Pertanika J. Sci. & Technol. 27 (4): 1803 - 1820 (2019)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Criminality Recognition using Machine Learning on Malay Language Tweets

Nurul Hashimah Ahamed Hassain Malim*, Saravanan Sagadevan and Nurul Izzati Ridzuwan

School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia

ABSTRACT

A large scale of investigation had been carried out to predict the personality, or in precise, the behaviour of online users through user-generated texts, such as Tweets and status messages. Nevertheless, only a handful of machine learning (ML) studies have applied the personality model to assess criminality behaviour, particularly within the context of Malay social network messages. Based on the concept of sentiment valence, this study annotated a list of Malay Tweets that might be subjected to crime or illicit messages from the stance of Psychoticism trait. Consequently, the supervised-based text classification method was conducted by using Naïve Bayes (NB), Sequential Minimal Optimisation (SMO), and Decision Tree (DT) on Tweets using several features determined via Chi Square (x^2). The analyses outcomes signified that SMO outperformed other classifiers insignificantly by achieving 92.85% of accuracy. Based on x^2, several swear terms, such as *bontot*, *melancap*, and *kote*, displayed significant correlation with Psychoticism Tweets due to the nature of the trait that has been subjected to criminality behaviour, for instance, aggressive and antisocial attributes. The findings illustrate the possibilities to adapt several personality aspects in order to enhance the effectiveness in detecting illicit social network messages.

Keywords: Machine learning, Malay tweets, personality detection, text mining

ARTICLE INFO

Article history:
Received: 07 February 2019
Accepted: 04 June 2019
Published: 21 October 2019

E-mail addresses:
nurulhashimah@usm.my (Nurul Hashimah Ahamed Hassain Malim)
saravanan_18@student.usm.my (Saravanan Sagadevan)
izzati.ucom13@student.usm.my (Nurul Izzati Ridzuwan)
* Corresponding author

INTRODUCTION

In earlier time, Sir Francis Galton had hypothesised that natural language terms might represent personality differences in humankind, while Allport et al. (1936) claimed that almost 18,000 English terms could represent personality, and Hofstee (1990) suggested that nouns, sentences,

ISSN: 0128-7680
e-ISSN: 2231-8526

and actions might mirror one's personality (Gerald et al., 2003). At present, the vast participation of online users in social network activities has contributed to unprecedented textual resources, such as tweets that are rich in personal discourses. For example, Twitter has become a source that generates a substantial amount of user data, wherein daily active users of the platform reached 157 million as of second quarter of 2017 and approximately 500 million tweets were shared each day among users (Kursuncu et al., 2018). The massive number of user activities on social networks offers valuable insights about one's behaviours, experiences, opinions, and interests. The insights are fundamentally directed via psychology, or in specific, personality that advocates both emotional aspects and characteristics of a human being.

As the domain of data science is expanding every passing day, interest towards recognising and understanding personality traits of virtual users via user-generated content, especially from text data, has grown rapidly in recent years. Since the study by Argamon et al., (2005) that used functional lexical features from student essays to automatically predict Extraversion and Neuroticism traits, many empirical investigations have been conducted to reckon human traits from digital data. Meanwhile, Oberlander and Nowson (2006) opened a new chapter in evaluating the personality amidst social media users through lexical analysis. The investigation revealed that the combination of attributes that derived from language models yielded good outcomes in classifying the personality of weblog users. Subsequently, many scholars have begun exploring the latent representation of personality characteristics embedded in texts.

One of the most prominent efforts to evaluate the personality of online users have been initiated by using *myPersonality* dataset and a common benchmark in a Workshop on Computational Personality Recognition (Shared Task) (Celli et al., 2013). The workshop refers to gold standard that evaluates the performance of Machine Learning (ML) algorithms using Big Five Personality traits, namely Openness to Experience, Conscientiousness, Extroversion, Neuroticism, and Agreeableness, to represent the nature of human being characteristics. The empirical review displayed that many personality recognition studies have embraced the description defined in Big Five model. For instance, based on the Big Five model, Peng et al., (2015) employed the Chi Square and Recursive Feature Elimination (RFE) selection methods to detect personality from Chinese texts, while Aalderks (2014) performed Latent Semantic Analysis to determine significant features so as to identify non-cognitive personality in post-secondary student essays. Shally (2014) substantially investigated the online behaviours of Facebook and LinkedIn users using Big Five traits, while Correa et al., (2010) and Guadagno et al., (2008) reported that users who scored highly on Neuroticism frequently used more social network platforms, such as Facebook and Twitter.

Additionally, several studies (Gerald et al., 2003; Oberlander & Nowson, 2006; Peng et al., 2015) have presented evidence of strong inter-correlation across personality-behaviour-linguistics areas. The strong inter-correlation among these areas should be further explored by incorporating relevant domain knowledge into other relevant aspects, such as linguistics-based criminality recognition. Criminality writing recognition within the context of personality may significantly assist the detection of language-based cybercrime activities, such as cyber bully and cyber harassment (Farshad et al., 2016). Nevertheless, only a handful of studies have adapted other personality models or have explored the dynamics of text messages from the perspective of crime.

