Fighting Disinformation Warfare with Artificial Intelligence

Identifying and Combatting Disinformation Attacks in 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
School of Criminology
Simon Fraser University
Burnaby, Canada
Email: rfrank@sfu.ca

***Abstract*—Following well-documented Russian interference in the 2016 U.S. Presidential election and in the Brexit referendum in the U.K., law enforcement, intelligence agencies and social network providers worldwide have expressed growing interest in identifying and interdicting disinformation warfare. This paper reports on a research project 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). The research project involves the development of a method for identifying hostile disinformation activities in the Cloud. Employing the ICCRC's Dark Crawler, Strathclyde's Posit Toolkit, and TensorFlow, we collected and analyzed nearly three million social media posts, examining "fake news" by Russia's Internet Research Agency, and comparing them to "real news" posts, in order to develop an automated means of classification. We were able to classify the posts as "real news" or "fake news" with an accuracy of 90.12% and 89.5%, using Posit and TensorFlow respectively.**

Keywords-Cybersecurity; Cloud-based social media platforms; disinformation; machine learning.

I. INTRODUCTION

Amongst the key challenges currently facing law enforcement agencies, intelligence agencies, cybersecurity personnel and business owners-operators around the world are how to monitor and efficiently respond to dynamic and emerging cybersecurity threats, with increasing attention being paid to hostile disinformation activities in Cloud-based social media platforms. To illustrate, on November 27, 2018, a senior executive from Facebook was grilled by a Parliamentary Committee in the U.K. regarding the (witting or unwitting) involvement of Facebook in the Russian hostile influence campaign during the run-up to the 2016 U.S. Presidential election. According to the CIA, the FBI, and the NSA, various other social media, including Twitter and Instagram, have also been implicated as (possibly unaware) participants in the hosting and dissemination of these disinformation attacks [1].

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 method for identifying hostile disinformation activities in the Cloud. The knowledge generated by this research will establish the foundation for more advanced work, eventually culminating in automatic tools which can rapidly and accurately pinpoint disinformation attacks in their very early stages.

II. RESEARCH CONTEXT

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 [2], [3], [4]. This has been accomplished in the past by applying a combination of technologies, including the Dark Crawler, SentiStrength, and Posit. For the present study, we have employed the Dark Crawler, Posit, and TensorFlow. Additional information on these research tools is provided below. From this background, we have a methodology that is applicable to the analysis and classification of data from Cloud-based social media platforms.

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 [5]. 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 automated textual analysis. Content retrieved by our 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 used this procedure to collect over 100 million forum posts from across a vast number of hacking and extremist forums for later analysis.

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, or 27 features in all [6]. This generates a detailed frequency analysis of the syntax, including multi-word units and associated part-of-speech components.

TensorFlow, originally developed by the Google Brain Team, is a machine learning system that deploys deep neural networks [7]. This is a machine learning technique inspired by real 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 [8].

III. METHODOLOGY

The research process commenced with an analysis of textual content from existing databases that had already assembled extensive materials from previously identified Russian disinformation attacks launched through social media platforms, including Twitter and Facebook. This paper reports on the analysis of textual content in Twitter, using Post and TensorFlow.

A. *Research Sample*

The research team downloaded a data set of 2,946,219 Twitter messages (tweets) from Github, which had been posted online by fivethirtyeight.com. This data set of tweets was collected and assembled by two professors at Clemson University, Darren Linvill and Patrick Warren [9]. These tweets were described as originating from the Internet Research Agency (IRA), also referred to in common parlance as the Russian troll factory, which was believed to have intentionally interfered in the 2016 U.S. Presidential election and the 2016 U.K. Brexit referendum.

A decision was made to extract only those entries that were labeled as being “English,” thereby excluding languages such as Albanian, Bulgarian, Catalan, Croatian, Dutch, Estonian, French, German, Italian, Russian, Ukrainian, Uzbek, Vietnamese. Thus, 13 new Excel spreadsheets were created, with 2,116,904 English-speaking tweets remaining in the data set following the removal of all non-English cases.

