An Empirical Study on Detecting Deception and Cybercrime Using Artificial Neural Networks

Alex V Mbaziira
Marymount University
1000 North Glebe Road
Arlington, VA 22207
+1 703 908 7508
ambaziir@marymount.edu

Diane R Murphy
Marymount University
1000 North Glebe Road
Arlington, VA 22207
+1 703 284 5958
dmurphy@marymount.edu

ABSTRACT

Ubiquity of the Internet and wide adoption of the computing and mobile devices is driving explosion of data. Interestingly, cybercriminals are also leveraging these popular technologies to cash in on cybercrime in form of scams, fraud and fake online reviews. Existing content filtering techniques, which have been successful in containing spam, are failing to filter these new types of cybercrime because cybercriminals generate text messages to bypass content filters. In this paper, we use natural language processing and a deception-detection discourse to build hybrid models for detecting these forms of text-based cybercrime. Since we have four datasets each of which contains deceptive text messages representing a specific type of cybercrime and truthful text messages, we combine 2 datasets and 3 datasets together to generate training sets for the hybrid models with more than one type of cybercrime. The hybrid cybercrime detection models are trained using Artificial Neural Networks (ANN), Naïve Bayes (NB), Support Vector Machines (SVM) and kth Nearest Neighbor (kNN). The models are then evaluated on test sets containing instances that were not part of the training sets. The results for model performance of NB, kNN and SVM classifiers are compared against those of ANN. Most the models generalize well in detecting cybercrime. ANN model performance on the test sets ranges from 70% to 90% accuracy compared to model performance range of 60% to 80% for the other three classifiers. The best performance is in detecting unfavorable fake reviews and fraud.

CCS Concepts

•Social and professional topics→Computer crime•Computing methodologies→Natural language processing•Computing methodologies→Supervised learning by classification

Keywords

Cybercrime, deception, supervised learning, artificial neural networks, natural language processing.

ICCDA 2018, March 23–25, 2018, DeKalb, IL, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6359-4/18/03…$15.00
https://doi.org/10.1145/3193077.3193080

Digital technology and the Internet have continued to drive explosion of text-based data. Cybercriminals continue to find ways to leverage digital technologies for text-based content dissemination systems to exploit their victims. Content filtering mechanisms have been successful in containing spam in email systems. However, cybercriminals use deception to trick and exploit their victims using text messages when committing cybercrime in form of fraud, scams and fake online reviews [1]. Most text-based cybercrime is committed in English since this is the most dominant language on the web [2]. We also need to point out the challenges arising from data explosion, also known as big data, are presenting new opportunities for investigating patterns of cybercrime[3]. In this paper, we demonstrate how artificial neural networks (ANN) can be applied in detecting and analyzing deception and cybercrime in text messages. We compare the results of models trained in ANN with models trained on well-studied classifiers like Naïve Bayes (NB), Support Vector Machines (SVM) and kth Nearest Neighbor (kNN). We use a deception discourse because since deception is the underlying attribute that links cybercrime in form of scams, fake reviews and fraud.

We evaluate performance of ANN against three well-studied binary text classification algorithms, that is NB, SVM and kNN which have been used to study cybercrime in form of spam [4]–[6]. ANN are machine learning algorithms comprising of interconnected nodes and directed links that are used to solve classification problems. NB discriminates between binary classes by computing posterior probabilities for each of the two classes in a given dataset [7]. SVM uses a maximal hyperplane to linearly discriminate instances into binary classes [8] while kNN uses a distance function to determine the closest known instances to a given unknown instance.

The remainder of the paper is organized as follows: Section 2 presents literature review on deception and cybercrime while Section 3 describes the methodology and experimental set up for data collection and analysis. In Section 4, we present the results of the experiments and discuss our findings. Section 5 has conclusions for the paper.

1.1 Key Contributions

We build generalizable natural language-based hybrid models for cybercrime detection. The paper investigates four types of cybercrime, which are: fraud, scam, unfavorable fake reviews and favorable fake reviews. Generally, hybrid models trained in ANNs perform better than the previous models trained in NB, SVM, kNN in detecting cybercrime [9].