Although no investigation has explored the correlation between crime and texts in the context of personality, several studies (Sagadevan et al., 2015; Saravanan, 2016) have assessed the representation of Psychoticism trait from the Psychoticism, Extraversion, and Neuroticism (PEN) Model, which is typically aggregated to criminal behaviours (Kamaluddin et al., 2015). As Psychoticism trait represents certain behaviours, such as aggressiveness and interpersonal hostility that are naturally linked with negativity and crimes (Kamaluddin et al., 2015), the integration of PEN framework with other pipeline concepts, such as sentiment valences and Automatic Personality Perception (APP), sheds light on the nature of criminality digital writings. The description of PEN Model traits and their specific characteristics are presented in Table 1.

Table 1
PEN dimensions (Allport, 1961)

Trait	Characteristics
Extraversion	Sociable, lively, active, assertive, sensation seeking, carefree, dominant, surgent and venturesome
Neuroticism	Anxious, depressed, guilt feelings, low self-esteem, tense, irrational, shy, moody and emotional
Psychoticism	Aggressive, cold, egocentric, impersonal, impulsive, antisocial, unempathetic, creative and tough-minded

Past studies have probed into the availability of open source English text messages from Facebook (Celli et al., 2013) and Twitter (Go et al., 2009) to assess proximal cues, which aggregated to a well-known criminal trait called Psychoticism by cross-validating the personality description provided in the psychological self-assessment report (derived from Facebook). Initially, Part-Of-Speech (POS) Tagging was implemented on cleaned datasets to identify adjectives, nouns, and verbs that brought sentiment polarity, wherein those modalities served as subject of study in the public questionnaire (Sagadevan et al., 2015) to gather general perception regarding sentiment valences and their association with PEN Model traits. The execution of questionnaire to identify the effects of words on

personality trait is a perception-based approach that falls under APP, which can determine the co-relationships between sentiment valences of a term and personality traits.

The APP refers to an Automatic Personality Detection approach that emphasises on predicting personality attributes based on observable behaviours, such as types of words used in writings. Albeit the prediction is based on other observations and perceptions of the public in influencing the personality in social interactions (Mohammadi & Vinciarelli, 2012), it is the predisposition of humans to comprehend the behaviours of others based on the observation of their behaviours in everyday life (Agarwal, 2014).

Valence refers to one of the Dimensional Views proposed by Osgood et al., (1957), which can be applied to measure the level of pleasantness and unpleasantness of each word (Bradley & Lang, 1999). Subsequently, the pre-processed dataset annotated using frequency (Schwartz, 2013) of sentiment words was categorised based on valences schema (Table 2) and applied feature selection prior to transformation into vectors that represented stream of strings. Feature selection is a method that reduces data dimensionality by selecting a subset of attributes from a bigger pool of inputs to devise a prediction system that can preserve the original information as much as possible.

Based on supervised learning, many ML classification experiments have been conducted on multiple types of language models using several prominent algorithms, such as Sequential Minimal Optimisation (SMO), Naïve Bayes (NB), K-Nearest Neighbour (KNN), and Decision Tree (DT). Saravanan (2016) reported that SMO and Unigram language model yielded promising outcomes, when compared to other classifiers, and forwarded suggestions that quadratic-based optimisation classifiers could perform well on high dimensional inputs and single-text attributes might channel valuable information towards automatic learning process.

To the best of the authors' knowledge, no study has used the Malay social media text messages to infer personality traits of users, especially within the context of criminality behaviours. Most of the past empirical studies that used Malay texts placed more focus on the typical linguistics domains, such as sentiment analysis (Chekima & Alfred, 2018; Al-Saffar et al., 2018; Al-Moslmi et al., 2017; Darwich et al., 2016) and machine translation (Wang et al., 2015).

Since the earlier study conducted by Saravanan (2016), which focused on English social network messages, yielded promising findings, the efforts of this present study are extended to study the representation of Malay Tweets in the context of Psychoticism and to measure the performance of ML algorithms in predicting traits-based instances. In this study, the authors assumed that perceptions towards semantic of lexical, which had been typically linked with sentiment polarity, might serve as an indicator to classify the sentences based on PEN Model traits. The following sections elaborate the methods, the outcomes, and the discussions of this study.

MATERIALS AND METHODS

The methodology in this study is composed of data collection and several pre-processing techniques, such as data cleaning, annotations, and eventually, automatic classification by three ML algorithms (NB, SMO & DT). Figure 1 illustrates the methodology of this study.

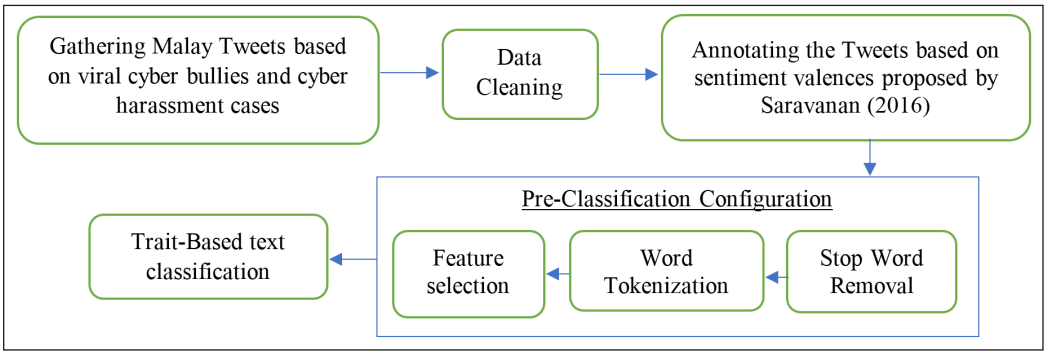

Figure 1. Methodology

Prior to data collection, the trending topics on Twitter that turned viral due to harassment or bully of other Twitter members were observed. Next, the types of words used by the account owners were randomly observed and studied. Since the lexicon resources for Malay language is limited and most of them have been readily available off-the-shelf classifiers designed for English texts (Darwich et al., 2016), the seed words applied in a previous study were translated to the Malay language and served as hints to identify the possible true cyber bullies and cyber harassment. The seed words and their sentiment valences are presented in Figure 2. Approximately one month was taken by the researchers to assess the