Having acquired the Russian IRA Twitter data, we sought a second Twitter data set that would allow us to develop a classification model based upon comparison between “real news” and what has frequently been referred to as “fake news” [10], [11]. 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, specifically nouns. The resultant “keyword” list was used with the International CyberCrime Research Centre’s Dark Crawler to retrieve a set of matching “real news” Twitter posts from legitimate news sites. The 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 (approximately the same time frame as the IRA tweets). Tweets from the “real news” data set 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, and the removal of 10,602 tweets that had been posted in late 2018-early 2019 left us with 80,003 individual cases or tweets that exemplified “real” or “legitimate” news sources. A research decision was made to random sample both data sets, creating two data sets of equal size, each consisting of 2,500 tweets, or roughly .001% of the larger “fake news” data set, and 3% of the “real news” data set. Unique identifiers were assigned to each of the data items, to ensure a means of fixed reference, and to permit future analysis of the data in NVivo and SentiStrength [12].

A somewhat different sample was assembled for the TensorFlow analysis. For TensorFlow to operate effectively, a larger data set 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 data set was supplemented with 3,000 Facebook messages posted by the IRA, plus an additional “real news” set of Twitter items. Thus, a large data set of 2,709,204 million tweets was analyzed in TensorFlow after the merging of these multiple data sets.

B. *Data Analysis*

1) *Posit*

Following the creation and cleansing of the data sets, 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 prove effective as a feature set for use in such text classification tasks [2], [4]. In the present study, the target textual data was made up of tweets. These have a limited maximum length of 280 characters, so they are inherently short and contain relatively few words. This was potentially an obstacle to Posit use.

Since Posit creates data on the basis of word-level information, the limited content of tweets means that many of the original features may have zero values. With this in mind, for the analysis of short texts, Posit has been extended to include analysis of character-level content. To this end, the system supplements the standard word-level statistics and generates an additional 44 character features for each instance of text data. These features include quantitative information on individual alphanumeric characters, and a subset of special characters, specifically, questions marks, exclamation marks, asterisks, periods and dollar signs. The extension of Posit to embrace character-level as well as word-level data maintains the domain-neutral nature of Posit analysis. As a result of this extended Posit analysis, each data item (tweet) is represented by a set of 72 features.

Thereafter, this list of tweet features was formulated as an arff file format, suitable for direct input to the Waikato Environment for Knowledge Analysis (WEKA) data analysis application [13]. In WEKA, we applied the standard J48 tree classification method and the Random Forest classification method [14], both with ten-fold validation. WEKA produced a measure of how many of the tweets were correctly classified.

2) *TensorFlow*

In this project, TensorFlow was adopted for processing the data with a Deep Neural Network (DNN). A large data set 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 data set of 2,709,204 million tweets was created by merging multiple data sets. 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 data set, we opted for Microsoft Access, which allowed for a large, unified database table. All of the data sets were merged into the 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 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 accuracy of results could be compared effectively.

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 takes longer for the model to converge at the local minima [15]. With a larger learning rate, the model gets closer to this convergence point more quickly. The values for these parameters— number of partitions, epochs, layers, learning rate, and regularization (L1 & L2)—were then tested to identify an optimal set of parameter values.

IV. RESEARCH RESULTS

A. *Posit*

The Posit analysis produced a feature set with corresponding values for each of the 5,000 tweets (2,500 "fake news” tweets and 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 Figure 1, below.

```
   a    b   ← classified as
2190  310 |   a = negative
 340 2160 |   b = positive
```

Figure 1. Confusion matrix for Posit: 27 features (Random Forest: default WEKA settings)

As indicated earlier, Posit was enhanced with an additional 44 character-based features. Using this extended feature set on the 5,000 tweets—and the default WEKA settings with 10-fold cross validation—the J48 and Random settings Forest classifiers gave 81.52% and 89.8% correctly classified instances, respectively. The confusion matrix for the latter performance is shown in Figure 2, below.

```
   a    b  ← classified as
2266  234 |   a = negative
 276 2224 |   b = positive
```

Figure 2. Confusion matrix for Posit: 71 features (Random Forest: default WEKA settings)

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 Figure 3, below.

```
   a    b  ← classified as
2269  231 |   a = negative
 263 2237 |   b = positive
```

Figure 3. Confusion matrix for Posit: 71 features (Random Forest: instances at 211 in WEKA settings)

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 Figure 4.