2. RELATED WORK

2.1 Artificial Neural Networks (ANN)

ANN comprise of inter-connected nodes, which are also called neurons to predict patterns in data. To predict patterns in data, ANN use input neurons for processing the data and determining the output [7]. There are different types of ANN, for this paper, however, we use a single layer perceptron, which is a type of feed forward ANN without hidden layers. Single layer perceptron ANN use linear functions to learn and detect patterns in the data. We therefore use single layer perceptron ANN to evaluate classifier performance in detecting deception and cybercrime on normalized datasets with linear patterns and compare the results of models trained with classifiers like NB, SVM and kNN which are studied in previous research [9], [10].

2.2 Deception and Cybercrime Detection

There is limited work on detecting deception and cybercrime in text messages using machine learning. Early work in [11] attempted to use computational linguistic processes and classification to detect deception in text messages. Similarly, there is also early research on psycholinguistic processes and deception detection in [12], [13]. We extend all that work by identifying computational linguistic processes linked to deception and cybercrime which we compliment with psycholinguistic processes to build models for detecting deception and cybercrime in text messages.

There is recent work on fake reviews and ANNs, which uses sentence representation and semantics [14]. The paper uses a publicly available dataset, which we also use to study favorable and unfavorable fake reviews. Our work differs from that paper because we only use both truthful favorable and unfavorable reviews with transaction data as ground truth to show that reviewers were guests at those hotels but not merely expressing opinions about the services offered.

There is also research on detecting favorable fake reviews using n-gram model with deeper syntactical features like: weight, location, price [15]. This paper does not use n-gram analysis because it is not robust and overfits [16], [17].

Similarly, there is also research on scam detection in Twitter that uses a semi-supervised learning approach with n-grams and component analysis [18]. This paper uses supervised learning to detect deception and cybercrime, however, we do not use n-grams because such models are not robust even when principal components are applied.

3. METHODOLOGY AND EXPERIMENT SET UP

3.1 Dataset Description

We use four datasets each of which contains a text messages with a specific type of cybercrime combined with truthful messages as shown in Table 1. The scam dataset was generated from Facebook accounts of publicly leaked email addresses of an actual cybercriminal network that was using an online data theft service [19]. All the messages obtained from these accounts were public. For the fraud dataset we used emails on the Enron case that were made public by the Federal Energy Regulatory Commission [20] as well as court evidence which was made public by Department of Justice due to public interest in the case. We also obtained two additional public datasets on favorable and unfavorable reviews [21].

We randomly sampled 100 instances from each dataset to generate our training models for detecting deception and cybercrime. For the test set we randomly selected 20 instances from each dataset that was not part of the training set. The size of our training and testing sets was limited by the ground truth data on fraud, favorable fake reviews and unfavorable fake reviews [9]. To generate the 2-dataset hybrid models, we combine any two of the four datasets while for the 3-dataset hybrid models, we combined any three of the four datasets.

All the instances in the four datasets were preprocessed to remove noise like non-English messages in Facebook posts, non-ASCII characters, and text-based emoticons, as well as email headers and email threads which were in the Enron and Department of Justice datasets. We also used 10-fold cross validation on the training sets when building the models and normalized the data using WEKA's normalization filter.

Table 1. Dataset and model description

Hybrid Model Type	**Model**	**Dataset Composition**	**Train Set Instances**	**Test Set Instances**
2-Dataset	EN + FB	Enron & Facebook	200	20
2-Dataset	EN + NR	Enron & Unfavorable Reviews	200	20
2-Dataset	EN + PR	Enron & Favorable Reviews	200	20
2-Dataset	FB + NR	Facebook & Unfavorable Reviews	200	20
2-Dataset	FB + PR	Facebook & Favorable Reviews	200	20
2-Dataset	NR + PR	Favorable & Unfavorable Reviews	200	20
3-Dataset	EN+FB+NR	Enron, Facebook & Unfavorable Online Reviews	300	20
3-Dataset	EN+FB+PR	Enron, Facebook & Favorable Reviews	300	20
3-Dataset	FB+NR+PR	Facebook, Unfavorable & Favorable Reviews	300	20
3-Dataset	EN+PR+NR	Enron, Favorable & Unfavorable reviews	300	20

