccination rate
			vaccination rollout
			wearing a mask

Figure 2

Hashtags, Keywords and Key Phrases Associated with COVID-19 "Misinformation"

a) Hashtags	b) Keywords	c) Key phrases
#antimask	anti-mask	big pharma
#covidhoax	chip	Bill Gates
#covidhoax2020	choice	corona hoax
#covidvaccinesideeffects	Chinavirus	Covid fraud
#Chinavirus	conspiracy	empty hospitals
#endthelockdown	depopulation	end the lockdown
#FilmYourHospital	draconian	false positives
#freecanada	fraud	great reset
#greatreset	freedom(s)	isolation camps
#hoax	Gates	male infertility
#hugsovermasks	globalists	mental illness
#LockddownsDon'tWork	hoax	mask mandate
#MasksDontWork	illuminati	naturally acquired immunity
#MasksOff	infertility	no vaccine
#NoMasksCanada	Kungflu	population control
#NoMoreLockdowns	lockdown(s)	oxygen deprivation
#nonewnormal	plandemic	vaccine industry
#NoVaccineForMe	rights	wake up
#plandemic	scam	world order
#scamdemic	scamdemic	
#thegreatreset	WuhanFlu	
#wedonotconsent		
#WuhanCoronavirus		

Figure 3

27 Features Extracted by the Posit Toolkit

adjective types	determiners	particle types	prepositions
adjectives	interjection types	particles	total unique words (types)
adverb types	interjections	personal pronoun types	total words (tokens)
adverbs	noun types	personal pronouns	type/token ratio
average sentence length	nouns	possessive pronoun types	verb types
average word length	number of characters	possessive pronouns	verbs
determiner types	number of sentences	preposition types	

Despite the targeted searches, manual inspection of random samples taken from the Reddit datasets resulted in 15.2% of the messages being classified as "not applicable," as they were not on the topic of COVID-19. Topics discussed in these messages included censorship, crimes, extremism, human rights, illegal cannabis, politics, protests, racial harmony, sexual harassment, and health-related issues such as how to spot the signs of a stroke, but not COVID-19. This is not regarded as a sampling problem, however, as it would be cumbersome to manually inspect 44,531 Reddit messages to ensure that they contained only COVID-19 content. Moreover, it was essential for the datasets to contain a certain amount of such information, as both manual assessment and automated machine-reading classification must be able to discern between real information about COVID-19, misinformation about COVID-19, and information that might be political or health-related but not about COVID-19.

3.2. Data-scraping with The Dark Crawler

The Twitter data for this study was harvested using The Dark Crawler (TDC), a web-crawling software tool developed by Richard Frank of Simon Fraser University's International CyberCrime Research Centre. TDC captures content from the open and Dark Web, as well as structured content from online discussion forums and social media platforms. TDC uses key words, key phrases, and other syntax to retrieve relevant pages; it analyzes them and recursively follows the links out of those pages. Statistics are automatically collected and retained for each webpage extracted, including frequency of keywords and the number of images and videos (if present), with the entire content of each webpage preserved for further manual and automated textual analysis (Mei & Frank, 2015; Zulkarnine et al., 2016).

The research team expanded its search for data by using ParseHub to harvest select datasets from Reddit sources. ParseHub is a web-scraping tool (Abiantoro & Kusumo, 2020) that is similar in some respects to TDC, albeit not quite as fast as TDC, nor as able to deal with large volumes of data (or at least, not in our experience).

3.3 The Posit Toolkit

The Posit Text Profiling Toolkit facilitates quantitative analysis of large text corpora. In the process, Posit applies a Part-of-Speech tagger and outputs statistical details in terms of individual words (tokens) and word types. Frequency data is provided for specific parts of speech, including frequency ordered details of each specific word in an analyzed text (Weir, 2007, 2009). For direct application in data classification, Posit's summary details can be employed as a feature set for input to existing data classification systems such as WEKA or Random Forest.

