algorithms

MDPI

Article

A Seed-Guided Latent Dirichlet Allocation Approach to Predict the Personality of Online Users Using the PEN Model

Saravanan Sagadevan, Nurul Hashimah Ahamed Hassain Malim * and Mohd Heikal Husin

School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia; saravanan_18@student.usm.my (S.S.); heikal@usm.my (M.H.H.)
* Correspondence: nurulhashimah@usm.my; Tel.: +60-4-6534645

Abstract: There is a growing interest in topic modeling to decipher the valuable information embedded in natural texts. However, there are no studies training an unsupervised model to automatically categorize the social networks (SN) messages according to personality traits. Most of the existing literature relied on the Big 5 framework and psychological reports to recognize the personality of users. Furthermore, collecting datasets for other personality themes is an inherent problem that requires unprecedented time and human efforts, and it is bounded with privacy constraints. Alternatively, this study hypothesized that a small set of seed words is enough to decipher the psycholinguistics states encoded in texts, and the auxiliary knowledge could synergize the unsupervised model to categorize the messages according to human traits. Therefore, this study devised a dataless model called Seed-guided Latent Dirichlet Allocation (SLDA) to categorize the SN messages according to the PEN model that comprised Psychoticism, Extraversion, and Neuroticism traits. The intrinsic evaluations were conducted to determine the performance and disclose the nature of texts generated by SLDA, especially in the context of Psychoticism. The extrinsic evaluations were conducted using several machine learning classifiers to posit how well the topic model has identified latent semantic structure that persists over time in the training documents. The findings have shown that SLDA outperformed other models by attaining a coherence score up to 0.78, whereas the machine learning classifiers can achieve precision up to 0.993. We also will be shared the corpus generated by SLDA for further empirical studies.

Keywords: machine learning; personality detection; PEN model; topic modeling

Citation: Sagadevan, S.; Malim, N.H.A.H.; Husin, M.H. A Seed-Guided Latent Dirichlet Allocation Approach to Predict the Personality of Online Users Using the PEN Model. *Algorithms* **2022**, *15*, 87. https://doi.org/10.3390/a15030087

Academic Editors: Ioannis E. Livieris and Panagiotis Pintelas

Received: 16 December 2021
Accepted: 3 March 2022
Published: 8 March 2022

1. Introduction

Research has found language to be a reliable mode of measuring and understanding personality, and various methods of analysis have been devised to explore how personality is encoded in the way that people use language. In this sense, the affective computing community started to explore the valuable information encoded in social networks (SN) platforms as the platforms continuously collect individual affective intelligence such as habits, social interactions, and interests. The Big 5 personality model comprised of five super-traits, namely Extraversion, Neuroticism, Agreeableness, Conscientious, and Openness, is predominantly employed in many affective computing studies [1]. On other hand, the lack of interest of the community in investigating other significant personality systems such as the Psychoticism-Extraversion-Neuroticism (PEN) model impoverishes the body of knowledge. The PEN model was widely applied in criminology to study criminals' behaviour, especially in the context of Psychoticism [2,3]. Nonetheless, it is tedious, expensive, and time-consuming to collect datasets aggregated to the PEN model due to the complexity of rules and regulations and the lack of confidence in the efficiency of feedback collection through self-reporting psychological instruments [4].

Because language and personality are strongly correlated, the use of seed words to categorize the topics according to human traits could be an alternative to overcome the

inherent problems of collecting the personality-based datasets as the efforts to collect meaningful seed words are much cheaper and easier [5]. The recent literature also showed that incorporating seed knowledge to guide the auto-modeling is distributed in many aspects such as long documents [6], event detection and mapping [7], and unsupervised error estimation on various natural language text corpora [8]. At the same time, past studies have shown that dataless techniques practically infer the underlying latent semantical structure of data automatically by applying the probabilistic model and representing the features in the training documents as a probability distribution over the heterogeneity topics [9]. To the best of our knowledge, no dataless study has modeled the topics according to personality traits using seed words despite the significant influence of such auxiliary knowledge in the psychological study [10]. Thus, the proof-of-concept by Toubia et al. [10] and the potential of Latent Dirichlet Allocation (LDA) in modeling the dataless problem [11–13], inspired our team to hypothesize that a small set of seed words is enough to decipher the underlying psycholinguistic states encoded in texts and could leverage the automatic model to categorizes the texts according to personality traits.

Therefore, our team experimented with the hypothesis by devising an unsupervised model called Seed-guided Latent Dirichlet Allocation (SLDA) to modeled the SN contents based on PEN model traits. To the best of our knowledge, this is the first PEN model textual corpus developed using an unsupervised model that can be used by other scholars for further empirical experiments. Because the affecting computing community has established many empirical studies on Extraversion and Neuroticism, this experiment also specifically focused on the representation of Psychoticism trait. Psychoticism is one of the canonical traits that conceptually defined the characteristic of bizarre thoughts and perceptions as well as typically correlated to criminal personalities and anti-social behaviour [3,14]. This type of data-driven experiment on Psychoticism tended to shed light on the prospect of personality aspects in the forensic area, especially in the context of user trustworthiness [15–17]. The SLDA algorithm outperformed other topic models such as LDA and Latent Semantic Analysis. We also conducted comparative analysis extrinsically using several prominent off-the-shelf supervised classifiers to predict the ground-truth topics generated by SLDA. The extrinsic evaluation showed that most of the classifiers can well predict the traits classes generated by SLDA.

Following this introductory section, the remainder of the paper is organized as follows. The literature about the personality model is described in Section 2. The discussions about our preliminary study, affective computing and dataless topic modeling are presented in Section 3. The problem and proposed methodology are described in Section 4, while Section 5 disclose the findings of this study. Furthermore, Section 6 described threats to the validity of this experiment. Finally, we discussed the limitation and future direction in Section 7 followed by the conclusion and discussion in Section 8.