3.2 Ground Truth

Text messages for the scam dataset were manually verified for work-at-home scams, lottery scams, advertisements for hacking services, and carding as well as truthful messages. The Enron emails, comprising of the fraud dataset, which were used as evidence to prosecute the former Enron corporation executives for securities and wire fraud were labelled as deceptive because since the charges for the case were not appealed [22]. However, other emails were made public by the Federal Energy Regulatory Commission were labeled as truthful. The public datasets for hotel reviews were already labelled [21]. Lastly, for the favorable and unfavorable review datasets, we considered truthful reviews with transactional data like deals prices for services, meals, valet services, etc. as assurance that reviewers were guests at the hotel [23].

3.3 Feature Selection

We identify CL features from CL processes linked to cybercrime and deception detection which are: verbs, modifiers, average sentence length, average word length, pausality, modal verbs, emotiveness, lexical diversity, redundancy, characters, punctuation marks, sentences, adjectives, adverbs, nouns and function words [11].

PL processes relevant to cybercrime are: analytical, words per sentence, 6-letter words, I, we, you, she/he, affect, positive emotion, negative emotion, insight, causation and certainty words.

We drop all features of hybrid models for non-native English cybercriminal networks in datasets whose normalized average was below 0.1 average.

3.4 Evaluating Classifier Performance

We use precision, recall, F-measure, and Receiver Operating Characteristic (ROC) curves. Precision measures the fraction of the relevant deceptive instances that the classifier declared deceptive, while recall measures the number of relevant deceptive messages that are correctly predicted. F-measure is the harmonic mean for precision and recall, and ROC curves illustrate the tradeoff between true positive rate and false positive rate of the classifier.

3.5 Feature Engineering

The features for the learning model are generated from two processes linked deception and cybercrime. We first extract some features of computational linguistic processes linked to deception and cybercrime, which are: verbs, modifiers, average sentence length, average word length, pausality, modal verbs, emotiveness, lexical diversity, redundancy, characters, punctuation marks, sentences, adjectives, adverbs, nouns and function words [11]. The rest of the features for our cybercrime detection models are generated from psycholinguistic processes linked to deception and cybercrime. These features are: analytical, words per sentence, 6-letter words, I, we, you, she/he, affect, positive emotion, negative emotion, insight, causation and certainty words.

3.6 Classifier Performance Evaluation

To evaluate performance of our classifiers, we use precision (P), recall (R), F-measure (F), and Receiver Operating Characteristic (ROC) curves. Precision measures the fraction of the relevant text-based instances that the classifier declared deceptive, while recall measures the number of relevant text-based messages with cybercrime that are correctly predicted. F-measure is the harmonic mean for precision and recall.

4. RESULTS AND DISCUSSION

Both the 2-dataset and 3-dataset hybrid models are trained using ANN and the results are compared against those of NB, SVM and kNN models for both types of hybrid models. Table 2 and Table 3 summarize the performance of the classifiers for the 2-dataset and 3-dataset hybrid models respectively. We only consider models whose predictive accuracies for both the truthful and deceptive class was over 50% because the human beings can only detect deception at an accuracy rate of 50% [23].

Table 4 shows results for predictive accuracy of 2-dataset hybrid models. Firstly, the EN+FB model is trained on deceptive text messages in form of fraud and scams as well as truthful text messages. Both the NB and ANN classifiers for the EN+FB model detect only favorable fake reviews with 60% accuracy. Secondly, the EN+NR model is trained on deceptive text messages comprising fraud and unfavorable text messages as well as truthful messages. The ANN and SVM classifiers for the EN+NR model detect favorable fake reviews with 70% and 60% accuracy respectively.

