Detecting Twitter Trolls Using Natural Language Processing Techniques Trained on Message Bodies

Courtney Falk
Ontological Semantic Technology Laboratory
Texas A&M University - Commerce
Commerce, Texas, USA
ontology@tamuc.edu

***Abstract*—Internet trolls based in Russia effectively shaped social media opinions during the 2016 U.S. presidential election. This paper is an empirical study of natural language processing techniques for identifying similar online trolls. The machine learning approaches used in this paper focus only on the text contained in the message itself and not with other meta-data associated with the posting accounts.**

Actual tweets attributed to the trolls were evaluated using 27 different machine learning models in a binary classification task. The results show that the support vector machine model, trained on a unigram language model, is the most effective choice.

We conclude that while natural language processing techniques are capable of successfully identifying online trolls, they alone are not a sufficient filtering mechanism to solve the problem. Rather, natural language processing can be a useful tool in a larger investigative effort.

***Index Terms*—natural language processing, machine learning, social media**

I. Introduction

The 2016 United States presidential election saw a sustained, concerted effort to shape social media opinions [1], [2]. This effort was orchestrated by a foreign organization based in Russia. Fake personas gathered real audiences to read their politically incendiary blog posts. But most of the activity was taking place on social media web sites such as Twitter and others.

This paper is an empirical study of using machine learning and natural language processing techniques to identify online trolls. 27 different combinations of settings were evaluated and compared to find the best possible combination.

Understanding these disinformation campaigns is important for a healthy democracy. Participating members of a civil society must practice good information hygiene, and being aware of deliberate manipulation of the information stream is one key component to that.

A. Background

The social media campaign to influence American political opinions was organized, sustained, and wide reaching. The trolls were timely and opportunistic in their posting of incendiary, politically-charged material. When Philando Castille was killed by police in Minneapolis, trolls stoked racial tensions and anti-police sentiment [3]. When there was a school shooting in Parkland, Florida, trolls invoked rhetoric from the gun control debate [4]. No issue was too sensitive or taboo to avoid manipulation. A sample of the affected social networks are listed below.

1) Twitter [5]
2) Facebook [6]
3) Reddit [7]
4) Tumblr [8]
5) Pinterest [9]

What was happening was an influence operation [10], [11]. Timothy Thomas of the Foreign Military Studies Office produced a literature review of Russian military thinking [12]. While the Russian concepts of information/psychological operations do not exactly map to those in Western military thinking, influence operations are understood to target civil society. The employment of trolls in addition to distributed denial-of-service (DDoS) attacks was a strategy observed in conjunction with events in Estonia and Georgia [13].

The trolls worked for a company, the Internet Research Agency, based in St. Petersburg, Russia. Russian Internet trolls were already a well-known phenomenon by 2016. Yevgeny Prigozhin is the owner of the Internet Research Agency. Prigozhin's nickname is "Putin's Chef," a nod to one of his other lines of business [14]. Concord Catering is another Prigozhin business with a lucrative contract with the Kremlin. Prigozhin, Concord, and the Internet Research Agency are all named in a February 2018 indictment submitted by Robert Mueller of the United States Department of Justice [15],

There were problems with operations security at the Internet Research Agency. Despite carrying out a task with possible international sensitivities, the Agency was infiltrated by investigative journalists [16]. The picture that is painted is of an environment similar to any other manpower-intensive technology operation like a help desk. The possible exception being an elevated level of physical security for the building itself.

Activist Russian trolls are a phenomenon that precedes the Internet Research Agency. In 2009, the Kremlin founded its so-called "school for bloggers." [17] On one hand, this could be viewed as Russia making a concerted effort to build their soft power. On the other hand, it could be viewed as an attempt to legitimize the "patriotic hackers" that Vladimir Putin blames for antagonizing Western nations [18]

Just as the trolls were hard at work before the US presidential election, so too do they continue to work to this day.

Activity traced to the trolls has touched several of the major, Western elections:

1) Brexit [19], [20]
2) Catalonia Independence Referendum [21]
3) French Presidential Election [22]
4) German Federal Election [23]

II. Prior Work

In testimony before the Senate Judiciary Committee [24], Twitter's Sean Edgett provided the group with a list of Twitter user account handles that belonged to what they assessed to be Russian Trolls. Democratic members of the House Permanent Select Committee on Intelligence then publicly released the list of those account handles [25].