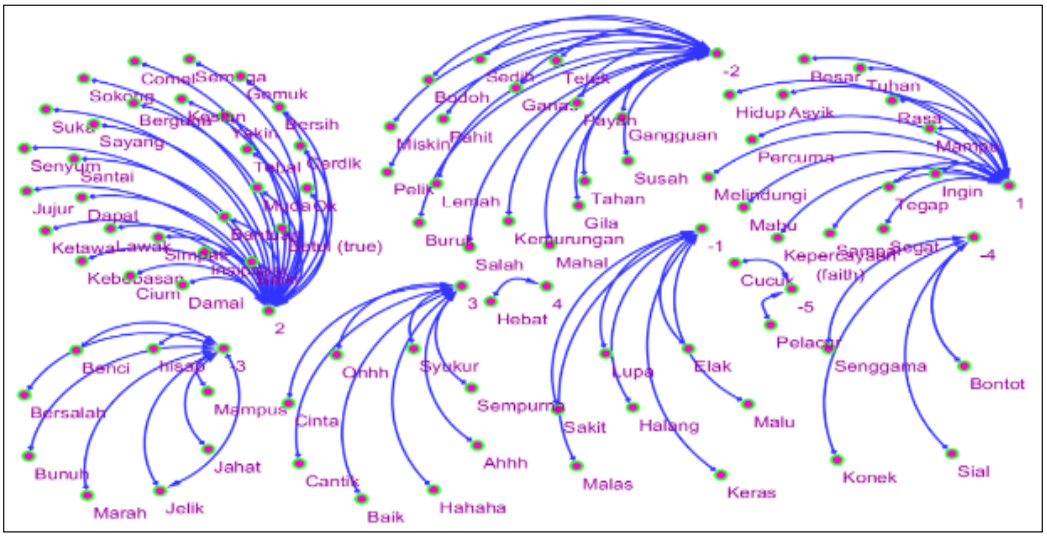

Figure 2. Sentiment valences-based seed words

messages posted in the selected accounts before analysing their tweets using *Tweepy* and labelling them in accordance to the schema proposed by Saravanan (2016). As for the non-criminal message corpuses, the researchers randomly crawled tweets from other trending topics, such as fashion, business, and news, apart from inspecting the context of the data in order to ensure that they were excluded from harassment and bully relationship. Due to time constrain, the data crawling merely yielded small scales of corpus that comprised 82,293 words, in which each instance consisted of about 2,939 words in average.

Due to the nature of the social networking tweets that is in the form of unstructured and unformatted (Salloum et al., 2017), data cleaning process was performed to normalise the structure of the tweets to meet standard forms. As the amount of the collected data had been small, the data cleaning process was manually conducted and particularly discarded meaningless strings, such as Uniform Resource Locator (URL), punctuation marks, symbols, numbers, and word standardisation from malleable to regulated forms (mls to malas). As a result, the data were annotated using the schema proposed by Saravanan (2016). Based on the schema, first, the words in the dataset were labelled as either positive or negative, and next, the words were matched with their sentiment by adhering to the rules presented in Table 2.

Table 2
Sentiment/valence categorisation

Trait	Sentiment	Valence
Extraversion	Positive	1 to 5
Neuroticism	Lower Negative	(-1) to (-3)
Psychoticism	Higher Negative	(4) to (-5)

Based on the PEN model, the Extraversion trait is significantly aligned to positive behaviours, whereas Neuroticism and Psychoticism lean towards negative characteristics (Kamaluddin et al., 2015). By adhering to this concept, all the sentiment terms were ranked based on frequency and were categorised according to the PEN model traits in identifying the sentiment strength of the collected tweets. As the term frequency method is a prominent technique applied in text mining to determine sentiment intensity of textual streams (Schwartz et al., 2013), this present study embraced the statistical approach to calculate both the intensity and the tendency of the users towards the relevant PEN model traits. As the goal of this study is to evaluate the representation of Psychoticism trait in Malay tweets, the term frequency ranking process was initiated from higher negative to positive as the nature of the online texts is more positive and negative, as well as the fact that the latter carry more information than the former (Garcia et al., 2012). Subsequently, the average word usage of each trait category was determined by using the formula given

in (1). The average word counting method (1) is a common technique employed to bridge the connection between language and psychological variables that belongs to manually-constructed classes of language (Schwartz, 2013).

$$Average\ word\ usage = \frac{Number\ of\ trait\ related\ words}{Total\ number\ of\ users\ in\ each\ traits} \quad (1)$$

Although an earlier study have reported that as minimum as five higher negative words usage in average matched the cross-check correlation between Psychoticism and Agreeableness + Neuroticism (Van Dam et al., 2005), this study revealed that the average usage of higher negative words was seven. Nonetheless, the instances were only labelled as Psychoticism if the document consisted as minimum as 14 higher negative words, mainly due to the small volume of dataset. The stipulated figure served as a guide to represent the data as structurally compact and to offer adequate volume of Psychoticism instances so as to hinder overfitting problem during ML classification. The annotation for Neuroticism instances adhered to formula (1), while the average Neuroticism words usage was set to 48. As the PEN Model merely consists of three global traits, the other instances were assumed to be more adjacent to positive polarity and were annotated as Extraversion. Eventually, the Psychoticism, Neuroticism and Extraversion classes contained 8696, 36,573, 37,024 number of words respectively. The examples of the Tweets representing each of the traits illustrated in Table 3.