2. Personality Model

In general, human beings are complex creatures with considerable heterogenous habitual patterns in the perspective of personality, specifically, emotions, memories, perceptions, thoughts, feelings, and actions [1]. Despite the intricacies of human nature, the central assumption of linking psychology with linguistics is that people's choice of words often reveals who they are [18] and typically represents the diverse aspects of psychological traits [19]. The reciprocal effects between psychology and linguistics initiated the affective computing community to conduct theory-driven experiments using personality models such as Big 5 and Myers Briggs Type Indicator (MBTI) to diagnose the behaviour of online users [1]. Apart from the constraints stated earlier, the lack of resources is also one of the factors that limited the efforts to investigate the representation of other personality frameworks such as the PEN model.

The PEN model was introduced in 1967 by Eysenck through factor analysis and has been widely applied in criminological research by linking the behaviours of criminals through the socialization process [2,3,20,21] and such traits manifestation inspired our

team to conduct our preliminary analysis [22] in the contexts of natural language. The model consists of three broad personality factors, namely Extraversion, Neuroticism, and Psychoticism [23]. Extraversion is based on cortical arousal that is related to social interest and interaction [23], whereas Neuroticism is a trait that measures emotional stability where it is assumed to be associated with the sympathetic nervous system and its activation threshold [23]. On the other hand, Psychoticism refers to the personality patterns that represent aggressiveness, interpersonal hostility or individuality and are strongly correlated with psychotic episodes [21]. Psychotic people typically exhibit their antisocial behaviour through their actions and can harm the emotional states of people [14]. The description of the PEN model traits and specific characteristics are shown in Table 1.

Table 1. Pen model traits and the corresponding characteristics [20].

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.

As mentioned earlier, the lack of resources such as domain knowledge limited the establishment of empirical research using other personality models. In this respect, we conducted our preliminary study to collect many reliable and meaningful seed words related to each PEN model trait so the domain knowledge could be integrated into the proposed SLDA algorithm and contribute the resources to the body of knowledge. A brief introduction of our preliminary study is stated in the Related Work section.

3. Related Work

Besides elaborating our preliminary study, this section also discussed previous experiments conducted in Affective Computing and Dataless Topic Modeling. Furthermore, the justifications and advantages of our proposed model are presented in the Dataless Topic Modeling section.

3.1. Overview of the Preliminary Study

This section briefly discusses the preliminary study conducted to compile the domain knowledge related to PEN model traits [22]. The motivation to compile the seed words aggregated to personality models was raised due to the representation power of such terms in revealing the demographical of the topical categories [5]. In this sense, we adopted a mechanism called Automatic Personality Perceptions (APP) and executed a survey to gather public perception towards the list of the sentiment words extracted from *myPersonality* using Part-of-Speech Tagging elements. Under affective computing, APP refers to the identification of personality based on perception and typically focuses on predicting personality attributes based on observable behaviours such as the usage of words in writing [1,23]. Even though APP concentrates on predicting the personality attributes based on observable behaviours, the general perception significantly indicates an individual's personality through social interaction [1,23].

In the context of natural language, the semantic and emotional spectrum of Extraversion is directed towards positive emotionality that indicates the tendency of positive conversation with other people, whereas Neuroticism and Psychoticism connotated negative emotions [14] diversified based on the level of negativity [22]. In this experiment, we defined Neuroticism to depict the low-medium negativity whereas Psychoticism portrays medium-high negativity based on premises of semantic valence. Therefore, our team defined the psycholinguistics emotional scope by categorizing the low-medium negative intensity words to Neuroticism whereas medium-high negative intensities words corre-

sponded to Psychoticism as psychotics people are prone to intimidate or humiliate other individuals. According to the linguistic universal concept, the intense negative words function as a medium to express the intentions to abuse or humiliate other people [24].

In this preliminary study, we embraced the concept of sentiment valence to bridge the gap between linguistics and personality by measuring the semantical dimensionality embedded in words and associated them with each PEN model trait. To identify the valence, the publicly available valence-based rated English words corpus AFINN was adopted [25], and at the same time, a survey was conducted using words unavailable in AFINN to enrich the collection of seed words. The survey was performed to gather public perception towards the filtered words using valence as a sentiment-measurement metric.

Throughout the survey, we collected 67 sentiment valence words such as mofo, cum, and Goddamnit and statistically associated them to each PEN model trait. In this experiment, Cronbach coefficient reliability scores of 0.951, 0.937, and 0.855 were achieved for Extraversion, Neuroticism, and Psychoticism, respectively [22]. The Cronbach coefficient measurement was adopted to statistically assess the internal consistency or reliability of sentiment valence metrics chosen by the respondents that assist summarization of overall perceptional dimensionality over the items (seed words) in the survey [26]. In this regard, the higher Cronbach coefficient values emphasized the validity of summarization and strengthen the correlation and degree of trustworthiness of public perceptions towards the items in the survey. Table 2 present some of the sentiment terms that will be used as seed words to model the topics according to PEN traits. The comprehensive information of our preliminary study was published in [22].

Table 2. Example of seed words from our preliminary study and AFFIN corpus [22].

Psychoticism	Extraversion	Neuroticism
Ass, Asshole, Assfucking, Cum, Bullshit, Wtf, Damn, Dick, Catastrophic, Fuck, Fucktard, Fuking, Piss, Shit, Bastard, Bitch, Cock, Cocksucker, Cunt, Nigger, Niggas, Mofo, Penis, Goddamnit, Motherfucker	Like, Good, Love, Happy, Fun, Great, Better, Lol, Please, Nice, Hope, Best, Awesome, Thank, Feeling, Pretty, Wish, Amazing, Cool, Wonderful, Wow, Beautiful, Care, Luck, Kind, Super, Funny, Yeah, Enjoy, Win, Hahaha, Glad, Peace, Excited	Bad, Stupid, Suck, Crap, Sad, Bore, Mad, Hurt, Kill, Stuck, Poor, Dead, Annoy, Sore, Sigh, Slap, Grrr, Worst, Disappoint, Fear, Weak, Weird, Fool, Difficult, Doubt, Upset, Idiot, Dumb, Lame, Hate, Shame, Afraid, Disgust, Sick, Arghhh, Foolish, Anxious, Hopeless

3.2. Affective Computing

In the past two decades, the interest to recognize and predict personality traits from digital sources has increased because of the growth of user-generated data that encapsulate key information about human characteristics. Oberlander et al. [27] have pioneered the study of personality via weblog data based on binary and multiclass classification using Naive Bayes and Support Vector Machine (SVM) classifiers [28]. Meanwhile, Iacobelli et al. [29] used unigram and bigram attributes along with some additional text mining techniques such as Inverse Document Frequency (IDF) to predict the personality of bloggers. Pereira et al. applied IBM Watson algorithms to predict the Big 5 personality of Twitter users based on the match words found in the standard LIWC psycholinguistic dictionary [30]. Instead of using a single classifier, ensemble-based boosting classifiers such as AdaBoost also have been used to predict the traits of the human being [31–34].

