Deploying Artificial Intelligence to Combat Disinformation Warfare

Identifying and Interdicting Disinformation Attacks Against Cloud-based Social Media Platforms

Barry Cartwright
School of Criminology
Simon Fraser University
Burnaby, Canada
Email: bcartwri@sfu.ca

George R. S. Weir
Department of Computer & Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
Email: george.weir@strath.ac.uk

Richard Frank
Karmvir Padda
School of Criminology
Simon Fraser University
Burnaby, Canada
Email: {rfrank; karmvir_padda}@sfu.ca

***Abstract*—Disinformation attacks that make use of Cloud-based social media platforms, and in particular, the attacks orchestrated by the Russian "Internet Research Agency," before, during and after the 2016 U.S. Presidential election campaign and the 2016 Brexit referendum in the U.K., have led to increasing demands from governmental agencies for technological tools that are capable of identifying such attacks in their earliest stages, rather than identifying and responding to them in retrospect. This paper reports on the interim results of an ongoing research project that was sponsored by the Canadian government's Cyber Security Directorate. The research is being conducted by the International CyberCrime Research Centre (ICCRC) at Simon Fraser University (Canada), in cooperation with the Department of Information and Computer Sciences at the University of Strathclyde (Scotland). Our ultimate objective is the development of a "critical content toolkit," which will mobilize artificial intelligence to identify hostile disinformation activities in "near-real-time." Employing the ICCRC's Dark Crawler, Strathclyde's Posit Toolkit, Google Brain's TensorFlow, plus SentiStrength and a short-text classification program known as LibShortText, we have analyzed a wide sample of social media posts that exemplify the "fake news" that was disseminated by Russia's Internet Research Agency, comparing them to "real news" posts in order to develop an automated means of classification. To date, we have been able to classify posts as "real news" or "fake news" with an accuracy rate of 90.7%, 90.12%, 89.5%, and 74.26% using LibShortText, Posit, TensorFlow and SentiStrength respectively.**

Keywords-Social media; disinformation warfare; machine learning.

I. INTRODUCTION

This paper elaborates on an earlier paper on the subject of fighting disinformation warfare through the use of artificial intelligence, presented at the Tenth International Conference on Cloud Computing, GRIDs, and Virtualization, held in Venice, Italy, in May 2019 [1]. As observed in our earlier conference paper, the key challenges facing law enforcement agencies, intelligence agencies, cybersecurity personnel and business owners-operators worldwide are how to monitor and effectively respond to dynamic and emerging cybersecurity threats, with increasing attention being paid to hostile disinformation activities in Cloud-based social media platforms [1]. To illustrate, Cambridge Analytica, through an app that it developed, managed to scrape data from over 80 million Facebook pages worldwide. This information was in turn used to micro-target voters through Facebook advertisements, which were premised upon the demographic profiles and known political leanings of those voters, in turn based upon information which had been extracted with the help of the Cambridge Analytica app [2], [3]. In July 2018, Facebook was fined £500,000—the maximum allowable under British law—for its mishandling of data in the Cambridge Analytica scandal [4]. In July 2019, the US Federal Trade Commission fined Facebook five billion USD for its failure to protect user privacy [5]. The nexus between Cambridge Analytica, WikiLeaks, and Russian interference in the 2016 U.S. Presidential election remained under investigation by the U.S. Congress as recently as the Summer of 2019 [6].

According to a 2017 Intelligence Community Assessment, prepared jointly by the Central Intelligence Agency (CIA), the Federal Bureau of Investigation (FBI) and the National Security Agency (NSA), a number of other Cloud-based social media, including Twitter and Instagram, have also been implicated as (possibly unaware) participants in the hosting and dissemination of disinformation attacks associated with the Russian "Internet Research Agency" (IRA) [7]. According to Special Counsel Robert Mueller's recently released report into Russian interference in the U.S. Presidential election [8], Facebook and Twitter accounts targeted certain groups, such as Blacks (through the Blacktivist Facebook page), Southern Whites (through the

Patriototus Facebook page), and the right-wing anti-immigration movement (through the Secured Borders Facebook page), as well as through Twitter feeds such as @TEN_GOP (which falsely claimed to have a connection to the Republican Party of Tennessee), and @America_1st (an anti-immigration account). In the U.K., the "fake news"—which primarily stoked Islamaphobic and anti-immigration passions—made extensive use of Twitter, employing Twitter handles such as ReasonsToLeaveEU, or #voteleave [9], [10], [11], [12]. Evidence also indicates that the Russian IRA maximized use of social media bots in their 2016 assaults on the U.S. Presidential election and the U.K. Brexit referendum [9], [10], [13], [14], thus amplifying the content in order to reach and influence a much wider audience. More will be said about Russian involvement in disinformation warfare in Section II of this paper, wherein we present our literature review.

Our research, sponsored by the Canadian government's Cyber Security Cooperation Program, and conducted by the International CyberCrime Research Centre at Simon Fraser University, in cooperation with the Department of Information and Computer Sciences at the University of Strathclyde, involves the development of a tool for identifying hostile disinformation activities in the Cloud. This research project commenced with a dataset of 2,946,219 "fake news" Twitter messages (tweets), identified as emanating from the Russian IRA. Later, our research came to include datasets that combined both Twitter and Facebook "fake news" messages, eventually including a number of comparator datasets of "real news" messages, plus a potential "training" dataset for machine learning that we have not yet explored fully, the latter consisting of a wide range of "real news" and "fake news" [15]. It is anticipated that the knowledge generated by this research will establish the foundation for more advanced work, eventually culminating in the construction of a "critical content toolkit," which will aid governmental agencies in the rapid and accurate pinpointing of disinformation attacks in their very early stages.

The research team has several years of collaborative experience in collecting and analyzing data from online extremist forums, child pornography websites, social media feeds and the Dark Web. Our previous experience in data classification has demonstrated that we are able, through automation, to achieve predictive accuracy in the 90-95% range when it comes to detecting the nuanced text found in extremist content on the Web [1], [16], [17], [18]. From this background, we have a methodology that is applicable to the analysis and classification of data from Cloud-based social media platforms. In the past, our predictive accuracy was accomplished by applying a combination of technologies, including the Dark Crawler, SentiStrength, and Posit [1]. For the present study, we have employed the Dark Crawler, Posit, SentiStrength, TensorFlow, and LibShortText. Additional information on these research tools is provided in Section III, wherein we set out our methodology. Our research results are reported in Section IV, and elucidated further in Sections V and VI, wherein we discuss our results, set out the directions that our future research endeavors are expected to take, and present our interim conclusions.

II. LITERATURE REVIEW

As noted in our introductory comments in Section I, Cloud-based social media platforms have come under increasing scrutiny for permitting hostile foreign actors to manipulate public opinion through the creation of fake social media accounts that disseminate false information, often referred to as "fake news" [19], [20], [21]. This false information, or fake news, can be broken down into two broader categories: misinformation and disinformation. The less sinister of the two, misinformation, is simply inaccurate or false information. Misinformation may be based upon a genuine misapprehension of the facts, as opposed to having been created with any particular intention of deceiving or manipulating people [22], [23], [24]. Disinformation, on the other hand, especially when employed by hostile foreign actors, is information that is created and spread intentionally, for the express purpose of deception and manipulation of public opinion [22], [24], [25].

The activities of Russia's IRA during the 2016 U.S. Presidential election would be a prime example of a disinformation campaign mounted by a hostile foreign actor [10], [13], [26], [27]. In February 2108, U.S. Special Counsel Robert Mueller, duly appointed to investigate Russian interference in the U.S. election, obtained a grand jury indictment against the IRA (which was bankrolled by Yevgeniy Prigozhin, often referred to as "Putin's chef"), plus Prigozhin's American-based companies Concord Management and Consulting LLC and Concord Catering as well as Prigozhin himself, along with a dozen Russian "trolls" who were employed by Prigozhin's IRA. The indictment stated that the accused had "operated social media pages and groups designed to attract U.S. audiences" in order to advance divisive issues and create dissension, falsely claiming that those pages and groups were controlled by American activists [9], [28].

The dozen Internet "trolls" who were described in the indictment obtained by Mueller belonged to an identifiable sub-group of a much larger workforce, comprised of 1,000 or more Russian trolls, all employed by Prigozhin's IRA [29], [30], [31]. These IRA employees, working in a building in the Russian city of St. Petersburg, toiled around the clock in two, 12-hour shifts (a day shift and a night shift), with the objective of fomenting division, distrust, dissent, and hostility within and between targeted groups in the American populace [32], [33], [34]. In particular, it has been said that these IRA trolls were instructed to spread disinformation that would buttress Donald Trump's campaign for the U.S. Presidency, and at the same time, undermine the campaign of Hillary Clinton [7], [30], [34], [35].

The Computational Propaganda Project, a multi-national project housed primarily in the Oxford Internet Institute, has reported that 19 million identifiable "bot" accounts tweeted in support of Trump or Clinton in the week leading up to the 2016 Presidential election, with 55.1% of those in favour of Trump, and only 19.1% in favour of Clinton [36], [37], [38]. The evident disparity in Twitter support would seem difficult

to account for, other than in terms of highly-orchestrated and deliberate political interference, given that Hillary Clinton received 65,844,954 votes, compared to 62,979,879 votes for Donald Trump [39].

A 2017 study by Zannettou et al. revealed that 71% of these "fake" accounts were created prior to the 2016 election [34]. In fact, the 2017 Intelligence Community Assessment, prepared jointly by the CIA, FBI and NSA, indicated that Russian operatives had begun researching U.S. electoral processes and election-related technology and equipment as early as 2014, two years prior to the election, and that the Prigozhin-led IRA had started advocating on behalf of Donald Trump's candidacy as early as 2015, one year prior to the election [7]. Zannettou et al. reported that 24 accounts were created on July 12, 2016, approximately one week before the Republican National Conference (at which Donald Trump was formally nominated as the Republican candidate for the 2016 Presidential election) [34]. The study also found that the Russian Internet trolls attempted to mask their disinformation campaign by adopting different identities, changing their screen names and profile information, and in some cases, deleting their previous tweets. In their examination of tweets posted between January 2016 and September 2017, for example, Zannettou et al. found that 19% of the accounts operated by IRA trolls changed their screen names as many as 11 times, and unlike other Twitter users, often deleted their tweets in large batches, in order to start again with a clean slate [34].

Much has been said about the use of social media bots in the U.S. Presidential election and the U.K. Brexit referendum [9], [10], [13], [14]. Briefly, the transfer and transformation of information on the Internet is not accomplished by people, but rather, by algorithms—scripts which convert mathematical expressions into instructions for the Internet [37]. The Internet Relay Chat System would be an early example of where bots were being used to manage and regulate social interaction on the Internet. These bots, which still comprise an integral part of the architecture of Cloud-based social media sites such as Twitter and Facebook, are capable of interacting with Internet users, answering simple questions, and collecting data. More sophisticated bots can also be deployed to crawl the Web, scrape social media sites for data, parse the information gleaned, and even manipulate political opinion [37]. Some online stores/companies, such as AliExpress, use these AI bots for managing the extensive help systems on their site. If you have an issue, you chat with the bot. The Cambridge Analytica app, which attracted so much negative attention to Facebook in the aftermath of the 2016 U.S. Presidential election and the 2016 U.K. Brexit referendum, would be an example of an algorithm that was designed for the express purpose of collecting and evaluating behavioral data such as the likes, dislikes and political proclivities of the Facebook users whose data it harvested [40].

To express it differently, the bots (robots) described herein are Cloud-based social media accounts that are controlled by software, rather than by real people. These social media bots are estimated to comprise between 5-9% of the overall Twitter population, and to account for approximately 24% of all tweets [41]. Users of social media may spend considerable time liking and disliking bots, sometimes arguing with (or even flirting with) bots, all the while thinking that they are interacting with a real person. Stories that "go viral"—i.e., that rise to the top of Twitter feeds—are often pushed there by these social media bots through manipulation of the social media platform's algorithms [41].

The main problem with "fake news" is that its consumers tend to accept what they read at face value. According to The Pew Research Center, 12% of Americans get their news from Twitter [42]. Of those who use the platform regularly, close to 60% depend on Twitter as their source of news [26], [42], [43]. With respect to the type of "fake news" that is the subject of this present study, it can be said that the frequent tweeting and re-tweeting by bots leads to ever-increasing exposure, resulting in an "echo chamber effect" [33]. To add to the mix, evidence suggests that many individuals are unable to distinguish between factual and non-factual content found on Twitter and Facebook [44], [45]. Indeed, according to a Stanford University study, far too many are inclined to accept images or statements that they come across on social media at face value, without questioning the source of those images or statements, or for that matter, asking whether they even represent what they purport to represent [9], [44], [45].

Russian interference in the U.S. Presidential election and the U.K Brexit referendum has been well documented, and has been the subject of considerable governmental and academic research, e.g., [7], [8], [9], [10], [12], [13], [14], [20], [27], [34]. However, such Russian interference is by no means restricted to the U.S. and the U.K. To illustrate, in 2019, the European Commission—along with the European External Action Service and other EU institutions and member states—released a progress report on its Action Plan Against Disinformation. According to the Commission's progress report, evidence gathered throughout 2018 and early 2019 confirmed ongoing disinformation activities originating from Russian sources, believed to be undertaken for the purpose of influencing voter preferences and suppressing voter turnout in the EU Parliamentary elections [23], [46].

