ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2019.1.157

PREDICTING THE SPREAD OF MALWARE OUTBREAKS USING AUTOENCODER BASED NEURAL NETWORK

Bhardwaj Gopika[✉], **Yadav Rashi**

Department of Information Technology, Indira Gandhi Delhi Technical University for Women, India

bhardwaj.gopika@gmail.com[✉]

Abstract

Malware Outbreaks are pervasive in today's digital world. However, there is a lack of awareness on part of general public on how to safeguard against such attacks and a need for increased cooperation between various national and international research as well as governmental organizations to combat the threat. On the positive side, cyber security websites, blogs and newsletters post articles outlining the working and spread of a malware outbreak and steps to recover from the same as well. In this project, an effective approach to predicting the spread of malware outbreaks is presented. The scope of the project is 15 Malware Outbreaks and the approach involves collecting these cyber aware articles from the web, assigning them to the 15 Malware Outbreaks using Topic Modeling and Similarity Analysis and along with Spread information of the Malware Outbreaks, this is input to auto encoder neural network for learning latent space representations which are further used to predict the spread of malware outbreak as either high or low spread outbreak, achieving a prediction accuracy of 75.56. This work can be used to process large amount of cyber aware content for effective and accurate prediction in the era of much-needed cyber security.

Keywords: *malware outbreaks, topic modeling, similarity analysis, auto encoders, prediction.*

Received: 14 April 2019
Accepted: 31 May 2019
Published: 24 June 2019

1 Introduction

1.1 Background

The world today is besieged with extortionate malware attacks. More than 165 out of the 195 countries of the world have found themselves at the mercy of ransomware invasions where the attacker encrypts users documents and secures ransom upward of hundreds of dollars to decrypt them and return to the user. The need of the hour is to develop adaptive, proactive systems that can predict the extent of such outbreaks so that sufficient safeguards and kill-switches can be kept in place and educate people. Given the vast amount of data available, Machine Learning can be employed to draw analytical insights and build proactive systems in place.

1.2 Motivation

However, there is a lack of awareness on part of general public on how to circumvent such attacks and a need for increased collaboration and understanding between national and international research as well as governmental organizations to reduce and mitigate these threats. On the positive side, cyber security websites, blogs and newsletters post information regarding the coding and dissemination of a malware attack and steps to recover stolen and encrypted data. These articles along with the ground truth associated with malware spread will be used to provide a basis for predicting the spread of a malware outbreak using neural network layers. This project is aimed at predicting the spread of malware outbreaks to aid the common man who is victim of such malicious attacks.

As the spread of the malware outbreak is known, awareness will be generated and along with that, common people will become more proactive about keeping softwares up to date and looking out for patches released since a malware outbreak can lead to loss of not just money, but also critical data. The secondary beneficiaries of the project may be the governing and administrative bodies that can develop adaptive and proactive solutions to mitigate the damages caused by such malware attacks.

In this project, an attempt has been made to develop such a system that predicts the proliferation of malware outbreaks by employing Topic Modeling, Similarity Analysis, Keyword Extraction and Deep Learning. Ground Truth encompassing spread of well-documented malware attacks from standardized vulnerability datasets has been collected and corpus from articles scraped from security websites, blogs and newsletters built. Ground

Truth is the historical information about the spread of a malware outbreak and within the scope of this project has been collected for up to 1 year of spread from the month of detection of the malware outbreak.

These articles in the corpus have been taken from blogs of security experts like Graham Cluley and Schneier who also came up with the Solitaire algorithm for Cryptography. The websites which contain newsletters like CNET and ZDNET are known for rapid reporting of information related to detection of malware outbreak as well as spread, actors involved, government response etc. The articles in the corpus have been scraped from authentic sources which are elucidated in Data Collection section. The articles in the corpus have been assigned a malware label and article along with its malware label and spread of the label malware outbreak has been fed to an auto encoder neural network. The latent space representations obtained in the form of output signal and weights of the encoder have been utilized for classification and prediction of 800 manually annotated malware outbreak articles using fully connected neural network layers with softmax activation.