Nonetheless, the most prominent effort to investigate the personality of users was initiated through *myPersonality* Facebook status message corpus, whereby the performance of machine learners was evaluated via heterogeneous machine learning techniques based on the standard benchmark defined by Workshop on Computational Personality Recognition (Shared Task) [28]. The organizer of the workshop provided gold standards based on the Big 5 traits, i.e., Openness to Experience (Openness), Conscientiousness, Extroversion, Neuroticism, and Agreeableness, to measure the machine classifiers performance. Since the

establishment of the workshop, many automatic personality recognition studies have been conducted using heterogeneous techniques and machine learning models on the various areas [28–30,35,36]. One of the notable studies had proposed a system to estimate happiness using the PEN model traits on WhatsApp messages; however, the investigation did not produce any conclusive results [21]. Besides the effort to recognize the psychological traits, Levitan et al. [34] also found that Random Forest performed well in identifying gender and native languages used by deceivers [33], although the classifier seems to be performed poorly compared to SVM in previous experiments using PEN model [21].

On the other hand, there is a limited number of studies that used topic modeling techniques to predict the personality of users based on natural language. For instance, Liu et al. used LDA to predict the traits and user behaviours from labeled documents using a multilabel classification mechanism [37]. In addition, Moreno et al. [38] applied the Non-Negative Matrix Factorization (NMF) and LDA models to reduce the dimensionality of Tweets to improve the prediction of personality traits. Kwantes et al. also demonstrated the applicability of Latent Semantic Analysis (LSA) in revealing the relationship between the scores obtained by participants in the Big Five questionnaire and the predictions of their essays [39]. To the extent of our knowledge, no study investigated the use of dataless techniques to categorize the unlabeled social network messages according to personality traits. This is interesting as categorizing the unlabeled contents automatically can eliminate the vulnerabilities caused by manual labeling. Furthermore, our experiments were conducted based on a single label instead of multilabel classification [35], as our objective is to automatically categorize the instances based on the most preferable single psychological trait rather than predicting the relational correlations among the traits. The following section will briefly discuss the literature about dataless topic modeling.

3.3. Dataless Topic Modeling

Topic modeling techniques such as LDA capture substantial inter–intra structures of documents using statistical distribution [9,40]. LDA had been acknowledged as one of the successful mechanisms to model the topics according to the document themes [40]. Dataless modeling is an unsupervised approach that does not require any labeled document, whereby the models penetrate the semantic similarity encapsulated between a given document and a set of predefined classes to determine the categories of the given document [5]. This technique has become an attractive approach as it can reduce the time and human effort needed to label the training document besides training without annotated data. Chen et al. [41] have proposed the Descriptive LDA (DescLDA) to model the topics by adjusting the Dirichlet Priors parameter (β) and other hyperparameters without injecting prior external knowledge. Li et al. [12] proposed the Laplacian Seed Word Topic Model (LapSWTM) that exploits local neighbourhood structures and enhances the capturing of discriminative features. The LapSWTM has revealed that the use of prior knowledge or seed words substantially improves the model to better converge in an unsupervised manner. On the other hand, Vendrow et al. [42] proposed the Guided Non-Negative Matrix Factorization (GNMF) model that could incorporate the external seed words to model the topics according to predetermined topics. Theoretically, GNMF is based on a linear-algebraic optimization algorithm called Non-Negative Matrix Factorization (NMF) that used the matrix factorization technique to explore the high dimensional data and automatically extracts sparse and meaningful features from a set of nonnegative data vectors. Although the explicit use of seed words to identify the topics has been performed previously [10,13], the tendency of using such a guided approach to model the topics in the contexts of psycholinguistics has not been investigated. Therefore, the research gap inspired our team to proposed a seed guided model to automatically label the social networks messages based on personality traits. This idea is practical because the seed words compiled earlier can be prior intelligence to model the topics in an unsupervised manner using dataless Seed Guided LDA (SLDA).

The are several reasons to proposed SLDA to categorize the documents based on PEN model traits. First, it is infeasible to use the conventional LDA that exploits the word co-occurrence on diversified datasets [11]. Second, the SLDA is more robust than conventional LDA as the guided model could assist the topic to discover secondary or nondominant statistical patterns in the document. Third, the type/number of the categories embedded in the document were known, and the availability of the corresponding domain knowledge could reduce the bottlenecks caused by manually collecting and labeling the document [5].

The premises of this study also differ from existing dataless studies where those experiments automatically generate the seed words based on statistics without considering the psychological themes embedded in the texts [43,44], guide the model to identify other relevant attributes or determined the number of topics [7,8,45], and merely applied to assist the topic outputs [42]. In Psychology, emotion and perception were quintessential characteristics that naturally formed the personality of a human being [10]. Thus, identifying psycholinguistics related seed words automatically may mislead the categorization and lead to the noisiness in the topic generation process as well as widen the semantics and contextual gap [42]. Another key difference between manual and automatic seed selection is the latter approach typically will grow the small set of seed words into a much larger set [43], whereas our experiments showed that the proposed algorithm performed well when the initial set of seed words reduced gradually on each run. The advantage of SLDA to incorporate a small set of seed words could accelerate the whole learning process by reducing the constraints of sampling space. In addition, most of the previous experiments were usually conducted on theme-known datasets like medical abstracts [5] and natural disasters [7], whereas our training documents that based on SN messages were typically heterogeneous and theme-eccentric. Hence, the following section elaborates the problem formulation of categorizing the SN messages automatically according to personality traits using a seed-guided approach.

4. Problem Formulation and Methodology

This section discusses the problem formulation and methodology devised to model the unlabeled SN user messages based on personality traits using our proposed seed-guided approach.