Table 3
Example of tweets each classes of trait

Class	Example of Tweets
Psychoticism	*Thanks* gemok pun dah tunang Hang Korek puki pakai jari ja ka tiap malam
Neuroticism	Susah jaga orang tua hampa tau ka
Extraversion	*Dear students* Belajar ibarat naik bukit Bila kau dah sampai puncak kau akan nampak permandangan yg sgt cantik Jangan mudah putus asa

The instances from the three traits were represented as a set of document *d,* wherein each d_i signified a sequence of sentence *s* and was classified based multi-classes classification, primarily because the PEN model is composed of three global traits. The training document *d* was transformed into vectors by using the Bag of Words (BOW) technique. In precise, the string to vector transformation involved removal of stop words, such as *ini, itu, sana, dari*, and *yang*, while string tokenisation was conducted on unigram attributes because the scale of the present dataset was small, along with the high possibilities of individual variances to articulate using single terms (Mairesse & Walker, 2011). After that, in the attempt of minimising the dimensionality and to enhance the detection performance,

several relevant significant attributes were selected based on Information Gain (IG). The IG refers to a function of probability distribution initiated from communication theory based on entropy mechanism that measures the uncertainty of random attributes, such as words in textual stream (Mitchell, 1997). The entropy measures the expected reduction caused by partitioning the samples based on the attributes. Simply put, IG calculates the number of bits of information gathered to predict classes by estimating the presence or absence of certain *words* in a document (Gao et al., 2014). In text mining, Entropy measures the information or knowledge encapsulated in vectors mapped from text contents. The lack of information or high entropy value will lead to poor prediction whereas low entropy will assists better prediction process. The IG for a *word (w)* is defined based on (2). Next, automatic classification was performed by using the algorithms stated in the following section based on 10-fold cross-validation resampling procedure due to the limited number of training instances (Refaeilzadeh et al, 2009).

$$IG: (D, w) = Entropy\ (D) - Entropy\ (D|w)\ ,$$

$$Entropy\ (D) = \sum_{c \in C} -p(w)\ log_2\ p\ (w) \qquad (2)$$

where, *D* refers to the training document where *w* is the set of all possible attributes (words).

P(w) is a word that is present in the instances that belongs to class $c \in C$.

Machine Learning Algorithms

This study measured the performances of three selected algorithms, namely NB, SMO, and DT, against the majority baseline. The three off-the-shelf algorithms were chosen in this study, since many text mining studies had employed these classifiers in their analyses (Nasa & Suman, 2012; Sujatha & Ezhilmaran, 2013; Kapur et al., 2017). These off-the-shelf algorithms are suitable for automatic classification with small volume of data (Ruparel et al., 2013). The following sub-sections describe each selected machine classifier.

Baseline

After taking example from a prior study (Celli et al., 2013), this present study employed a classifier called Zero Rules (ZeroR) as a majority baseline to measure the performances of each ML classifier applied in this study. The ZeroR predicts the mean (for numeric-type target attribute) or the mode (for nominal-type attribute) of the most commonly found attributes in the datasets and does not apply any rule that works on non-target attributes (Barber, 2012).

Naïve Bayes (NB)

The NB is a very popular probabilistic classifier based on Bayes rules with strong independent assumptions. In other word, a descriptive independent feature model» based on probability will make NB to assume that the presence or absence of a peculiar feature of a class is not related to the presence or absence of other features (Witten et al., 2011). The formula of NB illustrated in (3). Based on the model, the *P(c)* is a prior probability that initially set to *c* whereas *P(d)* indicated the initial probability of the training document *d*. *P(d|c)* referred to the appearance probability of document *d* when *c* is established. The *P(c|d)* is a calculation of posterior probability that represents the confidence of presence of *c* on *d.* As nature of text documents contains multiple attributed transformed through BOW that made the independent assumption on word orders, the *d* represented as word features $x_1, x_2 \ldots x_n$ and denoted as conditional independence $P(x_1,...x_n|c) = P(x_1|c) \cdot P(x_2|c) \cdot .....P(x_n|c)$. Eventually, NB made the decisions of the classes belongs to particular instances based on fractions of times words x_i appears among other *x* in *d* by estimating the maximum likelihood of *c* (Witten et al., 2011).

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \qquad (3)$$

*Where *c* referred to training classes and *d* is training document.