1.3 Outline

The paper is organized as follows. Section 2 contains a survey of the research literature regarding the techniques and tools used in the project. Section 3 describes the methodology on which the architecture has been built and Section 4 contains the implementation details of the architecture. Section 5 contains the results obtained and the conclusion and future work are detailed in Section 6. Section 7 contains the references for the literature survey.

2 Literature Survey

While a lot of research has been carried out on malware detection, analysis, clustering, classification, malware behavior as well as outbreak attacks, sadly, little work is available on the prediction of spread of a malware outbreak.

Kang et al[1] worked on predicting the number of malware infections in a country using historical data about detected malware outbreaks. They have presented an ensemble based approach which combines carefully designed domain-based features of both malware as well as machine host with a novel temporal non-linear model for malware spread and detection. However, their goal was to predict the percentage of hosts in a given population that will be infected by some malware while this project's problem statement involves predicting the spread of a malware outbreak.

For this reason, the scope of this study was defined as 15 malware outbreaks. The project began with extracting keywords related to these malware outbreaks after which regular expressions were applied for pattern matching and segregating malware-specific articles from the corpus. For the unlabeled articles, a combination of topic modeling and similarity analysis was used to label all the articles. Further autoencoder based neural network was used to get encoded representations of this entire labeled data for predictive analysis. A review of various techniques for implementing keyword extraction, topic modeling, similarity analysis and autoencoders has also been presented.

2.1 Keyword Extraction

In [2], Litvak et al introduce supervised and unsupervised approaches to identify keywords. For supervised, classification algorithms were trained on a collection of document summaries, to introduce keyword identification model. In the unsupervised approach, top-ranked nodes of a document graph represented keywords after execution of HITS algorithm. However, HITS does not improve the initial results of the degree ranking and accuracy of the obtained keywords is low. In [3], Ercan et al used Lexical chains to represent semantically related keyphrases and keywords from the text document. While lexical chains improve the precision of keyword extraction, the representation of lexical chains needs to be more accurate. Additionally, unsupervised approaches for automatic keyword extraction have been employed in [4], where the performance of TFIDF method using additional information of parts of speech and sentence salience score on speech transcripts has been improved but due to lack of structure in the data, graph based technique did not perform well. Hence, Rapid Automatic Keyword Extraction was employed in this paper.

2.2 Topic Modeling

In [5], Brants et al have used Probabilistic Latent Semantic Analysis model for topic-based document segmentation. While implementing PLSA allows for better representation of sparse information in a text block, the model cannot handle polysemy and uses a Gaussian distribution, while words in documents assume Poisson distribution hence for articles not segregated as belonging to a specific malware outbreak, topic modeling algorithm Latent Dirichlet Allocation was implemented which used Poisson as well as Dirichlet distribution for finding their individual topic probability distributions. LDA was employed because it allows choosing a fixed number of topics to be discovered and learning the topic representation of each unlabeled document along with words associated to each topic. Other topic modeling algorithms include Hierarchical Dirichlet Process which is proposed in [6] for topic discovery in document corpora. HDP is an extension of LDA but it is more complicated to implement and unnecessary in the case where a bounded number of topics is acceptable. Another technique is Non-negative Matrix Factorization proposed in [7] where in the latent semantic space derived by the NMF, each axis captures the base topic of a particular cluster of documents and each document is represented as an additive combination of the base documents. But NMF usually gives incoherent topics while LDA is more consistent and good in identifying coherent document topics.

2.3 Similarity Analysis

Various techniques for similarity analysis were studies before K L Divergence was decided upon. L2 divergence has been implemented in [8] for group-wise point-sets registration but since L2 is not suitable for probability distributions, the K L divergence metric was used. Other similarity metrics include Jensen-Renyi Divergence which has been used to measure the statistical dependence between consecutive ISAR (Inverse Synthetic Aperture Radar) image frames. Renyi Divergence is a bit more general as compared to K L Divergence, hence the latter was implemented.

