Predicting Cyber-Attacks Using Neural Language Models of Sociopolitical Events

A Senior Honors Thesis in Interdisciplinary Studies

Tufts University Spring 2019

Advisors: Megan Monroe, Tatyana Gassel-Vozlinskaya, Dilip Ninan

By Dan Pechi

Acknowledgments

In addition to my advisors, I would like to acknowledge the additional guidance provided by Shuchin Aeron in the Electrical Engineering and Computer Science Department for sparking my interest in neural language models and computational linguistics. The supervision and technical guidance of Amy Sliva at Charles River Analytics was also instrumental in my work.

This work was supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C0112. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government.

Table of Contents

Abstract

Cyber-attacks pose an existential threat to individuals, businesses, and democracies across the world. As such, it is necessary to develop systems capable of predicting and preventing the sorts of phishing and malware attacks used to influence elections and breach private email servers like those employed prior to the 2016 presidential election. Detecting these campaigns is made intentionally difficult by those perpetrating attacks, however there exists a growing tendency among cyber adversaries to perpetrate cyber-attacks in response to sociopolitical events. Prior work has explored using machine learning techniques to process text on social media to improve early warning systems for cyber vulnerabilities. In this paper, we propose a model which leverages news data to contextualize cyber warfare in patterns of sociopolitical events that have preceded past cyber-attacks. A variety of natural language processing and machine learning approaches are presented to model this relationship. Deep learning-based approaches proved most effective in modeling these indicators of cyber-attacks, underscoring the complex representational capacity necessary to effectively model the complex world of geopolitics.

1 Introduction

1.1 A Brief History of Cyber-Attack Campaigns

Cyber-attacks have been used by foreign adversaries for decades, however the past few years have underscored the serious threat these attacks pose to the United States and other governments. Arguably, the first instance of politically motivated cyber-attacks being used alongside traditional diplomatic, military, and intelligence operations occurred in 2007 when Russian-affiliated hackers retaliated against Estonia for removing a Soviet-era statue, targeting media, government, and banking websites and crippling the country's cyber systems (Traynor, 2007). More advanced techniques were employed by Russian operatives alongside traditional military attacks in the 2008 Russo-Georgian War (Gordon, 2015). In addition to defacing websites and employing Direct Denial of Service (DDoS) attacks employed in the 2007 attacks against Estonia, internet traffic from Georgia was re-routed to servers in Russia and Turkey, where access was blocked or diverted. Alongside government websites, attackers also targeted news agencies, disrupting Georgian civilians' media access during war-time. A similar phenomenon took place in Kyrgyzstan in 2009 during a time of relative peace in which the main Internet service providers were taken down, disrupting about 80% of Kyrgyz websites (Hodge, 2009).

1.1.1 DDoS Campaigns

Smaller organizations have also perpetrated politically-motivated cyber-attacks, most in the form of DDoS attacks. In response to the release of the film *Innocence of Muslims* in late 2012, a group known as the Cyber Fighters of Izz Ad-Din Al Qassam initiated a campaign that lasted a whole year (Fox-Brewster, 2016). The DDoS attacks brought down the websites of multiple American banks including J.P. Morgan Chase and PNC Financial Services. The lead up to the 2014 release of *The Interview*, a comedy film portraying an attempted assassination of Kim Jong-Un, prompted an attack on Sony Pictures, the studio producing the film (United States Department of Justice, 2018). Hackers associated with North Korea stole information about the studio's employees and copies of unreleased films, in addition to wiping the company's entire computer infrastructure. The same organization hacked into the SWIFT banking system, stealing a total of $101 million from the Bank of Bangladesh (Ibid.)

Other, non-aligned hacktivist organizations like Anonymous and LulzSec also use cyber-attacks to target state and non-state actors alike. Anonymous launched Operation Payback in 2010 in response to copyright laws in the United States (Laville, 2012). Organizations associated with these copyright laws and others in the United States, Australia, and Spain were targeted by Anonymous over the course of the year. In response to attempts to take down WikiLeaks that same year, Anonymous hacked multiple targets involved in the proceedings including PayPal, web-hosting companies, and U.S. Senator Joe Lieberman who had supported legislation against WikiLeaks (Ibid.). LulzSec has DDoS'd the organization responsible for handling cybercrime in the United Kingdom, websites belonging to the Brazilian government, Brazil's largest oil company Petrobras, and the government of Zimbabwe (Ward, 2012).

1.1.2 Phishing Campaigns

These earlier cyber-attack campaigns have given rise to cyber-attack efforts involving more elaborate, covert spear-phishing campaigns by foreign adversaries (FireEye, 2017). These attacks involve sending emails impersonating services or internal administrators that require the user to submit their information or download attached files. The Georgian Ministry of Defense was targeted in 2014 after signing an association accords alongside Ukraine and Moldova (FireEye, 2017). The Danish military was similarly hacked in March of 2015 a few days after the country joined NATO's missile defense system, prompting the Russian ambassador to claim the country had made itself a future target of nuclear attacks (Local, 2015). Energy policy changes prompted attacks against the Qatari Ministry of Defense in late 2015 after the government signed a liquid natural gas deal with Turkey, replacing Russian suppliers (Karagoz, 2015). Employing these attacks presents a new, cost-effective way for foreign adversaries to directly attack targets with a higher degree of plausible deniability than traditional military warfare.

Perhaps the most infamous usage of cyber-attacks were perpetrated against the Democratic National Committee (DNC) in the run-up to the 2016 Presidential Election. Alongside a novel campaign on social media to propagate fake news and influence voters in key swing states, Russian hackers used phishing attacks specifically targeting 108 accounts on the hillaryclinton.com domain, including campaign manager John Podesta's email (Secureworks, 2016) In the case of the phishing email that targeted Podesta, the hackers posed as Google, asking him to enter his credentials for his email account (Ibid.). The subsequent release of Podesta's emails prompted both conspiracy theories about DNC-supported child pedophilia rings and legitimate concerns over the DNC's bias against Democratic Primary candidate Bernie Sanders (Fisher, Cox, & Hermann, 2016; Blake, 2016). The coordinated release of these emails in what has now become known as 'the October surprise' proved detrimental to Americans' faith in the election system and politics at large. More importantly, the clear influence of these efforts on the democratic electoral process, although practically impossible to quantify, demonstrates a need to prevent foreign powers from using cyber-attacks to influence the political process in the United States and elsewhere.

1.2 Maskirovka

These efforts demonstrate an evolution in traditional diplomacy and warfare capabilities. Giles et al. (2015) describes "the holistic nature of the Russian information warfare approach, where cyber activity is not a separate discipline but is included implicitly in a much wider range of tools to affect 'information space'. This includes not only information technology but also the cognitive domain…" The covert nature of these intrusions falls neatly into Russia's historical usage of 'maskirovka.' The 1944 Soviet Military Encyclopedia describes employing measures "directed to mislead the enemy regarding the presence and disposition of forces..." (Hutchinson, 2004). These measures proved effective in the case of the Cuban Missile Crisis in which Soviet troops disguised as civilians effectively deployed nuclear weapons across the country (Central Intelligence Agency). These measures provide for a degree of plausible deniability that also characterizes more modern, cyber-attack campaigns. In coordination with traditional military actions like in Ukraine in 2014, these hybrid warfare approaches can be devastating. In Ukraine, Russia was able to synergize their cyberspace and traditional military strategies, hacking Ukrainian communications, severing telecommunication lines, and targeting government, financial, and military institutions using DDoS attacks, further contributing to the country's unrest (Johanson, 2018). The deployment in Ukraine of 'little green men,' Russian military personnel without official Russian military insignia, further obfuscated the nature of Russia's incursion into Ukrainian territory, making it difficult for observers to label these actions as a traditional military invasion. This approach to modern warfare has been termed by Western scholars as the 'Gerasimov

Doctrine’ after an article written by Chief of the Russian General Staff General Valery Gerasimov in *Military-Industrial Courier* (Galeotti, 2014). Therein, Gerasimov describes there no longer existing a clear distinction between war and peace. However, Galeotti (2018) clarifies that this strategy is not actually a distinct doctrine, rather a modern form of political war. The emphasis on informational and psychological warfare demonstrates the potential for actors without traditional military power to counter adversaries in a relatively low-cost, and effective manner.

The cyber-attacks perpetrated against the Democratic National Committee underscore this effectiveness. Although attacks were identified as targeting the committee, identifying the perpetrators of these attacks was made difficult due to active disinformation and difficulties in tracing particular attacks. Initially, separate Russian groups, APT29 and APT28, also known as FancyBear and PawnStorm, were identified as being behind the attacks (FireEye, 2017). This picture became even more distorted with the emergence of a Romanian hacker named Guccifer 2.0 who took responsibility for the attacks; it was not until 2018 that Guccifer 2.0 was confirmed to be a cover for APT29 and APT28, both of which operated under Russia’s Main Intelligence Directorate (Tucker, 2018). Thus, the detection of these attacks is further problematized due to the sophisticated deception employed by foreign adversaries. Despite this obfuscation, there exists a long history of applying unconventional sensors to identify and predict attacks.