NBC released their dataset in February of 2018 [5] with the stated goal of enabling the type of research presented in this paper. In the months since the dataset publication, multiple research projects with differing goals began. William Lyon of the graph database company, Neo4j, took the same dataset and used their database software to analyze the links between Twitter accounts [26]. This social network analysis approach is complementary to the natural language processing experiments found in this paper.

Badawy, Ferrara, and Lerman are political scientists [27]. Their research into the NBC dataset aimed to classify Twitter users based on their ideological bent.

Duh, Rupnik, and Korošak performed a temporal analysis of tweets during the Brexit referendum in the United Kingdom [20]. Their approach applies spin-glass models - an idea taken from the physics of electro-magnetism - and applies it to patterns of user activity.

Finnish researchers at the University of Turku conducted their own work in applying NLP to detecting Russian trolls [28], [29]. Finnish - a member of the Finno-Ugric language family - has a more complex morphology than English. Nouns in Finnish may be declined in one of fifteen grammatical cases. This means that a Finnish corpus may have a much higher number of unique strings in it as compared to an English corpus of the same size.

There are other problems in social media that might benefit from the solutions being sought for trolls. One area in which social media provides a forum for opinions is in online product reviews. The pseudo-anonymity of the Internet allows for unvalidated posting. Fake reviews are now a real business with its own Internet term to go with it, "astroturfing." The actual AstroTurf is a synthetic surface often used in athletic fields that imitates real grass, much the way that fake reviews imitate and blend in with real reviews.

Detecting astroturfed reviews using natural language processing techniques aligns closely with the goals of this paper. Ahmed, Traore, and Saad use n-gram features for language modeling to detect fake reviews [30]. They also cite the 2016 US election and fake news as other applicable domains.

III. Experimental Setup

The data used in this study is a collection of real world tweets attributed to accounts belonging to employees of the Internet Research Agency. Twitter removed these accounts and deleted the associated tweets. NBC collected the deleted tweets and freely distributed them as a database [5].

Plotting the frequency of tweets from the troll accounts over time, as seen in Figure 1, shows some distinct trends. Troll activity begins in late 2014 and continues at a relatively low level through mid-2016. This constant, low-level activity is the creation and backstopping of personas, the fake identities the trolls will eventually use to spread their message. Posting activity takes a large uptick in the last half of 2016. This is after both parties have nominated their candidates for president, simplifying the trolls' job of developing targeted messages. Peak activity coincides with the election itself then tapering off slightly, and making a big drop with the inauguration of Donald Trump as president.

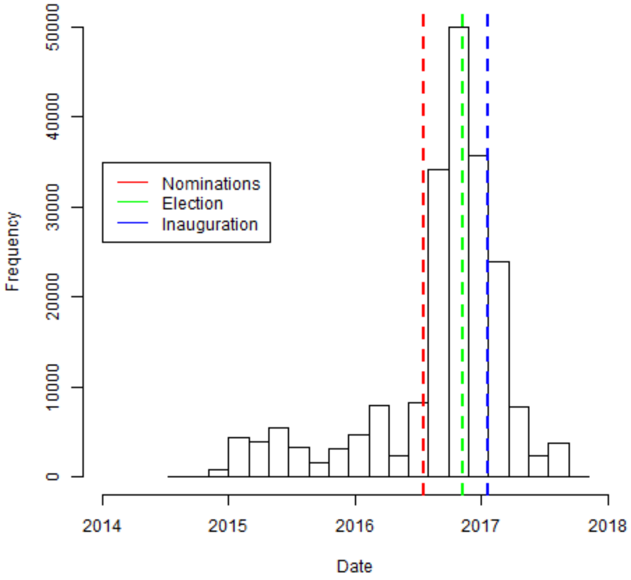

Fig. 1. Histogram of the tweets from the deleted accounts.

Posts attributed to trolls were largely retweets, comprising almost three quarters of all troll tweets. This makes sense for users with limited English language skills. It is simpler to read something and the retweet it is a kind of "me too" endorsement than crafting an original post. Because this research is solely focused on identifying trolls based on the language that they use, looking at retweets is not productive because this is language that *someone else* used. To that end, all retweets were excluded from the experimental corpus, leaving 56K tweets out of the previous 203K.