The summary data output from Posit includes values for 27 various properties of the text, as listed in Figure 3. Posit output is generated at several levels of detail. Of these, the summary level is the most general, for example, the total number of verbs, nouns, and adjectives. At the intermediate level, frequency data is provided for the contents of the analyzed text in terms of specific parts-of-speech, for example, verb types: the base form of verbs, the gerund form, the past tense form, the past participle form, the third person present form, the present tense (non-third person) form and the modal auxiliary form. At the fine detail level, frequency data is provided for each word in terms of part-of-speech type, such as the number of occurrences of every word that is a verb of gerund form.

3.4 WEKA, J48 and Random Forest

Following analysis in Posit, two common tree-based algorithms (J48 and Random Forest) in the Waikato Environment for Knowledge Analysis (WEKA) were used for modeling the pre-determined classifications with 10-fold cross validation.

With Random Forest, classification trees are independently constructed by employing a bootstrap sample of the entire dataset, and then relying on a simple majority vote for predictive purposes, rather than relying on earlier trees to boost the weight of successive trees (Breiman, 2001; Liaw & Wiener, 2002). The predicted label of Random Forest's input data is a vote by the trees in the forest, weighted by their probability

estimates. Thus, the prediction probabilities of Random Forest can be computed as the mean predicted class probabilities of the trees in the forest, and the class probability of a single tree is the fraction of samples of the same class in a leaf (Pedregosa et al., 2011).

3.5 LibShortText

LibShortText (LST) is an open-source software package, developed by the Machine Learning Group at National Taiwan University. LST is said to be more efficient and more extensible than other generalized text-mining tools, allowing for the conversion of short texts into sparse feature vectors, and for micro- and macro-level error analysis (Yu et al., 2013). On a typical computer, processing and training with 10 million short texts requires only half an hour or so, whereas Posit might require a day or more. LST includes an interactive tool for error analysis, and the program's default options usually work well, without tedious fine-tuning.

3.6 LibLinear

LibLinear (LBL) is a companion open-source software package to LibShortText (LST), developed by the same Machine Learning Group at National Taiwan University that developed LST (Fan et al, 2008). LBL is a classification program, whereas LST is a text analysis program. LBL predicts the accuracy of the classification performed by LST, much as WEKA predicts the accuracy of the classification performed by Posit. Another advantage to LBL is that it supports incremental and decremental learning, or in other words, the addition and removal of data to improve optimization and decrease run time. LST, on the other hand, does not readily support updating of the model.

3.7 Naïve Bayes

Naïve Bayes (NB) is a supervised machine learning model based on Bayes' theorem (Rish, 2001). Bayes' theorem is a probability theorem that works with conditional probabilities (the probability that something will happen if something else has already happened). The NB equation is used to determine the posterior probability for each class and the outcome of prediction is the class with the highest posterior probability.

In this algorithm, a variable's presence or absence has no bearing on the presence or absence of any other variable; however, this assumption of independent predictors can sometimes act as a flaw in NB, because with real data, the collection of predictors is usually not totally independent. Nonetheless, the technique works very well on data that contradicts this premise.

NB classifiers have a limited number of parameter tuning options, such as alpha = 1 for smoothing, fit prior = [True | False] to learn or not learn class prior probabilities, and a few others. There are three types of NB models, for example, the Gaussian (which assumes normal distribution), the Multinomial (used for discrete counts) and the Bernoulli (used if feature vectors are binary). The Gaussian (or Normal) distribution is the easiest to deal with as it just needs to estimate the mean and standard deviation from the training data (Singh et al., 2019). In this instance, we used the default parameter values.