Moreover, a recent study of Canadian Twitter data suggests that Russian trolls were behind "fake news" stories that attempted to stoke fear and distrust between Muslims and non-Muslims following the 2017 shooting deaths of six worshippers at a mosque in Quebec City, leading to renewed concerns that Russian trolls might attempt to interfere in the Fall 2019 Canadian federal election [47]. With this in mind, the research team recently collected a sample of 3,500 tweets from hashtags such as #TrudeauMustGoToJail, #TrudeauMustGo, and #TrudeauMustResign, some of which were suspected of containing "fake news" which was intended to influence the outcome of the 2019 Canadian federal election. In addition, we are currently focusing our efforts on collecting Canadian-specific "fake news" Facebook items, from *The Buffalo Chronicle-Canadian Edition*, Canadian Truth Seekers, Million Canadian March, The Canadian Defence League, The Silent Majority Canada, The Angry Cousin, Proud Canadians, and Canada Proud.

This Facebook dataset presently consists of 3,737 discrete data items.

Russian-orchestrated disinformation campaigns are long-standing in nature. The Kremlin reportedly founded a school for bloggers as far back as 2009, apparently foreseeing the long-range possibilities of utilizing Cloud-based social media in political influence campaigns [48]. In fact, Russian disinformation activities have been documented in the Czech Republic and Slovakia as far back as 2013 [49], and in the 2014 election in the Ukraine, which itself followed shortly after Russia's annexation of the Crimean Peninsula [50], [51]. This is not to suggest that all known disinformation campaigns have been launched by Russia, or that such campaigns have been restricted only to those countries mentioned above. Using a combination of qualitative content analysis, secondary literature reviews, country case studies and consultations with experts, a 2019 inventory compiled by the Oxford Internet Institute found evidence of disinformation campaigns in 70 different countries around the world, including but by no means limited to Armenia, India, Malaysia, Mexico, The Philippines, Saudi Arabia, The United Arab Emirates and Venezuela [52]. In many cases, however, the campaigns are spreading pro-party or pro-government propaganda, or attacking the political opposition, and in the absence of evidence to the contrary, they could well be mounted by local (rather than foreign) actors. That said, countries other than Russia, such as China and Saudi Arabia, are believed to be making increasing use of disinformation campaigns beyond their own borders [53]. In any event, the focus of this present paper is the Russian-orchestrated attacks on the 2016 U.S. Presidential election.

A number of researchers have mobilized artificial intelligence in an effort to counter the type of disinformation warfare employed by Russia during the 2016 U.S. Presidential election and the 2016 U.K. Brexit referendum. In 2017, Darren Linvill and John Walker (from Clemson University) gathered and saved vast numbers of Twitter and Facebook postings (prior to their removal from the Internet by the respective social media platforms), thereby preserving the evidence and making the data available to the academic, cyber-security and law enforcement communities for further study [54]. Linvill and Walker investigated the Twitter handles used by the Russian IRA, both qualitatively and quantitatively, breaking them down into troll accounts and bot accounts, and into right trolls, left trolls, fear mongers, news feeders and hash tag gamers. Our research team has made extensive use of the IRA's Twitter and Facebook postings that were gathered, saved and made available by Linvill and Walker. In 2017, William Yang Wang from the University of California at Santa Barbara released his LIAR dataset, which included 12,836 statements labeled for their subject matter, situational context, and truthfulness, broken down into training, validation and test sets, along with instructions for automatic fake news detection [15]. In addition, William Wang reported that the open source software toolkit, LibShortText, developed by the Machine Learning Group at National Taiwan University, had been shown to perform well when it came to short text classification [15], [55]. The dataset provided by Linvill and Walker, and the suggestion by William Wang about using LibShortText, have both been used by us to inform and refine the machine learning and automated analysis processes described in the following sections on Methodology and Research Results.

In his above-mentioned study using the LIAR dataset, William Wang found that when it came to automatic language detection, a hybrid, convoluted deep neural network that integrated both meta-data and text, produced superior results to text-only approaches [15]. We are employing a somewhat similar approach to that of William Wang, in that we are using a combination of deep neural networks (Tensor Flow) [56], a text-reading program (Posit) that also produces meta-data or mark-up [57], [58], and the LibShortText program developed by the Machine Learning Group at National Taiwan University [55]. Employing techniques of machine learning and natural language processing, a 2018 study of Twitter troll activity in the 2016 U.S. Presidential election found that a model blending measurements of "precision" and "recall" failed to accurately classify 34% of troll posts, suggesting that such models could not be relied upon to identify and screen out fake news [48]. However, a 2019 paper, entitled "Defending Against Neural Fake News," reports on the development of GROVER, a computer model that can both generate and detect neural fake news, premised on the notion that while most fake news is presently generated by humans, the fake news of the future may be generated by machines. The authors of this paper report additionally that they have been able to discriminate fake news with an accuracy of 92%, as opposed to the more standard 73% accuracy exhibited by other fake news discriminators [59]. Our research results, reported below, come much closer to approximating those described in this 2019 study.

Some researchers have sought to identify disinformation campaigns by employing "bot" detection, instead of relying upon automated text-reading software. Essentially, much of what may be regarded as "fake news" is thought to be spread and/or amplified by the use of bots [37], [40], [41]. Thus, the goal of bot detection is to discriminate accurately between bot-generated and human-generated activity on social media. Morstatter, Carley and Liu, for example, have proposed what they call "a new approach to bot detection," again blending measurements of "precision" and "recall" [41], similar to the measurements employed in the above-mentioned 2018 study of Twitter troll activity in the 2016 U.S. Presidential election. In their 2019 study, Morstatter et. al found that they could successfully classify bot activity in 76.55% of instances. Another approach, outlined by Gorodnichenko, Pahm and Talavera, identifies suspected bot activity in the Brexit referendum and the U.S. Presidential election, by measuring such variables as when the Twitter account was first created, the number of tweets per day, the timing of daily and hourly tweeting, and the number of tweets containing the same content. This is premised on the understanding that many of these bot accounts are created for the purpose of spreading disinformation, and that bots send more messages than humans, at all times of the day (even when human activity is much reduced), and that they re-send the same messages

over and over again [60]. Using this approach, Gorodnichenko et. al found that they could classify bots and non-bots with 90% accuracy. While we have not employed bot detection in this present study, it is an avenue that we plan to explore as our work progresses.

III. METHODOLOGY

Our analysis of "fake news" messages posted by the Internet Research Agency (IRA), before, during and after the 2016 U.S. Presidential election, employed a variety of approaches, including collection of IRA posts and "real news" datasets using the Dark Crawler, plus machine analysis of large samples of the posts using Posit, TensorFlow, SentiStrength and LibShortText [55].

Although this research was geared primarily toward machine learning and the development of an artificial intelligence tool to aid in the rapid and accurate pinpointing of disinformation attacks in their early stages, we also conducted qualitative, textual analysis of 1,250 of the IRA "fake news" Twitter posts, to probe into the alleged degree of Russian involvement in the disinformation campaign [8], [13], [26], [31], assess the veracity of claims that the posts were intended to support Donald Trump's campaign for the U.S. Presidency whilst simultaneously undermining the campaign of Hillary Clinton [7], [30], [34], [35], [36], [37], [38], and investigate the degree to which some of the posts were grounded in "real news," rather than in what is commonly referred to as "fake news" [9], [19], [20], [21], [22].

A. Research Tools

The Dark Crawler is a custom-written, web-crawling software tool, developed by Richard Frank of Simon Fraser University's International CyberCrime Research Centre. This application can capture Web content from the open and Dark Web, as well as structured content from online discussion forums and various social media platforms [61] [62], [63]. The Dark Crawler uses key words, key phrases, and other syntax to retrieve relevant pages from the Web. The Crawler 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 any are present). The entire content of each webpage is also preserved for further manual and automated textual analysis. Content retrieved by the Dark Crawler is parsed into an Excel-style worksheet, with each data element being identified and extracted. In previous studies of this nature, we have employed this same procedure to collect over 100 million forum posts from across a vast number of hacking and extremist forums, to be used for later analysis [61], [62].

The Posit toolkit was developed by George Weir of the Department of Computer and Information Sciences at the University of Strathclyde. Posit generates frequency data and Part-of-Speech (POS) tagging while accommodating large text corpora. The data output from Posit includes values for total words (tokens), total unique words (types), type/token ratio, number of sentences, average sentence length, number of characters, average word length, noun types, verb types, adjective types, adverb types, preposition types, personal pronoun types, determiner types, possessive pronoun types, interjection types, particle types, nouns, verbs, prepositions, personal pronouns, determiners, adverbs, adjectives, possessive pronouns, interjections, and particles, for a total of 27 features in all [9], [57], [58]. This process generates a detailed frequency analysis of the syntax, including multi-word units and associated part-of-speech components.

As it was configured for previous studies, the Posit toolkit created data on the basis of word-level information; thus, the limited content of the Russian IRA tweets that we were examining meant that many of the original features might have zero values. For this particular research project, Posit was extended to include analysis of character-level content, to assist with the analysis of short texts. To this end, the system supplemented the standard word-level statistics, generating an additional 44-character features for each instance of text data. These new features included quantitative information on individual alphanumeric characters, plus a subset of special characters—specifically, exclamation marks, question marks, periods, asterisks and dollar signs. The extension of Posit to embrace character-level as well as word-level data maintained the domain-neutral nature of Posit analysis. As a result of this extended Posit analysis, each data item (tweet) was represented by a set of 71 features, rather than the usual twenty-seven [1].

TensorFlow, originally developed by the Google Brain Team, is a machine learning system that employs deep neural networks [56], inspired by real-life neural systems. The learning algorithms are designed to excel in pattern recognition and knowledge-based prediction by training sensory data through an artificial network structure of neurons (nodes) and neuronal connections (weights). The network structure is usually constructed with an input layer, one or more hidden layers, and an output layer. Each layer contains multiple nodes, with connections between the nodes in the different layers. As data is fed into this neural system, weights are calculated and repeatedly changed for each connection [63].

Textual content was further analyzed using SentiStrength, which assigns positive or negative values to lexical units in the text [61], [62], [64]. This value is a measure that provides a quantitative understanding of the content of information being found online—specifically, the extent to which positive and negative sentiment is present. The program automatically extracts the emotions or attitude of a text and assigns them a value that ranges from "negative" to "neutral" to "positive."

In the case of Posit and SentiStrength, the resultant data were input to the Waikato Environment for Knowledge Analysis (WEKA) data analysis application [65]. For SentiStrength, the data, comprised of the noun keywords for each textual item, along with the associated sentiment score and the manual classification for that page, then employed WEKA's standard J48 tree classification method with ten-fold cross-validation. In this cross-validation, 10% of the data was hidden, and conditions were sought that would split the remaining 90% of the dataset in two, with each part having as many data-points as possible belonging to a single

class. Accuracy of the tree was then considered relative to the hidden 10% of the data. This process was repeated 10 times, each time with a different hidden 10% subset. WEKA produced a measure of how many of the pages were correctly classified.

For Posit, we applied the standard J48 tree WEKA classification method, plus the Random Forest classification method [65], [66], both with ten-fold validation (as described above). WEKA then produced a measure of how many of the text items were correctly classified. In the Random Forest method, classification trees (of the type found in WEKA) 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 [67].

Finally, to better enhance the machine learning process, and to improve our future classification accuracy, we turned our attention to the LibShortText toolkit, as William Yang Wang of the University of California at Santa Barbara had indicated that this tool produced superior results when it came to the accurate classification of shorter items of text, such as tweets or brief Facebook posts [15]. LibShortText, an open source software package developed by the Machine Learning Group at National Taiwan University, 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 [68].

B. Research Sample

At the beginning of the project, the research team downloaded a dataset of 2,946,219 Twitter messages (tweets) from git.hub, which had been posted online by fivethirtyeight.com. This dataset of tweets was collected and assembled by the aforementioned professors from Clemson University, Darren Linvill and Patrick Warren [54]. These tweets were described as originating from the Russian IRA, also referred to in common parlance as the Russian troll factory, a hostile foreign agency that was believed to have intentionally interfered in the 2016 U.S. Presidential election and the 2016 U.K. Brexit referendum [7], [8], [9], [10], [13], [14], [26], [27], [28], [29], [30], [31], [33].

As the various approaches used in our research (i.e., qualitative analysis, Posit, TensorFlow, SentiStrength and LibShortText) were designed to read English text, a decision was made to extract only those entries that were labeled as being "English," so in the process, we excluded languages such as Albanian, Bulgarian, Catalan, Croatian, Dutch, Estonian, French, German, Italian, Russian, Ukrainian, Uzbek, Vietnamese. As a consequence, 13 new Excel spreadsheets were created, with 2,116,904 English-speaking tweets remaining in the dataset following the removal of all non-English tweets.

Having acquired the Russian (IRA) Twitter data, we then sought a second Twitter dataset that would allow us to develop a classification model based upon comparison between "real news" and what has often been referred to simply as "fake news" [19], [20], [21], [22], [24], [25], [30]. To this end, we analyzed the textual content from the full set of IRA tweets (or "fake news") using Posit, in order to identify frequently occurring terms, and more specifically, nouns. The resultant "keyword" list was used by the International CyberCrime Research Centre's Dark Crawler, in order to retrieve a set of matching "real news" Twitter posts from legitimate news sites.