1.3 Conflict Early Warning Systems

Language encoded in news and other sources provides a means of analyzing politically-motivated cyber-attacks and more traditional warfare alike. Conflict early warning systems leverage the power of language to create ontologies of geopolitical events. The failures in addressing the 1994 Rwandan genocide and conflicts in the

Balkans in the 1990's motivated the creation of these systems to prevent future conflict and reduce loss of life (Nyheim, 2008). The creators of one of the most popular conflict early warning systems, Global Database of Events, Language, and Tone (GDELT) describe their work as "an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day" (The GDELT Project). However, this conflict early warning system only processes events related to violence against civilians and protests, which limits the system's sense of everything that is happening around the world. Conflict early warning systems with expanded event ontologies thus provide a higher resolution picture of geopolitics, and as a result are more robust in providing conflict early warnings.

1.3 Predicting Cyber-Attacks Using Sociopolitical Events

To address the challenge posed by cyber adversaries, I propose a model for predicting politically-motivated cyberattacks using news data for conflict early warning systems. Predicting cyber vulnerabilities is an emerging area of research that presents an improvement over traditional cybersecurity early warning systems that only detect intrusions either during or after infection using internal sensor data. Conflict early warning systems like GDELT (Leetaru & Schrodt, 2013) and ICEWS (Boschee et al., 2015) have traditionally been used to predict future conflict from sociopolitical events, but this model is the first to use conflict early warning systems for cyber vulnerability prediction. This model builds distributed representations of incoming news events, analyzing the feed for patterns that indicate the likelihood of triggering a cyberattack. In addition to exploring traditional machine learning models like support vector classifiers and linear regression models, various neural network architectures and state-of-the-art bidirectional language model embeddings were compared in their predictive capabilities.

To assess the efficacy of these distributed representations of sociopolitical events as predictors of cyber-attacks, I test the trained models on temporally-aligned ground truth data collected from targets of real world cyber-attacks. These results demonstrate this model is capable of predicting attacks in the ground truth data, exhibiting increased predictive accuracy in the days leading up to an attack. This paper contributes the following:

a) I propose using data from conflict early warning systems as a means of predicting the likelihood of socio-politically-motivated cyberattacks

b) I train a range of supervised machine learning models that learn the relationship between this language data and real-world cyber-attacks

c) I test these models against held-out ground truth data to assess their ability to predict future cyber-attacks

2 Related Work

The system presented in this paper is based on several rich bodies of existing research in machine learning, cybersecurity, and natural language processing.

2.1 Deep Learning for Cybersecurity

Prior research has applied deep learning to mitigate cyber-attacks. Xu et al. (2017) use a variety of features in order to detect malware infecting Android devices. Tobiyama et al. (2016) use a combination of convolutional and recurrent neural networks to analyze process behavior that can indicate malware infection. These models for malware detection post-infection present valid ways of preventing individual cybersecurity threats (Le et al., 2018). These systems which focus on particular strains of malware are far more prevalent in the deep learning-based cybersecurity literature, however less research exists on broader attack identification, let alone prediction. Hodo et al. (2016) use an artificial neural network to analyze patterns of packets being sent on the Internet of Things to detect possible Distributed Denial of Services attacks.

2.2 Machine Learning for Predicting Cyber Vulnerabilities

This early warning systems research stands alongside a body of research on machine learning systems that aim to both predict and warn users of cyber intrusions prior to infection. Edkrantz et al. (2015) used support vector machines on data gathered from the National Vulnerability Database and Exploit Database to predict future cyber-attack patterns based on previous attacks. These sorts of external data sources have been used in combination with machine learning techniques by other researchers as well (Bozorgi et al., 2010). (Liu et al., 2015) use a highly parameterized random forest classifier to predict cyber-attacks from observations of an organizations internal network activity, including network mismanagement symptoms and time series data about malicious activities like phishing on the network.

2.3 Natural Language Processing for Cybersecurity

These machine learning approaches have been augmented more recently by systems that use natural language processing techniques to analyze social media data for predicting cyber vulnerabilities (Mittal et al., 2016; Sabottke et al., 2015). Many of these systems aim to not only predict cybersecurity threats, but identify discussions on

Darkweb/Deepweb (D2web) hacking sites and other social media sites specifically relating to cyber vulnerabilities and cyber exploits (Almukaynizi et al., 2017; Lippmann et al., 2015). Ritter et al. (2015) similarly aims to extract useful social media data by discovering cybersecurity events on Twitter. Adaptive querying techniques have been used to predict DDoS attacks using Twitter data as well (Khandpur et al., 2017). Other researchers have used sentiment analysis on social media for predicting cyber-attacks. Hernandez et al. (2016) analyzed the collective sentiment of tweets responding to global events from Twitter users and hacktivist groups like Anonymous to predict cyber-attacks. Shu et al. (2018) use an unsupervised sentiment model that leverages emoticon and punctuations for aspect-based sentiment analysis on Twitter data to predict the likelihood of cyber-attacks against specific targets. More general neural language models have also been used to leverage the power of social media data for cyber-attack prediction. Tavabi et al. (2018) use the paragraph vector model proposed by Le and Mikolov (2014) to build feature representations of D2web conversations. These features are then passed to a secondary model along with other features like the discussed exploits Common Vulnerability Score System to predict cyber exploits.

3 Data Refinement and Representation

3.1 BBN Accent Encodings

BBNs ACCENT event coder presents a simple way to extract event data from global news sources. BBN ACCENT uses an extended version of the Conflict and Mediation Event Observations (CAMEO) event ontology (Leetaru & Schrodt, 2013) which consists of twenty top-level categories of events: Make Public Statement (01), Appeal (02), Express Intent to Cooperate (03), Consult (04), Engage in Diplomatic Cooperation (05), Material Cooperation (06), Provide Aid (07), Yield (08), Investigate (09), Demand (10), Disapprove (11), Reject (12), Threaten (13), Protest (14), Exhibit Military Posture (15), Reduce Relations (16), Coerce (17), Assault (18), Fight (19), Engage in Unconventional Mass Violence (20).

Each event type is further broken down into subtypes, and each event type has associated source and target actors, the initiator and recipient of the event action, respectively. For example, the news snippet ’Demonstrators in Ukraine called for the resignation of Prime Minister Mykola Azarov’ would be encoded as a ’1411 (Demonstrate for leadership change)’ with ’Protester (Ukraine)’ as the Source Actor and ’Mykola Azarov’ as the Target Actor. This ontology was expanded to include events related to cyber-attacks or possible trigger events that might motivate cyber-attacks including other cyber-attack events and election events. These new events are extracted using a hybrid of statistical and rule-based models that leverage syntactic parses, propositions, and within-document coreferences to build event encodings.

Taking all the ACCENT events that precede cyber-attacks and assuming causality adds too much noise to the relationship in the form of many false positive feature-label sets in the training data.

As such, I conducted an analysis of geopolitical events that logically preceded recorded cyber-attacks to limit the events that would be aligned with positive cyber-attacks. I analyzed open-source reports on phishing and other cyberattacks on military, defense, government, and private organizations attributed to hacking groups like PawnStorm and FancyBear. Potential trigger events for these attacks were also recorded in addition to the events ACCENT encodings, and the timeframe between the attack and trigger event. I built a frequency distribution for the ACCENT events corresponding to a total of the 91 cyberattacks from December 2013 to April 2017 (Figure 1). Only seven ACCENT event groups were used in the model based on these findings: cyberattacks, disapprove, exhibit force posture, reduce relations, reject, threaten, and elections.

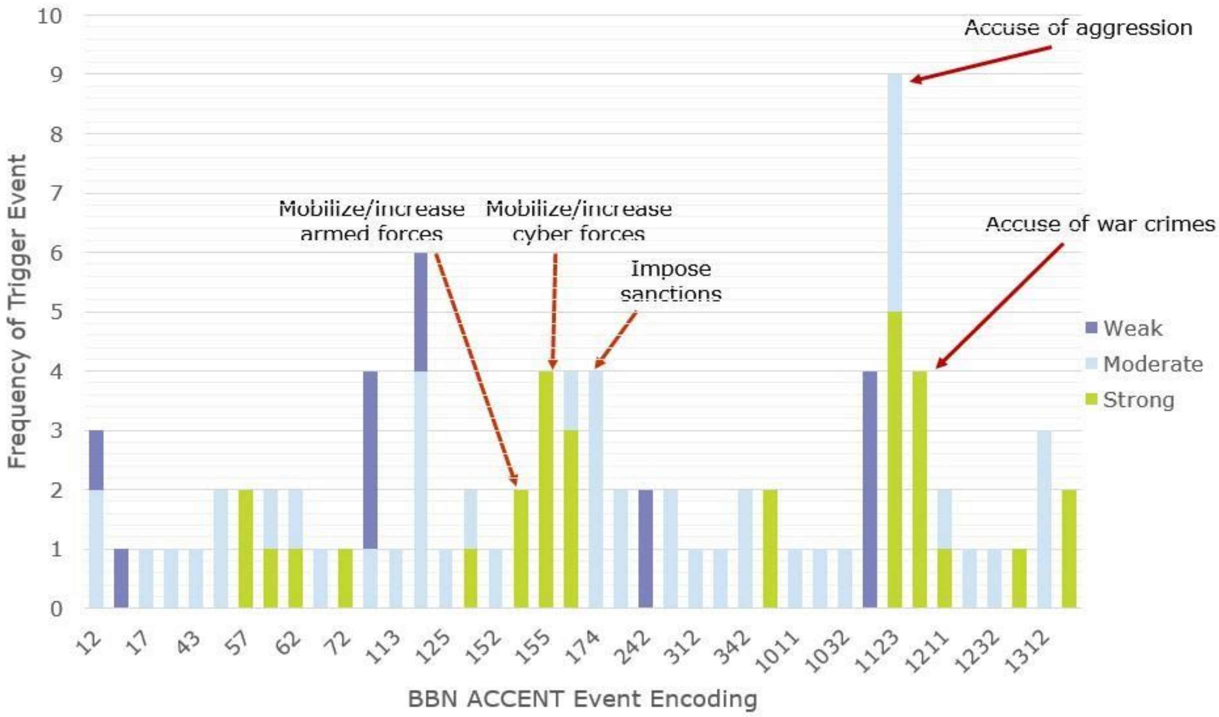


Figure 1: Histogram of BBN ACCENT Events Preceding Cyber-Attacks on Defense Sector Targets Indicating Strength of News Event Description

