Cybercriminal Personality Detection Through Machine Learning

Saravanan Sagadevan[1], Mohd Baqir Hakim[1], Nurul Izzati Ridzuwan[1],
Nurul Hashimah Ahamed Hassain Malim
[1]School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
nurulhashimah@usm.my

INTRODUCTION

Generally, the severe kinds of cyber criminal activities such as cyber bullying are executed through exploiting text messages and the anonymity offered by social networks such as Facebook and Twitter. However, linguistic cues such as patterns of writing and expression in the text messages often act as fingerprints in revealing the personality traits of the culprits who hide behind the anonymity provided by social networks[1]. Personality is referred to "combination of emotion, behavior, motivation and thinking patterns of human that often mirror the true characteristics of them through their activities that are conducted intentionally or unintentionally" [2]. In nature, each individual is different in terms of their talking and writing styles. The distinct styles of talking and writing are unique and there is a tendency to decipher the identity of writers by simply observing their patterns of writing especially the formation of words, phrases and clauses. Sir Francis Galton was identified as the first person who hypothesized natural language terms that might present the personality differences in humankind [3]. Furthermore, Hofstee suggested that nouns, sentences, and actions may have some kind of connotations towards personality [4].

Scholars from forensic psychology, behavioral sciences and the law enforcement agencies have also been working together to study and integrate the science of psychology to criminal profiling [5]. Through the review of literature related to psychology, linguistics and behavior, it can be affirmed that strong relationship is present between personality traits especially in relation to criminals and writing/language skills. Therefore, a curiosity was raised on whether the writing pattern in social networks by cyber criminals could be identified or detected by using automatic classifiers. If yes, how much better would be the performances of the classifiers be and what are the words or combination of words that may be frequently used by cyber predators?

Therefore, in order to find the answers to those questions, we conducted an empirical investigation [9] (main study) with two other small scale studies [10,11] (extension the main study) by using textual sources from Facebook and Twitter and exploiting the descriptions stated in Three Factor Personality Model, and sentiment valences. Three Factor Personality Model consists of three global traits namely Extraversion , Neuroticism, and Psychoticism where the third trait is commonly associated with criminal behavior [8].

In the main study [9], the open source data Facebook [6] and Twitter [7] were used as text input while the data for the other two small scale studies were harvested from Twitter using Tweepy, a Python library for accessing the Twitter API. The main study [9] and the second study [10] used data that was only written in English language while the third study [11] used tweets in Malay Language (Bahasa Malaysia).

In these studies, we employed four main classifiers namely Sequential Minimal Optimization (SMO), Naive Bayes (NB), K-Nearest Neighbor (KNN) and J48 with ZeroR as baseline from Waikato Environment for Knowledge Analysis (WEKA) Machine Learning Tool. The reasons for using the traits from Three Factor Model in this study are due to the widespread use of the model in criminology, less number of traits easing the characteristics categorization process and large number of empirical evidence that associate Psychoticism trait with criminal characteristics whereas sentiment valences was used to measure the polarity of sentiment terms.