Thirdly, EN+PR model is trained on deceptive text messages in form of fraud and unfavorable fake reviews as well as truthful messages. Both the SVM and kNN classifiers for the EN+PR model detect unfavorable fake reviews with 70% accuracy.

Fourthly, the FB+NR model is trained on deceptive text messages in form of scam and unfavorable fake reviews as well as truthful text messages. The NB and SVM classifiers for the FB+NR model detect only fraud with 70% and 80% accuracy, respectively, while the ANN classifier for the model detects both fraud and favorable fake reviews with 90% and 60% accuracy, respectively.

Fifthly, the FB+PR model is trained on deceptive text messages in form scam and favorable fake reviews as well as truthful messages. The NB, SVM and kNN classifiers for the FB+PR model detect both unfavorable fake reviews and fraud. The NB classifier detects both fraud and unfavorable fake reviews with 60% and 70% accuracy respectively. SVM classifier for the same model also detects unfavorable fake reviews and fraud with 80% and 60% accuracy. Furthermore, the kNN classifier detects unfavorable and fraud with 70% and 60% accuracy. However, the ANN classifier detects only fraud with 90% accuracy. Lastly, the NP+PR model is trained on deceptive messages comprising unfavorable and favorable fake reviews. All the four classifiers of the model detect unfavorable fake reviews with 60% accuracy.

Table 5 shows results for predictive accuracy of three 3-dataset hybrid models which are: EN+FB+PR, EN+NR+PR, and FB+PR+NR. Firstly, the EN+FB+PR model on deceptive messages comprising fraud, scam and favorable fake reviews. Both NB and ANN classifiers for the model detect unfavorable fake reviews with 60% accuracy while SVM and kNN classifiers detect favorable fake reviews 70% accuracy. Secondly, the FB+PR+NR model is trained on deceptive messages with scam, favorable and unfavorable fake reviews as well as truthful messages. Both the NB and SVM classifiers of the model detect fraud with 70% accuracy while the kNN and ANN classifiers detect fraud with 60% and 80% accuracy. Thirdly, the EN+FB+NR model is trained on fraud, scam and unfavorable fake reviews and only the kNN classifier detects favorable fake reviews with 60% accuracy.

Classifiers for both 2-dataset and 3-dataset hybrid models detect more patterns of deception and cybercrime within fraud and favorable fake reviews. Generally, the ANN 2-dataset models detect fraud and favorable fake reviews with higher predictive accuracies compared to models trained with NB, SVM and kNN. Furthermore, when more types of cybercrime are added to the datasets to generate 3-dataset models, we observe that ANN models generalize better in fraud detection compared to NB, SVM and kNN models. However, all hybrid 2-dataset and 3-dataset models fail to detect scams. This pattern was also observed in a previous paper since the scam dataset was collected from a non-native English-speaking cybercriminal network [5].

Table 2. Performance of the 2-dataset hybrid models

Model	**Classifiers**	**P**	**R**	**F**	**ROC**
EN + FB	NB	0.94	0.94	0.94	0.94
EN + FB	SVM	0.75	0.75	0.75	0.75
EN + FB	KNN	0.78	0.77	0.76	0.75
EN + FB	ANN	0.68	0.68	0.68	0.72
EN + NR	NB	0.67	0.67	0.66	0.73
EN + NR	SVM	0.78	0.77	0.77	0.77
EN + NR	KNN	0.75	0.75	0.75	0.75
EN+NR	ANN	0.84	0.84	0.84	0.89
EN + PR	NB	0.73	0.72	0.72	0.82
EN + PR	SVM	0.81	0.81	0.81	0.81
EN + PR	KNN	0.79	0.79	0.79	0.79
EN + PR	ANN	0.70	0.70	0.69	0.75
FB + NR	NB	0.6	0.6	0.59	0.59
FB + NR	SVM	0.65	0.65	0.65	0.65
FB + NR	KNN	0.61	0.61	0.61	0.6
FB + NR	ANN	0.68	0.68	0.68	0.72
FB + PR	NB	0.63	0.61	0.6	0.65
FB + PR	SVM	0.73	0.72	0.71	0.72
FB + PR	KNN	0.66	0.66	0.66	0.65
FB+PR	ANN	0.70	0.70	0.69	0.75
NR + PR	NB	0.67	0.66	0.66	0.71
NR + PR	SVM	0.76	0.76	0.76	0.76
NR + PR	KNN	0.68	0.68	0.67	0.74
NR + PR	ANN	0.76	0.76	0.76	0.83