The Dark Crawler harvested Twitter feeds maintained by more "traditional," mainstream news sources, such as the *Globe and Mail*, *CBC News*, *CTV News*, the *BBC*, the *New York Times*, the *Daily Telegraph*, the *Wall Street Journal*, *Asahi Shim-Bun*, *Times of India*, the *Washington Post*, the *Guardian*, and *Daily Mail Online*, collecting tweets posted between the beginning of January 2015 and the end of August 2018 (within the approximate time frame of the IRA tweets). Tweets from the "real news" dataset that were posted after August 2018 were removed, as the data from the IRA tweets did not extend beyond that time frame. We started with 90,605 tweets, but with the removal of 10,602 tweets that had been posted in late 2018 and early 2019, we were left with 80,003 individual cases or tweets that exemplified "real" or "legitimate" news sources. For the purpose of Posit, SentiStrength and LibShortText analysis, a further research decision was made to random sample both datasets, creating two datasets of equal size, each consisting of 2,500 tweets, or roughly .001% of the larger "fake news" dataset, and 3% of the "real news" dataset. Unique identifiers were assigned to each of the data items, to ensure a means of fixed reference.

A somewhat different sample was assembled for the TensorFlow analysis, because for TensorFlow to operate effectively, a larger dataset is desirable. To achieve this, we combined the 2,116,904 English-speaking "fake news" tweets that remained (following the removal of all non-English cases) with the 90,605 "real news" tweets that were downloaded by the Dark Crawler (prior to removal of tweets that extended beyond the time frame of the IRA activities). This dataset was supplemented with 2,500 Facebook messages posted by the IRA, plus an additional "real news" set of Facebook items. Thus, a large dataset of 2,709,204 million tweets and Facebook posts was analyzed in TensorFlow following the merging of these multiple datasets.

For SentiStrength analysis and LibShortText analysis, we consolidated four smaller, 2,500 item datasets into one larger, 10,000 item dataset. This larger, 10,000 item dataset consisted of the above-mentioned set of 2,500 randomly sampled "fake news" Twitter messages derived from the dataset of 2,946,219 Twitter messages collected by Clemson University professors Linvill and Warren, the above-mentioned set of 2,500 randomly sampled "real news" Twitter messages derived from the 90,605 tweets collected by the Dark Crawler from traditional, mainstream news sources, plus 2,500 "fake news" posts from Facebook and 2,500 comparator "real news" posts [9]. The 2,500 "fake news" Facebook messages that formed part of this larger, 10,000 item dataset were posted on Facebook by Russia's Internet Research Agency between 2015 and 2017, and were

again collected and made available by Clemson University professors Linvill and Warren [54]. To secure a source of "real news" data for our comparison with the Facebook "fake news," we obtained a second "real news" dataset, this time of actual Facebook posts made available at github.com by data scientist Max Woolf. The data that we retrieved was originally comprised of 164 sets of publicly accessible Facebook status posts. From these status posts, we manually selected Facebook IDs that appeared to be associated with traditional news sources, such as *USA Today*, the *New York Times*, and *CNBC*. From these, we randomly selected 2,500 "real news" Facebook posts to serve as our comparator dataset [9].

C. Data Analysis

1) Qualtiative Textual Analysis

Qualitative textual analysis was conducted on the first 1,250 messages appearing in the above-mentioned set of 2,500 randomly sampled "fake news" Twitter posts, these 2,500 posts having been derived (winnowed down) from the dataset of 2,946,219 Twitter posts collected by Clemson University professors Linvill and Warren [54]. To express it differently, one half of the 2,500 randomly sampled "fake news" Twitter posts were read and classified manually. This process involved two experienced qualitative researchers, sitting side-by-side, reading each of the posts together, in many cases several times, until agreement on an appropriate classification was reached. Where there was disagreement, or where there was insufficient information upon which to arrive at a conclusion, the classification was designated as "undetermined." The classification for each of the 1,250 Twitter posts was recorded carefully in an Excel spreadsheet, with both researchers watching over each other's shoulder, to ensure the integrity of the data entry.

In a number of cases, the qualitative classification process included a Google search, to determine whether or not the content of the post was entirely fictional, partially true, or mostly true (i.e., grounded in "real news"). The two researchers were already familiar with some of the "real news" events that appeared and re-appeared in these posts, having conducted previous qualitative research on a different "fake news" dataset of messages emanating from the Russian Internet Research Agency, in this other case investigating fake Facebook accounts, rather than Twitter hashtags [9].

2) Posit

Following the creation and cleansing of the datasets, we extracted features from the texts using Posit, which is designed to generate quantitative data at the level of word and part-of-speech content of texts. Posit analysis was applied to each of the 5,000 tweets in order to produce a 27-item feature list for each tweet. This was supplemented by an additional feature, to indicate the "real" or "fake" characteristic of each tweet.

Previous research has indicated that Posit's domain-independent meta-data can be effective as a feature set for use in such text classification tasks [16], [17], [18]. In the present study, however, the target textual data was made up entirely of tweets. These have a limited maximum length of 280 characters, so they are inherently short and contain relatively few words. To illustrate, one of the tweets said only: "@realDonaldTrump True," while another said only: "Stay strong! #MAGA." With this shorter content in mind, Posit was extended such that the system supplemented the standard word-level statistics by generating an additional 44-character features for each instance of text data. As noted above, the result of this extended Posit analysis was that each data item (tweet) was represented by a set of 71 features, rather than the standard 27 features [1], [9].

The list of tweet features generated by Posit was formulated as an arff file format, suitable for direct input to the Waikato Environment for Knowledge Analysis (WEKA) data analysis application [65]. In WEKA, we applied the standard J48 tree classification method and the Random Forest classification method [66], [67], both with ten-fold validation. WEKA produced a measure of how many of the tweets were correctly classified.

3) TensorFlow

In this project, TensorFlow was used for processing the data with a Deep Neural Network (DNN) [56], [63]. A large dataset was initially fed into TensorFlow, in order to conduct DNN learning. The DNN results either updated an existing model or created a new model. TensorFlow then compared the same data against the constructed DNN model, and utilized that model to predict the category for each data entry.

In order to build an initial TensorFlow model, a large dataset of 2,709,204 million tweets was created by merging multiple datasets. The more data that could be collected for training a model, the better the accuracy should be. However, the individual data files were inconsistent, since they were collected from various online resources, and were formatted in very different ways. Thus, in the process of combining them into a single dataset, we opted for Microsoft Access, which allowed us to create a large, unified database table. All of the datasets were merged into this Access database, after which a class label column "*category*" was defined, denoting whether the data represented "fake" or "real" news.

The model was evaluated for its accuracy in predicting class values for the "fake" or "real" news category. To simplify the analysis, we decided to build our DNN model based on the content of the 2,709,204 tweets, without any further pre-processing. The DNN model used was a TensorFlow Estimator.DNNClassifier.

In the early stages of experimentation, we employed TensorFlow with default settings for the parameters pertaining to the number of partitions, epochs, layers, learning rate, and regularization. With respect to regularization, data was partitioned into groups according to the order in which it appeared in the dataset. Thus, if the majority of "fake news" appeared in the beginning of the dataset, it would be difficult to maintain consistent accuracy when conducting X-fold cross validation. To overcome this issue, the data was randomized as it became partitioned. Furthermore, each partition maintained the same data across all X-fold cross validation tests, so that the accuracy of the results could be compared properly.

With TensorFlow, epochs refer to the number of times the dataset is processed during training. The greater the number of epochs, the higher the accuracy tends to be. The learning rate determines the rate at which the model converges to the local minima. Usually, a smaller learning rate means it that it would take longer for the model to converge at the local minima [69]. With a larger learning rate, the model would get closer to this convergence point more quickly. The values for these parameters—i.e., the number of partitions, epochs, layers, learning rate, and regularization (L1 & L2)—were then tested to identify an optimal set of parameter values.

4) SentiStrength

For the SentiStrength analysis [61], [64], the general sentiment of the consolidated, 10,000 item dataset was first calculated without using any keywords. As there were no immediate trends identified between the "fake news" and "real news" items, keywords were generated using the top 100 nouns that appeared in the 10,000 posts. This produced a 100 x 10,000 matrix, against which we ran various algorithms in WEKA [56], again in an effort to distinguish between the "fake news" and "real news" items. This analysis included an examination of WEKA's decision trees, Naïve Bayes, BayesNet, and Multilayer Perceptron, the latter being a deep neural net algorithm, similar to that found in TensorFlow, in that it employs neurons, weights, and hidden layers [56], [63], [70], [71].

5) LibShortText

As noted earlier, LibShortText is an open source software package, developed by the Machine Learning Group at National Taiwan University. The use of LibShortText was recommended in a 2018 paper by William Yang Wang of the University of California at Santa Barbara, wherein he also described (and provided access to) his benchmark LIAR dataset. This LIAR dataset, which included 12,836 statements labeled for their subject matter, situational context, and truthfulness, was broken down into training, validation and test sets, and accompanied by instructions for automatic fake news detection [15]. For this particular research project, we employed LibShortText, but did not make use of William Wang's LIAR dataset. We plan to return to the LIAR dataset for purposes of additional machine training on short text items, as we progress in the development of our critical content toolkit.

LibShortText 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 also for micro- and macro-level error analysis [59]. For our research project, we built a model using the default settings that came with the LibShortText software. We employed the "$ python text-train.py trainfile" command which generated a "trainfile.model" for our given training file ("trainfile"). Working with this previously built model, we set out to predict the classification labels of the test set, or "trainfile" using the instructions: "$ python text-predict.py -f testfile trainfile.model predict_result," followed by "Option -f" to overwrite the existing model file and predict_result. The LibShortText software is available for free download from the National Taiwan University at: https://www.csie.ntu.edu.tw/~cjlin/libshorttext/.

IV. RESEARCH RESULTS

A. Qualitative (Textual) Analysis

As discussed in Section III (above), qualitative textual analysis was conducted on the first 1,250 messages that appeared in the set of 2,500 randomly sampled "fake news" tweets posted by the Internet Research Agency (IRA). These tweets were read and classified manually by two experienced qualitative researchers, who read the posts together, and jointly assigned an appropriate classification for each individual tweet. The classification for each of these 1,250 tweets was recorded in an Excel spreadsheet.

One of the patterns that became apparent early in the process was that close to one-third (31.92%, n = 399) of the 1,250 tweets that were read manually consisted of what could best be described as "apolitical chatter" (see Table I, below). Tweets that were classified as apolitical chatter did not appear to be re-circulating "real news," either to targeted or untargeted audiences. Moreover, they did not appear to be supporting the candidacy of either Donald Trump or Hillary Clinton, nor were they overtly attempting to advance divisive issues, create dissension, or otherwise undermine democratic processes. Examples of apolitical chatter would include tweets such as: "#TerribleHashTagIdeas MostRomanticKissAfterVomiting," "#ToFeelBetterI get high," "#ThingsYouCantIgnore Christmas sales," and "#DontTellAnyoneBut I prefer sex with the lights on."

There are a number of possible explanations when trying to account for the presence of so much apolitical chatter. One explanation could be that the Russian IRA simply did not get its money's worth when hiring some of these Internet trolls. To illustrate, whichever troll (or group of trolls) was responsible for the IRA hashtag BOOTH_PRINCE generated a disproportionate number of apolitical tweets, for example: "#ThereIsAlwaysRoomInMyLifeForDrake," "#tofeelbetteri think about Iphone 7S" and "#MyAmazonWishList FEMBOTS." On the other hand, the hashtag BOOTH_PRINCE also produced some Anti-Clinton tweets, such as: "A plastic fork too cut a steak #ThingsMoreTrustedThanHillary," and another referring sarcastically to then-Democratic President Barack Obama, and to Hillary Clinton's opponent in the Democratic primaries, Bernie Sanders: "#ObamasWishList Bernie, actually." Thus, it seems more likely that the political messaging was intentionally interspersed with a lot of apolitical chatter, in an effort to make these IRA-sponsored hashtags and tweets appear more akin to the type of discourse typically found on social media.

Another possible explanation for the high number of messages that we found necessary to classify as "apolitical chatter" would be the difficulties we encountered when trying to retrieve the videos or twitter feeds that were linked to these IRA tweets. While the tweets themselves seemed relatively innocuous, at least on the surface, it is conceivable

that they may have been targeted toward specific, pre-identified groups, and may have included links to political messaging and political advertising, as has been suggested by various other observers [7], [8], [9], [10],[11], [12], [13], [14], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38].

TABLE I. RESULTS OF QUALITATIVE, TEXTUAL ANALYSIS

Classification	Frequency	%
Apolitical Chatter	399	31.92
Pro-Trump/Anti-Clinton	328	26.24
Undetermined	171	13.68
Real News	152	12.16
Pro-Clinton/Anti-Trump	81	6.48
Racist	57	4.56
Helpful Advice	35	2.80
Anti-Racist	27	2.16
Total	**1250**	**100.00**

A similar explanation might apply to the 171 tweets (13.68%) where we could not arrive at a classification decision, and where the intent of those tweets thus ended up being classified as "undetermined" (see Table I, above). To illustrate the difficulties in the classification process, amongst the 1,250 tweets that we read manually, there were seven that contained a link to "#RejectedDebateTopics," and three that contained a link to "#BetterAlternativesToDebates," both of which were reportedly subsidiary hashtags created by the Russian troll army [72], [73]. An examination of what little remained of the Internet content from these two hashtags suggested that they were mostly pro-Trump or anti-Clinton, but there were embarrassing images of—and embarrassing statements about—Trump as well as Clinton; therefore, it was impossible to determine with certainty to which variety of Internet content the readers were being directed.