4.1. Problem Formulation

In this experiment, we treated each of the training instances d as a collection of D. Based on (1), let $D = \{d_1, d_2, \ldots, d_i\}$ be a set of the documents where m is a number of the instances more than zero. The D were represented as a finite-dimensional vector, in which $\underset{di}{\rightarrow} = \left(a_{i1}, \ldots, a_{ij}\right)$, where a_{ij} was the weight that represented the proportion of the attributes in the d_i. Each document $d_i \in D$ represented as word tokens a, where $a_v \in V(k \in \{1, 2, 3, \ldots, j_d\})$, k was the size of the vocabulary and j_d denoting the number of token a in d_i. Based on notion, the data matrix $X \in \mathbb{R}^{v \times a}$ representing the words in the topics distributed along the column c in v and each category of $T = \{t_1, \ldots, t_n\}$ associated with a small set of seed words of vocabulary S. Supposed the n representing the number of categories and S consisted a number of seed word s, $s^{(1)}, \ldots, s^{(z)} \in \mathbb{R}^v$, then, the equation can be formulated as $S = \left(s^{(1)}, \ldots, s^{(z)}\right) \in \mathbb{R}^{v \times c}_{\geq 1}$. In a real-world problem, the S is expected to be very sparse as the number of seed words that belong to each trait is quite smaller compared to the number of attributes in V. We made the assumptions that each s has one-to-one relationships among those traits in order to associate the semantic information of a corresponded to each of the traits.

In this experiment, we embraced the concepts of multiclass (3-class) and one-vs-all (2-class) distributions and formulated the vocabulary S according to the respective distributions. The vocabulary S for multiclass consists of a set of seed words s and labeled exactly according to each PEN model trait as identified in our preliminary study (Table 2). In contrast, the vocabulary S for one-vs-all distributions merely consisted of a list of s labeled as Psychoticism and non_Psychoticism (interchangeably used as not_Psychoticism) where

all the Extraversion and Neuroticism seed words turned to non_Psychoticism category. For instance, the seed words "like" and "stupid" that associated to Extraversion and Neuroticism respectively in Table 2 was labeled as non_Psychoticism whereas the word "motherfucker" and "fuck" was maintained the association to Psychoticism category. The primary goal of this automatic modeling implementation is to incorporate the S into the LDA architecture to automatically categorize each of the documents d_i based on the most relevant trait category without using the labeled document. Hence, the following section shows the devised methodology of this study to experiment with the premise defined earlier.

$$D = \left\{ \underset{d}{\rightarrow}_{m} \right\}^{j}_{m > 0} \tag{1}$$

4.2. Proposed Methodology

The methodology of this study is illustrated in Figure 1. This investigation began with the collection of two textual corpora, namely *myPersonality* [46], and *Sentiment140* [47], from their respective sources. Although the *myPersonality* corpus provider [46] had decided to stop providing the dataset since April 2018 for empirical investigations, permission was acquired from them to publish this work since the experiment was conducted in 2015, and the thesis was already published [48]. On the other hand, *Sentiment140* were only used for cross-domain analysis to prove the concept of automatic labeling of SN texts based on human traits.

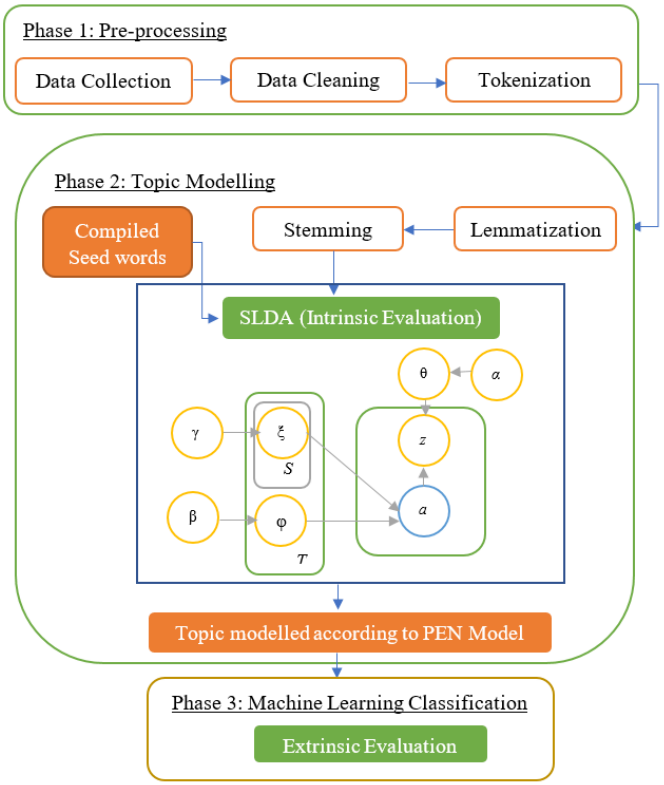

Figure 1. Research Methodology.

The *myPersonality* corpus consisted of 9917 Facebook status messages with 144,041 words from 250 selected users along with additional information such as Authid, status messages, personality traits based on the Big 5 model, betweenness, and density. Nonetheless, only the raw status messages were used for further investigation as the focus of this study was on categorizing natural texts according to PEN model traits. The information regarding the users' personality traits was used in the cross-examination process and was exhaustively excluded from further analysis due to the divergences in terms of the application of the

personality model. Meanwhile, *Sentiment140* was a text corpus comprising 1.6 million Tweets with more than 30 million words. Nevertheless, due to the limitation of resources, this study removed redundant tweets and randomly extracted the tweets that contained more than four words. The *myPersonality* corpus was the primary dataset for this study where the seed words to annotate the classes of the instances were derived, whereas the *Sentiment140* was used as a case study to assess the effectiveness of SLDA topic modeling and the performances of machine learning classifiers. The *myPersonality* corpus was used as the primary dataset due to the availability of self-assessment reports along with datasets that would facilitate the recognition of instances that may correlate to Psychoticism by cross-checking the traits correlation between Conscientiousness and Agreeableness and Psychoticism [14].