Sequential Minimal Optimisation (SMO)

SMO is a variant of SVM that optimize a problem iteratively by splitting the problem into a series of sub-problems (Huang & Yan, 2014). The selection of smallest optimization problem involved two Lagrange Multipliers where both must follow the equality constraints and jointly optimized to find the optimal values as well as update the vector machine to reflect the new optimal values. This process was repeated until the convergence criteria was met. To speed up the convergence, SMO use heuristics to select both Lagrange Multipliers to jointly optimize the problem. The problem has been solved when all the Lagrange Multipliers satisfy the Karush–Kuhn–Tucker (KKT) conditions (Huang & Yan, 2014). The working procedure of SMO illustrated in Figure 3. In the first place, the selection of dual Lagrange Multipliers m_1 and m_2 were based on heuristic methods. The first heuristic provides the outer loop to iterate over the entire training set to determine the violations of KKT conditions by each instance. The determination evaluated based on computations of m_2's upper bound, *u,* and lower bound. *l.* If any of the instances violate the KKT condition, then it is eligible for optimization. The second heuristics intended to support another multiplier to maximize the size of the step taken during joint optimization. Consequently, the m_2 need to be update based on Δm_2. The updating process fails if the Δm_2 is smaller than threshold, otherwise, m_1 should be update along with all of the objective functions, f_i, values.

Further, computation conducted to determine the deviation between the output of function and classification target, *t*. SMO will terminate if the value of *t* below than threshold. The pair-wise classification mechanism was applied in this study as the training data consisted of instances from three classes (multi-class classification). In order to prevent inter-class generalisation problem, Polynomial kernel functions were applied due to the popularity of the technique in NLP (Goldberg & Elhadad, 2008).

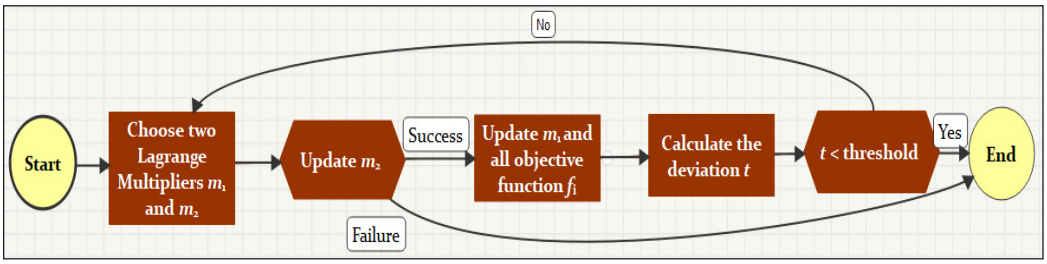

Figure 3. SMO training process

Decision Tree (DT)

Theoretically, DT refers to a tree-based classifier, wherein each branch of nodes denotes a selection between a list of alternatives and each leaf depicting the decisions (Dunham, 2006). In precise, the foundation of the classifier is to determine the structure of the vectored-attributes behaviour for several instances in the classes and newly-generated instances (Korting, 2006). In particular, one of the DT classifiers called J48 was employed to predict the classes in this study. The J48 adhered to the implementation of C4.5 classifier and incorporated several additional functions, such as tackling missing values, DT pruning, and rules derivation (Kaur & Chhabra, 2014). Generally, DT algorithm follows the following steps to classify the instances.

(a) First, J48 develops a DT based on the attribute values of the available training data or from the 10-fold cross-validation.
(b) Second, J48 identifies the attributes that discriminate the various instances with the highest information gain.
(c) Third, the non-ambiguous attributes branches terminate and assign the attributes to the target value.

Evaluation Metric

The accuracy evaluation metrics were applied to measure the performance of ML algorithms in this study. Accuracy is a widely applied evaluation metrics in ML classification (Hossin & Sulaiman, 2015). Typically, ML studies use statistical metrics called confusion matrix to measure the strength of an algorithm in solving the given problem automatically. The confusion matrix comprised four elements of conditions, namely True Positive (TP),

False Positive (FP), True Negative (TN), and False Negative (FN). The TP refers to the proportion of positive cases that are correctly identified, FP is the proportion of negative cases that are incorrectly classified as positive, while TN is the proportion of negative cases that are classified correctly, and FN is the proportion of positive cases that are incorrectly classified as negative (Hossin & Sulaiman, 2015). The accuracy of evaluation metrics is determined based on the mathematical operation on the elements of confusion matrix, as illustrated in formula (4).

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

RESULTS AND DISCUSSION

Based on the methodology described, this study investigated the performance of ML algorithms using supervised learning mechanism and assessed the existing correlation between Neuroticism and Psychoticism, apart from extracting several statistically significant terms linked with Psychoticism trait. The following subsections present the findings of each investigation.

Machine Learning Evaluation

Three ML classifiers, namely SMO, NB and DT, had been employed in this study to evaluate the prediction accuracy on the supervised-based constructed inputs that represented the PEN model traits instances. Based on the outcomes displayed in Figure 4, all the classifiers performed better than the baseline did, with SMO leading the prediction and followed by NB and DT. The outperformance exhibited by SMO is in agreement with the findings reported in previous studies concerning personality recognition (Verhoeven et al., 2013; Saravanan, 2016), which exploited the representation of English texts. Basically, SMO learnt the representation via analytic quadratic programming that decomposed the overall inputs into sub-problems and iteratively optimised the smaller units using Lagrange multipliers until the problem was solved. The goal of SMO is to forward the Lagrange multipliers or alphas that satisfy the actual inherent learning process by identifying the support vectors (Platt, 1998), whereas the role of kernel function is to transform the inputs into a higher dimension and to optimise the sub-problems until convergence prevents large matrix computation (Urmaliya & Singhai, 2013). Hindering computation on large matrix enables SMO to exploit massive volume of sparse text dataset that contains a huge number of zero vector elements.