Despite the pro-Trump bias of "#RejectedDebateTopics," one of the Russian IRA tweets that we examined that was linked to this hashtag could safely be classified as being Pro-Clinton/Anti-Trump, because it irreverently asked: "Which Eastern European country will Trump's next wife come from?" However, another one from the same hashtag, #RejectedDebateTopics asked: "which Kardashian is least likely to have an STD?" The Kardashians were supporters of Hillary Clinton during the 2016 U.S. Presidential election, which allowed us to classify this second tweet as pro-Trump, anti-Clinton. On the other hand, one tweet that was linked to #RejectedDebateTopics asked simply: "who killed the Kennedy's ?" [sic], presumably referring to the Kennedy family, famous for producing Democratic President John F. Kennedy, Democratic Presidential candidate Robert Kennedy, and Democratic Senator Ted Kennedy. Another tweet, this time linked to #BetterAlternativesToDebates, talked about "smoke signaling using Bill's special cigars," perhaps referring to Bill Clinton, the former Democratic President of the United States, and the husband of Hillary Clinton. It is conceivable that both of the above-mentioned tweets were pro-Trump or anti-Clinton, but it was agreed that there was insufficient information to arrive at a decision in this regard. Thus, to err on the side of caution, both were assigned to the "undetermined" category.

Many of the 1,250 tweets that were read and classified manually were actually engaged in the re-circulation (or regurgitation) of "real news" stories, and thus ended up being classified as "real news." These "real news" tweets accounted for 12.16% (n = 152) of the dataset. This is comparable to our findings in a companion research project that involved the analysis of a sample of 2,500 Russian-generated Facebook posts, wherein we learned that 13.5% of the Facebook posts were based to one extent or another on recognizable, named entities, such as people, places, and specific dates or events [9]. Many of the "real news" tweets that are reported in this present paper were innocuous news stories and did not appear to be either pro-Trump or pro-Clinton. Examples of such tweets include: "The Latest: Sister says crash victim was retired from FBI," "San Antonio loses another popular radio star after on-air announcement #art," "University of Texas-Arlington police consider roaming robot" and "Texas appeals court overturns ex-Baylor player's conviction." Again, there are a couple of possible explanations, one being that the Russian IRA did not get its money's worth from these trolls, the other being that there was a concerted effort to make these IRA-sponsored hashtags and tweets look more like the typical discourse found on social media, the latter being the more likely of the two. To express it differently, they could be described as "background noise," intended to obfuscate the real motivation behind this online activity.

There were also 35 tweets that appeared to be providing "helpful advice." Examples of such "helpful advice" tweets would include: "Free And Cheap Things To Do In #London 27-28 January 2017 More Info Here," "How to Get Magazines to Review Your Music," and "Q&A: "What are trans fats and why are they unhealthful? #news." Again, it is entirely possible that there was a concerted effort to make these IRA-sponsored hashtags and tweets look more like the typical discourse found on social media, by throwing in some "chaff" with the "wheat," or that they were put in simply to create background noise. In any event, tweets that were classified as providing "helpful advice" were few and far between, comprising only 2.8% of the 1,250 "fake news" messages that were read and classified manually.

Of the 1,250 tweets analyzed manually, the 328 tweets that overtly supported the presidential candidacy of Donald Trump, or that were blatantly anti-Clinton, comprised the second largest group overall (after "apolitical chatter"), and vastly outnumbered the 81 tweets that supported the candidacy of Hillary Clinton (or in the alternative, were anti-Trump), by a ratio of four to one (see Table I, above). An example of a pro-Trump tweet that attempted to cover all of the main talking points of Trump and his supporters in one shot would be: "OUR MAN–He will get us out of the last 8 year mess against our Religion, Jobs, Illegal & Refugee Overkill, Homeless Vets & more–NEED HIM!." Other exemplars of unabashedly Pro-Trump tweets would be: "I just spoke to @realDonaldTrump and he fully supports my plan to replace Obamacare the same day we repeal it. The time to act is now," "Because all legal citizens vote Trump!

#VoteTrump," and "Trump is making manufacturing great again."

Tweets that were intended to undermine the campaign of Hillary Clinton, whilst simultaneously buttressing the campaign of Donald Trump, were in abundance: "Hillary Clinton to Fundraise with Anti-Christ (No, not Obama a different one)," "@SheriffClarke: If Trump made me his FBI Director I would be arresting Hillary Clinton today. #Comey," and "BREAKING: Julian Assange Is Back! And He Just Put The Nail In Hillary's Coffin." The latter tweet was clearly referring to the hack of the Democratic National Committee's email server by Russia's General Main Staff Intelligence Unit (the GRU), and to the subsequent leak of potentially embarrassing internal emails on WikiLeaks [7], [8]. Another tweet, again related to WikiLeaks, stated that: "WikiLeaks CONFIRMS Hillary Sold Weapons to ISIS... Then Drops Another BOMBSHELL! Breaking News." One anti-Clinton tweet targeted her daughter, saying: "Chelsea Clinton has received another award, this time for a day's worth of work." Yet another targeted Clinton's husband, saying: "Remember when Trump got a $1 million birthday gift from Saudi Arabia? Oh wait, that was Bill Clinton!"

There were 81 tweets (6.48% of the 1,250 tweets that were manually classified) that arguably supported Hillary Clinton, or in the alternative, talked negatively about Donald Trump, but most were not so blatantly in favour of one candidate over the other as the tweets supporting Donald Trump, or those attacking Hillary Clinton. To illustrate, one tweet that was classified as being pro-Hillary, anti-Trump, announced that: "Keith Ellison Plays Race Card, Claims Trump Brings White Supremacy to the White House." This tweet could also have been classified as "real news," in that it was reporting about Democrat Congressman Keith Ellison, who was running in 2018 for the Attorney General position in the state of Minnesota [74]. This was evidently some time after the 2016 Presidential campaign, but it has been widely reported that the Russian IRA carried on with its pro-Trump, pro-Republican, anti-Clinton, anti-Democrat agenda throughout 2017 and 2018 [75], [76]. The above tweet was "generously" classified as being anti-Trump, because it mentioned "Trump" and "White Supremacy" in the same breath. However, a closer examination of the news of the day might suggest that it could have been classified as pro-Trump, given that Ellison's invocation of the "race card" was viewed by some as a sign of desperation on his part. But to err on the side of caution, and in view of the oft-repeated claims by Donald Trump that alleged Russian interference in the 2016 Presidential election is "a hoax" or "fake news" [77], [78], this tweet was classified as being anti-Trump.

Another example of erring on the side of caution would be the following tweet: "BEHNA: ABSURD! Secret Service Agent Declares She Wouldn't Take A Bullet For Trump." This too was generously classified as being "anti-Trump," because it talked about a secret service agent who apparently would not perform her obligatory duties to protect the President, due to her personal animosity toward Donald Trump. It could just as well have been classified as "real news," because the story also appeared in mainstream news sources [79]. And it could arguably have been classified as "pro-Trump," because the sender of the tweet appeared to be saying that the behaviour of the secret agent was "absurd." However, we sought at all times to maintain a neutral stance in our qualitative textual analysis. In any event, if we had taken in-between messages such as these, that seemed at least on the surface to be saying something negative about Donald Trump, and classified them instead as "pro-Trump," then the four-to-one ratio of tweets in favour of Trump would have widened measurably.

Indeed, many of the tweets that were classified as pro-Clinton and/or anti-Trump could have gone either way or could have been classified as "undetermined" in their intent. The following tweet serves to illustrate this classification conundrum: "Donald Trump's Frog Meme 'SINISTER,' Clinton Campaign Warns." This tweet talked about a warning from the Clinton campaign concerning "sinister" activity on the part of Donald Trump and was thus classified as being pro-Clinton. However, it may well have been intended as a sarcastic "dig" toward Hillary Clinton and her team, or could even have been intended to direct the followers of the tweet to a video in which Donald Trump was "poking fun" at Clinton (this was not possible to verify, as the attachment has since been removed from the Internet, presumably because of the political fallout and furor following the detection of the Russian disinformation campaign).

This is not to say that there was a total dearth of pro-Clinton, anti-Trump messages. One example of a clearly pro-Clinton tweet would be: "#ImStillWithHer; She's #MyChoice #MyPresident #MyHero." Another example of a pro-Clinton tweet would be: "I think people also assume that folks who may vote for HRC won't push her. That couldn't be further from the truth." There were also a number of tweets that were clearly anti-Trump, such as: "The Latest: GOP senator says party has gone 'batshit crazy' #Texas," "Protesters in Texas seek release of Trump tax returns," and "Designer of Make America Great Again dress is an immigrant," not to mention "#anderr LOL : Mad Max Reveals THE EXACT MONTH Trump Will be Impeached." But such overtly pro-Clinton or anti-Trump messages were comparatively few and far between, and in many cases, had to be "teased out" of the dataset.

There were quite a few blatantly racist messages in the first 1,250 "fake news" tweets (4.56%, n = 57), some of which could arguably have been categorized as pro-Trump and anti-Clinton, as they mimicked Donald Trump's portrayal of Mexicans as criminals, drug traffickers and rapists [80], favoured his "Muslim ban" [81], supported his anti-immigration stance, and generally concurred with his description of Haiti, El Salvador and certain African nations as "shithole countries" [82], all of which was reportedly intended by Trump—and by the Russian IRA—to foster an atmosphere of distrust, divisiveness and fear with respect to immigrants and racial minorities, in order to "rile up" Trump's voter base [7], [8], [32], [33], [34]. Examples of anti-Mexican or Anti-Central American tweets include: "11 dumped from Rio Grande raft rescued by Border Patrol," and "A mayor was just shot dead in Mexico on the day after she took office." Messages targeting African-Americans were

also in evidence, with exemplars including such tweets as: "When a tall ass nigga, sees a short ass nigga w/ a tall girlfriend," or "Young Black folks keep saying they're not like the ancestors. And I keep saying that's the problem," or "They steal everything. Black folks have to be wiser." Anti-Muslim messaging could be found in abundance, in tweets such as: "'The Koran is a fascist book which incites violence. This book, just like Main Kampf [sic], must be banned.'- G. Wilders," or "5-year-old girl was raped by muslim immigrants and nobody's talking about that! #IsalmIsTheProblem," or "Did you know that Muslims are now allowed to have sex with slave woman even after their death?! #BanIslam." The overall thrust of these anti-Muslim, anti-immigration, anti-refugee messages is best encapsulated in the following tweet: "See those countless women and children? Neither do I https://t.co/tZOkWo7OjZ #banIslam #Rapefugees https://t.co/XoETkXAidV."

The above-mentioned racist messages, many of which clearly supported the Trump political agenda, were counter-balanced by approximately half as many anti-racist messages (2.16%, n = 27). Anti-racist tweets included the following: "We deserve to feel safe in our cars, our businesses, our parks, our homes and our churches. #BlackSkinIsNotACrime," "New Mexico Store in Trouble for Controversial Obama, Anti-Muslim Signs," and "34-year-old African-American man in Wisconsin brought 3 different documents to DMV & still couldn't get voter ID." Again, we imagine that the inclusion of this comparatively small number of anti-racist tweets was likely intended to offset the overtly pro-Trump, anti-immigration bias that was in evidence throughout the dataset, and to make these Russian-generated hashtags and tweets look more like the typical discourse found on social media. Quite apart from that, ostensibly anti-racist tweets such as these could actually have been crafted in such a way as to stoke fear and distrust among immigrants and racial minorities (thereby suppressing their vote), with comments about their overall lack of safety, and the difficulties that they could expect to experience when attempting to register to vote.

The findings of our qualitative textual analysis of the first 1,250 messages that appeared in the set of 2,500 randomly sampled "fake news" tweets posted by the Russian IRA strongly support the oft-reported conclusion that these Twitter feeds were intended to buttress the Presidential campaign of Donald Trump, and to stoke dissension, distrust, anger and fear in the American voting populace [7], [8], [9], [13], [26], [28], [35], [36], [37], [38]. Although it was sometimes difficult to tease out, we also adduced evidence that the hashtags and tweets were indeed generated by Russian sources, which runs counter to the White House narrative about "the Russian hoax" [77], [78]. To illustrate, one tweet announced: "Nikolai Nikolaevich Ge - _ Russian realist painter famous for his works on historical and religious motifs - was born today." Another noted that "The Russian band Leningrad is bringing its smashing program titled '20 Years for Joy' to the US," while still another reported that "For the first time since 2010, the MoscowState University has returned to the top 100 of QS World University Rankings global ranking." While such news stories might have held some interest for the Russian Internet trolls, it seems unlikely that they would have been of particular interest to American users of Twitter.

B. Posit Results

As noted earlier, the Posit toolkit generates frequency data and Part-of-Speech (POS) tagging while accommodating large text corpora. The Posit analysis produced a feature set with corresponding values for each of the 5,000 tweets, that is, the 2,500 "fake news" tweets and the 2,500 "real news" tweets. The feature set was loaded into WEKA as a basis for testing the feasibility of classification against the predefined "fake" and "real" news categories. Using the "standard" set of 27 Posit features—and the default WEKA settings with 10-fold cross validation—the J48 and Random Forest classifiers gave 82.6% and 86.82% correctly classified instances respectively. The confusion matrix for the latter performance is shown in Table II, below.

TABLE II. CONFUSION MATRIX FOR POSIT: 27 FEATURES (RANDOM FOREST: DEFAULT WEKA SETTINGS)

n=5,000	**Predicted: NEGATIVE**	**Predicted: POSITIVE**	
Actual: NEGATIVE	2,190	310	2,500
Actual: POSITIVE	340	2,160	2,500

As indicated previously, Posit was enhanced with an additional 44 character-based features, resulting in a total of 71 features, rather than the standard 27 features [1]. This was done in order to address the fact that tweets have a limited maximum length of 280 characters; thus, they are inherently short, and contain relatively few words. Using this extended feature set on the 5,000 tweets—and the default WEKA settings with 10-fold cross validation—the J48 and Random Forest settings classifiers gave 81.52% and 89.8% correctly classified instances respectively. The confusion matrix for the latter performance is shown in Table III, below.