4.2.1. Data Cleansing and Linguistic Marker Identification

Next, the data cleaning process was performed on *myPersonality* and *Sentiment140* datasets to remove all the irrelevant textual attributes such as Uniform Resource Locators (URLs), punctuations, and symbols as well as non-English strings. Then, spelling correction and lowercase transformation were conducted to reduce the size of the training attributes and enhance the prediction tasks. After pre-processing, lemmatization using the Natural Language Toolkit (NLTK) and the stemming process were performed for each word in *myPersonality* and *Sentiment140* instances so that each morphological form was mapped into its root term. In this study, both datasets were stemmed using the Porter Stemmer Algorithm [49] to increase the density of the training data. This stemming process was different from the stemming process performed in the preliminary study, which was carried out to identify the root terms for valence identification. Subsequently, the Part-of-Speech (POS) tagging was conducted to extract the adjectives, nouns, verbs, and adverbs as these grammatical elements often served as hints to indicate sentiments [12]. For instance, the status message "motherfuck..(ing) norwegian bureaucratic bullshit. yet again" was tagged as "motherfuck/VB norwegian/JJ bureaucratic/JJ bullshit/NN./yet/RB again/RB". The abbreviations VB, JJ, NN, and RB refer to the grammatical attributes of verb, adjective, noun, and adverb, respectively. Afterwards, the relevant sentiment words labeled as JJ, NN, and VB (e.g., motherfuck) were extracted, and training documents were tokenized using regular expression where all the text attributes except word and white space were removed. Then, the word attributes were transformed using the Bag-of-Words (BoW) mechanism, where the probabilities of each distinct word attribute a represented as $p(a_1 a_2, \ldots, a_n) = p(a_1), p(a_2), \ldots, p(a_n)$.

4.2.2. Topic Modeling

Recap the earlier premise, the general idea of a seed-dependent mechanism is to modeled the textual contents based on the given document collection of D, where the D contained a number of topics, $T = \{t_1, \ldots, t_n\}$, and each of the t contained a number of instances d, and defined by a small set of seed words, s. Specifically, the distribution of D mixture over T topics, where each of the t is a mixture of a regular topic (φ_t^r) and seed topic (φ_t^s). The model made desirable output during the inference, meaning that s and other attributes a that occur in the same context eventually have a high probability in the same topic. In this sense, SLDA inferred the most relevant traits by penetrating the explicit word co-occurrence structures between s and other regular attributes a in the D. The model employed the given s and maximized the relevancy of t using cooccurrence correlation without employing any predefined labels in training datasets. Thus, SLDA made assumptions that each s and its corresponded t associated with a single trait category based on the semantical connotation and word distribution of the training instances in D. The attributes a and seed words s were mapped to the dictionary of V and S, respectively, and all the attributes in D tokenized at word level and transformed using the Bag-of-Word (BOW) mechanism. In SLDA, the degree of co-occurrence probability between the

attributes a and s in D were computed using (2), where $df(s)$ were the number of instances that contained s and $df(a|s)$ were the volume of d that contained both a and s.

Because our preliminary collected seed words were statistically correlated with PEN model traits [22], it was reasonable to train SLDA to converge and categorize the relevancy between d and t via (3) and (4). Equation (3) was applied to determine the relevancy of each attribute to be associated for each t based on the probability of a and s (formula (2)), where the S_t is a set of seed words aggregated to each t. As the conditional probability of $U(a,t)$ dominated by the statistical properties of s, we determined the maximum value for the attributes by normalizing the weights for a and rid out the weights of noisy attributes by subtracting the average score and $\frac{1}{T}$ for each distribution of t using (4). Based on (4), the higher co-occurrences of a with s could resulted larger value of $U_n(a,t)$, where the bigger value indicates a was more likely to be a salient attribute of t. To simplify, the s acted as a guide to identify the latent a using cooccurrence probability to indicate the relevancy of the d to be categorized under prespecified t. The chance of choosing cooccurrence distribution of an a aggregated to a trait φ_t, relies on the domain knowledge supplied by S. However, the nature of the probabilistic model that made the higher-order cooccurrence structures increase the chances of frequent attributes a co-occurred in the same topic. By considering the weights from (4), we can refine the probability of attribute a being a feature for topic t based on (5), i.e., the probability of $\delta_{a,t}$, that an a is a latent attribute for a topic t. Although the proposed modeling procedure is similar to the previous nonpersonality study [11], our modeling process used fewer parameters that reduced the computation complexity on exploiting the underlying structure of short texts from SN compared to the existing experiment that used news datasets.

$$p(a|s) = \frac{df(a|s)}{df(s)} \tag{2}$$

$$U(a,t) = \frac{1}{|S_t|} \sum_{s \in St} p(a|s) \tag{3}$$

$$U_n(a,t) = max(\frac{U\ (a,t)}{\sum_{t\prime} U\ (a,t\prime)} - \frac{1}{T},\ 0) \tag{4}$$

$$a,t = \frac{Un(a,t)p}{1 - p + Un(a,t)} \tag{5}$$

$$\frac{n^{(d)}_{-i,t} + \alpha}{\sum_u^t n^{(d)}_{-i,u} + \alpha} \times \frac{n^{(w_i)}_{-i,t} + \beta}{\sum_{w\prime}^W (n^{(w\prime)}_{-i,t} + \beta)} \tag{6}$$

Subsequently, the SLDA algorithm generated the topics using the Gibbs sampler based on the generative process stated in Algorithm 1. In the Gibbs sampling process, the conditional probability of a term a in d is computed based on (6), where $n^{(d)}_{-i,t}$ is the number attributes in d assigned to topic t, and $n^{(w_i)}_{-i,t}$ is the number attributes in whole corpus assigned to topic t. Although the application of the Gibbs sampler is prevalent in LDA-based topic modeling experiments [7], our implementation is unique on the part of the documents where they are unlabeled and restricted to assess the topics in the perspective of psycholinguistics. Although the experiment using unlabeled documents will accelerate the constraints of sampling space [6], we were optimistic that the Gibbs sampler could improve the whole learning process and approximately estimate the t by leveraging the concealed knowledge and topical decomposition channeled through s and its associated t from respective S. Intuitively, our team made the inference as the social network texts usually shorter and succinct with less contextual information [7,50], would be structurally and semantically less complex than long texts [6].