It should be noted that these scalar values limit the representational capacity of ACCENT encodings. In being unidimensional, these encodings fail to embed other dimensions of semantics that might transcend scalar values. This thus motivated the usage of an alternative encoding scheme using word vectors. Two different approaches were employed in this vein: using pre-trained Word2Vec embeddings (Mikolov et al., 2013) trained on billions of tokens of news data, and a pre-trained bidirectional transformer language model (BERT) (Devlin et al., 2018).

3.2 Pre-Trained Word Vectors and Geopolitics

Word embeddings have greatly advanced natural language processing by providing a means of quantitative language representation that can be integrated with machine learning systems. These models leverage the power of deep learning and distributional semantics as a means of representing language. Word2Vec, one of the most popular models for distributional word embeddings uses words contexts as a means of representing each word. These vectors are now standard for training deep learning models for sentiment analysis (Socher et al., 2013), machine translation (Gehring, Auli, Grangier, & Dauphin, 2016; Vaswani et al., 2017), and language generation (Graves, 2013).

The particular Word2Vec vectors used in this project were trained on English news data. The resulting embeddings encode semantic information about the words they represent, as demonstrated in Figure 2 from Mikolov et al. (2013).

To further investigate what pre-trained word vectors learn about geopolitics, I assessed what vectors are closest to countries' word vectors, visualized these country vectors in relation to one another, and ran analogy comparisons in Word2Vec to assess what countries had spatially similar relationships to the United States and its allies.

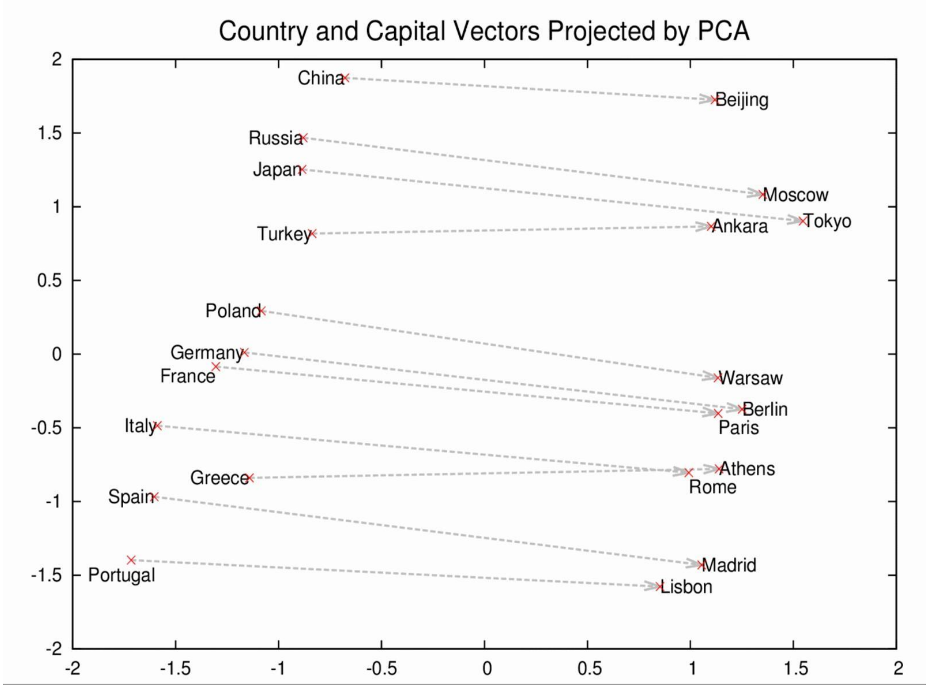
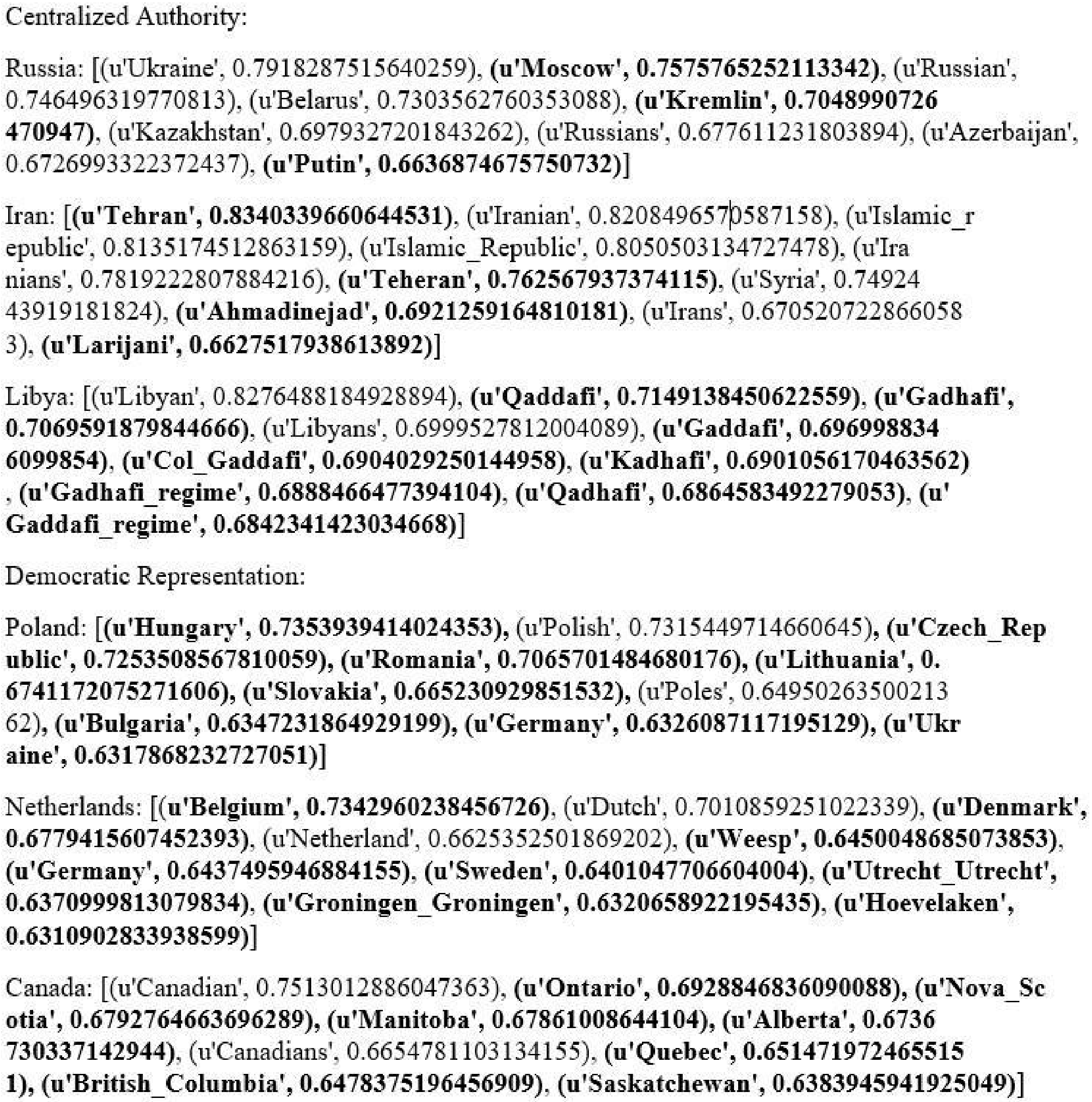


Figure 2: Visualization of Country Word Vectors in Relation to Their Capitals Projected into Two Dimensions (Mikolov et al., 2013)

3.2.1 Authoritative and Democratic Country Descriptions in English Language News

Country vectors cosine similarity with other vectors provided valuable insights into the information contained in these vectors. While countries aligned with the United States and its foreign policy interests were most similar to regions in the country or other countries the United States is allied with, political adversaries of the United States like North Korea and Iran had vectors more similarly aligned to either the country's leader or country's capital. This finding suggests these adversarial countries are discussed in the news in ways that evoke notions of a centralization of power or authoritarianism. For example, while German foreign policy may be described in terms of Germany or German parliament undertaking some action, Russian foreign policy may be described in terms of Putin, the Kremlin, or Moscow undertaking some action. Similar vectors suggestive of democratic and authoritative tendencies are in bold (Figure 3).