Changing the number of instances (trees) from the default value of 100 to 211 in Random Forest provided a boost to the level of correctly classified instances to 90.12%. The confusion matrix for this performance is shown in Table IV, below.

TABLE III. CONFUSION MATRIX FOR POSIT: 71 FEATURES (RANDOM FOREST: DEFAULT WEKA SETTINGS)

n=5,000	**Predicted: NEGATIVE**	**Predicted: POSITIVE**	
Actual: NEGATIVE	2,266	234	2,500
Actual: POSITIVE	276	2,224	2,500

TABLE IV. CONFUSION MATRIX FOR POSIT: 71 FEATURES (RANDOM FOREST: INSTANCES AT 211 IN WEKA SETTINGS)

n=5,000	**Predicted: NEGATIVE**	**Predicted: POSITIVE**	
Actual: NEGATIVE	2,269	231	2,500
Actual: POSITIVE	263	2,237	2,500

Our best performance results (90.12%) were obtained from the Posit classification using the 71-feature set with Random Forest (instances at 211). The "detailed accuracy by class" for this result is shown in Table V.

TABLE V. DETAILED ACCURACY BY CLASS FOR BEST POSIT RESULT

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
NEGATIVE	0.908	0.105	0.896	0.908	0.902
POSITIVE	0.895	0.092	0.906	0.895	0.901
Weighted Avg.	0.901	0.099	0.901	0.901	0.901

Following these classification efforts using Posit, the two datasets (the real and fake tweets) were subjected to further analysis. The aim at this point was to determine whether any obvious characteristics in the data might skew the classification results. Several checks were made on the complexion of the two sets of data, focusing particularly on their relative content in terms of words and characters—since these features are the focus of the Posit analyses.

A comparison was made of the length of tweets in the two datasets. This revealed some differences in the distribution of tweets according to their length measured in words (Figure 1). Generally, distribution by length in words for the real news tweets rose above the curve for distribution by length in words for the fake tweets. Conceivably, this would ease the challenge of discriminating between the two datasets.

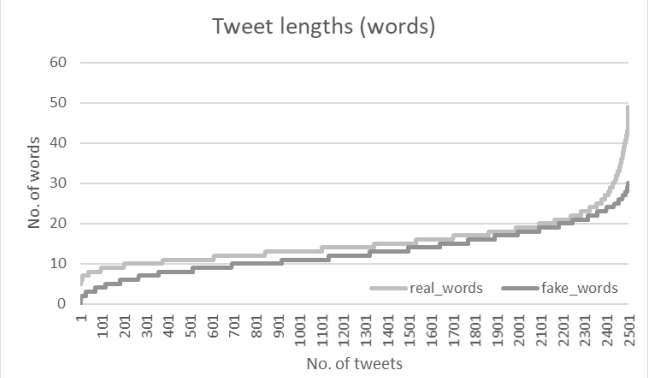

Figure 1. Comparison of Tweet Lengths (Words)

Since tweets are limited to 280 characters in length, a natural contrast was to consider the relative lengths of tweets by number of characters. This comparison (Figure 2), revealed a further distinctive trend in the real as opposed to the fake tweet content.

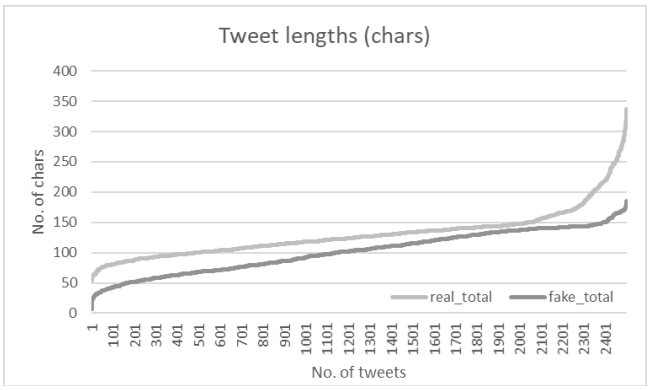

Figure 2. Comparison of Tweet lengths (Chars)

Figure 2 indicates that, as with length measured in number of words, length as measured by number of characters showed a distinctive trend for the real tweet content above the fake tweet content. As before, this may reasonably ease the task of differentiating real from fake tweets.

This post-classification analysis revealed one further notable insight on the character-lengths of real tweets. As shown in Figure 2 (above), some tweets with real content exceeded the 280-character maximum size permitted on Twitter. In total, twelve tweets in the real category of tweets were found to exceed 280 characters. Upon further investigation, this was found to be due to the presence of appended URLs in these tweets that had not been removed during the data cleaning stage. While this accounted for only 0.48% of the total real tweets, the excessive length of these tweets single them out as different from every example of fake tweet.

Additional insight on data complexion was derived from comparison of average and median values for length by words and length by characters (Table VI). This showed little difference in average and median tweet lengths in words and a wider separation in terms of characters.

TABLE VI. AVERAGE AND MEDIAN TWEET LENGTHS

	Real	**Fake**
Average tweet length (words)	15	13
Median tweet length (words)	14	12
Average tweet length (chars)	130	102
Median tweet length (chars)	125	104

A final contrast was made across the real and fake tweet datasets in terms of the use of specific characters. Two factors were considered: the presence of 'special characters' and the number of character types (i.e., unique characters) in the tweets.

The character-level Posit analysis generates several features based upon use of special characters for each data item. In this case, the special characters are full-stop, question mark, exclamation mark, dollar sign and asterisk, i.e., five possible special characters.

The contrast between real and fake tweet content in terms of how many different special characters appear in each tweet is illustrated in Figure 3 (below). This reveals notable

differences between the two types of tweet. Fake tweets avoid all special characters more commonly than real tweets. While many of both types deploy one special character, many more of the fake tweets deploy two special characters. There is less difference between the varieties of tweet at the three special character level while no tweets combine use four or five of these special characters.

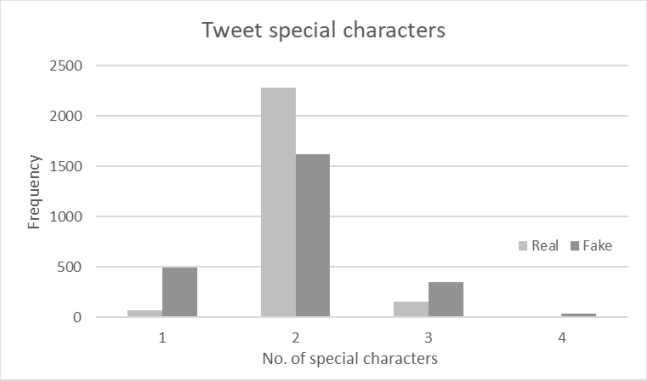

Figure 3. Comparison of Special Character Usage

Our comparison of the number of unique characters present across the tweet datasets, surveyed the presence of the alphanumeric set of characters (not case-sensitive) and the special characters noted above. This contrast is illustrated in Figure 4 (below) and further indicates a subtle difference between real and fake tweet content.

As noted earlier, the classification performance using Posit as a basis for feature generation gave a best performance match to the manual classification of 90.12%. While balanced in sample size, classification performance on this relatively small data subset of 2,500 real and 2,500 fake tweets, may have been influenced positively by the data characteristics described above. As a step toward eliminating such a potential anomaly, we deployed a much larger dataset when classifying with Tensorflow.

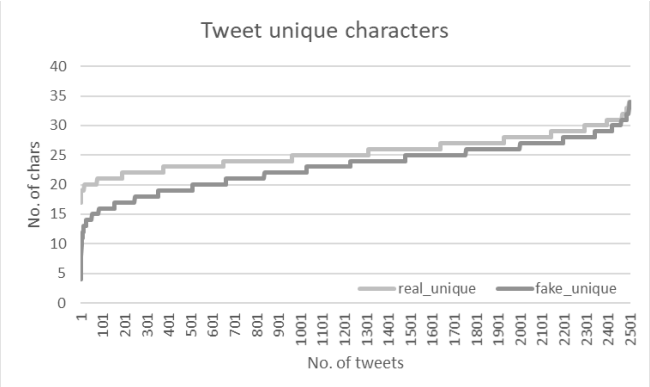

Figure 4. Comparison of Unique Character Usage

C. TensorFlow Results

Recall that in this project, TensorFlow (developed by Google Brain) was used for processing the data with a Deep Neural Network (DNN) [56], [63]. Posit analyzed a randomized sample of 5,000 tweets, that is, the 2,500 "fake news" tweets, and the 2,500 "real news" tweets. A much larger dataset, consisting of 2,709,204 million tweets and Facebook posts, was fed into TensorFlow upon commencement, in order to conduct DNN learning. Then, the DNN results either updated an existing model, or created a new model. TensorFlow next compared the same data against the constructed DNN model, and utilized that model to predict the category for each data entry. In the early stages of experimentation, using default TensorFlow parameters for number of partitions, epochs, layers, learning rate, and regularization, the accuracy results yielded an average of around 60%. Many parameter values (for each parameter: number of partitions, epochs, layers, learning rate, and regularization) were then tested to identify an optimal set of parameter values. This resulted in an increase in accuracy to 89.5%, a substantial improvement from the earlier results. These parameters are described below, with the post-training optimal values shown in Table VII.

To be able to run large numbers of experiments, we wrapped all code into a standalone function, so that large numbers of various scenarios could be designed, set up, and tested continuously. These batch jobs allowed us to evaluate different combinations of parameters. The parameters of each run, and the corresponding results, are also shown below. Tests were run using 10 partitions, with training on the first 5 partitions, and testing on the last 5 partitions.

D. SentiStrength Results

SentiStrength assigns positive or negative values to lexical units in the text [61], [64]. Recall that this value is a measure that provides a quantitative understanding of the content of information—specifically, the extent to which positive and negative sentiment is present. The program automatically extracts the emotions or attitude of a text and assigns a value that ranges from "negative"' to "neutral" to "positive." For SentiStrength analysis, we consolidated four smaller, 2,500 item datasets into one larger, 10,000 item dataset. This larger, 10,000 item dataset consisted of the set of 2,500 randomly sampled "fake news" Twitter messages, the set of 2,500 randomly sampled "real news" Twitter messages, the set of 2,500 "fake news" posts from Facebook, and the set of 2,500 comparator "real news" posts.

For initial SentiStrength analysis, the "general sentiment" was calculated (i.e., without keywords), but all scores were negative, without any apparent distinguishing trends—between "fake news" and "real news," or between Twitter items and Facebook items. We then proceeded to use keywords, by calculating the top 100 nouns out of the 10,000 posts, and running sentiment analysis again, this time with respect to the 100 identified nouns. This produced a 10,000 x 100 matrix (4 x 2,500 = 10,000 rows, one for each post, and 100 columns for each noun, or keyword). On this matrix, we ran various algorithms using WEKA [65] and TensorFlow [56], in an effort to differentiate between the four classes, that is, "fake news" Twitter messages, "real news" Twitter messages, "fake news" posts from Facebook, and the comparator "real news" posts. This too proved to be futile, as

there were too many missing values for the decision trees to handle properly, and given that the reported predictive accuracy was not much better than a random guess.

TABLE VII. TENSORFLOW PERFORMANCE RESULTS

Layers	Learn Rate	Partition	Size	Time	Accuracy
[500, 500]	0.003	0	674941	44.683	0.873
[500, 500]	0.003	1	675072	48.102	0.873
[500, 500]	0.003	2	674613	45.654	0.873
[500, 500]	0.003	3	675109	45.638	0.873
[500, 500]	0.003	4	9479	2.562	0.871
[700, 700]	0.003	0	674941	217.444	0.873
[700, 700]	0.003	1	675072	57.929	0.874
[700, 700]	0.003	2	674613	59.508	0.873
[700, 700]	0.003	3	675109	58.923	0.873
[700, 700]	0.003	4	9479	3.020	0.872
[500, 500]	0.03	0	674941	128.865	0.882
[500, 500]	0.03	1	675072	59.551	0.882
[500, 500]	0.03	2	674613	60.684	0.881
[500, 500]	0.03	3	675109	61.396	0.882
[500, 500]	0.03	4	9479	3.205	0.895

Finally, as were primarily interested in distinguishing "fake news" from "real news," we collapsed the four datasets into two classes, "real news" and "fake news," each consisting of 5,000 items. The results of this final sentiment analysis are shown in Table VIIII (below). While the BayesNet and Naïve Bayes indicated 56.85% and 58.06% of correctly classified instances respectively, these would be considered barely better than random guesses, at 50.00%. However, the MultiLayer Perceptron, a deep neural net algorithm, similar to that found in TensorFlow, in that it employs neurons, weights, and hidden layers [56], [63], [69], [71], yielded a classification accuracy of 74.26%. This would be considered "acceptable," or at least more acceptable than barely better than random guesses, but not up to the standards that we are presently seeking.