In order to integrate S into the Gibbs Sampler, we defined ξ as the distribution of seeds S and applied Symmetric Dirichlet prior for θ, and φ with hyperparameters α, β and

Dirichlet smoothing γ = 0.01. During the sampling process, we set the Markov Chain states for the latent topics t using multinomial distribution, $\theta = p(t|\ \alpha, \gamma)$, where the computation controlled by α in the $d, \alpha = p(t|d)$. On the other hand, the conditional distribution of a and s derived through $p\ (a|t,\ \beta, \gamma) = \pi_{\mathrm{ai}}|T,\ \pi_{\mathrm{Si}}|T$, which controlled by β that determines the number of times a is sampled from the t, prior to observing the words in d, $\beta = p(a|t)$ [49]. The parameter π controls the probability of drawing a word from the seed topic distribution versus the regular topic distribution.

We used the parameter values of 1.0 and 0.1 for α and β, respectively, as they showed insignificantly better modeling than the other values based on weights (5). The Gibbs sampler will approximately estimate the parameters and assign the high probable predefined topics from S by assuming that each a_i has their topic distribution of $T = \{t_1, \ldots, t_n\}$, where the $n \in \mathbb{N}$ restricted to 2 and 3 denoting the number of t in one vs. all and multiclass distribution, respectively. The procedure to generate the topics depicted in Algorithm 1 and the descriptions about the notion stated in Table 3.

Algorithm 1: Topic Modeling with SLDA

For each topic $t = 1 \ldots n$, choose θ_t :
Draw the regular topic $\varphi_t^r \sim \mathrm{Dir}\ (\beta_r)$.
Draw the seed topic $\varphi_t^r \sim \mathrm{Dir}\ (\beta_s)$
Select $\pi_t \sim \mathrm{Beta}\ (1, 1)$
For each seed set $S \in \{1, \ldots, K\}$:
Draw a distribution over seeds $\xi t, S_k \sim \mathrm{Dirichlet}\ (\gamma)$
For each document $d \in \{1 \ldots d\}$:
Draw $\theta_d \sim \mathrm{Dir}\ (\alpha)$.
For each token $a_i \in \{1 \ldots |d|\}$:
Select a topic $z_i \sim Mult\ (\theta_d)$
Draw $x_i^d \sim \mathrm{Bernouli}(\pi_{zi})$
If $x_i = 0$:
Select the regular topic r *from* $d = a_i \sim Mult(\varphi_{zi}^r)$
If $x_i = 1$:
Select the seed topic from $d = a_i \sim Mult(\varphi_{zi}^s)$

Table 3. List of notions.

Notion	Description
D	Total number of documents in each dataset
T	Total number of topics.
V	The vocabulary of attributes
S	The vocabulary of seed words
A	A regular attribute in the document
S	A seed word in the document
Θd	The topic distribution of document d
Φt	The word distribution of topic t
$\delta_{a,t}$	The probability of attribute a being a latent feature for category t
α, β, γ	Dirichlet Priors

Since discriminative power carried by the s practically determined the classes of training instances, it was noted that the diversity of the datasets, which was not focused on a certain topic, would divert the correlations between the semantics of the instances and personality traits. Initially, all the seed words collected during our preliminary study were incorporated to guide the modeling process. However, the initial training yielded awkward outcomes where the majority of the data points were oriented to either Neuroticism or Extraversion. This is due to the variation in the number of seed words that belong to each t in our training. Thus, our team decided to tune the number of seed words in S by repeating the training until a global equilibrium is optimized to the optimal coherence score, and at the same time, the generated topics are interpretable and meaningful to the category of the seed words.

To minimize the wrong topics categorization, the SLDA was run multiple times where each run used instances without replacement. In this study, the analysis was run twice and thrice for *myPersonality* and *Sentiment140,* respectively, to improve the categorization. The variation in the number of running times was due to the variation in terms of the size of datasets. The multiple iterations of SLDA showed the improvement in terms of reliability of modeling the topics according to aggregated categories by minimizing arbitrary post hoc selection of topics. Typically, multiple iterations using Gibbs Sampling on LDA variant models may cause inconsistency because of random sampling [34]. This inherent problem also persists in our experiment, but the effects are almost insignificant for the categorization tasks described here. Consequently, once SLDA generated the relevant topics for each training instance, the instances that contained fewer than five words were eliminated because the categorization seems to be irrelevant to the predicted classes.

4.2.3. Cross Validation Criteria

Prior to analyzing the output produced by SLDA, it was pertinent to determine whether the seed words effectively revealed the personality insights of social media users. As mentioned earlier, even though the personality traits are not mutually exclusive in real-world cases [21], the reliability of SLDA in generating the topics still can be determined using the traits correlation between the Big 5 (Conscientiousness and Agreeableness) and PEN (Psychoticism) [14,21,51]. Meanwhile, personality theorists seem to have accepted that Extraversion and Neuroticism are two fundamental personality dimensions of human beings and are strongly convergent [52]. Therefore, to bridge the gap between the two personality systems, the information about personality scores provided by *myPersonality* provider was used to cross-examine the topics generated by SLDA aggregated to the PEN traits based on the following criteria:

(1) Any training instances labeled as Psychoticism by SLDA must be correlated to Conscientiousness or Agreeableness scores provided in *myPersonality*. The two traits seem to be correlated with antagonism characteristics and Psychoticism [14]. Texts that were labeled as Psychoticism also may be correlated to Neuroticism due to the negative coverage;
(2) Any training instances labeled as Extraversion or Neuroticism by SLDA must be directly correlated to the Extraversion or Neuroticism scores, respectively, as provided in *myPersonality*;
(3) Any training instances labeled as Psychoticism or Neuroticism by SLDA may have the element of Extraversion due to boundary ambiguities among the traits [21] and positivity biases in the human language [53].