2.4 Autoencoders and Prediction

Using Topic Modeling and Similarity Analysis, all the articles in the corpus have been labelled. Now a large dataset has been built consisting of all the articles, their malware outbreak labels and ground truth associated with the malware outbreak. Autoencoder based neural network was trained as described by Hinton et al in [9] to convert the high dimensional data to low-dimensional codes that are further used for classification and prediction. It has been shown in [10] that using denoising autoencoders which are trained to denoise corrupted versions of their input give very low classification error. However the input was not corrupted and a simple stacked autoencoder neural network was used. As shown in [11], variational autoencoders produce state-of-the-art results for learning the latent representations of data in an unsupervised manner to provide relevant features which enhance the performance of classifiers which in this case was a speech recognition classifier while this project entailed textual data. In [12], dimensionality reduction has been proposed using manifold learning. Here, relations within the data have been explored iteratively and used to build the manifold structure. Evaluation done on three experimental datasets have shown promising results. Autoencoder based neural network has been employed in this research study to perform dimensionality reduction as implemented in [13] where low level features have shown significant improvement in correctly classifying malware.

3 Problem Formulation

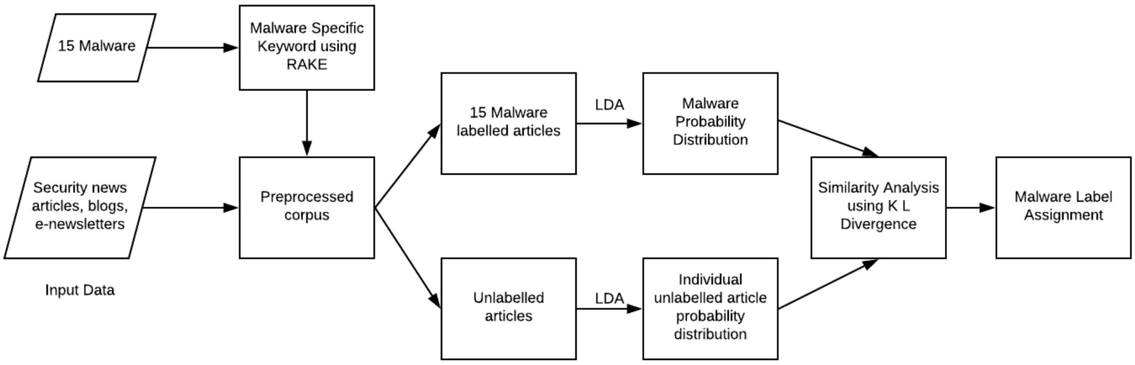
15 Malware
Malware Specific Keyword using RAKE
Security news articles, blogs, e-newsletters
Input Data
Preprocessed corpus
15 Malware labelled articles
LDA
Malware Probability Distribution
Unlabelled articles
LDA
Individual unlabelled article probability distribution
Similarity Analysis using K L Divergence
Malware Label Assignment

Figure 1: Malware Label Assginment

Predicting the Spread of Malware Outbreak is a unique research problem. While traditional research has centered on Intrusion Detection as well as Malware Detection, Spread Prediction is a novel problem statement and Deep Learning has been chosen for the same since Deep Learning Techniques like Artificial Neural Networks, Recurrent Neural Networks, Convolutional Neural Networks so on and so forth have given fantastic results in fields of Classification and Prediction and are employed by top tech companies like Google, Amazon, Facebook etc as well. However, while there is a huge amount of authentic data available on the web with respect to Malware Attacks, there is a need to develop an NLP-centered approach to extract actionable information from the same. Hence, before application of Autoencoder neural network, Topic Modeling, Similarity Analysis and Keyword Extraction for the same was applied.

3.1 Data Collection

The project began with collection of data from various sources. First, 15 malware outbreaks were identified which would be the scope of this study. Then the ground truth related to the spread of these malware outbreaks was researched and alongside, a corpus of 40000 cyber security related articles was amassed from write-ups of vulnerabilities, exploits, mitigation of various attacks, elucidation of spread of various malware outbreaks and general information regarding malware and security.