To build a binary classifier using supervised machine learning there must be two different classes of data. The NBC corpus provided the "Troll" class, but there must also be a "Control" class, which are tweets not belonging to troll sources. The researchers registered their own app with Twitter in order to obtain the necessary API key to download posts from the Twitter stream. Using this approach, 100K tweets were sampled. This produced a ratio of roughly 2:1 in terms of Control to Troll tweets in the corpus.

All subsequent processing of the data was done within the Weka machine learning toolkit [31]. However, the resulting combined corpus was too large to fit in the heap space of the available hardware. The decision was made to produce a random subsample of the corpus that would be representative of the whole and also able to be processed. Weka includes the Resample module to perform just such a task. The final, subsampled corpus is 10% of the size of the original corpus.

Nine different data sets for the language models were constructed by varying the length of the n-grams and the type of statistic they measured. The list below specifies the two parameters and their values. The motivation for examining nine different combinations was to see if the length of n-grams, or the type of data captured in the statistic, affected the performance of the model.

- N-grams: Unigrams, bigrams, trigrams
- Statistic: Boolean presence, word count, and term frequency-inverse document frequency

Reducing the number of unique strings was important to finding useful features. The strings use for n-grams were also forced into the lower case, and stemmed, in order to reduce the overall number of unique strings. English stop words were filtered out due to their high frequency and low information content [32]. Finally, certain other kinds of tokens received special treatment in order to reduce the string count. Twitter-specific strings were cast into special tags:

- Twitter handles became ⟨HL⟩
- URLs became ⟨URL⟩
- Retweet preambles became ⟨RT⟩
- Emoji became ⟨EM⟩

Each of the nine data combinations were used to train and evaluate models using three different algorithms. These three algorithms could be considered “shallow” machine learning because they differ from the multi-layered neural network architectures seen in deep learning approaches. All three of these algorithms, as shown in Table I, feature different core attributes [33].

TABLE I
SELECTED ALGORITHMS

Algo.	Gen./Disc.	Type
Naïve Bayes	Generative	Probabilistic
SVM	Discriminative	Linear function
C4.5	Discriminative	Decision tree

Another benefit of choosing these three algorithms is that they are commonly available across a variety of machine learning toolkits. Table II provides a brief list of non-deep learning machine learning toolkits that support some or all of the algorithms listed in Table I.

All models were trained using K-fold cross-validation in order to avoid over-fitting. For these experiments the number of folds was 10.

TABLE II
TOOLKITS WITH SUPPORT FOR SHALLOW MACHINE LEARNING ALGORITHMS.

Toolkit	Language	Naïve Bayes	SVM	C4.5
Weka	Java	Y	Y	Y
Scikit-Learn	Python	Y	Y	Y
R	R	Y	Y	
ML.NET	C#	Y	Y	

IV. RESULTS

Twenty-seven different combinations of models and parameters were tested. Figure 2 shows the F_1 scores for all of these tests. F_1 is the harmonic mean between the precision and the recall of a model, making it a good statistic for giving a concise summarization of model performance. In this application, the precision is the ratio of correctly identified troll tweets to the number of all tweets identified as belonging to trolls. Similarly, the recall is the ratio of correctly identified troll tweets to the number of all actual troll tweets.

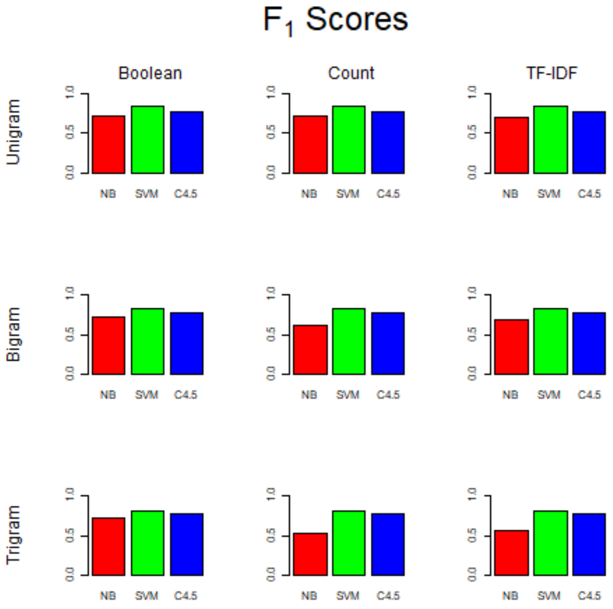

Fig. 2. F_1 scores for 27 different binary classifier models.