TABLE VIII. DETAILED ACCURACY BY ALGORITHM FOR BEST SENTISTRENGTH RESULT

Algorithm	Accuracy
Random Guess	50.00%
Decision trees	50.33%
BayesNet	56.84%
Naïve Bayes	58.06%
Multilayer Perceptron	74.26%

E. LibShortText Analysis

For LibShortText analysis, we again consolidated four smaller, 2,500 item datasets into one larger, 10,000 item dataset, identical to the one used for the SentiStrength analysis (see above). This 10,000 item dataset was split into two randomly sorted 5,000 item datasets, one for training purposes, and the other for testing purposes. For our research project, we built a model using the default settings that came with the LibShortText software [68]. On the first attempt, our classification accuracy was 80.56%, substantially better than the accuracy yielded by the SentiStrength analysis. Our second attempt resulted in a classification accuracy of 90.2%, comparable to the classification accuracy yielded by Posit, at 90.12%, albeit using a larger and more diverse dataset than the one input to Posit.

V. DISCUSSION

We were disappointed with the SentiStrength analysis, given that when we combined SentiStrength with the WEKA standard J48 decision-tree classification method in an earlier study of online extremist content, we were able to correctly classify 80.51% of the webpages [16]. In fact, with our earlier extremism study, the binary anti-extremist and pro-extremist categories had even higher degrees of correctly identified pages, with 92.7% of the pro-extremist cases and 88% of the anti-extremist cases correctly identified. This indicated to us that the decision tree worked well when it came to classifying extremist content [16]. In this present study, the MultiLayer Perceptron (a deep neural net algorithm) yielded a classification accuracy of 74.26%, which is comparable to the results of other studies that have employed sentiment analysis on tweets [59], [83]. We are hoping that further machine training, perhaps enhanced by an expanded list of keywords provided by the ongoing qualitative analysis, will improve upon these SentiStrength results.

TensorFlow epitomizes machine-learning and artificial intelligence, in that it gradually teaches itself, once provided with sufficient data and the requisite training/learning epochs. It is anticipated that the predictive accuracy of the TensorFlow component will ultimately exceed 90% once it is fully trained and fully operational. In a current trial experiment, we demonstrated that the predictive accuracy of TensorFlow does indeed improve with the amount of inputted data. For the first round of analysis, we randomly selected 10,000 Facebook items from another "real news" dataset and 10,000 items from another "fake news" dataset that we had recently generated using the Dark Crawler, next merging and shuffling the two files to create one file containing 10,000 Facebook items. In this case, the predictive accuracy of TensorFlow was only 48.65% when analyzing the content alone, and 50.4% when analyzing the content along with tagged text generated by Posit. On the other hand. TensorFlow's predictive accuracy increased to 79.84% and 79.94% (with the Posit features) when we used 90,000 Facebook items from our "real news" dataset and 10,000 items from our "fake news" dataset to create a larger file containing 100,000 Facebook items.

That said, TensorFlow requires big data and significant processing times. Thus, while TensorFlow will be instrumental in analysing the massive amount of data to be harvested, it will likely not be capable of providing the type of near-real-time alerts on hostile information activities required for our anticipated "critical content toolkit." Rather, we expect that it will provide ongoing, deep-level analysis of all of the data as it is collected, and assist in the building of new models in response to any changes in the strategies and tactics of hostile foreign actors. As a consequence, we anticipate that we will be turning to other (companion) models to enhance the prospects for near-real-time alerts.

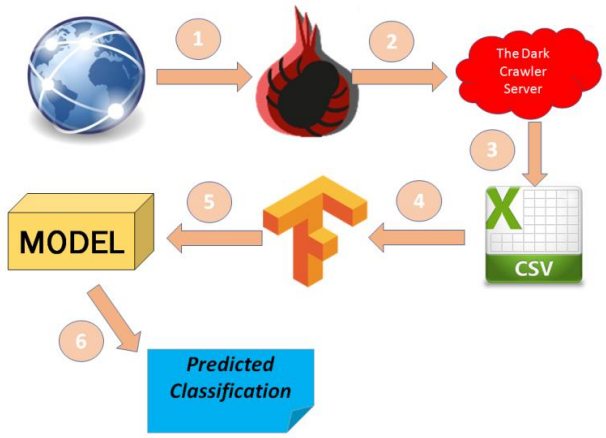

Figure 5. The TensorFlow Model

The TensorFlow model on which we are presently working (see Figure 5, above) commences with The Dark Crawler searching the Internet and downloading all relevant content onto The Dark Crawler server. The data from the stored content is then converted into an Excel file containing all of the pertinent information for each individual data item (e.g., the time and date of the message, text, or post; the hashtag, Facebook page or publication source; the forum and subforum, if taken from a forum; the Internet address, if available; the title of the text or message, if any; the body of the text or message; the number of likes, re-posts or re-tweets; etc.). This data is input to TensorFlow for deep neural network analysis, leading to the generation of a model for measuring the presence of hostile information activities on the Web—a tool which will then predict/classify social media messaging and other sources of online news as "fake" or "real."

Given the limited number of words and word varieties in most tweets, the performance of the Posit analysis using the default 27 word-level features proved to be better than expected, with 86.82% correctly classified instances using Random Forest. The addition of character-level information enhanced this performance to a creditable 90.12% correctly classified instances, again using Random Forest. This result was somewhat surprising, given that alphanumeric details seem far removed from tweet content-level [1].

The Posit toolkit is limited by the speed at which it can read and analyze large volumes of text. Posit is not as slow as TensorFlow, and when combined with WEKA, it has an initial classification accuracy that exceeds that of TensorFlow and in some cases matches that of LibShortText. Nevertheless, while Posit does not excel in reading and analyzing large text corpora as quickly as LibShortText, or in analyzing the vast amounts of data that can be input into TensorFlow for machine learning purposes, it does bring an entirely different dimension to the model that we are building, in that the Posit toolkit generates frequency data and Part-of-Speech (POS) tagging, with data output including values for total words (tokens), total unique words (types), type/token ratio, number of sentences, average sentence length, number of characters, average word length, noun types, verb types, adjective types, adverb types, preposition types, personal pronoun types, determiner types, possessive pronoun types, interjection types, particle types, nouns, verbs, prepositions, personal pronouns, determiners, adverbs, adjectives, possessive pronouns, interjections particles. As Posit recognizes and records individual words and characters, it can aid significantly in the adaptation of the overall model to the changing strategies and tactics of hostile foreign actors, and at the same time, glean unique keywords or key phrases from incoming data so that the activities of hostile foreign actors can be identified quickly and targeted more precisely.

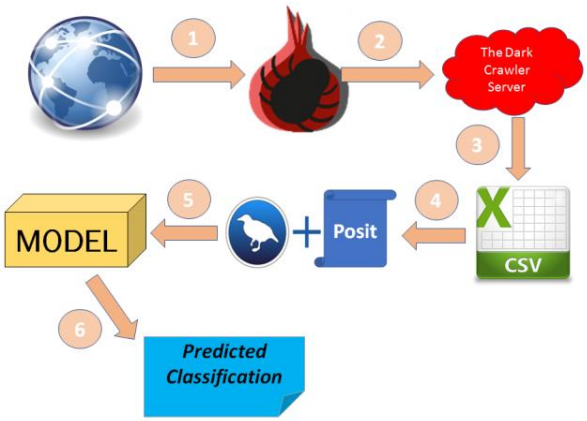

Figure 6. The Posit/WEKA Model

The Posit model that we envision (see Figure 6, above) differs from the TensorFlow Model, in that once the data is harvested, organized, and ready for input, it first goes into Posit for analysis, and then into WEKA for secondary assessment of classification accuracy. Posit (in combination with WEKA) has at times generated classification accuracy in the 98-99% range when it comes to processing various of the recently generated data sets that we have on hand.

The LibShortText results were very encouraging, with a creditable classification accuracy of 90.2%, comparable to the 90.12% classification accuracy yielded by Posit. Recently, in conjunction with our work with LibShortText, we downloaded and configured LibLinear, a companion open source software package, again developed by the same Machine Learning Group at National Taiwan University that developed LibShortText [84]. LibShortText is a text analysis program, while LibLinear is a classification program. LibLinear predicts the accuracy of the classification performed by LibShortText, much like WEKA predicts the accuracy of the classification performed by Posit. Another advantage to LibLinear is that is supports incremental and decremental learning, or to express it differently, the addition and removal of data in order to improve optimization and decrease run time. LibShortText, on the other hand, does not readily support updating of the model.

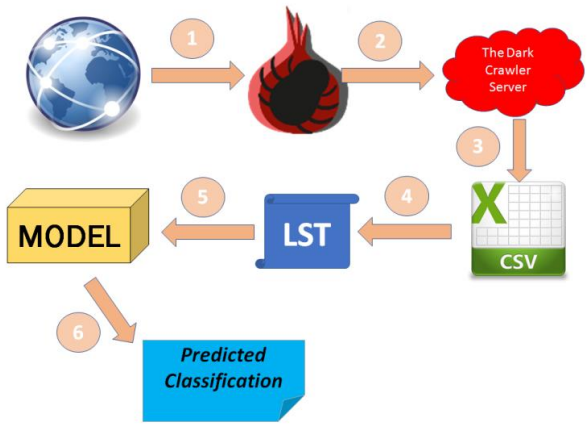

Figure 7. The LibShortText/LibLinear Model

Generally speaking, LibShortText and LibLinear have been outperforming TensorFlow and Posit in a number of our current trial experiments. To illustrate, when analyzing 1,000 randomly selected data items taken from our more recently generated "real news" datasets, contrasted with 1,000 randomly selected data items taken from our more recently generated "fake news" datasets, we found that LibShortText and LibLinear exhibited classification accuracies of 93% and 92% respectively, as opposed to Posit and WEKA at 72.7%, TensorFlow (using Posit-generated .arff content at 54.5%, TensorFlow (using content only) at 52.5%, and TensorFlow (using tagged text) at 48%. We would consider these TensorFlow numbers to be no better than tossing a coin, but these results were not entirely unexpected, as TensorFlow thrives on large data, and this experiment was conducted using only 2,000 discrete data items.

This LibShortText/LibLinear Model (see Figure 7, above) is essentially the same as the TensorFlow Model set out in Figure 5 (above), in that it commences with The Dark Crawler searching the Internet and downloading all relevant content onto The Dark Crawler server. The data from the stored content is then converted into an Excel file containing all of the pertinent information for each individual data item. This data is input to either LibShortText or LibLinear for the generation of a model for measuring the presence of hostile information activities on the Web—a tool which will then predict/classify social media messaging and other sources of online news as "fake" or "real," much more quickly than TensorFlow or Posit.

In this model, LibShortText and LibLinear can be used almost interchangeably, in most cases without unduly affecting the processing times or predictive accuracy. We did, however, encounter limitations with LibShortText on the training and testing file sizes when using only 4GB of RAM. We received "memory exhausted" notifications, and instructions to "restart python." After upgrading to 32GB of RAM, this problem was resolved. That said, the required RAM size is an issue to be borne in mind as we design the final model.

VI. CONCLUSION

Through the research process outlined above, we have: 1) developed typologies of past and present hostile activities in Cloud-based social media platforms; 2) identified indicators of change in public opinion (as they relate to hostile disinformation activities); 3) identified the social media techniques of hostile actors (and how best to respond to them); and 4) undertaken cross-cultural analyses, to determine how hostile actors seek to fuel tensions and undermine social cohesion by exploiting cultural sensitivities.

Our current research will ultimately generate an algorithm that can automatically detect hostile disinformation content. In the longer term, we will use the knowledge generated by this research project to further expand and integrate the capabilities of the Posit toolkit and the Dark Crawler, in order to facilitate near-real-time monitoring of disinformation activities in the Cloud. Further, we plan to add a feature that will permit us to capture disinformation messages prior to their removal by social media organizations attempting to delete those accounts, and/or their removal by actors seeking to conceal their online identities. Ideally, this integrated, "critical content toolkit" will be able to recalibrate itself when confronted with ever-changing forms of disinformation.

During the research process, we also downloaded 2,500 "fake news" Facebook messages that had been posted by the IRA on Facebook pages known variously as Blacktivist, Patriototus, LGBT United, Secured.Borders, and United Muslims of America. (These 2,500 Facebook messages were included in our TensorFlow, SentiStrength and LibShortText analysis). All 2,500 of these messages have been subjected to a preliminary review in the qualitative research tool, NVivo, and also, to preliminary review in Posit [9]. Early insights from this companion study revealed that many of the allegedly "fake news" items were founded to one degree or another in contemporaneous "real news" events.

Following the initial rounds of data collection described earlier in this paper, we broadened and enriched our selection of data sources, focussing primarily on Facebook, Twitter, and other web-based news sources. A "fake news" list of Facebook pages was generated by searching for Facebook pages that belonged to websites described by MediaBiasFactCheck.com as coming from "questionable sources." MediaBiasFactCheck was founded and is edited by Dr. David Van Zandt—a professor, lawyer, and current president of The New School—along with his team of volunteers. In all, we harvested 96,219 Facebook "fake news" items, posted between January 2014 and September 2019. This was recently supplemented by a set of 3,736 Canadian Facebook "fake news" items, posted from May 2014 up to the present.

Data for the expanded Twitter dataset, specifically assembled by the research team for this ongoing project, were also extracted the same way as the set of "fake" Facebook posts, that is, by using the list of 530 "questionable sources" published by MediaBiasFactCheck.com. From this, 181 Twitter accounts were identified for data collection,

accounting for 43,193 data items posted between March 2009 up to the present. Only Twitter accounts that contained a link to the websites identified as suspect by MediaBiasFactCheck.com were included in this sample.