Meanwhile, NB significantly performed well in this present study, when compared to that reported by Saravanan (2016) that predicted below the baseline upon use of larger instances as the input. The significant prediction displayed by NB is attributable to the

small volume of the instances applied in this study, which adjusted the weights for decision boundary, minimised the bias effects, and normalised the magnitude of the weights for strong and weak word dependency classes (Rennie et al., 2003). The exceptional performance exerted by NB contradicts that stated by Raschka (2014), who denoted that the probability-based classifier tended to perform poorly on non-linear classification problems. The significant performance signifies that NB can perform well on non-linear problems if the volume of the training instances can normalise the weights and when the magnitude of the weights is assigned to the decision boundary and the attributes of the classes.

The DT classifier merely achieved 70.96% of accuracy. As the nature of DT is greedy, each separation in the tree was determined based on isolation without considering the possible biases in future tree that may poorly capture the underlying characteristics of training instances. The poor pruning mechanism due to small volume of dataset may cause weak generalisation and potentially directed to underperformance in predicting future points (Bertsimas & Dunn, 2017).

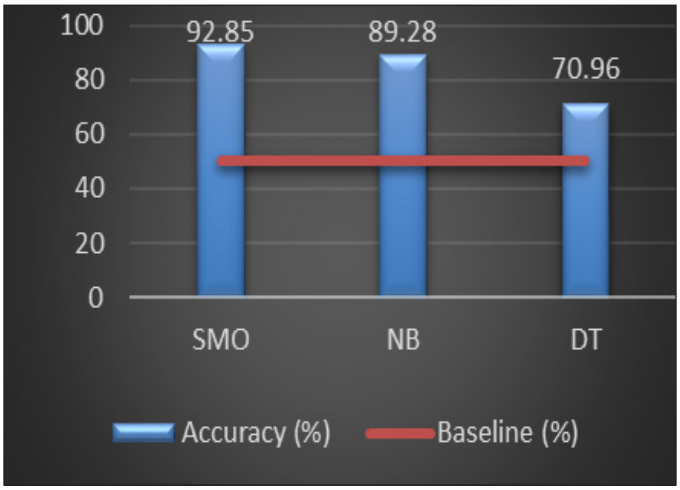

Figure 4. Machine learning classification result

Projection Plot

The analysis was extended to determine the correlations present between Psychoticism, Neuroticism, and Extraversion instances. Projecting the instances in the form of plot graph illustrated an interesting pattern, whereby strong correlation was discovered between the language structures of Psychoticism and Neuroticism (Figure 5). To recap, although both types of respective traits characterised negative behaviours, it was detached based on higher and medium+lower negative by integrating the concept of sentiment valences. The grouping of Psychoticism and Neuroticism instances on the nearest/adjacent plotting points suggested that users from the respective categories of traits might apply similar

syntactical and semantical structures of Malay texts to describe their emotions, feelings, and opinions in a negative manner. To the best of the authors' knowledge, no study has evaluated the representation of lower and higher negative words in the Malay Language. From the stance of psychology-personality, particularly Big Five Personality Model, Psychoticism characteristics were viewed as a trait that strongly correlated to Neuroticism and Agreeableness (Van Dam et al., 2005). Although general psychologists have employed factor analysis to determine the relationships between the traits, the strong correlation between Psychoticism and Neuroticism illustrated in plotting of the text instances seems to be in line with that reported by Van Dam et al., (2015). Within the context of text mining, the evidence retrieved from the plotting indicates that FP and FN classification errors may be minimised if personality recognition mechanisms are embedded into the criminality text detection model.

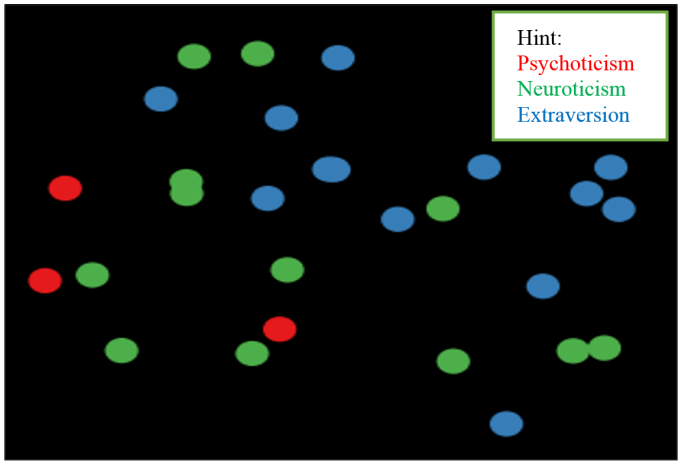

Figure 5. Plotting of PEN model instances

Significant Terms Extraction

Further analysis was performed specifically on Psychoticism instances to seek the nature of writing structures that advocated higher negativity. The Chi Square (X^2) feature identification method was applied to extract the significant terms correlated to Psychoticism. The analysis was conducted from single to triple language models. The highly correlated language models are presented in Table 4, whereby most of the significant terms that featured highly negative/cursing referred to sexuality, sensuality, and sexual expressions. The use of abusive English curse words, such as *fuck* and *shit*, displayed bilingual cursing expressions. Bakar et al., (2018) found that Malaysians prevalently used bilingual or *Bahasa Rojak*, especially a mixture of Malay and English words in social network posts to express their emotions, thoughts, and feelings. Nevertheless, in the context of tweets linked with Psychoticism, the use of profanity English terms became significant and substantially

mirrored the latent characteristics of the trait, such as being aggressive and antisocial. Apart from the terms listed in Table 4, other terms, such as *(pu)kimak*, *puki*, *kafir*, *jadah*, and *sial*, were statistically insignificant, but were frequently mentioned by many users.