Centralized Authority:

Russia: [(u'Ukraine', 0.7918287515640259), **(u'Moscow', 0.7575765252113342)**, (u'Russian',
0.746496319770813), (u'Belarus', 0.7303562760353088), **(u'Kremlin', 0.7048990726
470947)**, (u'Kazakhstan', 0.6979327201843262), (u'Russians', 0.677611231803894), (u'Azerbaijan',
0.6726993322372437), **(u'Putin', 0.6636874675750732)**]

Iran: [**(u'Tehran', 0.8340339660644531)**, (u'Iranian', 0.8208496570587158), (u'Islamic_r
epublic', 0.8135174512863159), (u'Islamic_Republic', 0.8050503134727478), (u'Ira
nians', 0.7819222807884216), **(u'Teheran', 0.762567937374115)**, (u'Syria', 0.74924
43919181824), **(u'Ahmadinejad', 0.6921259164810181)**, (u'Irans', 0.670520722866058
3), **(u'Larijani', 0.6627517938613892)**]

Libya: [(u'Libyan', 0.8276488184928894), **(u'Qaddafi', 0.7149138450622559)**, **(u'Gadhafi',
0.7069591879844666)**, (u'Libyans', 0.6999527812004089), **(u'Gaddafi', 0.696998834
6099854)**, **(u'Col_Gaddafi', 0.6904029250144958)**, **(u'Kadhafi', 0.6901056170463562)**
, **(u'Gadhafi_regime', 0.6888466477394104)**, **(u'Qadhafi', 0.6864583492279053)**, **(u'
Gaddafi_regime', 0.6842341423034668)**]

Democratic Representation:

Poland: [**(u'Hungary', 0.7353939414024353),** (u'Polish', 0.7315449714660645)**, (u'Czech_Rep
ublic', 0.7253508567810059), (u'Romania', 0.7065701484680176), (u'Lithuania', 0.
6741172075271606), (u'Slovakia', 0.665230929851532),** (u'Poles', 0.64950263500213
62)**, (u'Bulgaria', 0.6347231864929199), (u'Germany', 0.6326087117195129), (u'Ukr
aine', 0.6317868232727051)**]

Netherlands: [**(u'Belgium', 0.7342960238456726)**, (u'Dutch', 0.7010859251022339), **(u'Denmark',
0.6779415607452393)**, (u'Netherland', 0.6625352501869202), **(u'Weesp', 0.6450048685073853)**,
(u'Germany', 0.6437495946884155), **(u'Sweden', 0.6401047706604004)**, **(u'Utrecht_Utrecht',
0.6370999813079834)**, **(u'Groningen_Groningen', 0.6320658922195435)**, **(u'Hoevelaken',
0.6310902833938599)**]

Canada: [(u'Canadian', 0.7513012886047363), **(u'Ontario', 0.6928846836090088), (u'Nova_Sc
otia', 0.6792764663696289), (u'Manitoba', 0.67861008644104), (u'Alberta', 0.6736
730337142944)**, (u'Canadians', 0.6654781103134155), **(u'Quebec', 0.651471972465515
1), (u'British_Columbia', 0.6478375196456909)**, **(u'Saskatchewan', 0.6383945941925049)**]

Figure 3: Vectors with High Cosine Similarity to Relevant Countries' Vectors Suggest Democratic and Authoritative Descriptions in English Language News

3.2.2 Implicit Cold War Alliances' Representation in Pre-Trained Word Vectors

By reducing the dimensionality of these vectors to 2-dimensional space, one can observe geopolitical relationships in the embedding space. The embeddings of former Soviet bloc countries are clustered together around Russia, with a separate cluster of embeddings representing allies of the United States (Figure 4). The implicit learning of Cold War alliances by these distributional word vectors demonstrates their semantic representational capacity. The similarities in the spatial relationship between Russia and its allies in the Commonwealth of Independent States, and the United States and its allies in NATO and the EU, and geopolitical groupings like the Baltic states, are also visible in these clusters. Countries in the United States' or Russia's spheres of influence are situated to the country embedding's right; a similar relationship exists between China and North Korea, and Iran and Syria, although this is poorly visualized in 2-dimensional space compared to the 300-dimensional space in which these vector relationships were analyzed. This underscores the encoding of semantic information valuable for developing better representations of these countries.

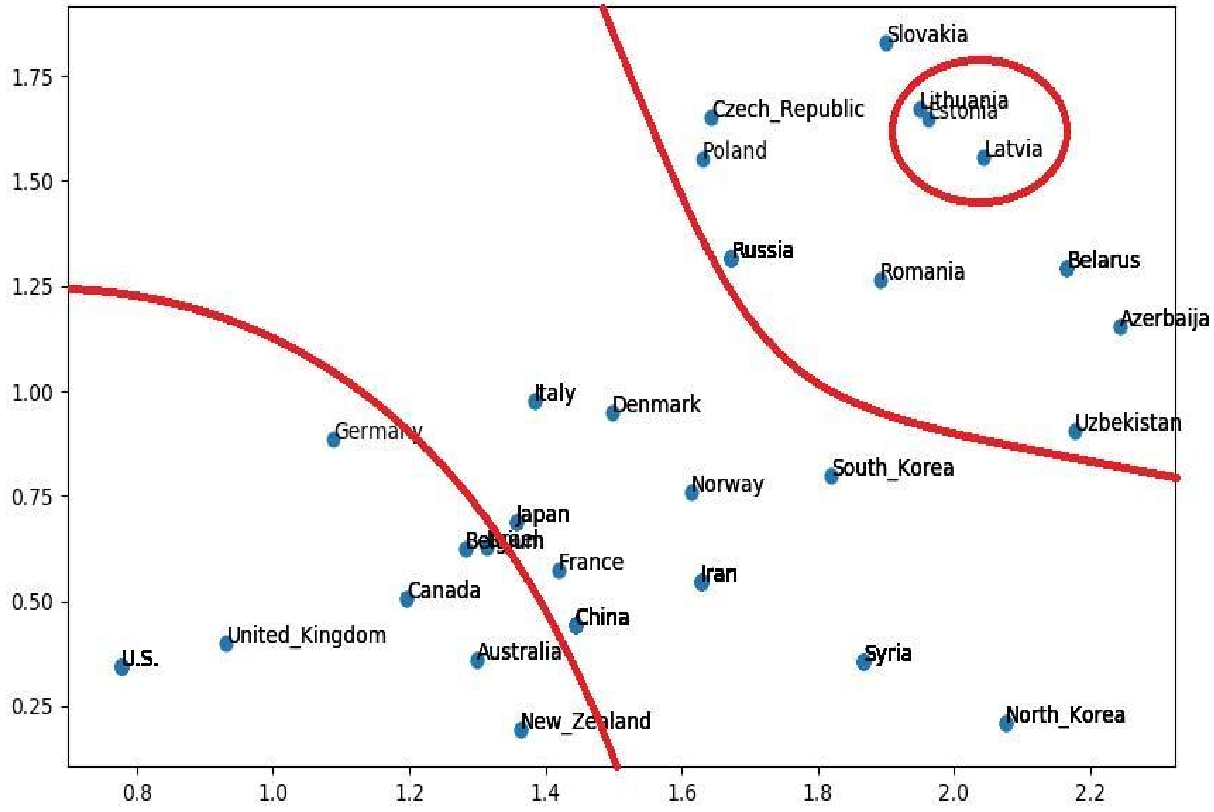


Figure 3: Clusters Representing Cold War Alliances in Two-Dimensional Word2Vec embeddings

3.2.3 Country Alliances' Representation in Distributional Word Vectors

To further examine the semantic representation of Word2Vec embeddings, an analogy task was devised to examine the geospatial relationship of the United States and its allies, and foreign governments that have used cyber-attacks targeting the United States and its allies (Figure 5).