What is immediately clear from Figure 2 is that the SVM models dominate. In every parameter configuration, SVM performs the best. But the confusion matrix of this model, as seen in Table III, gives more insight into the model's performance.

The upper-left corner of the matrix are the true positives, actual troll messages that were correctly classified as troll. The lower-right corner contains the true negatives, control messages correctly classified as control. In a perfectly performing system, these two cells would account for the totality of the corpus. But the matrix shows an imperfect system. The lower-left corner contains the false positives, control messages that

TABLE III
CONFUSION MATRIX FOR THE UNIGRAM SVM MODEL WITH BOOLEAN STATISTIC.

		Classified As	
		Troll	**Control**
Actual	**Troll**	3699	1937
	Control	482	9347

the model classified incorrectly as troll messages. And the upper-right corner contains the false negatives, troll messages that were passed over in the thinking that they were benign. While there are relatively few false positives, the confusion matrix shows that the model lets through about 34% of all troll posts.

An interesting finding is that Naïve Bayes performs best with the simplest, unigram language model, and performs worse with bigrams and worse yet with trigrams. This result seems counterintuitive to the idea that having more complex features should enable improved performance. Table IV below gives the true positive rate, false positive rate, precision, recall, and F_1 score for all 27 models. This table shows how performance decreased with the more sophisticated models.

TABLE IV
RESULTS OF ALL THE MACHINE LEARNING MODELS.

Algo.	**N**	**Freq.**	**TP**	**FP**	**Prec.**	**Rec.**	F_1
NB	1	Bool.	0.730	0.355	0.723	0.730	0.722
SVM	1	Bool.	0.844	0.236	0.849	0.844	0.837
C4.5	1	Bool.	0.779	0.307	0.776	0.779	0.771
NB	1	WC	0.719	0.331	0.718	0.719	0.719
SVM	1	WC	0.842	0.238	0.848	0.842	0.836
C4.5	1	WC	0.781	0.303	0.778	0.781	0.773
NB	1	TF-IDF	0.701	0.354	0.699	0.701	0.700
SVM	1	TF-IDF	0.838	0.227	0.839	0.838	0.834
C4.5	1	TF-IDF	0.779	0.306	0.777	0.779	0.772
NB	2	Bool.	0.730	0.338	0.725	0.730	0.726
SVM	2	Bool.	0.829	0.367	0.833	0.829	0.822
C4.5	2	Bool.	0.780	0.302	0.777	0.780	0.773
NB	2	WC	0.618	0.282	0.718	0.618	0.617
SVM	2	WC	0.827	0.257	0.831	0.827	0.820
C4.5	2	WC	0.776	0.304	0.773	0.776	0.769
NB	2	TF-IDF	0.689	0.316	0.706	0.689	0.694
SVM	2	TF-IDF	0.823	0.247	0.822	0.823	0.818
C4.5	2	TF-IDF	0.776	0.303	0.772	0.776	0.769
NB	3	Bool.	0.729	0.340	0.724	0.729	0.725
SVM	3	Bool.	0.826	0.259	0.830	0.826	0.819
C4.5	3	Bool.	0.778	0.301	0.774	0.778	0.771
NB	3	WC	0.552	0.302	0.705	0.552	0.536
SVM	3	WC	0.825	0.260	0.829	0.825	0.818
C4.5	3	WC	0.776	0.302	0.772	0.776	0.770
NB	3	TF-IDF	0.570	0.306	0.695	0.570	0.562
SVM	3	TF-IDF	0.819	0.253	0.819	0.819	0.814
C4.5	3	TF-IDF	0.776	0.302	0.772	0.776	0.769

V. CONCLUSIONS

Twenty-seven different machine learning models were trained and tested on the same data set. The result was that using a unigram language model with a boolean statistic and the SVM algorithm produced the best results. Increasing the length of the n-gram model, or changing the statistic, did not measurably improve results. We assess that the overall performance of the models presented in this paper are insufficient to justify their use in implementing an automated account suspension algorithm. Twitter is already suffering from negative backlash related to its banning of 70 million fake accounts [34]. Customer reactions would be much worse if an automated algorithm began banning legitimate accounts. People are rushing to apply machine learning to applications without a full appreciation of the consequences.