3.1.1 15 Malware Outbreaks Ground Truth

The 15 malware outbreaks studied in the research as follows - Zeus botnet, Stuxnet worm, Red October malware, Sony Pictures hack, Cryptolocker ransomware, Teslacrypt and Alphacrypt ransomware, Angler exploit kit, Nuclear exploit kit, Duqu2.0 worm, TrickBot trojan, Mirai botnet, Locky ransomware, Cerber ransomware, WannaCry ransomware and Petya.

Information regarding the working and spread of these malware outbreaks was collected based on which further information was extracted. Ground Truth related to a malware outbreak represents the extent of dissemination of a particular malware attack. In this research, Ground Truth was collected based on the following parameters - the number of countries to which the malware has spread, the number of companies which has been affected, the number of machines compromised by the malware spread, customer base loss due to the malware and monetary damages incurred by people as well as the companies. Ground Truth has been amassed from the time of when malware outbreak has been detected till 1 year of spread. While spread as per severity score of few malware outbreaks is available on National Vulnerability Database, most of the information regarding the working, detection, spread and removal of malware has been obtained from various popular blogs of cyber security and anti virus firms - Avast, Kaspersky Labs, Symantec, CISCO etc.

3.1.2 Keyword Extraction

The data collected for the 15 malware outbreaks was processed to extract relevant keywords that define the malware outbreak in question. Employing the use of Stanford Topic Modeling Toolbox was considered as described by Buyukkokten et al in [14], however, while the toolbox extracts significant keywords from text units, it is ineffective if significant words have been discarded from the dictionary either during the dictionary pruning phase or they were not crawled in the first place. Hence, Rapid Automatic Keyword Extraction, the algorithm described in [15] was used. The algorithm first extracts all possible words, phrases or terms that can be a keyword - these are all called candidates and then the algorithm calculates properties that indicate whether a particular candidate is a keyword or not, after which all candidates are scored based on the properties and keywords are selected by setting a score threshold. The properties in consideration are keyword's frequency of appearance and its co-occurrence with other words in the document. Using this technique, relevant keywords were extracted for each malware outbreak. These are elucidated in Appendix.

3.1.3 Corpus

A corpus of 41884 articles from various cyber security blogs, websites and newsletters was collected. These were scraped using python library beautiful soup and stored in CSV. Most of the data has been collected from popular website CNET and its sister website ZDNET. Both websites publish reviews, news, blogs with information about various technologies and articles were extracted for the corpus from the security topic of the news section. Other sources include the blogs of famous security professionals - Bruce Schneier, Marcus Hutchins - MalwareTech, a computer security researcher known for temporarily stopping the WannaCry ransomware attack as well as Brian Krebs, investigative reporter best known for his coverage of for-profit cyber criminals and Graham Cluley, another cyber security expert. Other sources include popular online resources that report various malware attacks, vulnerabilities and measures to keep the systems secure. The various data sources can be found in Table 1.

3.2 Malware Label Assignment

The research corpus entailed 41884 cyber security articles containing information related to various malware outbreaks as well as keywords related to the 15 malware outbreaks that form the scope of this study. Now

Table 1: Data Collection Sources

Source	Frequency
CNET	25000
ZDNET	13884
BrianKrebs	500
BruceSchneier	500
MalwareTech	500
GrahamCluley	500
SecurityWeek	200
DarkReading	300
InfoSecurity Magazine	1000
Security Week	400

each article in the corpus was assigned to the 15 malware outbreaks using regular expressions, topic modeling and similarity analysis. Flowchart Fig. 1 depicts the pipeline implemented to accomplish Malware Label Assignment.

3.2.1 Creation of Malware Outbreak Buckets

Taking the keywords extracted using Rapid Automatic Keyword Extraction, they are used to segregate malware specific articles from the entire corpus using regular expressions. Regular expressions have been used as a specific text string for a search pattern implemented on the entire corpus and if a particular malware keyword is found in an article in the corpus, the article is assigned to that malware outbreak. Malware specific articles belong to a particular malware outbreak and hence, 15 malware csv buckets were created containing articles that contain information specifically about a particular malware outbreak due to presence of the relevant keyword for the particular malware outbreak. The articles which were overlapping and present in more than one malware bucket since they contained more than one malware outbreak related keyword were removed to reduce any kind of ambiguity.