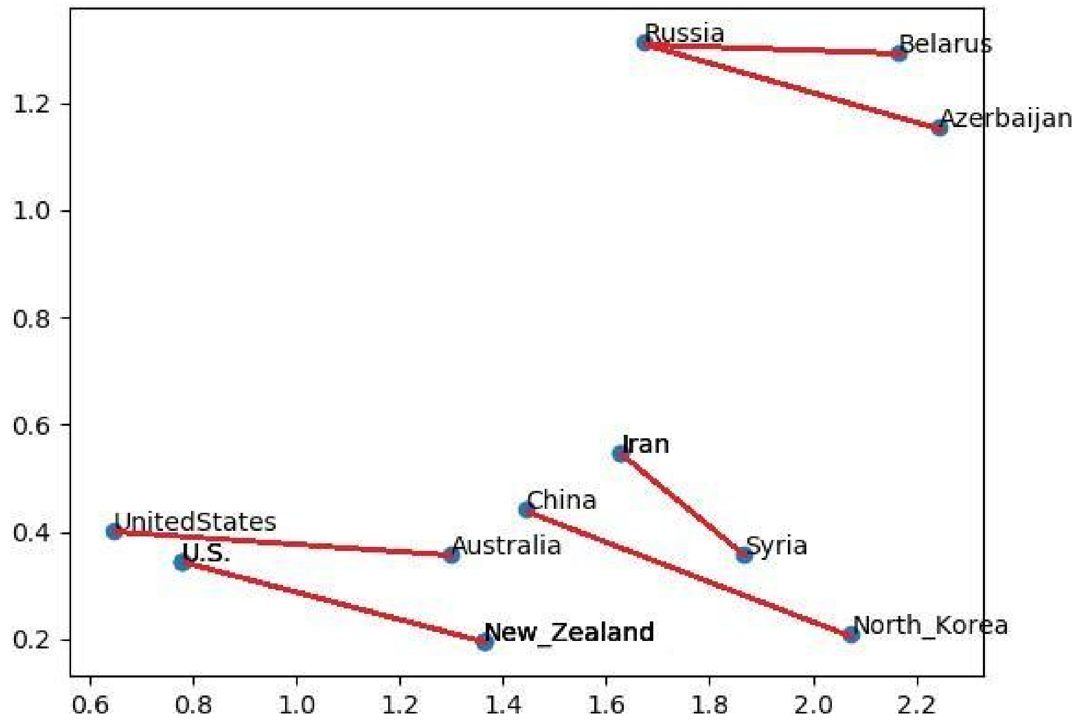


Figure 5: Word Vectors Compressed to Two-Dimensional Space Demonstrate Spatial Relationships between Countries and their Allies

As can be seen, these representations manage to encode information that roughly corresponds to alliances in international relations. It is important to note that these representations, because they are compressed to two-dimensional representations, are limited in their visual interpretability. Higher dimensions of the original 300-dimensional vectors encode semantic features that increase the accuracy of these analogies, resulting in some of the compressed relationships shown above. In 300-dimensional space, the geospatial relationship between the United States and Belgium is analogous to the relationship between Russia and both Iran and China, and to the relationship between Iran and Syria. The United States relationship to its allies in NATO and the EU was frequently analogous to Russia's relationship with allies in the Commonwealth of Independent States like Azerbaijan, Uzbekistan and Turkmenistan. The respective alliances between these countries corresponds to these geospatial representations, demonstrating the encoding of semantic information valuable for developing better representations of these countries. This is important as it improves the salience of features used to predict cyber-attacks.

3.3 Country Word Vector-Augmented ACCENT Event Encodings

These country encodings were used alongside scalar ACCENT event encodings and scalar encodings of source and target actors, increasing the dimensionality of the input features. Due to these vectors being embedded in 300-dimensional space, there was concern these features would drown out the signal from the features representing the scalar ACCENT encoding, and source and target actors. To combat this, I performed Singular Value Decomposition on these vectors to embed them in 5-dimensional space. This resulted in 13-dimensional features: 5 for the target and source country each, 2 for the source and target actors, and 1 for the ACCENT number. The other features were not embedded as word vectors due to the infrequency and wide variability of their word vector representations. ’Barack Obama’ for instance may be referred to as ’President Obama,’ ’Commander in Chief,’ or ’POTUS’. Similarly, the phrasing of ACCENT events like 154: mobilize/increase cyber forces may lack a well-defined word vector representation. Spelling variability is also problematic in some cases, exemplified by Muammar Gaddafi’s name in Figure 3.

3.4 Transferring Pre-Trained Bidirectional Transformer Language Model

These representational limitations in addition to advances in natural language processing research prompted further exploration of how to best represent geopolitical events. A growing body of work has explored transferring general purpose language models to alternative tasks (Devlin et al., 2018; Howard & Ruder, 2018; Peters et al., 2018; Radford et al. 2018). By fine-tuning language models to improve downstream performance, these models are capable of handling polysemy as embeddings are conditioned on the context of words around them. This means that there is no global representation of any one word. This adaptability motivated us to apply these models to predicting cyber-attacks from news data, specifically a bidirectional transformer language model (BERT) developed by Devlin et al. (2018). Instead of using ACCENT events, I used the supporting snippets used to generate ACCENT events as input to BERT. This means these models have access to semantic information not represented in the aforementioned event ontologies. However, having access to more information about the events also presents the possibility of adding a considerable amount of noise from the supporting snippets, so stop words were removed to mitigate this interference.

3.5 Feature-Label Sets and Training Data Segmentation

Labels were constructed using ground-truth data from private sector and government sources targeted by cyber-attacks. These attacks had been classified into four categories: malicious destination, endpoint malware, malicious emails, and attacks on internet-facing services (atoifs). Separate models were constructed for each of these attack types. To survey the full range of timeframes between attacks, separate models were constructed for timeframes ranging from 1 day in advance of an attack to 14 days in advance. This means that if a major geopolitical event were to precede an attack in the ground-truth data by 7 days, it would be labeled as a 1 for the 14 to 7 day models, but labeled as a 0 for the 6 to 1 day models. Data points labeled with a 1 were filtered so they only included geopolitical events between countries that are allied with the United States or known cyber adversaries. This effectively filtered out noisy news data involving countries whose sociopolitical events coincided with, but did not necessarily cause cyber-attacks. Training data was further segmented based on ACCENT event groupings such that different event types were used for each model. For example, the three subtype events that indicated exhibiting a force posture were used to train a model separate from other event types. These segmentations based on these 4 attack types, 14 day ranges, and 7 different ACCENT groups resulted in a total of 392 models being trained. Given the sparsity of positive feature-label sets in the data, I upsampled positively-labeled data points, effectively instituting a 4:1 ratio of negative to positive attacks. Prior to this upsampling procedure, most models were training on data that consisted of 3 percent or fewer positive labels.

4 Model Architectures for Predicting Cyber-Attacks

Because the problem of predicting cyber-attacks is particularly challenging given the low signal to noise ratio, I explored several different machine learning techniques to develop predictive models of cyber-attacks based on sociopolitical events. In this section I briefly describe each of these models and compare their performance.

4.1 Support Vector Classifiers

The first models I tested to predict cyber-attacks using the 13-dimensional features described in the previous section were support vector classifiers (SVC). SVC's are a class of supervised learning algorithms restricted to binary classification problems. The training algorithms optimizes what is referred to as a hinge loss, defined as $l(t) = (0, 1 - t \cdot y)$ where t is a predicted output of ±1, and y is the classifier score. This loss is minimized when $t = y$, and maximized when $t =- y$. By optimizing over the hinge loss, the SVM constructs a decision boundary between the two classes of data in the form of a multi-dimensional hyperplane, maximizing the distance between the points and the decision boundary.

It is important to note that SVM's attempt to maximize this margin in addition to accurately classifying data points. This tradeoff is shown in Figure 6.

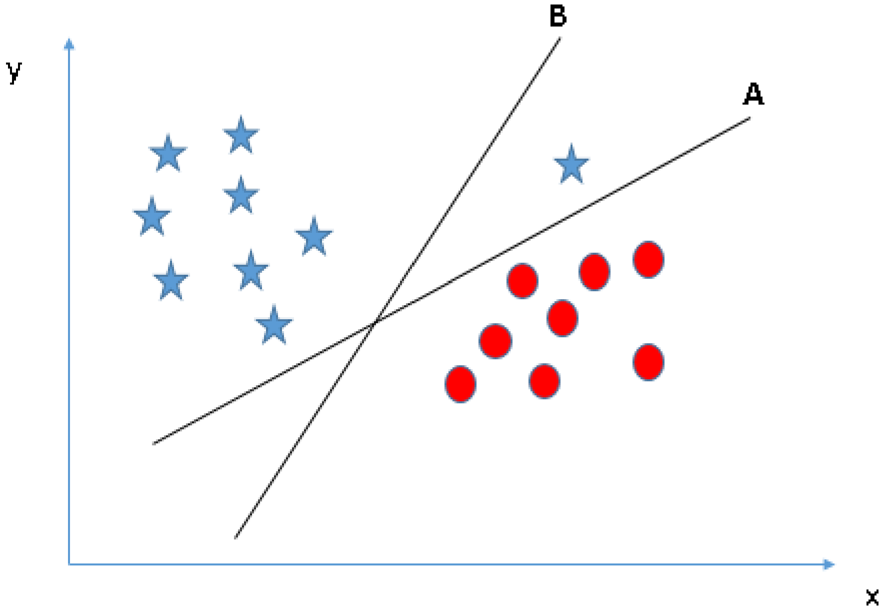


Figure 6. A Decision Boundary Tradeoff Between Classification and the Hinge Loss (Ray, 2017)

In this case, the SVC algorithm would choose line A as it more accurately classifies the training data, although it does not maximize the distance between the points and the decision boundary. This robustness motivated the implementation of an SVC classifier for the first iteration of the model. An inverse regulation strength of 5 was used, and class weights were balanced to prevent overfitting. All models used a 4:1 train/test split of the data. The results for the model were impressive, as seen in Table 1. It should be noted that these models had very low Kappa values, which suggests a lack of confidence in the models' predictions. These shortcomings are further reflected in the models' low recall rates, motivating an investigation of other model architectures.