3.8 Support Vector Machines (SVM)

The Support Vector Machine (SVM) algorithm has come to play a significant role in pattern recognition and data classification and is said to achieve superior results compared to other supervised machine-learning algorithms (Durgesh & Lekha, 2010). SVM seeks to find a "line" or "boundary" that separates different classes from each other, referred to as a "hyperplane" when analyzing more than three dimensions. Then classification is accomplished by locating the hyper-plane that distinguishes the two classes as best as possible and has the maximum margin. The margin is the distance between the nearest data point (of either class) and the hyperplane. The higher the margin, the higher the robustness of the hyperplane. Thus, the hyperplane that best separates the two classes and that maximizes the margins from both classes is chosen.

If one of the coordinates is an outlier in the region of the other class, a straight line cannot be used to separate the two classes. In these situations, the SVM algorithm offers a feature that allows it to disregard outliers and select the linear hyperplane with the greatest margin. However, in cases of large data collection, or where the dataset contains more noise, SVM tends not to perform especially well (Suthaharan, 2014).

4. Research Findings

The Posit approach has been used in a variety of studies wherein Posit classification was compared with classification of identical datasets by other machine learning algorithms where the characteristics of those datasets (e.g., real information and misinformation) had been pre-established through manual inspection or targeted harvesting (cf., Weir et al., 2016; Owoeye & Weir, 2018). We add the pre-classification value ("real information" or "misinformation") to the 27 Posit features described in our methodology section. The resultant set of 28 feature values is formatted and input to the data classification tool WEKA.

A TP (True Positive) occurs when the sample value is true and model prediction value is positive, i.e., when the pre-determined value matches the model's classification (correct prediction). An FP (False Positive) occurs when the sample value is false, but the model prediction value is true, i.e., when the pre-classified value is false, and the model prediction is positive (incorrect prediction). An FN (False Negative) occurs when the sample value is false and the model prediction value is negative, i.e., when pre-classified as false and the prediction is true (incorrect prediction). Finally, a TP (True Positive) occurs when both the sample and model prediction values are negative.

4.1 Reddit Datasets

The 44,531 data items scraped by ParseHub and pre-sorted as "real information" or "misinformation" were analyzed using Posit and the resultant values for the textual features were input to WEKA, along with the pre-classification for each data item (22,559 real and 21,972 misinformation). Two tree-based algorithms (J48 and Random Forest) were the basis for modelling the pre-determined classifications, with 10-fold cross validation. The Posit/J48 algorithm achieved an overall F1 score of 0.863 compared to the pre-determined classification. The Posit/Random Forest (RF) combination achieved a higher overall weighted F1 score of 0.931 (see Table 1 and Table 2).

Table 1

Posit/J48 Measures for Reddit Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.866	0.140	0.864	0.866	0.865
Mis-info	0.860	0.134	0.862	0.860	0.861
Weighted Avg.	0.863	0.137	0.863	0.863	0.863

Table 2

Posit/Random Forest Measures for Reddit Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.925	0.064	0.937	0.925	0.931
Mis-info	0.936	0.075	0.924	0.936	0.930
Weighted Avg.	0.930	0.069	0.931	0.930	0.931

Except for LibShortText (LST), none of the other machine-learning algorithms that we applied to this Reddit dataset for comparison purposes outperformed Posit. The weighted average F1 for Naïve Bayes (NB) was 0.644 (Table 3) compared to the weighted average F1 of 0.931 when the Posit results were input to RF, while the weighted average F1 for SVM was somewhat better, at 0.718 (Table 4).

Table 3

Naïve Bayes Measures for Reddit Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.746	0.365	0.746	0.513	0.608
Mis-info	0.635	0.254	0.635	0.829	0.720
Weighted Avg.	0.690	0.309	0.690	0.673	0.664

Table 4

SVM Measures for Reddit Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.921	0.336	0.921	0.503	0.651
Mis-info	0.664	0.079	0.664	0.958	0.784
Weighted Avg.	0.791	0.206	0.791	0.733	0.718