The above criteria (i.e., (1–3)) were applied to develop the multiclass-based distribution. On other hand, the development of one vs. all distribution followed the rules in (1) and (3) as well as slight changes for rule (2) where the instances will be representing both Extraversion and Neuroticism traits as well as neutral instances. This implied that *myPersonality* and *Sentiment140* transformed into a distribution termed one vs. all [54], where "one" referred to Psychoticism instances and "all" constituted Extraversion, Neuroticism, and neutral instances and was labeled as "not_Psychoticism" (also referred to as "non-Psychoticism"). There were two reasons to employ the one vs. all analysis. First, our intention is to evaluate the performance of SLDA in the context of Psychoticism and non-Psychoticism where this experiment can be a stepping-stone to highlight the potential of psychological representation in forensic areas. Second, except for the sentiment coverage, the pilot study had shown the non-existence of unique structural relationships between the instances of Extraversion and Neuroticism despite the two traits being rooted in opposite sentiment polarities. The following sections will disclose the findings of our experiments.

5. Findings of the Study

This section discussed the findings of this study. Three types of evaluation were conducted, namely performance comparison where we compare the performance of SLDA

against other topic modeling techniques, an intrinsic evaluation that disclosed the nature of data produced by SLDA, and third is an extrinsic evaluation that disclosed the performance of several machine learning classifiers in classifying the topics generated by SLDA.

5.1. Performance Comparison

Prior to evaluating the topics generated by SLDA, we used three nonseeded models, namely LDA [40], NMF [55], and LSA [55] as well as a seed-guided model called GNMF [42] to make comparisons with our proposed model. We also conducted an analysis to determine the sensitivity of seed words against the performance of SLDA and GNMF. The LDA, NMF and LSA were chosen due to their popularity in topic modeling and affective computing [31,38,42,55]. The parameter settings for LDA and NMF, respectively, followed the same settings applied in SLDA and GNMF [42], whereas the settings applied in [55] were adopted for LSA. For GNMF, we followed the settings applied in [42] and evaluated the performance based on a number of high probable seed words in S as the vocabulary played an integral part in the seeded algorithms. We made some changes to the original GNMF code so that the model can produce the results based on perplexity and coherence.

Theoretically, perplexity is an intrinsic metric that captures how well a model has performed on unseen data and is measured through the normalized log-likelihood mechanism. Meanwhile, the coherence score measures the degree of semantic similarity between high scoring words in the d and helps distinguish the semantical interpretation of topics based on statistical inference [56]. The measurement using coherence score is widely applied in topic model experiments [7,50,57,58] due to the weaknesses of perplexity to serve the correlation between predictive likelihood and human judgement, and topic categorization through unsupervised manner does not guarantee the interpretability of their output [59]. Thus, to measure the coherence score, the cooccurrence for the given a was evaluated using the sliding window mechanism. The computation was executed using Normalized Pointwise Mutual Information between the top words, and the similarities were measured via cosine similarity [42,59]. Given a topic t and its top N attributes $At = \{a_1^t, \ldots, a_N^t\}$, the coherence score computed through (7), where $D(a)$ is the document frequency of attributes a, and $D(a_i, a_j)$ is the count of co-occurrences of a_i and a_j. Based on (7), the higher coherence score by SLDA would indicate better quality and reliability of the modeling process.

$$C(t : A^t) = \sum_{n=2}^{N} \sum_{l=1}^{n} log \frac{D(a_n^t, a_l^t) + 1}{D(a_l^t)} \tag{7}$$

Table 4 unveils the outcome of this experiment. This experiment showed that the coherence score of nonseeded LDA, NMF and LSA for both *myPersonality* and *Sentiment140* is below 50. Furthermore, the nonseeded LDA outperformed other nonseeded models by attaining a coherence score of 0.4976 on *myPersonality* multiclass distribution. The outperformance of LDA compare to NMF also can be seen in Moreno et al. [38], who focused on predicting the Big 5 traits. Our literature review also showed that standard LDA yielded better results compare to LSA in nonpersonality topic modeling study [55]. However, as expected, our experiment indicated that the seed-guided model performed better than nonseeded topic models where the coherence scores of SLDA and GNMF are more than 50 in all the analyses disclosed in Table 4. In addition, the higher perplexity obtained by LDA, NMF and LSA compared to SLDA and GNMF indicates that nonseeded models poorly interpret the topics aggregated to personality. The negative perplexity was due to the log space computation by the Gensim LDA package.

Table 4. Performance evaluation.

Non−Seeded Topic Model						
			myPersonality		*Sentiment140*	
	Model	**Distribution**	**Perplexity**	**Coherence**	**Perplexity**	**Coherence**
	LDA	Multiclass	9.85	0.4976	7.14	0.4621
		One vs. All	12.65	0.4643	12.65	0.4465
	NMF	Multiclass	10.56	0.4839	7.35	0.4328
		One vs. All	11.86	0.4601	13.81	0.4483
	LSA	Multiclass	15.61	0.4543	13.34	0.4254
		One vs. All	18.96	0.4471	14.46	0.4136
Seed−Guided Topic Model						
			myPersonality		*Sentiment140*	
Number of seed words	**Model**	**Distribution**	**Perplexity**	**Coherence**	**Perplexity**	**Coherence**
50	SLDA	Multiclass	−3.21	0.5112	−3.43	0.5274
		One vs. All	−3.23	0.5287	−3.54	0.5443
	GNMF	Multiclass	−3.23	0.5087	−3.46	0.5254
		One vs. All	−3.27	0.5293	−3.23	0.5467
40	SLDA	Multiclass	−3.13	0.5441	−3.20	0.5751
		One vs. All	−3.25	0.5824	−3.46	0.6164
	GNMF	Multiclass	−3.20	0.5465	−3.27	0.5673
		One vs. All	−3.17	0.5831	−3.49	0.5877
30	SLDA	Multiclass	−2.87	0.6331	−2.93	0.6775
		One vs. All	−2.31	0.6539	−3.05	0.6643
	GNMF	Multiclass	−2.88	0.6231	−2.99	0.6621
		One vs. All	−2.32	0.6321	−3.09	0.6712
20	**SLDA**	**Multiclass**	**−2.78**	**0.7293**	**−2.85**	**0.7824**
		One vs. All	**−2.03**	**0.7739**	**−2.27**	**0.7412**
	GNMF	Multiclass	−2.98	0.6854	−3.01	0.7061
		One vs. All	2.29	0.6935	−3.05	0.7276
10	SLDA	Multiclass	−3.12	0.6634	−2.99	0.7012
		One vs. All	−2.78	0.6645	−3.01	0.6943
	GNMF	Multiclass	−3.17	0.6212	−3.02	0.6273
		One vs. All	−2.76	0.6572	−2.98	0.6632