Threat Type	**Day**	**Attack Type**	**Recall Rate**	**False Positive Rate**	**Logistic Accuracy**	**Kappa Value**
malware	4	Make Public Statement	0.113812	0.099848	0.99095	0
malicious_destination	2	Reject	0.124585	0.099825	0.99398	0
malware	11	Exhibit Force Posture	0.107503	0.099798	0.97721	0
malicious_destination	3	Cyberattack	0.12531	0.099768	0.99194	0

Table 1. Best Support Vector Classifier Model Results on Knox Datasets

4.2 Deep Neural Networks

A deep neural network (DNN), or multilayer perceptron, is similar to an SVM insofar as it is also a supervised learning algorithm. However, its mathematical underpinnings are quite different, as seen in this representation of the perceptron and multilayer perceptron in Figure 7.

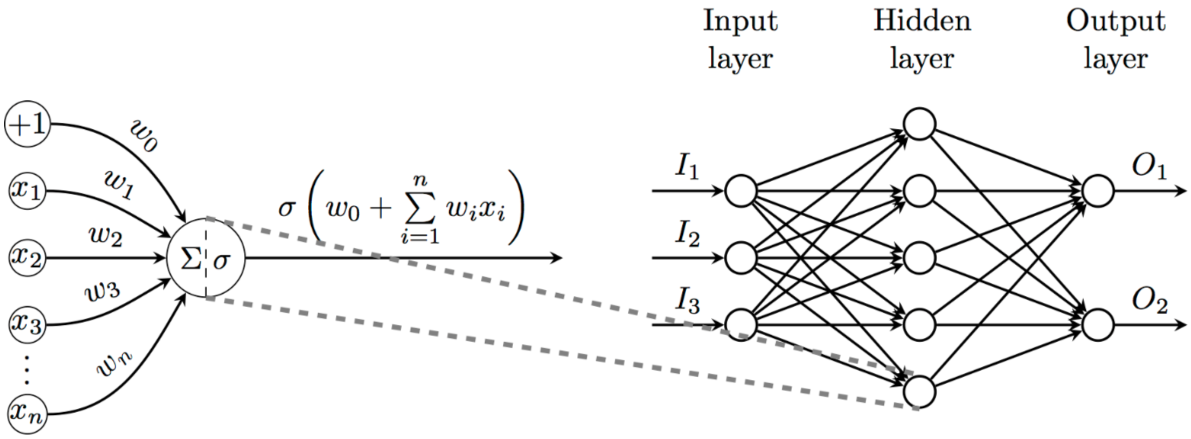


Figure 7. Mathematical Representation of a Perceptron and Its Ensembling into a Multilayer Perceptron Model (Veličković, 2016)

The DNN model was trained to optimize cross entropy which, because it's a binary classification task, takes the form of $l(t) = y \log log\ (t)\ + (1 - y) \log log\ (1 - t)$, where y is the true label and t is the predicted label. If $y = 0$, then the first term goes to 0, while the second term will be just $\log log\ (1 - t)$, meaning $l(t) = 0$ when $t = 0$ and $y = 0$. The second term cancels out when $y = 1$, meaning $l(t) = \log log\ (t)$, so $l(t) = 0$ when $t = 1$ and $y = 1$ in this case. The loss is thus maximized when the classification is the opposite of the ground truth label. The weights in the model are updated according to this loss function, reducing the algorithm's loss over time.

In addition to their loss function, DNN's differ heavily from SVC's in their representational capacity. Cybenko (1989) demonstrated the capacity of DNN models to act as universal function approximators. This is only true under the conditions that the architecture of the DNN has a sufficient number of hidden units (Hornik, 1991). In order to account for this, I performed a randomized hyperparameter search over the network architecture to test different hidden layer sizes. Hidden layer sizes of 10, 20, 30, 40, 50 and 100 were tested using 2 hidden layers, giving a total architecture search space of 36 configurations. I also manipulated additional hyperparameters alongside the architecture such as activation function, learning rate, and learning rate schedule. It is important to note that these searches in being randomized do not necessarily test all configurations; for any one model, the search may not cover the full range of hidden unit sizes or other hyperparameters. Important trends were still observed in these randomized searches. Optimal architectures tended to have smaller learning rates with a total of 49 of the 70 optimal models trained for Endpoint-Malware attacks against one data provider having the 3 lowest learning rates in the hyperparameter search (Figure 8). Larger hidden layer sizes were also far better for prediction. The 7 most commonly-selected hidden layer sizes included 50 or 100 neurons in one of the layers (Figure 9). This underscores the complexity of the functions represented by the neural networks used to predict cyber-attacks.

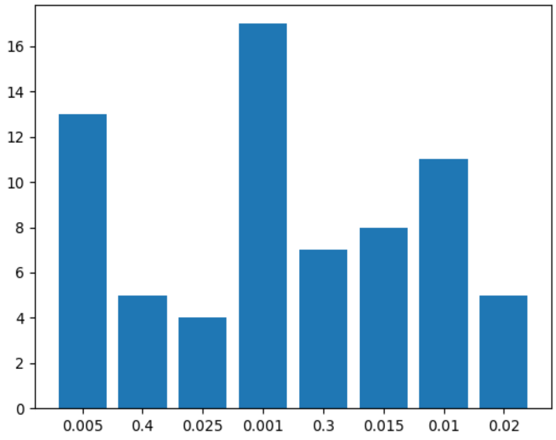


Figure 8. Knox Endpoint Malware Model Learning Rates

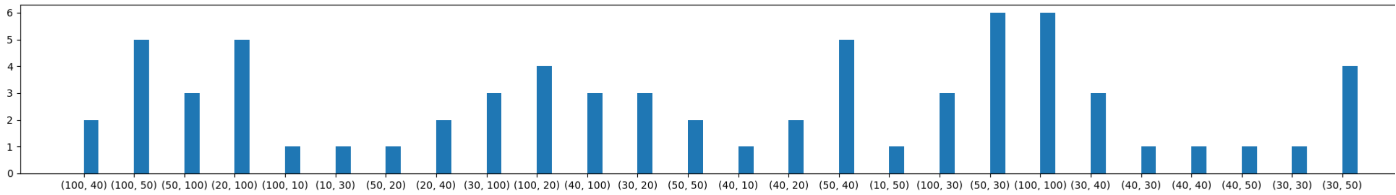


Figure 9. Armstrong Malicious-Email Model Hidden Layer Sizes

4.2 Bidirectional Transformer Language Model (BERT)

Bidirectional Transformer Language Models (BERT) are a variant of neural networks that have achieved state-of-the-art results on natural language processing tasks. The bidirectionality of these models refers to their ability to process language sequentially from both the beginning of a sentence to the end, and from the end to the beginning. Bidirectionality has also been applied to LSTM networks so as to reduce the biases induced by conditioning sequences on earlier tokens in the sequence (Huang, 2015). This is important given the hierarchical composition of language; words may be conditioned on other words and phrases in the sentence that may appear later in the sequence. A caveat to this approach is that words may indirectly be conditioned on themselves given the multilayer representations induced by the neural network. As such, BERT randomly masks tokens in the input to predict them as output, making it possible to predict arbitrary tokens in the sentence conditioned on all the other tokens.

This bidirectionality builds on top of a Transformer encoder-decoder architecture introduced in Vaswani et al (2017). Traditional, long short-term memory (LSTM) networks suffered from an inability to model long-term dependencies in which the model struggled to pass information from earlier time steps into the future. The Transformer architecture involves attention mechanisms which allow the model to focus on particularly relevant sequences of text when producing the next word in the sequence, even if they are at the opposite end of a sentence. In Vaswani et al (2017), they pass word embeddings through a neural network using a system of queries, keys and values corresponding to tokens in the sentence. By taking the softmax of dot products of keys and queries, the model can selectively attend to other embeddings of tokens in the sentence in constant time. Due to this constant path length, the model overcomes the limitations of traditional encoder-decoder architectures; this is because the model satisfies equality for the Data Processing Inequality (Tishby & Zaslavsky, 2015) between any two tokens. This means that the information for each token forms a fully-connected Markov Random Field with all other tokens. This is incredibly important as it means each token can access information from other tokens in the sequence without any noise in the channel between the tokens. This is pertinent to real world language generation, as the process of language generation often involves dependencies on arbitrary tokens already produced. This capacity to pay attention to any token also enables the model to resolve ambiguous references that often require semantic knowledge.

BERT is in a class of recent NLP models that use generative pre-training to perform better on downstream tasks. These algorithms involve training language models that build up intermediate representations of language that encode valuable information for downstream tasks. How these encodings are actually applied to these downstream tasks varies. In the case of ELMO and BERT, the intermediate representations of words in the language model's hidden layers are used to represent different features of the words. This would suggest that a general sense of language syntax is implicitly learned by the model as a sort of feature extraction. Peters et al. (2018) suggest this is why for their POS tagging model, "accuracies using the first biLM layer are higher than the top layer , consistent with results from deep biLSTMs in multi-task training (Søgaard & Goldberg, 2016; Hashimoto et al., 2017) and MT (Belinkov et al., 2017).” For a Word Sense Disambiguation model, later layers are preferred, presumably because they have more complex representations. This makes sense; understanding the difference between a noun and a verb tends to require analyzing very high-level, semantic and pragmatic structure in the form of phonological representations. That’s not the case for word-sense disambiguation which would require some sort of semantic knowledge. Because these models are based off of pre-trained language models, fine-tuning on a supervised learning problem takes far less time than training from scratch.