LITERATURE REVIEW

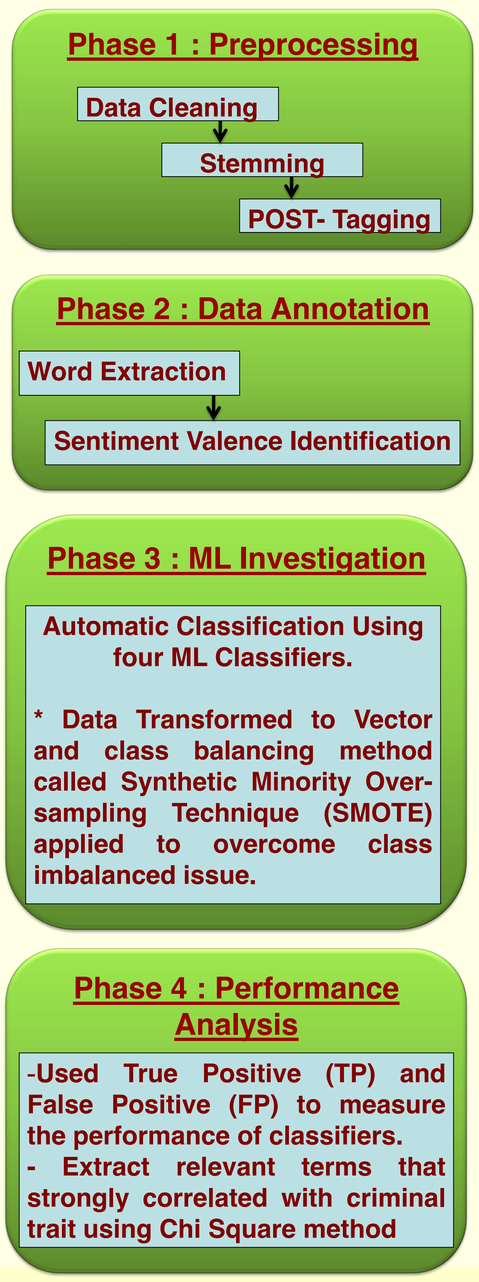
As pioneers of personality detection study using digital texts, Argamon and his colleagues[12] investigated the personalities of undergraduate students in the context of Neuroticism and Extraversion by analyzing functional lexical and appraisal related features from essays written by the participants. The experiments were conducted using Sequential Minimal Optimization (SMO) classifier from Waikato Environment for Knowledge Analysis (WEKA). Through linear kernel and 10-fold cross validation, they found that functional lexical features assisted the identification of Neuroticism but were unsupportive of the detection of Extraversion. Another study [13] investigated the effects of classification levels and language models towards the accuracy of personality classification through Weblog data and claimed that adding classification instances from multiple classes made the automatic classification process harder.

Throughout the literature review, none of the digital investigation studies that examined the relationship between linguistics and personalities of cyber criminal suspects were found. However, there are some related works that considered the linguistics used by criminals in digital conversations. For instance, the study of [14] proposed a method to cluster anonymous emails based on stylistic features and unique styles of writing using Expectation Maximization (EM), K-Means, and Bisecting K-Means algorithms. The study found that Bisecting K-Means is more scalable than EM and K-Means.

METHODOLOGY

The three studies [9,10,11] used similar research framework as the following:- Phase 1 : Data Collection & Preprocessing (Data Cleansing, Stemming, Part-Of-Speech Tagging), Phase 2 : Data Annotations, Phase 3 : Machine Learning (ML) Investigation (Automatic Classification by the four Classifiers), Phase 4 : Performance analysis, criminal related terms identification (using Chi-Square method). The Figure 1 illustrates the research methodology of this study.

Figure 1 : Research Methodology



RESULTS

Table 1 : Performance of classifiers based on measuring the effect of class measuring[11].

Performance measurement based on True Positive (TP)and False Positive (FP)						
Cross Validation	3		5		10	
Classifier	TP	FP	TP	FP	TP	FP
ZeroR	53.3	46.7	53.3	46.7	53.3	46.7
NB	80.0	20.0	90.0	10.0	90.0	10.0
KNN	63.3	36.7	56.7	43.3	56.7	43.3
SMO	73.3	26.7	70.0	30.0	86.3	16.7
J48	50.0	50.0	63.3	36.7	70.0	30.0

Table 2 : Performance of classifiers based on with/without SMOTE class balancing methods[10].

Performance measurement based on True Positive (TP)and False Positive (FP)				
	Without SMOTE		With SMOTE	
Classifier	TP	FP	TP	FP
ZeroR	47.27	52.73	40.63	59.38
NB	58.18	41.82	68.75	31.25
KNN	47.27	52.73	53.13	46.88
SMO	72.73	27.27	73.44	26.56
J48	78.18	21.82	75.00	25.00

Table 3 : Terms from Facebook that highly associated with criminal behavior [9].

Facebook		
Unigram	Bigram	Trigram
Damn	The hell	I want to
Shit	Damn it	Damn it I
Fuck	Hell i	Is a bitch
Hell	My Fuck	What the Fuck
Ass	The shit	What the hell
Suck	Damn you	I feel like
Bad	The fuck	The hell I
Feel	A bitch	
Hate	Fuck yeah	

Table 4 : Terms from Twitter that highly associated with criminal behavior [9].