LibLinear (LBL), the companion software for LST, attained a weighted average F1 of 0.827 for the Reddit dataset (Table 5), a noticeable improvement upon NB and SVM, but still below that attained when the Posit results were input to J48 (F1 = 0.863) and RF (F1 = 0.931). LST, on the other hand, continues to be an exceptionally solid performer when it comes to the automated machine classification of social media posts (Yang, 2017), as reported elsewhere in other dis/misinformation studies conducted by the ICCRC and Strathclyde (Cartwright et al, 2019; Cartwright et al., 2022). In this instance, with the Reddit dataset, LST attained a weighted average F1 of 0.982 (Table 6), better than Posit at 0.931, and much better than the weighted averages of the other machine-learning algorithms, which ranged from a low of 0.64 to a high of 0.827.

One possible conclusion to draw here is that the features extracted by Posit are critical to the differentiation between mis- and real-information. This could be a property of the misinformation itself, in that, for example, mis-information tends to use more emotional language and thus might contain more adjectives than real-information. This nuance would not be incorporated into the SVM or NB models but would be used by Posit, and probably LST, as it is designed specifically for short texts.

Table 5

LibLinear Measures for Reddit Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.864	0.202	0.864	0.772	0.815
Mis-info	0.798	0.136	0.798	0.881	0.838
Weighted Avg.	0.831	0.169	0.831	0.827	0.827

Table 6

LibShortText Measures for Reddit Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.986	0.022	0.986	0.978	0.982
Mis-info	0.978	0.014	0.978	0.986	0.982
Weighted Avg.	0.982	0.018	0.982	0.982	0.982

4.2 Twitter Datasets

All 131,337 of the tweets that were harvested by The Dark Crawler (TDC) were input to NB, SVM, LBL and LST for machine classification. Due to performance reasons, for Posit analysis of the Twitter data, a subset of 2,500 real and 2,500 misinformation data items was extracted from the dataset of 131,337 tweets harvested by TDC. As noted earlier, Posit takes much longer than NB, SVM, LBL and LST when it comes to the processing of large datasets.

These 2,500 real and 2,500 misinformation data items were then analyzed using Posit to produce the relevant feature set based upon the textual characteristics of the input data. Once again, these feature details were entered into WEKA as a basis for matching the pre-determined classifications. In WEKA, we employed J48 and RF in turn, with 10-fold cross validation, to evaluate the scope for match from textual features to classification.

Again, except for LST, none of the other machine learning algorithms that we applied to the Twitter dataset for comparison purposes approached the classification accuracy of Posit. The weighted average F1 score for NB was only 0.585 (Table 7), slightly better than a coin toss, compared to 0.613 for SVM (Table 8), and 0.695 for LBL (Table 9). It is noteworthy that (albeit lower for Twitter than for Reddit), the performance for these algorithms appears in the same order as for Reddit, with NB the lowest, SVM in between, and LBL above those, but nowhere close to the accuracies attained by Posit and LST.

Table 7

Naïve Bayes Measures for Twitter Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.421	0.333	0.421	0.219	0.288
Mis-info	0.667	0.579	0.667	0.839	0.743
Weighted Avg.	0.581	0.493	0.581	0.623	0.585

Table 8

SVM Measures for Twitter Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.673	0.315	0.673	0.157	0.255
Mis-info	0.685	0.327	0.685	0.960	0.800
Weighted Avg.	0.681	0.323	0.681	0.684	0.613

Table 9

LibLinear Measures for Twitter Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.632	0.262	0.632	0.395	0.486
Mis-info	0.738	0.368	0.738	0.881	0.803
Weighted Avg.	0.702	0.332	0.702	0.716	0.695

That said, Posit/J48 did not perform as well with the Twitter data as it did with the Reddit data (a weighted average F1 of 0.789 for Twitter, as opposed to 0.863 for Reddit). Similarly, the Posit/Random Forest algorithm also suffered, achieving a weighted average F1 of only 0.829 for Twitter, as opposed to 0.931 for Reddit. On the other hand, LST again outperformed all of them with the Twitter data, including Posit, with a weighted LST average F1 of 0.943. However, as was the case with Posit, LST did not perform as well with the Twitter data as it did with the Reddit data (a weighted average F1 of 0.944 for Twitter, as opposed to 0.982 for Reddit). We will continue to investigate the reasons for this in our ongoing research.