Table 2: Malware Label Assignment Statistics

Malware	Articles
Angler Exploit Kit	1712
Alphacrypt Teslacrypt	4611
Cerber Ransomware	3510
Cryptolocker Ransomware	2919
Duqu	847
Locky Ransomware	1501
Mirai Botnet	2154
Nuclear Exploit Kit	14419
Petya	817
Red October Malware	831
Sony Pictures Hack	983
Stuxnet Worm	1589
Trickbot Trojan	1068
Wannacry	5981
Zeus Bot	1093

3.2.2 Topic Modeling

This phase began with a corpus of 41884 cyber security articles, out of which 10000 have been allocated to their respective malware outbreak buckets. To assign the remaining articles to their malware outbreak buckets, topic modeling algorithm Latent Dirichlet Allocation(LDA) was used. LDA was chosen because the unlabeled corpus size is large and manual reading of articles is circumvented by this technique. LDA has been used for web spam filtering in [17], where they employed the novel multi-corpus technique for supervised web spam classification by creating a bag-of-words document for every site. It has also been used in [18] for automatically categorizing software systems in open-source repositories.

LDA was applied on each individual CSV malware outbreak bucket first. The algorithm was used to find 50 topics for each malware outbreak and the topics were bigrams. Topic probability distribution for the 50 topics was computed and stored as individual malware outbreak topic probability distribution. Additionally, LDA was applied on each article not assigned to a malware outbreak bucket for 50 bigram topics and stored topic probability distributions for each individual unlabeled article as well.

3.2.3 Similarity Analysis

After having obtained topic probability distributions, similarity analysis was performed to compare and allocate unlabeled articles to labeled malware buckets. To assign the unlabeled articles to their corresponding malware, similarity analysis was carried between individual malware outbreak bucket topic probability distributions and individual unlabeled articles topic probability distributions. K L divergence was used for performing Similarity Analysis. K L divergence has been used by Scheinder et al [19] for defining a new feature selection score for text classification by applying K L divergence between distribution of words in training documents and their classes.

K L divergence was used to find out how much the topic probability distribution of an individual unlabeled article in the corpus varied from each malware outbreak csv bucket topic probability distributions. In the case where the divergence metric gave the minimum value of divergence, the article was assigned to that malware outbreak bucket. Iterating this technique over the topic probability distributions of individual unlabeled articles and topic probability distributions of 15 malware outbreak csv buckets, each article was assigned in the corpus to a malware outbreak csv buckets and this concludes malware label assignment. The statistics for the same are given in Table 2.

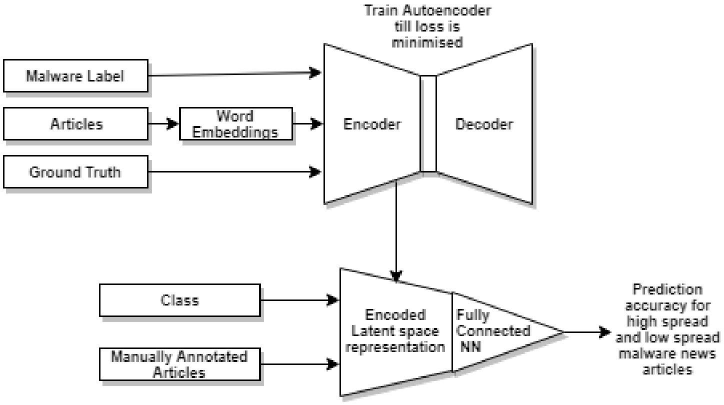
Train Autoencoder till loss is minimised
Malware Label
Articles
Word Embeddings
Ground Truth
Encoder
Decoder
Class
Manually Annotated Articles
Encoded Latent space representation
Fully Connected NN
Prediction accuracy for high spread and low spread malware news articles

Figure 2: Autoencoder based Spread Prediction

3.3 Spread Prediction

The corpus was expanded by appending the 41884 cyber security articles with their assigned malware label and the ground truth associated with malware outbreak spread. This was fed as input to an auto encoder based neural network. The individual article along with malware outbreak label and ground truth was input as a sequence of 1000 length with a maximum word length of 2000.