A better application for the machine learning models presented in this paper would be as lead generators for human analysts. The automated lead generators would identify suspicious accounts and queue them up for review. In such a proposed system, it would be the final decision of the human as to whether or not to actually ban the account that was flagged by the algorithm. This human analyst would supplement the evidence provided by the algorithm with other tools that can fill in gaps such as temporal data and social network analysis.

The problem of online trolls continues [35]. In reaction, individuals and organizations are working to inoculate the public against toxic trolling. The non-governmental organization, Alliance for Democracy, launched the Hamilton 68 web site that tracks Russian influence operations in real time [36]. Meanwhile, the United States government issued a new indictment against Russian intelligence officers for hacking that took place during the presidential election [37]. It will take the concerted efforts of public, private, and government sectors in order to keep democracy alive and healthy.

ACKNOWLEDGMENT

Thank you to ValpoHacks and CircleCityCon for providing venues for experimenting with these ideas.

REFERENCES

[1] Office of the Director of National Intelligence, "Assessing Russian activities and intentions in recent US elections," January 2017. [Online]. Available: https://www.dni.gov/files/documents/ICA_2017_01.pdf

[2] Senate Select Committee on Intelligence, "The intelligence community assessment: Assessing Russian activities and intentions in recent U.S. elections," July 2018. [Online]. Available: https://www.burr.senate.gov/imo/media/doc/SSCIICAASSESSMENT_FINALJULY3.pdf

[3] D. O'Sullivan, "Her son was killed then came the Russian trolls," *CNN*, June 2018. [Online]. Available: https://www.cnn.com/2018/06/26/us/russian-trolls-exploit-philando-castiles-death/index.html

[4] B. Popken and J. L. Kent, "Russian trolls flood Twitter after Parkland shooting," *NBC News*, February 2018. [Online]. Available: https://www.nbcnews.com/tech/social-media/russian-trolls-flood-twitter-after-parkland-shooting-n848471

[5] B. Popken, "Twitter deleted 200,000 Russian troll tweets. Read them here." *NBC News*, February 2018. [Online]. Available: https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

[6] A. Stamos, "Authenticity matters: The ira has no place on facebook," Online, April 2018. [Online]. Available: https://newsroom.fb.com/news/2018/04/authenticity-matters/

[7] S. Huffman, "Reddits 2017 transparency report and suspect account findings," Online, April 2018. [Online]. Available: https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/

[8] I. Lapowsky, "Tumblr finally breaks its silence on russian propaganda," Online, March 2018. [Online]. Available: https://www.wired.com/story/tumblr-russia-trolls-propaganda/

[9] E. Dwoskin, "How Russian content ended up on Pinterest," *The Washington Post*, October 2017. [Online]. Available: https://www.washingtonpost.com/news/the-switch/wp/2017/10/11/how-russian-content-ended-up-on-pinterest/

[10] E. V. Larson, R. E. Darilek, D. Gibran, B. Nichiporuk, A. Richardson, L. H. Schwartz, and C. Q. Thurston, "Foundations of effective influence operations: A framework for enhancing army capabilities," RAND Corporation, Tech. Rep., 2009.

[11] United States Air Force, "Air force doctrine document 2-5, information operations," 2005. [Online]. Available: http://www.dtic.mil/dtic/tr/fulltext/u2/b311353.pdf

[12] T. L. Thomas, *Recasting the Red Star: Russia Forges Tradition and Technology through Toughness*. Foreign Military Studies Office, 2011.

[13] A. Soldatov and I. Borogan, *The Red Web: The Struggle between Russia's Digital Dictators and the New Online Revolutionaries*. PublicAffairs, 2015.