Table 3. Performance of the 3-dataset hybrid models

Model	**Classifier**	**P**	**R**	**F**	**ROC**
EN+FB+NR	NB	0.66	0.66	0.65	0.68
EN+FB+NR	SVM	0.71	0.7	0.7	0.7
EN+FB+NR	KNN	0.71	0.7	0.7	0.7
EN+FB+NR	ANN	0.75	0.75	0.75	0.82
EN+FB+PR	NB	0.67	0.66	0.65	0.71
EN+FB+PR	SVM	0.75	0.75	0.75	0.75
EN+FB+PR	KNN	0.73	0.72	0.72	0.81
EN+FB+PR	ANN	0.76	0.76	0.76	0.84
FB+NR+PR	NB	0.62	0.62	0.61	0.63
FB+NR+PR	SVM	0.64	0.64	0.64	0.64
FB+NR+PR	KNN	0.63	0.62	0.62	0.66
FB+NR+PR	ANN	0.69	0.69	0.69	0.74

Table 4. Predictive accuracy of the 2-dataset hybrid models

Model	Classifier	Unfav. fake reviews	Fav. fake reviews	fraud
EN + FB	NB	50%	**60%**	
EN + FB	SVM	50%	30%	
EN + FB	KNN	40%	50%	
EN + FB	**ANN**	50%	**60%**	
EN + NR	NB		50%	
EN + NR	SVM		**60%**	
EN + NR	KNN		50%	
EN + NR	**ANN**		**70%**	
EN + PR	NB	50%		
EN + PR	SVM	**70%**		
EN + PR	KNN	**70%**		
EN + PR	ANN	50%		
FB + NR	NB		50%	**70%**
FB + NR	SVM		50%	**80%**
FB + NR	KNN		50%	50%
FB + NR	**ANN**		**60%**	**90%**
FB + PR	NB	**60%**		**70%**
FB + PR	SVM	**80%**		**60%**
FB + PR	KNN	**70%**		**60%**
FB + PR	**ANN**	50%		**90%**
NR + PR	NB	**60%**		
NR + PR	SVM	**60%**		
NR + PR	KNN	**60%**		
NR + PR	**ANN**	**60%**		

Table 5. Predictive accuracy of the 3-dataset hybrid models

Model	Classifier	unfav. fake reviews	fraud	fav. fake reviews
EN+FB+PR	NB	**60%**		
EN+FB+PR	SVM	**70%**		
EN+FB+PR	KNN	**70%**		
EN+FB+PR	**ANN**	**60%**		
FB+NR+PR	NB		**70%**	
FB+NR+PR	SVM		**70%**	
FB+NR+PR	KNN		**60%**	
FB+NR+PR	**ANN**		**80%**	
EN+FB+NR	NB			50%
EN+FB+NR	SVM			50%
EN+FB+NR	KNN			60%
EN+FB+NR	**ANN**			50%

5. CONCLUSION

In this paper, we demonstrate that it is possible to build natural language processing-based hybrid models for detecting and analyzing deception and cybercrime in text messages, using ANN. The ANN models generalize well in detecting deception and cybercrime. Future work will explore improvements in detection of cybercrime in non-native English-speaking cybercriminal networks.