These advancements in sequential processing, attention mechanisms, and fine-tuning language models underscore BERT’s applicability to this prediction problem. The standard BERT model used in this work consists of 12 attention heads and 12 hidden layers, each consisting of 64 hidden units, bringing the model to a total of 110 million parameters. Fine-tuning on a single Tesla V100 GPU took approximately 4 hours to train all models.

5 Results

Table 2 summarizes the relative performance of the SVC, DNN and BERT models on their respective test-sets. The models’ accuracy, recall and false-positive rates are averaged across all the attack type and day range models of that particular architecture. The DNN models perform relatively well compared to the SVC models. In certain cases, SVC models were able to achieve accuracies comparable to the DNN models, however Kappa scores were consistently far lower for SVC models than DNN models. The vast improvements BERT presents are particularly impressive given the input features were higher-dimensional than those used for the SVC and DNN models. This suggests the architecture was capable of leveraging features beyond those presented in standard word vectors and ACCENT event encodings.

For many attack types, models predicting attacks in the near future performed better than those making long range predictions (Figure 6). This suggests that these models were able to build a stronger sense of geopolitical circumstances in the days leading up to attacks, leading to better predictive power.

Model	Accuracy	Recall	False Positive Rate	Kappa Score
SVC	65.6%	69.5%	22%	.24
DNN	89.1%	73.8%	7.0%	.66
BERT	96.4%	92.2%	2.6%	N/A

Given the generally poor performance of SVC models, more detailed results are omitted. Models using trigger events describing elections and acts of disapproval were among the best-performing DNN models. These models were particularly effective at predicting attacks on internet-facing services and endpoint-malware attacks. These particular event types occurred far more frequently than the other event types, making up about 65% of all the event types collected through ACCENT.

Table 2: Test-Set Performance Metrics Across Model Types

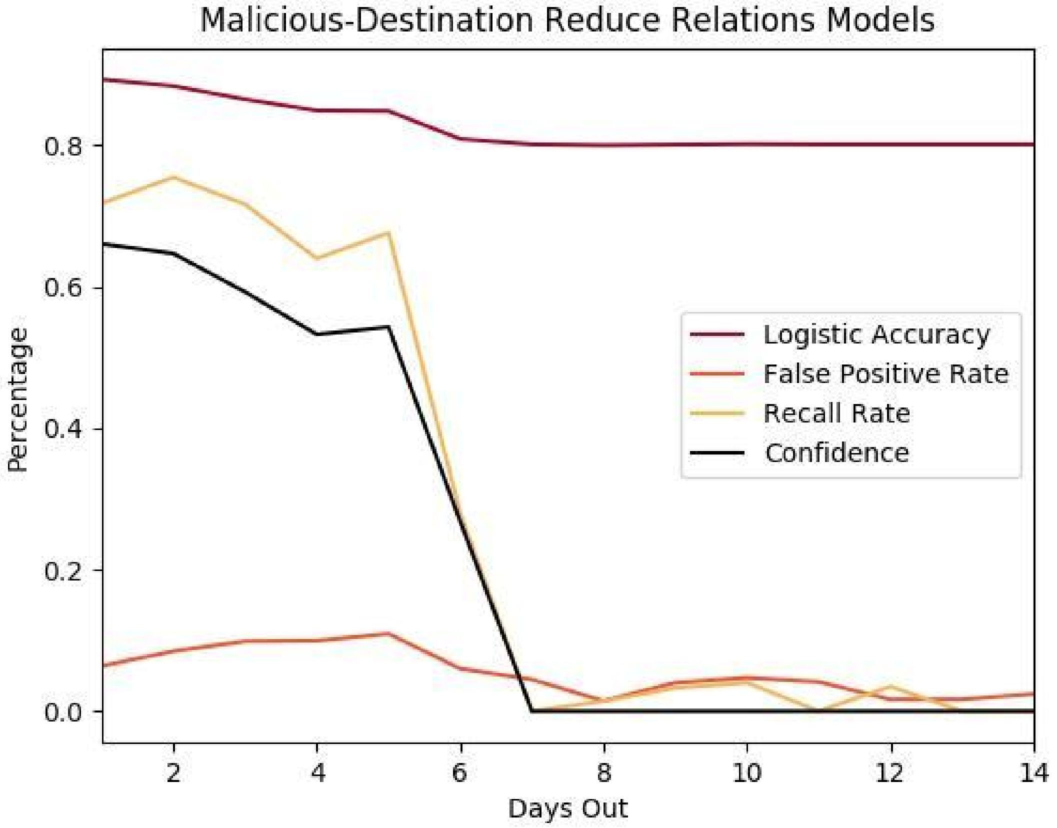


Figure 6: Prediction Metrics Worsen for Many DNN Models Predicting Attacks Further Into the Future

6 Discussion

These results suggest that neural language models with sufficient representational capacity are capable of predicting cyber-attacks from sociopolitical events. Simpler machine learning models without that representational capacity are unable to learn the associations between cyber-attacks and complex sociopolitical events.

6.1 Implications for Conflict Early Warning Systems

The effectiveness of BERT on raw language data as opposed to word vector-augmented ACCENT encodings further suggests that neural language models produce more salient representations of complex sociopolitical events than traditional early warning system encoding schemes. Other researchers have also acknowledged these shortcomings in conflict early warning systems. Wang (2018) uses a multi-instance convolutional neural network to perform the same event identification and extraction as ACCENT and GDELT, with significantly improved results over baseline models. This suggests that the translation of supporting snippets into elements of an event ontology like ACCENT has serious limitations. The improved natural language processing capabilities of neural networks over these systems underscores the need to integrate these algorithms into existing conflict early warning system pipelines. However, the capability of BERT to use raw, supporting snippets instead of feature representations induced by ACCENT suggests these early warning systems may have increasingly limited use in conflict prediction. Because BERT is powerful enough to build meaningful intermediate representations of the sociopolitical events represented in the supporting snippets, there is no need to augment these linguistic features with representations created by ACCENT. While this is the case for predictive tasks, conflict early warning systems like GDELT and ACCENT are likely to remain commonplace due to the divide between research on these systems and deep learning. It is important to note that these systems still present a convenient, human-interpretable means of condensing the nature of sociopolitical events. These representations may fail to preserve some of the semantics present in their relevant supporting snippets, however, it is important to assess just how much of a relative improvement BERT presents in semantic preservation.

6.2 Representational Capacity in Machine Learning Models

BERT, like the majority of modern natural language models, has a sense of language grounded in distributional semantics; words with similar contexts have similar meanings. Using this notion of language alone has demonstrated capabilities ranging from sentiment analysis to machine translation, tasks that necessitate not only a sense of lower level linguistic features like phonology, morphology and syntax, but more complex, socialized features like semantics and pragmatics. It is clear that modern natural language processing models, especially those based off of pre-trained language models, have a sense of these lower features, and can use them to hierarchically build compositionality. Goldberg (2019) analyzes BERT's understanding of syntax relative to models using LSTM networks, finding that the models perform similarly in word completion tasks involving long-range dependencies. This is despite the fact that LSTM models explicitly encode word order, whereas BERT has to implicitly encode this information by passing additional positional scalar values as input to the model. Attention effectively allows the model to identify other tokens in the sentence that are important for a particular classification task.

Syntactic knowledge is certainly beneficial and sufficient in many tasks; this knowledge provides a valuable basis for building more complex linguistic features, however it is unclear whether these capabilities are sufficient to build the notions of semantics necessary to analyze sociopolitical events necessary for this task. Distributional semantics certainly encodes some basis of semantics, but there exist gaps in this semantic knowledge given these systems' lack of grounding in the real world. These models' performances using the Winograd Schema demonstrates the lack of real, semantic knowledge they possess. The anaphora resolution task requires identifying which particular antecedents are being referred to by ambiguous referents, thus necessitating grounded, semantic knowledge. An example sentence and query in the task might require the agent identifying what 'it' refers to in the sentence "the trophy did not fit in the suitcase because it was too big.” Although untested with BERT, similar state-of-the-art pre-trained language models have demonstrated an inability to perform comparably to humans in this task (Radford et al., 2018). This gap in semantic understanding suggests that distributional semantics is not enough to learn word meaning. Future work grounding agents in environments where meaning may be constructed alongside traditional language models presents an avenue for learning more meaningful representations of words. For the time being, the failure of state-of-the-art language models to encode semantics suggests they are limited in their application to downstream predictive tasks. What these language models do learn about lower level syntactic features and distributional semantics was clearly enough to make accurate predictions in the context of this experiment.

This is particularly interesting given the distributional semantic knowledge of these models is constructed from news sources that aim to represent real-world events. There is thus an intermediate compression of sociopolitical events from the real-world into linguistic representations that may be biased. Although only traditionally reliable news sources were used in this research, these news sources are nonetheless limited in their capacity to represent the reality of sociopolitical events through language. More generally, this is of concern with the rise of news sources that aim to intentionally misrepresent events as they really happened. Natural language processing systems built using this data may then integrate biases that do not reflect reality and make subsequent, misinformed decisions based on those biases. It is thus critical to consider relevant social and political biases of language data so as to build systems that are more grounded in the nuance and complexity of the world.