Table 10

Posit/J48 Measures for Twitter Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.792	0.213	0.788	0.792	0.790
Mis-info	0.787	0.208	0.791	0.787	0.789
Weighted Avg.	0.789	0.211	0.789	0.789	0.789

Table 11

Posit/Random Forest Measures for Twitter Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.848	0.191	0.816	0.848	0.832
Mis-info	0.809	0.152	0.842	0.809	0.825
Weighted Avg.	0.829	0.171	0.829	0.829	0.829

Table 12

LibShortText Measures for Twitter Classification

Class	TP Rate	FP Rate	Precision	Recall	F1
Real	0.946	0.057	0.946	0.886	0.915
Mis-info	0.943	0.054	0.943	0.974	0.958
Weighted Avg.	0.944	0.055	0.944	0.944	0.943

As with previously considered data, reducing the required feature set derived from Posit, whilst still attaining a reasonable degree of correlation to the pre-determined classification, is desirable as offering greater efficiency in modelling such classifications. For the Twitter data we found that a small subset of two textual features on their own (average word length, and particles) could achieve an overall classification match of 73.64%. While Posit takes longer than NB, SVM, LBL and LST to process large quantities of data, it remains highly useful to us when pre-classifying smaller datasets, or cross-validating the classifications assigned by the machine-learning algorithms and/or by manual inspection.

5. Limitations and Future Research

The datasets that we collected for this DCCP-funded COVID-19 research project were Canadian-

specific (due to our research mandate), and of limited size, due to time restrictions. In fact, we had access in the early stages to a 146,002,957-tweet dataset provided to us by the Canadian-based Media Ecosystem Observatory, collected by the Observatory in connection with other research projects pertaining to social media messaging about COVID-19 (Bridgman et al., 2020). However, qualitative analysis of a random sample of 1,000 of those tweets indicated that only 43% were from Canadian sources, and that all had been collected before COVID-19 was declared to be a public health emergency, rendering it largely unsuitable for our study. In future, we will harvest larger datasets using TDC, and strive to compare our Canadian results to COVID-19 misinformation from other countries.

During this study, we found that much of the social media misinformation content consisted primarily or exclusively of memes, images and videos, thus necessitating a much different collection and analytical process than that employed for typical textual content. We also observed that a significant amount of the misinformation was being spread and/or amplified by bot activity.

We are developing a proof-of-concept system that incorporates automated searching and analysis of textual content, images, and videos. Going forward, we plan to expand the scope of the social media content that our technology can monitor and analyze by combining natural language processing (NLP), robust optical character recognition (OCR) and Mask R-CNN to extract information from images, videos, and memes.

To help navigate these models and misinformation feeds, we will amend our existing web-interface to our crawler to provide a summary of feeds, and the type of content within them, based on our machine-learning models. This interface could then be provided to governments, law enforcement agencies and perhaps social media platforms after the project.

6. Conclusion

We have customized TDC to monitor selected social media and online news sources, acquired datasets that are representative of COVID-19 "misinformation" and "real information" on social media, and demonstrated our ability to classify "misinformation" and "real information" with accuracy, using machine-learning and complementary automated text-reading/classification programs. During this project, we combined multiple technologies, and applied them to real-world data (Reddit posts and tweets), demonstrating our ability to discern measurable differences between "misinformation" and "real information" pertaining to the COVID-19 pandemic.