Table 4
Extracted twitter terms using x^2

Unigram	Bigram	Trigram
Anjing	Gila Babi	Nak Main *Game*
Babi	Anak / Makan Babi	*Fuck Off* la
Tetek	Tetek Kecil	Pakai Kondom tak
Bontot	*Fuck* Orang	*Shit* kalau aku
Melancap	Pakai Kondom	Tetek Size A/kecil
Kote	Lubang Buntut	
Cilaka	Nak Main / *Sex*	
Lancau	Pancut Dalam	
Senggama		

CONCLUSIONS

The role of personality to represent individual characteristic patterns of thoughts, emotions, and behaviour, along with psychological mechanisms, portrays an essential role in every action executed by human. Many researchers have suggested that the human language has rather strong and positive reflections towards their behaviours and intentions, especially within the context of delinquent behaviours. Based on this concept, this study annotated a list of tweets in the Malay language based on a trait that has been frequently used in criminology called Psychoticism by employing three ML algorithms to automatically predict the training instances. To the extent of the authors' knowledge, this appears to be the first study that has measured Malay illicit messages from the stance of personality. The promising accuracy displayed by the algorithms illustrates the ability of automatic classifiers to detect the presence of criminality element in texts. The findings of x^2 based on significant terms exhibited the possibilities of adapting those attributes as complimentary features to improve detection of illicit messages automatically. The analyses outcomes signified that SMO outperformed other classifiers insignificantly by achieving 92.85% of accuracy. Based on x^2, several swear terms, such as *bontot*, *melancap*, and *kote*, displayed significant correlation with Psychoticism Tweets due to the nature of the trait that has been subjected to criminality behaviour, for instance, aggressive and antisocial attributes. The Malay swear words seem more like cursing and taboo, when compared to such words in the English Language. Perhaps, this is due to the curse words in the English language that have been used commonly or the preference amongst the local people in practicing unofficial terms/language (Bahasa Selanga/Pasar), such as *tetek* or *bontot*, in which the

standard terms refer to *payu dara* and *punggung* for the respective words. In future, more data will be employed to predict the criminality texts by incorporating the concepts of Artificial Intelligence and Deep Learning.

ACKNOWLEDGEMENT

This study is funded by Universiti Sains Malaysia (USM) RUI grant (1001. PKOMP.8011035).

REFERENCES

Aalderks, D. R. (2014). *Using Latent Semantic Analysis to Detect Non-Cognitive Variables of Academic Performance* (Master Dissertation). New Jersey Institute of Technology, Newark, New Jersey.

Agarwal, B. (2014). Personality detection from text: A review. *International Journal of Computer System, 1*(1), 1-4.

Al-Moslmi, T., Omar, N., Albared, M., & Alshabi, A. (2017). Enhanced Malay sentiment analysis with an ensemble classification machine learning approach. *Journal of Engineering and Applied Sciences, 12*(20), 5226-5232.

Al-Saffar, A., Awang, S., Tao, H., Omar, N., Al-Saiagh, W., & Al-bared, M. (2018). Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLoS One, 13*(4), 1-18.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*(1), i-171.

Allport, G. W. (1961). *Pattern and growth in personality*. Oxford, England: Holt, Reinhart & Winston.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. (2005, July 24-28). Lexical predictors of personality type. In *Proceedings of Interface and the Classification Society of North America* (pp. 1-16), Cincinnati, USA.

Bakar, M. S. A., & Mazzalan, A. M. (2018). Aliran pertuturan bahasa rojak dalam kalangan pengguna facebook di Malaysia [Speech Flow of Mixed Language among Facebook Users in Malaysia]. e-*Academia Journal, 7*(1), 61-71.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge, UK: Cambridge University Press.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning, 106*(7), 1039-1082.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical Report C-1, The Centre for Research in Psychophysiology, University of Florida.

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013, July 11). Workshop on computational personality recognition: shared task. In *Seventh International AAAI Conference on Weblogs and Social Media* (pp. 1-5). Boston, USA.

Chekima, K., & Alfred, R. (2018). Sentiment analysis of Malay social media text. In R. Alfred, H. Iida, A. A. A. Ibrahim & Y. Lim Eds.), *Computational science and technology* (pp. 205-219). Singapore: Springer.

Correa, T., Hinsley, A. W., & De Zuniga, H. G. (2010). Who interacts on the Web? The intersection of users' personality and social media use. *Computers in Human Behavior, 26*(2), 247-253.

Darwich, M., Noah, S. A. M., & Omar, N. (2016). Automatically generating a sentiment lexicon for the Malay language. *Asia-Pacific Journal of Information Technology and Multimedia, 5*(1), 49-59.

Dunham, M. H. (2006). *Data Mining: Introductory and advance topic*. New Delhi, India: Pearson Education.

Farshad, R., Arefeh, K. S., & Kathayoun, M. (2016). Forensic linguistics in the light of crime investigation. *Pertanika Journal of Social Sciences and Humanities, 24*(1), 375-384.