6. REFERENCES

[1] P. Engel, “ISIS has mastered a crucial recruiting tactic no terrorist group has ever conquered,” *Business Insider*, 09-May-2015. [Online]. Available http://www.businessinsider.com/isis-is-revolutionizing-international-terrorism-2015-5. [Accessed: 16-Mar-2016].

[2] H. Young, “The digital language divide,” 2014. [Online]. Available: http://labs.theguardian.com/digital-language-divide/. [Accessed: 27-Nov-2016].

[3] G. C. | in T. D. Maker, September 16, 2013, and 8:43 Am Pst, “Use big data to fight cybercrime,” *TechRepublic*. [Online]. Available: https://www.techrepublic.com/blog/tech-decision-maker/use-big-data-to-fight-cybercrime/. [Accessed: 12-Nov-2017].

[4] L. Firte, C. Lemnaru, and R. Potolea, “Spam detection filter using KNN algorithm and resampling,” in *2010 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2010, pp. 27–33.

[5] L. Pearl and M. Steyvers, “Detecting authorship deception: a supervised machine learning approach using author writeprints,” *LLC*, vol. 27, pp. 183–196, 2012.

[6] S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali, “Detecting deceptive reviews using lexical and syntactic features,” in *2013 13th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2013, pp. 53–58.

[7] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to* Data *Mining*, 2nd ed. South Asia: Dorling Kindersley, 2014.

[8] C. Chang and C.-J. Lin, *LIBSVM: a Library for Support Vector Machines*. 2001.

[9] A. V. Mbaziira and J. H. Jones, “Hybrid Text-based Deception Models for Native and Non-Native English Cybercriminal Networks,” in *Proceedings of the International Conference on Compute and Data Analysis*, New York, NY, USA, 2017, pp. 23–27.

[10] A. Mbaziira and J. Jones, “A Text-based Deception Detection Model for Cybercrime,” *Int. Conf. Technol. Manag.*, Jul. 2016.

[11] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker, “A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication,” *J. Manag. Inf. Syst.*, vol. 20, no. 4, pp. 139–165, 2004.

[12] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, “Lying words: predicting deception from linguistic styles,” *Pers. Soc. Psychol. Bull.*, vol. 29, no. 5, pp. 665–675, May 2003.

[13] Y. R. Tausczik and J. W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods,” *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[14] Y. Ren and D. Ji, “Neural networks for deceptive opinion spam detection: An empirical study,” *Inf. Sci.*, vol. 385–386, no. Supplement C, pp. 213–224, Apr. 2017.

[15] V. W. Feng and G. Hirst, “Detecting Deceptive Opinions with Profile Compatibility.,” in *International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013, pp. 338–346.

[16] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety,” in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012, pp. 71–80.

[17] K. Reynolds, A. Kontostathis, and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” in *2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, 2011, vol. 2, pp. 241–244.

[18] X. Chen, R. Chandramouli, and K. P. Subbalakshmi, “Scam detection in Twitter,” in *Data Mining for Service*, Springer, 2014, pp. 133–150.

[19] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. McCoy, “Constructing and Analyzing Criminal Networks,” *IEEE Secur. Priv. Workshop*, p. 8, 2014.

[20] W. Cohen, “Enron Email Dataset,” 08-May-2015. [Online]. Available: http://www.cs.cmu.edu/~enron/. [Accessed: 29-Mar-2016].

[21] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding Deceptive Opinion Spam by Any Stretch of the Imagination,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Stroudsburg, PA, USA, 2011, pp. 309–319.

[22] DoJ, “Former Enron CEO Jeffrey Skilling Resentenced to 168 Months for Fraud, Conspiracy Charges,” 21-Jun-2013. [Online]. Available: https://www.justice.gov/opa/pr/former-enron-ceo-jeffrey-skilling-resentenced-168-months-fraud-conspiracy-charges. [Accessed: 02-Apr-2017].

[23] E. Fitzpatrick, J. Bachenko, and T. Fornaciari, “Automatic Detection of Verbal Deception,” *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 3, pp. 1–119, Sep. 2015.