A multilayer feed-forward network neural network was trained with backpropagation algorithm using ReLu activation. The input vector was reproduced onto the output layer, hence the number of input and output neural units are same. Sincethe autoencoder was used for dimensionality reduction, the input was passed through the encoder layer and output of the hidden layer was saved as the compressed record. Then, to reconstruct the original vector, the compressed record is passede through the decoder and the output values of the output layer are saved as the reconstructed vector. During deployment, the network was applied to new manually annotated data, the network output values were denormalised and the chosen error metric - root mean square error (RMSE) was calculated. The aim was to train the network to achieve an acceptable performance.

The auto encoder was trained till the loss was minimized to less than 1. A deep auto encoder with 6 layers starting with 1000 hidden units, 500 hidden units, 250 hidden units, 250 hidden units, 500 hidden units and 1000 hidden units was trained. The neural network was implemented using keras with tensorflow as backend.

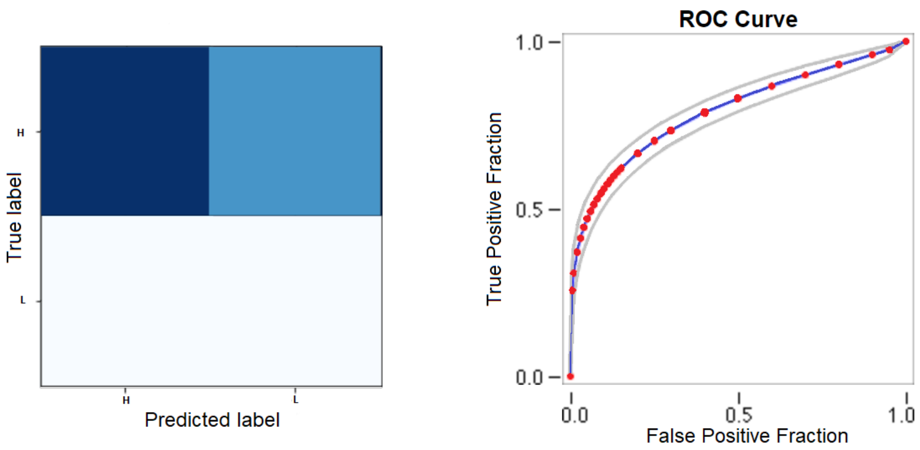

Figure 3: Confusion Matrix and ROC curve

After training the auto encoder and minimizing the loss, the decoder part of the auto encoder was removed and the encoder output - the encoded latent space representations of the input - was connected to a fully connected feed-forward artificial neural network which consisted of 3 layers of 100, 50 and 1 neuron each.

This neural network along with the encoder output was trained and tested on a manually annotated set of 800 cyber security articles. These articles were read and annotated as high spread or low spread based on the information contained within them, as articles with high spreading malware outbreak details or low spreading malware outbreak data. The spread criteria was based on the number of countries involved, number of companies mentioned, number of machines infiltrated as well as customer or monetary loss, if any, on the same lines as ground truth collection.

The annotated articles were divided into training and testing set in an 80:20 ratio and after training, the class labels for the testing set were predicted as well. The results for this are attached in the next section. Fig. 2 depicts the procedure in a flowchart.

4 Problem Solution

The results of the encoder output based fully connected neural network were Model Accuracy of 80 and Model loss of 0.4. However, this Classification accuracy alone would be misleading since there are an unequal number of observations in each class given more malware outbreak outbreaks tend to be high spread especially recently. Hence a confusion matrix is calculated to get a better idea of the accuracy of this classification model and the errors it is making.

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. The confusion matrix shows the ways in which the classification model gets confused when it makes predictions due to which it makes errors but more importantly the types of errors that are being made. It is this breakdown that overcomes the limitation of using classification accuracy alone.