Our third category of "fake news" was recently derived from Web sites presenting themselves as legitimate sources of real news but considered "fake." News articles were collected from four publicly available datasets: (1) *ISOT Fake News*, (2) *Getting Real About Fake News*, (3) *Fake News Corpus*, and (4) *FA-KES: A Fake News Dataset around the Syrian War*. *ISOT Fake News* was created by the Information Security and Object Technology (ISOT) research lab at the University of Victoria [85]. The dataset contains both fake and real news. The former was obtained from websites considered unreliable by Politifact, a website dedicated to fact-checking U.S. news. Real news was obtained from the website Reuters.com. In total, there were 21,417 real news and 23,481 fake news items. *Getting Real About Fake News* was created in 2016 by Megan Risdal, a Product Lead at Kaggle (an online data science community). This dataset contains 12,999 news articles from 244 sources obtained from the BS Detector Chrome extension. The articles are labeled according to their credibility as fake, conspiracy, hate, bias, satire, junk science, and "bullshit."

Fake News Corpus is an open source dataset from 2018 that contains 9,408,908 news articles, created by GitHub user "several27." News articles were obtained from a list of 745 domains from www.opensources.com, as well as the *New York Times* and webhose English news articles. For the current project, after cleansing the dataset by removing unlabelled items, we have retained 779,882 fake news items and 1,783,529 credible news items. Finally, the *FA-KES* dataset, created at the American University of Beirut with the intention of helping train machine learning models, contains 805 news articles about the conflict in Syria, of which 46 are labelled as "fake," with the remaining 378 labelled as "real" [86].

Comparator "real news" Facebook and Twitter data sets have been collected from official news sources representing the top 24 Canadian newspapers in accordance with their known circulation in 2016. We also included *Huffington Post Canada* and two TV News sources with large online followings—*CBC News* and *CTV News*. Apart from the *CBC*, *CTV* and the Canadian edition of the *Huffington Post*, we obtained data from 24 sources, for example, *The Globe and Mail*, *The National Post*, *The Toronto Star*, *Le Journal de Montreal* (French), *Le Journal de Quebec* (French), *Le Soleil* (French), *The Vancouver Sun*, *The Toronto Sun*, *The Calgary Herald*, *The Winnipeg Free Press*, *The Ottawa Citizen*, and *The Montreal Gazette*, to mention a few of the sources. In total, we recently collected 31,557 "real news" Facebook data items from these "trustworthy" news sources, dating from July 2018 to the present. We also collected 253,936 "real news" Twitter data items from these "trustworthy" news sources, dating from December 2013 through September 2019.

This vast databank of recently acquired "real news" and "fake news" (and everything in between real and fake) has been assembled for use in conjunction with our ongoing qualitative analysis, as well as to provide a basis for our ongoing quantitative analysis and machine-learning-based classification. In fact, data drawn from these new datasets were used in our recent comparison tests involving Posit, TensorFlow, LibShortText and LibLinear, as outlined above in our Discussion section. The data collection and data analysis processes are in progress and robust. We anticipate developing a "proof-of-concept" model of our "critical content toolkit" in the near future.

ACKNOWLEDGMENTS

This research project would not have been possible without funding from the Cyber Security Cooperation Program, operated by the National Cyber Security Directorate of Public Safety Canada. We would also like to thank our research assistants, Soobin Rim (TensorFlow) and Aynsley Pescitelli (NVivo).

REFERENCES

[1] B. Cartwright, G. R. S. Weir and R. Frank, "Fighting Disinformation Warfare with Artificial Intelligence: Identifying and Combatting Disinformation Attacks in Cloud-based Social Media Platforms," *Tenth International Conference on Cloud Computing, GRIDs, and Virtualization*, pp. 73-77, May 2019. URL: http://thinkmind.org/index.php?view=article&articleid=cloud_computing_2019_5_30_28006 [Last accessed: 2019.07.28]

[2] D. Ebner and C. Freeze, "AggregateIQ, Canadian data firm at centre of global controversy, was hired by clients big and small," *Globe and Mail*, April, 2018. URL: www.theglobeandmail.com/canada/article-aggregateiq-canadian-data-firm-at-centre-of-global-controvery-was [Last accessed: 2019.04.8]

[3] R. Rathi, "Effect of Cambridge Analytica's Facebook ads on the 2016 US Presidential Election," *Towards Data Science*, 2019. URL: https://towardsdatascience.com/effect-of-cambridge-analyticas-facebook-ads-on-the-2016-us-presidential-election-dacb5462155d [Last accessed: 2019:07.20]

[4] J. Russell, "UK watchdog hands Facebook maximum £500K fine over Cambridge Analytica data breach," *TechCrunch*, 2018. URL: https://techcrunch.com/2018/10/25/uk-watchdog-hands-facebook-500k-fine/ [Last accessed: 2019.07.18]

[5] M. H. McGill and N. Scola, "FTC approves $5B Facebook settlement that Democrats label 'chump change,'" *Politco*, July 12, 2019 URL: https://www.politico.com/story/2019/07/12/facebook-ftc-fine-5-billion-718953 [Last accessed: 2019.07.18]

[6] I. Lapowsky, "House Probes Cambridge Analytica on Russia and Wikileaks," *Wired*, 2019. URL: https://www.wired.com/story/congress-democrats-trump-inquiry-cambridge-analytica/ [Last accessed: 2019.07.20]

[7] Office of the Director of National Intelligence, "Assessing Russian Activities and Intentions in Recent US Elections," 2017. URL: www.dni.gov/files/documents/ICA_2017_01.pdf [Last accessed: 2019.07.28]

[8] R. S. Mueller III, "Report on the Investigation into Russian Interference in the 2016 Presidential Election," pp. 1-448, 2019. URL: www.justsecurity.org/wp-content/uploads/2019/04/Muellller-Report-Redacted-Vol-II-Released-04.18.2019-Word-Searchable.-Reduced-Size.pdf [Last Accessed: 2019.07.28]

[9] B. Cartwright, G. R. S. Weir, L. Nahar, K. Padda and R. Frank, "The Weaponization of Cloud-Based Social Media: Prospects for Legislation and Regulation," *Tenth International Conference on Cloud Computing, GRIDs, and Virtualization*, pp. 7-12, May 2019. URL: http://thinkmind.org/index.php?view=article&articleid=cloud_computing_2019_2_10_28021 [Last accessed: 2019.07.28]

[10] M. T. Bastos and D. Mercea, "The Brexit botnet and user-generated hyperpartisan news," *Social Science Computer Review*, 0894439317734157, 2017. URL: https://journals-sagepub-com.proxy.lib.sfu.ca/doi/pdf/10.1177/0894439317734157 [Last Accessed: 2019.07.28]

[11] M. Field and M. Wright, "Russian trolls sent thousands of pro-Leave messages on day of Brexit referendum, Twitter data reveals: Thousands of Twitter posts attempted to influence the referendum and US elections," *The Telegraph*, 2018. URL: www.telegraph.co.uk/technology/2018/10/17/russian-iranian-twitter-trolls-sent-10-million-tweets-fake-news/ [Last accessed: 2019.04.8]

[12] G. Evolvi, "Hate in a Tweet: Exploring Internet-Based Islamophobic Discourses," *Religions*, *9*(10), pp. 37-51, 2018. URL: https://www.mdpi.com/2077-1444/9/10/307 [Last accessed: 2019.07.21]

[13] A. Badawy, E. Ferrara and Lerman, K., "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," *arXiv*, 2018 URL: https://arxiv.org/abs/1802.04291 [Last accessed: 2019.07.28]

[14] C. Shao, P. M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer and and G. L. Ciampaglia, "Anatomy of an online misinformation network," PloS one, 13(4), e0196087, 2018. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196087 [Last Accessed: 2019.07.28]

[15] W. Y. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," *arXiv preprint arXiv:1705.00648*, 2018. URL: https://arxiv.org/abs/1705.00648 [Last accessed: 2019.07.15]

[16] G. Weir, R. Frank, B. Cartwright and E. Dos Santos, "Positing the problem: enhancing classification of extremist web content through textual analysis," *International Conference on Cybercrime and Computer Forensics (IEEE Xplore)*, June 2016. URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/7740431[Last accessed: 2019.08.13]

[17] G. Weir, K. Owoeye, A. Oberacker and H. Alshahrani, "Cloud-based textual analysis as a basis for document classification," *International Conference on High Performance Computing & Simulation (HPCS)*, pp. 672-676, July 2018. URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/8514415 [Last accessed: 2019.08.13]

[18] K. Owoeye and G. R. S. Weir, "Classification of radical Web text using a composite-based method, *IEEE International Conference on Computational Science and Computational Intelligence*, December 2018. URL: https://pure.strath.ac.uk/ws/portalfiles/portal/86519706/Owoeye_Weir_IEEE_2018_Classification_of_radical_web_text_using_a_composite_based.pdf [Last accessed: 2019.08.13]

[19] H. Berghel, "Lies, damn lies, and fake news," *Computer*, 50(2), pp. 80-85, 2017. URL: https://www.computer.org/csdl/magazine/co/2017/02/mco2017020080/13rRUzp02jw [Last accessed: 2019.07.29]

[20] N. W. Jankowski, "Researching fake news: A selective examination of empirical studies," *Javnost-The Public*, 25(1-2), pp. 248-255, 2018. URL: https://www-tandfonline-com.proxy.lib.sfu.ca/doi/full/10.1080/13183222.2018.1418964 [Last accessed: 2019.07.29]

[21] E. C. Tandoc Jr, Z. W. Lim and R. Ling, "Defining 'fake news': A typology of scholarly definitions," *Digital Journalism*, 6(2), pp. 137-153, 2018. URL: https://www.researchgate.net/publication/319383049_Defining_Fake_News_A_typology_of_scholarly_definitions [Last accessed: 2019.07.29]

[22] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer and M. Schudson, "The science of fake news," *Science*, *359*(6380), pp. 1094-1096, 2018. URL: https://science-sciencemag-rg.proxy.lib.sfu.ca/content/359/6380/1094 [Last Accessed: 2019.07.21]

[23] M. de Cock Buning, L. Ginsbourg and S. Alexandra, *Online Disinformation ahead of the European Parliament elections: toward societal resilience*, European University Institute, School of Transnational Governance, April 2019 URL: https://cadmus.eui.eu/bitstream/handle/1814/62426/STG_PB_2019_03_EN.pdf?sequence=1&isAllowed=y [Last accessed: 2019.07.15]

[24] S. Desai, H. Mooney and J.A. Oehrli, *"Fake News," Lies and Propaganda: How to Sort Fact from Fiction*, 2018. URL: https://guides.lib.umich.edu/fakenews [Last accessed: 2019.07.15]

[25] N. Kshetri and J. Voas, "The Economics of 'Fake News,'" *IEEE Computer Society*, pp. 8-12, (2017). URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/stamp/stamp.jsp?tp=&arnumber=8123490&tag=1 [Last accessed: 2019.07.15]

[26] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 31(2), pp. 211-236, 2017. URL: https://web.stanford.edu/~gentzkow/research/fakenews.pdf [Last Accessed: 2019.07.28]

[27] W. L. Bennett and S. Livingston, "The disinformation order: Disruptive communication and the decline of democratic institutions," *European Journal of Communication*, 33(2), pp. 122-139, 2018. URL: https://journals-sagepub-com.proxy.lib.sfu.ca/doi/pdf/10.1177/0267323118760317 [Last accessed: 2019.07.28]

[28] *United States v. Internet Research Agency LLC*, Case 1:18-cr-00032-DLF, The United States District Court for the District Of Columbia, February 26, 2018. URL: www.justice.gov/file/1035477/download [Last accessed: 2019.04.8]

[29] J. J. Green, "Tale of a Troll: Inside the 'Internet Research Agency' in Russia," *WTOP*, 2018. URL: https://wtop.com/j-j-green-national/2018/09/tale-of-a-troll-inside-the-internet-research-agency-in-russia/ [Last accessed: 2019.07.15]

[30] L. Reston, "How Russia Weaponizes Fake News: The Kremlin's influence campaign goes far beyond Trump's victory. Their latest unsuspecting targets: American conservatives," *The New Republic*, 2017. URL: https://newrepublic.com/article/142344/russia-weaponized-fake-news-sow-chaos [Last accessed: 2019.07.20]

[31] K. Wagner, "Facebook and Twitter worked just as advertised for Russia's troll army: Social platforms are an effective tool for marketers — and nation states that want to disrupt an election," *Recode Daily*, 2018. URL: https://www.vox.com/2018/2/17/17023292/facebook-twitter-russia-donald-trump-us-election-explained [Last accessed: 2019.07.20]

[32] A. Marwick and R. Lewis, *Media Manipulation and Disinformation Online*, New York: Data & Society Research Institute, pp. 1-106, 2017. URL: https://datasociety.net/output/media-manipulation-and-disinfo-online/ [Last accessed: 2019.07.29]

[33] K. Shu, A. Silva, S. H. Wang, J. Tang and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," pp. 1-15, 2017. URL: https://arxiv.org/abs/1708.01967 [Last accessed: 2019.07.29]

[34] S. Zanettou, T. Caulfied, E. de Cristofaro, M. Sirivianos, G. Stringhini and J. Blackburn, "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web," pp. 1-11, 2019. URL: https://arxiv.org/pdf/1801.09288.pdf [Last accessed: 2019.07.29]

[35] M. Papenfuss, "1,000 Paid Russian Trolls Spread Fake News On Hillary Clinton, Senate Intelligence Heads Told," *Huffington Post*, March 2017. URL: https://www.huffingtonpost.ca/entry/russian-trolls-fake-news_n_58dde6bae4b08194e3b8d5c4 [Last accessed: 2019.07.29]