There is a fine line between the monitoring of social media and the potential abrogation of the right to privacy, to the extent that such privacy rights are believed to exist in the public domain. Notwithstanding arguments in favor of respecting personal privacy and obtaining informed consent, it is now widely accepted by courts and ethics review bodies in most countries that individuals posting messages in these public venues are seeking public visibility or are at least aware of the public nature of their behavior (Heilferty, 2011; Moreno et al., 2008). With this study, it was evident that the individuals and groups posting messages on Twitter and Reddit were aware that the messages were being read, and that many were striving to promote their views about COVID-19 to a wider audience.

Large social media platforms such as Facebook and Twitter have been criticized in the past for refusing to remove misinformation from their sites (Durkee, 2021; Tsesis, 2017). Recently, Twitter and Facebook have instead come under fire for removing misinformation, because both platforms have fact-checked, warned about, and then proceeded to remove what they considered to be COVID-19 misinformation (Rupar, 2021). Such fact-checking, warnings, and removal of misinformation by the platforms are often met with public outcries about censorship and the abuse of power (Biggs, 2021; Radu, 2020). Indeed, governments themselves are facing increasing pressure to address social media misinformation through legislation and regulation, raising issues of governmental restrictions on freedom of expression (Bayer et al., 2021) and the possibility of setting a legal precedent for censorship of political dissent (Bleyer-Simon, 2021).

As we have also seen, however, there have been over a half billion COVID-19 cases and over six million COVID-19-related deaths since the pandemic began (WHO, 2022). Beyond the harm created by the virus itself, COVID-19 has had massive negative impacts on government budgets and economies at the local, national, and international levels, as well as serious negative impacts on social cohesion and social well-being (Gilmore, 2022; O'Connell, 2022; Prasad, 2020). Thus, it could be argued that the harms associated with COVID-19 misinformation—e.g., the social and economic damage, vaccine hesitancy, and the resultant deaths and severe illnesses—outweigh the right to personal privacy and freedom of speech considerations. We anticipate that the AI technology we are developing will permit governments, law enforcement agencies and possibly social media platforms (should governments and law enforcement agencies choose to share this technology with them) to identify medical misinformation on social media in near-real-time, and to take action to counter such misinformation where appropriate and legal for them to do so.

References

Abiantoro, D., & Kusumo, D. S. (2020). Analysis of Web Content Quality Information on the Koseeker Website Using the Web Content Audit Method and ParseHub Tools. *8th International Conference on Information and Communication Technology (ICoICT)*, 2020, 1-6.

Bayer, J., Katsirea, I., Batura, O., Holznagel, B., Hartmann, S., & Lubianiec, K. (2021). *The fight against disinformation and the right to freedom of expression*. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf

Berghel, H. (2017). Lies, damn lies, and fake news. *Computer*, 50(2), 80-85.

Bernard, R., Bowser, G., Sullivan R., & Gibson-Fall, F. (2020). Disinformation and Epidemics: Anticipating the Next Phase of Biowarfare. *Health Security*, 19(1), 1-12.

Bester, J. C. (2016). Measles and Measles Vaccination: A Review. *JAMA Pediatrics*, 170(12), 1209-1215.

Biggs, T. (2021). Twitter expands efforts in AI-assisted war on COVID fake news. *Sydney Morning Herald*. smh.com.au/technology/twitter-s-expands-efforts-in-ai-assisted-war-on-covid-fake-news-20210714-p589oa.html

Bleyer-Simon, K. (2021). Government repression disguised as anti-disinformation action: Digital journalists' perception of COVID-19 policies in Hungary. *Intellect Limited 2021 Journal of Digital Media & Policy*, *12*(1), 159–176. https://doi.org/10.1386/jdmp_00053_1

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5-32. https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Special Issue on COVID-19 and Misinformation, 1018. https://misinforeview.hks.harvard.edu/article/the-causes-and-consequences-of-covid-19-misperceptions-understanding-the-role-of-news-and-social-media/

Cartwright, B., Frank, R., Weir, G. *et al.* (2022). Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks. *Neural Computing and Applications*, 1-23. https://doi.org/10.1007/s00521-022-07296-0

Cartwright, B, Weir, G. R. S., & Frank, R. (2019). Fighting disinformation warfare with artificial intelligence: Identifying and combating disinformation attacks. *Tenth International Conference on Cloud Computing, GRIDS, and Virtualization*, May 2019, 67-72.