ROC Curve summarizes the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0. The shape of the curve contains a lot of information, including the errors made by the model, the expected false positive rate, and the false negative rate.

The confusion matrix obtained - [125, 45], [0, 0] along with the ROC curve in Fig. 3 shows that the model sometimes misclassifies high spread malware as low spread but most truly classifies high spread as high spread. This can be fixed by using more data processed with targeted Natural Language Processing and a deep learning based Neural Network like an LSTM for classifying and making predictions based on time-series data like Malware Outbreak spreads over decades and more.

5 Conclusion

In this paper, it has been demonstrated how techniques for Keyword Extraction, Topic Modeling, Similarity Analysis and Deep Learning can be implemented to form an architecture that can predict the spread of a Malware Outbreak. The implementation results obtained showed that the accuracy metrics are relatively high given probabilistic approach towards Malware Label Assignment and Supervised Learning based Approach towards Spread Classification have been applied. In the future, this approach can be extended to recent malware outbreaks by expanding ground truth collection as well as the corpus and implementing the methodology for the new outbreaks. While this paper represents a supervised learning approach, in the future, a plan to adopt a reinforcement learning paradigm for better results is also being studied.

References

[1] Kang, C., Park, N., Prakash, B. A., Serra, E., and Subrahmanian, V. S. 2016. Ensemble models for data-driven prediction of malware infections, In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pp. 583–592.

[2] Litvak, M. and Last, M. 2008. Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*. Association for Computational Linguistics, pp. 17–24.

[3] Ercan, G. and Cicekli, I. 2017. Using lexical chains for keyword extraction. *Information Processing & Management* 43, 6, pp. 1705–1714.

[4] Liu, F., Pennell, D., Liu, F., and Liu, Y. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics, pp. 620–628.

[5] Brants, T., Chen, F., and Tsochantaridis, I. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, pp. 211–218.

[6] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems (NIPS 2004, December 13-18)*. Vancouver, British Columbia, Canada, pp. 1385–1392.

[7] Xu, W., Liu, X., and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM pp. 267–273.

[8] Wang, F., Vemuri, B., and Syeda-Mahmood, T. 2009. Generalized L2-divergence and its application to shape alignment. In *International Conference on Information Processing in Medical Imaging*. Springer, Berlin, Heidelberg, pp. 227–238.

[9] Hinton, G. E. and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786, pp. 504–507.

[10] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, Dec, pp. 3371–3408.

[11] Latif, S., Rana, R., Qadir, J., and Epps, J. 2017. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. arXiv:1712.08708. Retrieved from https://arxiv.org/abs/1712.08708

[12] Wang, W., Huang, Y., Wang, Y., and Wang, L. 2014. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. IEEE, pp. 490–497.

[13] De Paola, A., Favaloro, S., Gaglio, S., Re, G. L., and Morana, M. 2018. Malware Detection through Low-level Features and Stacked Denoising Autoencoders. In *Proceedings of the Second Italian Conference on Cyber Security, Milan, Italy, February 6–9, 2018*. CEUR Workshop Proceedings.

[14] Buyukkokten, O., Garcia-Molina, H., and Paepcke, A. 2001. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *WWW '01 Proceedings of the 10th international conference on World Wide Web* . ACM New York, NY, USA, pp. 652–662.

[15] Rose, S., Engel, D., Cramer, N., and Cowley, W. 2010. Automatic keyword extraction from individual documents. In *Text mining: applications and theory*. John Wiley & Sons, pp. 1–20.

[16] Bíró, I., Szabó, J., & Benczúr, A. A. 2008. Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. ACM, pp. 29–32.

[17] Tian, K., Revelle, M., and Poshyvanyk, D. 2009. Using latent dirichlet allocation for automatic categorization of software. In *2009 6th IEEE International Working Conference on Mining Software Repositories*. IEEE pp. 163–166.

[18] Schneider, K. M. 2004. A new feature selection score for multinomial naive Bayes text classification based on KL-divergence. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain, Article No. 24.