7 References

Mohammed Almukaynizi, Eric Nunes, Krishna Dharaiya, Manoj Senguttuvan, Jana Shakarian, and Paulo Shakarian. 2017. Proactive identification of exploits in the wild through vulnerability mentions online. *2017 International Conference on Cyber Conflict (CyCon U.S.)*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. 2017. What do neural machine translation models learn about morphology? In ACL.

A. Blake. (2016, July 25). Here are the latest, most damaging things in the DNC's leaked emails. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/07/24/here-are-the-latest-most-damaging-things-in-the-dncs-leaked-emails/

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. Icews coded event data.

Mehran Bozorgi, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2010. Beyond heuristics. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 10*.

Central Intelligence Agency. Soviet Deception in the Cuban Missile Crisis. (2008, June 27). Retrieved from https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol46no1/article06.html

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. abs/1810.04805.

Michel Edkrantz, Staffan Truve, and Alan Said. 2015. Predicting vulnerability exploits in the wild. *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*.

FireEye. 2017. Apt28: A window into russias cyber espionage operations?

Fisher, M., Cox, J. W., & Hermann, P. (2016, Dec. 06). Pizzagate: From rumor, to hashtag, to gunfire in D.C. Retrieved from https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html

T. Fox-Brewster. (2016, March 24). U.S. Accuses 7 Iranians Of Cyberattacks On Banks And Dam. Retrieved from https://www.forbes.com/sites/thomasbrewster/2016/03/24/iran-hackers-charged-bank-ddos-attacks-banks/

M. Galeotti. (2018, March 05). I'm Sorry for Creating the 'Gerasimov Doctrine'. Retrieved from https://foreignpolicy.com/2018/03/05/im-sorry-for-creating-the-gerasimov-doctrine/

The GDELT Project. (n.d.). Retrieved from https://www.gdeltproject.org/

Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2016. *A Convolutional Encoder Model for Neural Machine Translation*. eprint: arXiv:1611.02344.

The 'Gerasimov Doctrine' and Russian Non-Linear War. (2017, September 17). Retrieved from https://inmoscowsshadows.wordpress.com/2014/07/06/the-gerasimov-doctrine-and-russian-non-linear-war/

Keir Giles, Philip Hanson, Roderic Lyne, James Nixey, James Sherr and Andrew Wood. (June 2015). The Russia Challenge. Chatham House Report. Retrieved from https://www.chathamhouse.org/sites/files/chathamhouse/field/field_document/20150605RussianChallengeGilesHansonLyneNixeySherrWood.pdf

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. eprint: arXiv:1901.05287.

Max Gordon. (2015 Dec.). Lessons from the Front: A Case Study of Russian Cyber Warfare. *Air Command and Staff College*

Alex Graves. Generating Sequences With Recurrent Neural Networks. 2013. eprint: arXiv:1308.0850.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In EMNLP 2017.

Aldo Hernandez, Victor Sanchez, Gabriel Sanchez,

Hector Perez, Jesus Olivares, Karina Toscano, Mariko Nakano, and Victor Martinez. 2016. Security attack prediction based on user sentiment analysis of twitter data. *2016 IEEE International Conference on Industrial Technology (ICIT)*.

Hodge, N. (2009, Jan. 28). Russian 'Cyber Militia' Takes Kyrgyzstan Offline? *Wired*. Retrieved from https://www.wired.com/2009/01/cyber-militia-t/

Elike Hodo, Xavier Bellekens, Andrew Hamilton,

Pierre-Louis Dubouilh, Ephraim Iorkyase, Christos Tachtatzis, and Robert Atkinson. 2016. Threat analysis of iot networks using artificial neural network intrusion detection system. *2016 International Symposium on Networks, Computers and Communications (ISNCC)*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991.

Hutchinson, W. 2004. The Influence of Maskirovka on Contemporary Western Deception Theory and Practice. *Proceedings of the 3rd European Conference on Information Warfare and Security*. Academic Conferences Limited, p. 166.

Ben Johanson. 2018. Asymmetric Advantage in the Information Age: An Australian Concept for Cyber-Enabled 'Special Information Warfare'. *Australian Army Journal Cyber-Warfare Edition 2018*, *14*(2).

Serdar Karagoz. 2015. Turkey, qatar sign liquefied natural gas agreement.

Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Crowdsourcing cybersecurity. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM 17.

S. Laville. (2012, Nov. 22). Anonymous cyber-attacks cost PayPal £3.5m, court told. Retrieved from https://www.theguardian.com/technology/2012/nov/22/anonymous-cyber-attacks-paypal-court

Quan Le, Oisn Boydell, Brian Mac Namee, and Mark Scanlon. 2018. Deep learning at the shallow end: Malware classification for non-domain experts. Digital Investigation, 26.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. CoRR, abs/1405.4053.

Kalev Leetaru and Philip A. Schrodt. 2013. Gdelt: Global data on events, location, and tone. ISA Annual Convention.

R. P. Lippmann, David Weller, and Carol Mensch. 2015. Finding malicious cyber discussions in social media.

Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. 2015. Cloudy with a chance of breach: Forecasting cyber security incidents. In 24th USENIX Security Symposium (USENIX Security 15), pages 1009–1024, Washington, D.C. USENIX Association.

The Local. 2015. Russia delivers nuclear threat to denmark. Retrieved from https://www.thelocal.dk/20150321/russia-threatens-denmark-with-nuclear-attack

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26.

R Socher, A Perelygin, J.Y. Wu, J Chuang, C.D. Manning, A.Y. Ng, and C Potts. (Jan. 2013). “Recursive deep models for semantic compositionality over a sentiment treebank”. In: EMNLP 1631, pp. 1631–1642.

Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. 2016. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

David Nyheim. 2008. Can Violence, War and State Collapse be Prevented? - The Future of Operational Conflict Early Warning and Response Systems. 10TH MEETING OF THE DAC FRAGILE STATES GROUP AND CONFLICT, PEACE AND DEVELOPMENT CO-OPERATION.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openaiassets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf, 2018.

Sunil Ray, and Business Analytics. "Understanding Support Vector Machine Algorithm from Examples (along with Code)." Analytics Vidhya, 11 Mar. 2019, www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly supervised extraction of computer security events from twitter. Proceedings of the 24th International Conference on World Wide Web - WWW 15.

Carl Sabottke, Octavian Suciu, and Tudor Dumitras. 2015. Vulnerability disclosure in the age of social media: Exploiting twitter for predicting realworld exploits. In 24th

USENIX Security Symposium (USENIX Security 15), pages 1041–1056, Washington, D.C. USENIX Association.

Secureworks. Threat Group 4127 Targets Hillary Clinton Presidential Campaign. (2016). Retrieved from https://www.secureworks.com/research/threat-group-4127-targets-hillary-clinton-presidential-campaign

Kai Shu, Amy Sliva, Justin Sampson, and Huan Liu. 2018. Understanding cyber attack behaviors with sentiment information on social media. Social, Cultural, and Behavioral Modeling Lecture Notes in Computer Science, page 377388.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In ACL 2016.

Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. Darkembed: Exploit prediction with neural language models. In AAAI.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In Information Theory Workshop (ITW), 2015 IEEE, pages 1–5. IEEE, 2015.

Shun Tobiyama, Yukiko Yamaguchi, Hajime Shimada, Tomonori Ikuse, and Takeshi Yagi. 2016. Malware detection with deep neural network using process behavior. 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC).

E. Tucker (2018, July 14). 12 Russians indicted for meddling in 2016 US election. Retrieved from https://www.apnews.com/1ddb174446a34785becd670275fedcbf

Ian Traynor. 2007. Russia accused of unleashing cyberwar to disable Estonia. *The Guardian*. Retrieved from https://www.theguardian.com/world/2007/may/17/topstories3.russia

United States Department of Justice (September 6, 2018). North Korean Regime-Backed Programmer Charged With Conspiracy to Conduct Multiple Cyber Attacks and Intrusions. Retrieved from https://www.justice.gov/opa/pr/north-korean-regime-backed-programmer-charged-conspiracy-conduct-multiple-cyber-attacks-and

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 2017. eprint: arXiv:1706.03762.

Petar Veličković. Multilayer perceptron. 2016. GitHub repository, https://github.com/PetarV-/TikZ/tree/master/Multilayer%20perceptron

Wei Wang. 2018. Event Detection and Encoding from News Articles. *Virginia Tech*. Retrieved from https://vtechworks.lib.vt.edu/bitstream/handle/10919/82238/Wang_W_D_2018.pdf?sequence=1&isAllowed=y

Ward, M. (2012, March 30). Anti-Sec: Who are the world's most wanted hackers? Retrieved from https://www.bbc.com/news/technology-17548704

Lifan Xu, Dongping Zhang, Nuwan Jayasena, and John Cavazos. 2017. Hadm: Hybrid analysis for detection of malware. Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016 Lecture Notes in Networks and Systems, page 702724.