Centers for Disease Control and Prevention (2022a). *COVID Data Tracker*. https://covid.cdc.gov/covid-data-tracker/

Centers for Disease Control and Prevention (2022b). *Estimates of vaccine hesitancy for COVID-19*. https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw/

Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. *Journal of theoretical and applied information technology*, *12*(1), 1-7.

Durkee, A. (2021). Biden Says Facebook, Tech Platforms Are "Killing People" By Spreading Misinformation on Covid Vaccines. *Forbes*, Jul 16, 2021. https://www.forbes.com/sites/alisondurkee/2021/07/16/biden-says-facebook-tech-platforms-are-killing-people-by-spreading-misinformation-on-covid-vaccines/?sh=68b969e439f0

Gilmore, R. (2022). 'Freedom convoy' forums find new focus: disinformation about Russia-Ukraine war. *Global News*, 8 Mar 2022. https://globalnews.ca/news/8659667/ukraine-russia-convoy-misinformation-conspiracy/

Gisondi, M. A., Barber, R., Faust, J. S., Raja, A., Strehlow, M. C., Westafer, L. M., & Gottlieb, M. (2022). A Deadly Infodemic: Social Media and the Power of COVID-19 Misinformation. *Journal of Medical Internet Research*, *24*(2), 1–13. https://doi.org/10.2196/35552

Government of Canada (2022). *COVID-19 daily epidemiology update*. https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.htmlHall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Heilferty, C. M. (2011). Ethical considerations in the study of online illness narratives: A qualitative review. *Journal of Advanced Nursing*, *67*(5), 945-953. doi: 10.1111/j.1365-2648.2010.05563.x

Jankowski, N. W. (2018). Researching fake news: A selective examination of empirical studies. *Javnost-The Public*, 25(1-2), 248-255.

Kata, A. (2010). A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine*, 28(7), 1709-1716.

Krishnan, N., Gu, J., Tromble, R., & Abroms, L. C. (2021). Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, *2*(6), 1-25.

Krause, N. M., Freiling, I., Beets, B., & Brossard. D. (2020). Fact-checking as risk communication: the multilayered risk of misinformation in times of COVID-19. *Journal of Risk Research*, 23 (7-8), 1052-1059.

Lavoie, K., Gosselin-Boucher, V., Stojanovic, J., Gupta, S., Gagné, M., Joyal-Desmarais, K., ... & Bacon, S. (2022). Understanding national trends in COVID-19 vaccine hesitancy in Canada: results from five sequential cross-sectional representative surveys spanning April 2020–March 2021. *BMJ open*, *12*(4), e059411.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Loebach, J., & Madigan, R. (2015). Collecting Social Media Data for Qualitative Research (Research Short Series #2). Charlottetown, PE: Young Lives Research Lab, University of Prince Edward Island. URL: https://younglivesresearch.ca/wp-content/uploads/2018/10/Collecting-Social-Media-Data-for-Qualitative-Research.pdf

Mei, J., & Frank, R. (2015). Sentiment crawling: Extremist content collection through a sentiment analysis guided webcrawler. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, August 2015. Accessed: 25 July 2021

Moreno, M. A., Fost, N. C., & Christakis, D. A. (2008). Research ehtics in the MySpace era. *Pediatrics, 121*(1), 157-160. doi: 10:1542/peds.2007-3015

Posetti, J., & Bontcheva, K.(2020). DISINFODEMIC: Deciphering COVID-19 disinformation. The United Nations Educational, Scientific and Cultural Organization, 1-17. Downloaded from: https://en.unesco.org/covid19/disinfodemic/brief1