[14] J. S. Tim Lister and M. Ilyushina, "Exclusive: Putin's 'chef,' the man behind the troll factory," *CNN*, October 2017. [Online]. Available: https://www.cnn.com/2017/10/17/politics/russian-oligarch-putin-chef-troll-factory/index.html

[15] "United States v. Internet Research Agency et.al." *United States District Court for the District of Columbia*, February 2018. [Online]. Available: https://www.justice.gov/file/1035477/download

[16] B. Popken and K. Cobiella, "Russian troll describes work in the infamous misinformation factory," *NBC News*, 2017. [Online]. Available: https://www.nbcnews.com/news/all/russian-troll-describes-work-infamous-misinformation-factory-n821486

[17] N. Hodge, "Kremlin launches 'school of bloggers'," *Wired*, May 2009. [Online]. Available: https://www.wired.com/2009/05/kremlin-launches-school-of-bloggers/

[18] K. Calamur, "Putin says 'patriotic hackers' may have targeted u.s. election," *The Atlantic*, June 2017. [Online]. Available: https://www.theatlantic.com/news/archive/2017/06/putin-russia-us-election/528825/

[19] P. Wintour, "Russian bid to influence Brexit vote detailed in new US Senate report," *The Guardian*, January 2018. [Online]. Available: https://www.theguardian.com/world/2018/jan/10/russian-influence-brexit-vote-detailed-us-senate-report

[20] A. Duh, M. Slak Rupnik, and D. Korošak, "Collective behavior of social bots is encoded in their temporal Twitter activity," *Big Data*, vol. 6, no. 2, pp. 113–123, 2018.

[21] T. White, "Russian hackers fueled Catalan separatism, Madrid institute says," *Bloomberg*, November 2017. [Online]. Available: https://www.bloomberg.com/news/articles/2017-11-08/russian-hackers-fueled-catalan-separatism-madrid-institute-says

[22] L. Daniels, "How russia hacked the French election," *Politico*, April 2017. [Online]. Available: https://www.politico.eu/article/france-election-2017-russia-hacked-cyberattacks/

[23] S. Shuster, "How Russian voters fueled the rise of Germany's far-right," *Time*, September 2017. [Online]. Available: http://time.com/4955503/germany-elections-2017-far-right-russia-angela-merkel/

[24] United States Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism, "Testimony of Sean J. Edgett," October 2017. [Online]. Available: https://www.judiciary.senate.gov/imo/media/doc/10-31-17EdgettTestimony.pdf

[25] Democrats of the U.S. House of Representatives Permanent Select Committee on Intelligence, "Twitter accounts," Online, 2017. [Online]. Available: https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf

[26] W. Lyon, "The story behind russian twitter trolls: How they got away with looking human and how to catch them in the future," *Neo4j*, March 2017. [Online]. Available: https://neo4j.com/blog/story-behind-russian-twitter-trolls/

[27] E. F. Adam Badawy and K. Lerman, "Analyzing the digital traces of political manipulation: the 2016 Russian interference Twitter campaign," *arXiv*, February 2018. [Online]. Available: https://www.arxiv.org/pdf/1802.04291.pdf

[28] J. Paavola and H. Jalonen, "An approach to detect and analyze the impact of biased information sources in the social media," in *ECCWS2015-Proceedings of the 14th European Conference on Cyber Warfare and Security 2015: ECCWS 2015*. Academic Conferences Limited, 2015.

[29] J. Paavola, T. Helo, H. Jalonen, M. Sartonen, and A. Huhtinen, "Understanding the trolling phenomenon: The automated detection of bots and cyborgs in the social media," *Journal of Information Warfare*, vol. 15, no. 4, 2016.

[30] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, 2018.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[32] H. S. Christopher D. Manning and P. Raghavan, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[33] C. Falk, "Knowledge modeling of phishing emails," Ph.D. dissertation, Purdue University, West Lafayette, IN, 2016.

[34] A. Newcomb, "Twitter is purging millions of fake accounts and investors are spooked," *NBC News*, July 2018. [Online]. Available: https://www.nbcnews.com/tech/tech-news/twitter-purging-millions-fake-accounts-investors-are-spooked-n889941

[35] D. O'Sullivan, "American media keeps falling for russian trolls," *CNN*, June 2018. [Online]. Available: https://money.cnn.com/2018/06/21/technology/american-media-russian-trolls/index.html

[36] A. W. Clint Watts, J.M. Berger and J. Morgan, "Hamilton 68 dashboard," 2018. [Online]. Available: https://dashboard.securingdemocracy.org/

[37] "United States v. Viktor Borisovich Netyksho et.al." *United States District Court for the District of Columbia*, July 2018. [Online]. Available: https://www.justice.gov/file/1080281/download