Gao, Z., Xu, Y., Meng, F., Qi, F., & Lin, Z. (2014, May 11-14). Improved information gain-based feature selection for text categorization. In *2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace and Electronic Systems (VITAE)* (pp. 1-5). Aalborg, Denmark.

Garcia, D., Garas, A., & Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science, 1*(3), 1-12.

Gerald, M., Ian, J. D. & Martha C. W. (2003). *Personality traits* (2nd Ed.). Cambridge, UK: Cambridge University Press.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1*(12), 1-6.

Goldberg, Y., & Elhadad, M. (2008, June 16-17). splitSVM: fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In *Proceedings of ACL-08: HLT, Short Papers* (pp. 237-240). Columbus, USA.

Guadagno, R. E., Okdie, B. M., & Eno, C. A. (2008). Who blogs? Personality predictors of blogging. *Computers in Human Behavior, 24*(5), 1993-2004.

Hofstee, W. K. (1990). The use of everyday personality language for scientific purposes. *European Journal of Personality, 4*(2), 77-88.

Hossin, M. & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process (IJDKP), 5*(2), 1-11.

Huang, F., & Yan, L. (2014). Combined kernel-based BDT-SMO classification of hyperspectral fused images. *The Scientific World Journal, 2014*, 1-13.

Kamaluddin, M. R., Shariff, N. S., Othman, A., Ismail, K. H., & Saat, G. A. M. (2015). Linking psychological traits with criminal behaviour: A review. *ASEAN Journal of Psychiatry, 16*(2), 13-25.

Kapur, B., Ahluwalia, N., & Sathyaraj, R. (2017). Comparative study on marks prediction using data mining and classification algorithms. *International Journal of Advanced Research in Computer Science, 8*(3), 394-402.

Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications, 98*(22), 13-17.

Korting, T. S. (2006). *C4. 5 Algorithm and Multivariate Decision Trees*. São Paulo, Brazil: National Institute for Space Research–INPE.

Kursuncu, U., Gaur, M., Lokala, U., Thirunarayan, K., Sheth, A., & Arpinar, B. (2018). Predictive analysis on twitter: techniques and applications. In N. Agarwal, N. Dokoohaki & S. Tokdemir (Eds.), *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 67-104). Cham, Switzerland: Springer.

Mairesse, F., & Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics, 37*(3), 455-488.

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw Hill Publishers.

Mohammadi, G., & Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing, 3*(3), 273-284.

Nanda, M. A., Seminar, K. B., Nandika, D., & Maddu, A. (2018). A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information, 9*(1), 1-14.

Nasa, C., & Suman, S. (2012). Evaluation of different classification techniques for web data. *International Journal of Computer Applications, 52*(9), 34-40.

Oberlander, J., & Nowson, S. (2006, July 17-21). Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 627-634). Sydney, Australia.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, Illinois: University of Illinois Press.

Peng, K. H., Liou, L. H., Chang, C. S., & Lee, D. S. (2015, October 23-24). Predicting personality traits of Chinese users based on Facebook wall posts. In *2015 24th Wireless and Optical Communication Conference (WOCC)* (pp. 9-14). Taipei, Taiwan.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *MSRTR: Microsoft Research, 3*(1), 88-95.

Raschka, S. (2014). *Naive bayes and text classification: Introduction and theory*. Ithaca, NY: Cornell university library.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. Liu & Ozsu, M. T. (Eds.), *Encyclopedia of database systems* (pp. 532-538). Boston, MA: Springer.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 616-623). Menlo Park, USA: AAAI Press.

Ruparel, N. H., Shahane, N. M., & Bhamare, D. P. (2013). Learning from small data set to build classification model: A survey. *International Journal of Computer Applications, 975*(8887), 23-26.

Sagadevan, S., Malim, N. H. A. H., & Husin, M. H. (2015). Sentiment valences for automatic personality detection of online social networks users using three factor model. *Procedia Computer Science, 72*, 201-208.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Advances in Science, Technology and Engineering Systems Journal, 2*(1), 127-133.

Saravanan, S. (2016). *Personality Detection in Online Social Networking by using Three Factor Personality Model* (Master dissertation). Universiti Sains Malaysia, Malaysia.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One, 8*(9), 1-16.

Shally, B. (2014). *Personality Assessment Using Multiple Online Social Networks* (Master dissertation). University of Ottawa, Canada.

Sujatha, R. & Ezhilmaran, D. (2013). Evaluation of classifiers to enhance model selection. *International Journal of Computer Science and Engineering Technology (IJCSET), 4*(1), 16-21.

Urmaliya, A., & Singhai, J. (2013, December 9-11). Sequential minimal optimization for support vector machine with feature selection in breast cancer diagnosis. In *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)* (pp. 481-486). Shimla, India.

Van Dam, C., Janssens, J. M., & De Bruyn, E. E. (2005). PEN, Big Five, juvenile delinquency and criminal recidivism. *Personality and Individual Differences, 39*(1), 7-19.

Verhoeven, B., Daelemans, W., & De Smedt, T. (2013, July 11). Ensemble methods for personality recognition. In *Seventh International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media* (pp. 35-38). Boston, USA.

Wang, P., Bojja, N., & Kannan, S. (2015, June 5). A language detection system for short chats in mobile games. In *Proceedings of the third International Workshop on Natural Language Processing for Social Media* (pp. 20-28). Denver, Colorado.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.