O'Connell, O. (2022). Soup kitchen harassment to confederate flags: Controversial moments at Canadian trucker convoy protest. *Independent*, 31 Jan 2022. https://www.independent.co.uk/news/world/americas/canada-trucker-protest-convoy-freedom-b2004613.html

Owoeye, K., & Weir G.R.S. (2018). Classification of radical Web text using a composite-based method. *IEEE International Conference on Computational Science and Computational Intelligence.* https://pure.strath.ac.uk/ws/portalfiles/portal/86519706/Owoeye_Weir_IEEE_2018_Classification_of_radical_web_text_using_a_composite_based.pdf. Accessed: 13 August 2020

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pennycook, G., McPhetres, J., Zhang, Y.H., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7), 770-780.

Prasad, R. R. (2020). *Media Freedom and Covid-19*. https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/human_rights-droits_homme/policy-orientation-covid-19.aspx?lang=eng

Radu, R. (2020). Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation. *Social Media and Society*, *6*(3). https://doi.org/10.1177/2056305120948190

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Rupar, A. (2021). Trump's Twitter and Facebook ban is already working. Here's how we know. *Vox*, 9 Jul 2021. https://www.vox.com/2021/1/16/22234971/trump-twitter-facebook-social-media-ban-election-misinformation-zigna

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593-596).

Shin, T. (2021). A Mathematical Explanation of Support Vector Machines: Develop a deeper understanding of one of the most popular machine learning models. *Towards Data Science*. https://towardsdatascience.com/a-mathematical-explanation-of-support-vector-machines-e433ffe04362

Suarez-Lledo, V., & Alvarez-Garcia, J. (2021). Prevalence of Health Misinformation on Social Media: Systematic Review. *Journal of Medicine Internet Research*, 23(1), 1-17.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *Performance Evaluation Review*, 41(4), 70-73.doi: 10.1145/2627534.2627557

StatCan COVID-19: Data to Insights for a Better Canada. (2021). COVID-19 vaccine willingness among Canadian population groups. Downloaded from: https://www150.statcan.gc.ca/n1/en/pub/45-28-0001/2021001/article/00011-eng.pdf?st=vpGfk0q0

Tsesis, A. (2017). Social media accountability for terrorist propaganda. *Fordham Law Review*, *86*(2), 605–632.

Weir, G. R. S. (2007). The posit text profiling toolset. *12th Conference of Pan-Pacific Association of Applied Linguistics*, pp. 106-109.

Weir, G. R. S. (2009). Corpus profiling with the Posit tools. *Proceedings of the 5th Corpus Linguistics Conference,* July 2009. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.9606&rep=rep1&type=pdf

Weir, G., Frank, R., Cartwright, B., & Dos Santos, E. (2016). Positing the problem: enhancing classification of extremist web content through textual analysis. *International Conference on Cybercrime and Computer Forensics (IEEE Xplore).* https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/7740431. Accessed: 13 July 2021

Weir, G., Owoeye, K., Oberacker, A., & Alshahrani, H. (2018). Cloud-based textual analysis as a basis for document classification. *International Conference on High Performance Computing & Simulation (HPCS)*:672-676. https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/8514415. Accessed: 13 August 2019

Wilson, S. L., & Wiysonge, C. D. (2020). Social media and vaccine hesitancy. BMJ Global Health, 5(10), 1-7.

World Health Organization (2022). *Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yu, H.-F., Ho, C.-H. , Juan, Y.-C., and Lin, C.-J. LibShortText: A Library for Short-text Classification and Analysis (2013). Retrieved from https://www.csie.ntu.edu.tw/~cjlin/LibShortText/

Zulkarnine, A.T., Frank, R., Monk, B., Mitchell, J., & Davies, G. (2016). Surfacing collaborated networks in dark web to find illicit and criminal content. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, September 2016. https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/7745452 Accessed: 4 August 2019