[36] The Computational Propaganda Project, "Resource for Understanding Political Bots," 2016. URL: https://comprop.oii.ox.ac.uk/research/public-scholarship/resource-for-understanding-political-bots/ [Last accessed: 2019.07.29]

[37] P. N. Howard, S, Woolley and R. Calo, "Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration," *Journal of Information Technology & Politics*, *15*(2), 81-93, 2018. URL: https://www-tandfonline-

com.proxy.lib.sfu.ca/doi/full/10.1080/19331681.2018.1448735 [Last accessed: 2019.07.18]

[38] G. Resnick, "How Pro-Trump Twitter Bots Spread Fake News," *The Daily Beast*, July 2017. URL: https://www.thedailybeast.com/how-pro-trump-twitter-bots-spread-fake-news [Last accessed: 2019.07.29]

[39] G. Krieg, "It's official: Clinton swamps Trump in popular vote," *CNN Politics Data*, December 2016, URL: https://www.cnn.com/2016/12/21/politics/donald-trump-hillary-clinton-popular-vote-final-count/index.html [Last accessed: 2019.07.29]

[40] L. Stark, "Alorithmic psychometrics and the scalable subject," *Social Studies of Science*, 48(2), pp. 204-231, 2018 URL: https://journals-sagepub-com.proxy.lib.sfu.ca/doi/pdf/10.1177/0306312718772094 [Last accessed: 2019.08.03]

[41] F. Morstatter, L. Wu, T. H. Nazer, K. N. Carley and H. Liu, "A new approach to bot detection: Striking the balance between precision and recall," *IEEE/ACM Conference on Advances in Social Networks Analysis and Mining*, pp. 553-540, August 2016. URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/7752287 [Last accessed: 2019.08.03]

[42] E. Shearer and K. E. Matsa, "News Use Across Social Media Platforms 2018: Most Americans continue to get news on social media, even though many have concerns about its accuracy," Pew Research Center, 2018. URL: www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ [Last accessed: 2019.07.30]

[43] J. Gottfried and E. Shearer, "News Use Across Social Media Platforms 2016," Pew Research Center, URL: https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/ [Last accessed: 2019.07.30]

[44] N. Kshetri and J. Voas, "The Economics of 'Fake News,'" *IEEE Computer Society*, pp. 8-12, (2017). URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/stamp/stamp.jsp?tp=&arnumber=8123490&tag=1 [Last accessed: 2019.07.15]

[45] S. Wineburg, S. McGrew, J. Breakstone and T. Ortega, "Evaluating Information: The Cornerstone of Civic Online Reasoning," Stanford Digital Repository, 2016. URL: https://stacks.stanford.edu/file/druid:fv751yt5934/SHEG%20Evaluating%20Information%20Online.pdf [Last accessed: 2019.07.15]

[46] European Commission, *A Europe that protects: EU reports on progress in fighting disinformation ahead of European Council*, June 2019. URL: https://ec.europa.eu/commission/commissioners/2014-2019/ansip/announcements/europe-protects-eu-reports-progress-fighting-disinformation-ahead-european-council_en [Last accessed: 2019.06.15]

[47] A. Al-Rawi and Y. Jiwani, "Trolls Stoke Fear: Russian disruption a concern in Fall vote," *Vancouver Sun*, p. G2, August 2016.

[48] C. Falk, "Detecting Twitter Trolls Using Natural Language Processing Techniques Trained on Message Bodies," July 2018. URL: http://www.infinite-machines.com/detecting-twitter-trolls.pdf [Last accessed: [2019.07.15]

[49] I. Smoleňová, "The pro-Russian disinformation campaign in the Czech Republic and Slovakia," *Prague: Prague Security Studies Institute*, 2015. URL: http://www.pssi.cz/download/docs/253_is-pro-russian-campaign.pdf [Last accessed: 2019.07.21]

[50] I. Khaldarova and M. Pantti, "Fake news: The narrative battle over the Ukrainian conflict, "*Journalism Practice*, 10(7), pp. 891-901, 2016. URL: https://www-tandfonline-com.proxy.lib.sfu.ca/doi/full/10.1080/17512786.2016.1163237 [Last accessed: 2019.07.21]

[51] U. A. Mejias and N. E. Vokuev, "Disinformation and the media: the case of Russia and Ukraine. Media," *Culture & Society*, 39(7), pp. 1027-1042, 2017. URL: https://journals-sagepub-com.proxy.lib.sfu.ca/doi/full/10.1177/0163443716686672 [Last accessed: 2019.07.21]

[52] S. Bradshaw and P. N. Howard, "The Global Disinformation Order: 2019 Global Inventory of Organized Social Media Manipulation." URL: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf [Last Accessed: 2019.11.16]

[53] D. Alba and A. Satariano, "At Least 70 Countries Have Had Disinformation Campaigns, Study Finds," *New York Times*, September 2019. URL: https://www.nytimes.com/2019/09/26/technology/government-disinformation-cyber-troops.html [Last Accessed: 2019.11.16]

[54] D. L. Linvill and P. L. Warren, "Troll factories: The Internet Research Agency and state-sponsored agenda-building," Resource Centre on Media, 2018. URL: https://www.google.com/search?q=Troll+factories%3A+The+Internet+Research+Agency+and+state-sponsored+agenda-building&oq=Troll+factories%3A+The+Internet+Research+Agency+and+state-sponsored+agenda-building&aqs=chrome..69i57j69i60l3.354j0j7&sourceid=chrome&ie=UTF-8 [Last accessed: 2019.07.21]

[55] H. F. Yu, C. H. Ho, Y. C. Juan and C. J. Lin, "LibShortText: A Library for Short-text Classification and Analysis", Department of Computer Science, National Taiwan University, 2013. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/libshorttext.pdf [Last accessed: 2019.08.4]

[56] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis,J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker,V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, "TensorFlow: A system for large-scale machine learning," *12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265-283, November 2016. URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf [Last accessed: 2019.08.4]

[57] G. R. S. Weir, "The posit text profiling toolset," *12th Conference of Pan-Pacific Association of Applied Linguisitics*, pp. 106-109, 2007. URL: https://www.researchgate.net/publication/228740404_The_Posit_Text_Profiling_Toolset [Last accessed: 2019.08.4]

[58] G. R. S. Weir, "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 [Last accessed: 2019.08.4]

[59] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner and Y. Choi, "Defending Against Neural Fake News," *arXiv preprint arXiv:1905.12616*, 2019. URL: https://arxiv.org/abs/1905.12616 [Last Accessed: 2019.11.17].

[60] Y. Gorodnichenko, T. Pham and O. Talavera, *Social media, sentiment and public opinions: Evidence from# Brexit and# USElection*, No. w24631, National Bureau of Economic Research, 2018. URL: https://www.nber.org/papers/w24631.pdf [Last Accessed: 2019.11.19].

[61] J. Mei and R. Frank, "Sentiment crawling: Extremist content collection through a sentiment analysis guided webcrawler," 2015 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1024-1027, August 2015. URL: esearchgate.net/publication/301444687_Sentiment_Crawling_Extremist_Content_Collection_through_a_Sentiment_Analysis_Guided_Web-Crawler [Last accessed: 2019.08.4]

[62] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell and G. Davies, "Surfacing collaborated networks in dark web to find illicit and criminal content," *2016 IEEE Confernce on Intelligence and Security Informatics (ISI)*, pp. 109-114, September 2016. URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/7745452 [Last accessed: 2019.08.4]

[63] T. C. Kietzmann, P. McClure and N. Kriegeskorte, "Deep neural networks in computational neuroscience," *bioRxiv*, pp. 133504-133527, 2018. URL: https://www.biorxiv.org/content/10.1101/133504v2 [Last accessed: 2019.08.4]

[64] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the*

American Society for Information Science and Technology, 61(12), 2544–2558, 2010. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.3863&rep=rep1&type=pdf [Last accessed: 2019.08.4]

[65] M. Hall, E. Frank, H. Geoffrey, B. Pfahringer, P. Reutemann and I. Witten, "The Weka data mining software: an update," *SIGKDD Explorations*, vol. 11, pp. 10-18, 2009. URL: https://www.kdd.org/exploration_files/p2VD:\BTSync\Ricsi\Academic Work\2019\20190508 - Venice Cloud Conference11n1.pdf [Last accessed: 2019.08.4]

[66] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf [Last accessed: 2019.08.4]

[67] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, pp. 18-22, 2002. URL: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf [Last accessed: 2019.08.4]

[68] H. F. Yu, C. H. Ho, Y. C. Juan and C. J. Lin, "LibShortText: A Library for Short-text Classification and Analysis", Department of Computer Science, National Taiwan University, 2013. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/libshorttext.pdf [Last accessed: 2019.08.4]

[69] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," *Advances in neural information processing systems*, pp. 2933-2941, 2014. URL: https://papers.nips.cc/paper/5486-identifying-and-attacking-the-saddle-point-problem-in-high-dimensional-non-convex-optimization [Last accessed: 2019.08.4]

[70] D. R. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley and B. W. Suter, "The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function," *IEEE Transactions on Neural Networks*, 1(4), pp. 296-298, 1990. URL: https://ieeexplore-ieee-org.proxy.lib.sfu.ca/document/80266 [Last accessed: 2019.08.6]

[71] W. S. Sarle, "Neural Networks and Statistical Models," *Nineteenth Annual SAS Users Group International Conference*, April 1994. URL: https://people.orie.cornell.edu/davidr/or474/nn_sas.pdf [Last accessed: 2019.08.6]

[72] S. Chadha, "Text Analytics on Russian Troll Tweets-Part 1," n.d., URL: https://www.kaggle.com/chadalee/text-analytics-on-russian-troll-tweets-part-1 [Last accessed: 2019.08.7]

[73] *CNN*, "Big Tech braces for first presidential debates, a target of Russian trolls in 2016," June 2019. URL: https://m.cnn.com/en/article/h_8d2922458b2b4c80698925b2be1ab879 [Last accessed: 2019.08.7]

[74] M. Choi, "Keith Ellison reeling after abuse allegations: The No. 2 at the Democratic National Committee is running behind in his bid for Minnesota attorney general," *Politico*, October 2018. URL: https://www.politico.com/story/2018/10/27/keith-ellison-abuse-allegations-minnesota-ag-2018-943086 [Last accessed: 2019.08.10]

[75] E. Nakashima, "U.S. Cyber Command operation disrupted Internet access of Russian troll factory on day of 2018 midterms," *Washington Post*, February 2019. URL: https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html?noredirect=on [Last accessed: 2019.08.10]

[76] T. Starks, L. Cerulus and M. Scott, "Russia's manipulation of Twitter was far vaster than believed," *Politico*, June 2019. URL: https://www.politico.com/story/2019/06/05/study-russia-cybersecurity-twitter-1353543 [Last accessed: 2019.08.10]

[77] M. Lander, "Trump Says He Discussed the 'Russian Hoax' in a Phone Call With Putin," *The New York Times*, May 2019. URL: https://www.nytimes.com/2019/05/03/us/politics/trump-putin-phone-call.html

[78] *Baltimore Sun*, "Even if Trump is right about collusion, Russia story is big (not fake) news," October 2017. URL: https://www.baltimoresun.com/opinion/editorial/bs-ed-1101-trump-russia-20171031-story.html [Last accessed: 2019.08.10]

[79] R. Dicker, "Secret Service Agent Says She Wouldn't Take A Bullet For Trump: Agency says it is taking quick action, " *Huffington Post*, January 2017. URL: https://www.huffingtonpost.ca/entry/secret-service-agent-says-she-wouldnt-take-a-bullet-for-trump_n_58887d4be4b0441a8f71e671 [Last accessed: 2019.08.10]

[80] M. Y.H. Lee, "'Rapists?' Criminals? Checking Trump's facts," *The Philadelphia Inquirer*, July 2015. URL: https://www.inquirer.com/philly/news/politics/20150709__Rapists__Criminals__Checking_Trump_s_facts.html [Last accessed: 2019.08.11]

[81] J. Hing, "This Is the Beginning of Donald Trump's Muslim Ban: Friday's executive order extended to seven countries—but that list could grow," *The Nation*, January 2017. URL: https://www.thenation.com/article/this-is-the-beginning-of-donald-trumps-muslim-ban/ [Last accessed: 2019.08.11]

[82] L. Gambino, "Trump pans immigration proposal as bringing people from 'shithole countries'," *The Guardian*, January 2018. URL: https://www.theguardian.com/us-news/2018/jan/11/trump-pans-immigration-proposal-as-bringing-people-from-shithole-countries [Last accessed: 2019.08.11]

[83] M. Bouazizi and T. Ohtsuki, "Multi-Class Sentiment Analysis on Twitter: Classification Performance and Challenges," *Big Data and Mining Analytics*, vol. 2, pp. 181-194, 2019. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8681053 [Last Accessed: 2019.11.24]

[84] C. H. Tsai, C. Y. Lin, and C. J. Lin, "Incremental and decremental training for linear classification" *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 343-352, 2014. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/ws/inc-dec.pdf [Last Accessed: 2019.11.24]

[85] H. Ahmed, I. Traore and S. Saad, "Detecting opinion spams and fake news using text classification," *Journal of Security and Privacy*, vol. 1, pp. 1-15. URL: https://www.uvic.ca/engineering/ece/isot/assets/docs/SPY_Detecting%20opinion%20spams%20and%20fake%20news%20using%20text%20classification.pdf [Last Accessed: 2019.11.24]

[86] F. K. A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber and M. Farah, "FA-KES: A Fake News Dataset around the Syrian War," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 573-582, 2019. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/3254/3122 [Last Accessed: 2019.11.24]