Twitter		
Unigram	Bigram	Trigram
Suck	Damn It	A big ass
Adore	The hell	A bit more
Annoy	A bitch	A bitch and
Asshole	A damn	A damn good
Shit	A fuck	All fuck up
Fuck	A hell	A great fuck
Hell	Damn you	A great night
Cute	A shit	A pain in
Damn	My ass	A fuck off

Table 5 : Additional findings gathered from the study [9].

Additional Findings	
Binary Classification VS Multi Classification	Binary classification yielded better classification results than multi-classification.
Language Models	Classification using unigram language model yielded better classification results in majority of classification run.
Creativity	There is the possibility that cyber criminals writings may contains creativity elements due to the present of creativity terms on Psychoticism instances.

DISCUSSION

Several issues have been identify in the end of the study. First, it is essential to detect the entities and their relationships in the conversation/text so that the effectiveness of class labeling in the supervised classification could be increased. In other words, it is necessary to identify the frame semantics prior to the class labeling process. Moreover, rather than solely depending on the lexical, it is necessary to consider the effects of other strings such as punctuations, emoticons, text cases, and et-cetera in the classification process. Nowadays, these types of strings are frequently used to show feelings or emotions that can be used to identify the characteristics of people especially cyber criminals. There are some studies claimed that cyber terrorists often used these types of emotional words to influence ordinary people. Therefore, there is a possibility to extend this study to detect cyber terrorists based on the words they use to recruit innocent people.

CONCLUSIONS

Eventually, our investigation showed that J48 performed better than other classifiers with and without the application of SMOTE class balancing technique and the effect of cross validation vary for each classifier. However, in an overall view, Naïve Bayes performed better on each cross validation experiments. This investigation also produced a list of the words that could be used by cyber criminals based on analysis on language models using Chi Square. For future study, we plan to use deep learning methods to analyze the contents related to cyber terrorism and welcome collaborations particularly in the areas of financial and social networks cyber terrorism textual data investigation.

REFERENCES

1. Olivia Goldhill.. Digital detectives: solving crimes through Twitter, 2013. *The Telegraph.*
2.Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning based document modeling for personality detection from text. *IEEE Intelligent Systems* 32(2):74–79.
3. Sapir, Edward. Language: An Introduction to the Study of Speech. New York: Harcourt, Brace, 1921.
4..Matthews, G, Ian, J. D., & Martha, C. W. Personality Traits (2nd edition). *Cambridge University Press*, 2003.
5.Gierowski, J. K. Podstawowa problematyka psychologiczna w procesie karnym. Psychologia w postępowaniu karnym, Lexis Nexis, Warszawa 2010.
6.Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. Workshop on Computational Personality Recognition (Shared Task). *In Proceedings of WCPR13, in conjunction with ICWSM-2013*.
7.Alec, G., Richa, B, & Lei, H.. Twitter Sentiment Classification using Distant Supervision, 2009.
8.Coleta, V. D, Jan, M. A., Janssens, M., & Eric E. J. PEN, Big Five, juvenile delinquency and criminal recidivism. *Personality and Individual Differences, 39, (2005) 7–19.* DOI:10.1016/j.paid.2004.06.016.
9.Saravanan Sagadevan. Thesis : Comparison Of Machine Learning Algorithms For Personality Detection In Online Social Networking, 2017.
10. Muhd, Baqir. Profiling Online Social Network (OSN) User Using PEN Model and Dark Triad Based on English Text Using Machine Learning Algorithm, 2017 (In Review).
11.Nurul Izzati Binti Ridzuwan. Online Social Network User-Level Personality Profiling Using Pen Model Based On Malay Text (In Review), 2017.
12.Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W. Pennebaker. 2005. Lexical predictors of personality type. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America.
13. Oberlander, J., and Nowson, S. Whose thumb is it anyway? (2006). Classifying author personality from weblog text. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics ACL. pp. 627–634.
14. Farkhund Iqbal, H. B., Benjamin C.M. Fung, Mourad Debbabi. (2010). Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1-2), 56-64