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

Figure 4. Detailed Accuracy By Class for best Posit result

B. *TensorFlow*

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 of to 89.5%, a substantial improvement from the earlier results. These parameters are described below, with the post-training optimal values shown below in Table I.

To be able to run large numbers of experiments, we wrapped all code into a standalone function, so 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 and testing on the last 5.

V. DISCUSSION

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 at 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 may be surprising, given that alphanumeric details seem far removed from tweet content-level.

The natural presumption may be that establishing objective truth is the primary goal of such research. This could be an inaccurate assumption. Since the principal basis for any automated judgment will be secondary sources, objective truth is not always readily discernible. Facts as they pertain to the real world are present in first-order reports, but when confronted solely with such reports, we can only resort to authority, provenance and inherent credibility as bases for judgement.

Despite working solely from available sources, we have aimed to discriminate between several significant classes of report: 1) plausible and probably correct reflections of the facts; 2) likely, based upon the facts with evident 'observer' influence (such as colour or bias or prejudice); 3) largely 'interpreted' with some factual basis; 4) almost entirely devoid of factual content.

TABLE I. 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

VI. CONCLUSION

Through the research process outlined above, we are: 1) developing typologies of past and present hostile activities in

Cloud-based social media platforms; 2) identifying indicators of change in public opinion (as they relate to hostile disinformation activities); 3) identifying the social media techniques of hostile actors (and how best to respond to them); and 4) undertaking 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 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.

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 analysis.) All 2,500 of these messages have been subjected to a preliminary review in the qualitative research tool, NVivo. Early insights revealed that many of the allegedly “fake news” items were founded to one degree or another in contemporaneous “real news” events. We are presently devising a process for capturing “real news” stories that align as closely as possible with the “fake news,” to better address the spectrum between “real” and “fake” news, and the nexus between them. Apart from informing ongoing NVivo analysis, we anticipate that this spectrum of “real” and “fake” news stories will serve as a basis for further discrimination in Posit, with the likely addition of sentiment analysis [12].

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 and Karmvir Padda (NVivo).

REFERENCES

[1] Background to “Assessing Russian Activities and Intentions in Recent US Elections”: The Analytic Process and Cyber Incident Attribution, ICA 2017-01D, January 2017. URL: https://www.dni.gov/files/documents/ICA_2017_01.pdf [accessed: 2019.04.05]

[2] 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.

[3] 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.

[4] 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.

[5] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell and G. Davies, “Surfacing collaborated networks in dark web to find illicit and criminal content,” *Intelligence and Security Informatics (ISI)*, pp. 109-114, September 2016.

[6] G. R. S. Weir, “Corpus profiling with the Posit tools,” *Proceedings of the 5th Corpus Linguistics Conference*, July 2009.

[7] 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.

[8] T. C. Kietzmann, P. McClure and N. Kriegeskorte, “Deep neural networks in computational neuroscience,” *bioRxiv*, pp. 133504-133527, 2018.

[9] D. L. Linvill and P. L. Warren, “Troll factories: The Internet Research Agency and state-sponsored agenda-building,” 2018. URL: http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf [accessed: 2019.04.05]

[10] L. Reston, “How Russia Weaponizes Fake News,” *New Republic*, pp. 6-8, 2017.

[11] N. W. Jankowski, "Researching fake news: A selective examination of empirical studies," *Javnost-The Public* 25.1-2, pp. 248-255, 2018.

[12] 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.

[13] 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.

[14] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[15] 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.