Overall, SLDA achieved the highest coherence for both *myPersonality* and *Sentiment140* on multiclass and one vs. all distributions compared to the nonseeded topic models and GNMF when the vocabulary S contained 20 topical seed words. Apart from the best performance on 20 seed words, SLDA also seems to perform better than GNMF when the vocabulary contained 10 seed words. However, GNMF also insignificantly performed better than SLDA on a certain analysis. For instance, GNMF performed better on *myPersonality* one vs. all distribution when the number of seed words is 40 and 50. On other hand, this experiment showed that reducing the number of seed words in SLDA and GNMF's vocabularies to a certain extent improved the coherence score of the predictive models. This indicates that both of the models lose their discriminative power to distinguish the psycholinguistics effects encoded in texts if more seed words are supplied to the modeling process, and the small number of seed words is sufficient to accelerate the learning process as the small volume could reduce the constraints of sampling space.

To summarize, the SLDA and GNMF outperformed the nonseeded models as the seed guided models have the advantage of implicitly leveraging the information from seed

words to gain better statistical inference about the data. This investigation also showed that SLDA performed better than GNMF in identifying the coherence of texts in the context of PEN model traits. Henceforward, we extended our investigation to gain additional insights about the nature of data modeled by SLDA because the GNMF has limitations to established labels for the training instances [42], and the nonseeded models poorly interpret the subjects of this study.

5.2. Intrinsic Evaluation

The intrinsic evaluations were conducted to gain insights about the nature of topics generated by SLDA and to determine whether the generated topics match the criteria defined earlier. The following subsection will discuss the findings of intrinsic evaluation.

5.2.1. Descriptive Statistics

Statistically, SLDA generated *myPersonality* multiclass distribution consisting of 10,893 tokens with 4288 unique tokens, whereas *Sentiment140* was comprised of 67,658 unique tokens from a total of 591,262 tokens. Meanwhile, the one-vs-all corpuses consisted of 12,670 tokens with 3357 unique tokens and 7,131,433 tokens with 944,592 unique tokens for *myPersonality* and *Sentiment140*, respectively. The analysis of the top 30 words aggregated for the three traits showed Extraversion terms leading in both *myPersonality* and *Sentiment140* multiclass distributions (Figure 2). This finding supports the Pollyanna Hypothesis that asserts, conceptually, that humans' natural language tends toward universal positivity biases [53] resembled in the Extraversion language which promotes positive emotions and sentiments. This domination effect was also seen in the one vs. all distribution as the not_Psychoticism classes also consisted of the Extraversion instances. Therefore, the probability distribution of the top 30 words for multiclass and one vs. all indicated the nature of the human language where there was more positivity than negativity in the data.

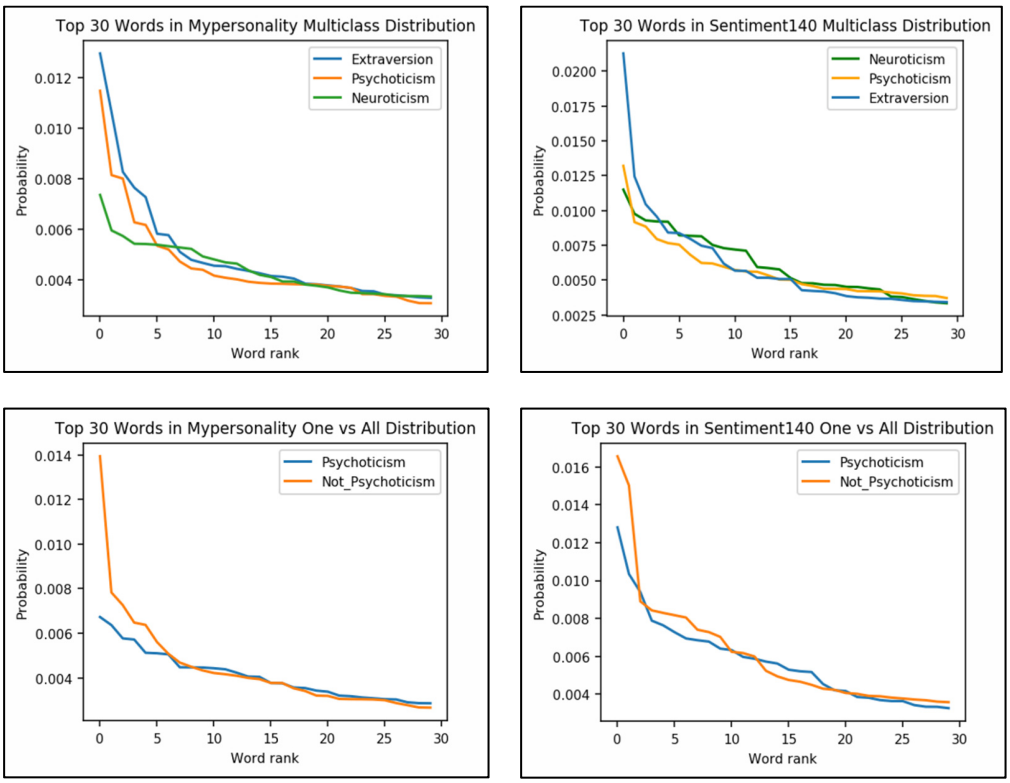

Figure 2. Probability distribution of the top 30 words.

5.2.2. Cosine Similarity

As an alternative to human judgement, cosine similarity has been applied by Towne et al. [58] to measure the similarity of topics modeled by LDA variants especially in the contexts of psychology. This type of metric-based validation enables the researchers to seek the ground truth and reliability of topics generated by the automatic models [58]. Therefore, we used cosine similarity as a metric to infer the intra-similarities of topics modeled by SLDA. As presented in Table 5, the cosine intra-similarity for *myPersonality* was 0.832 and 0.827 for multiclass and one vs. all distribution, respectively, whereas for *Sentiment140* it was 0.772 (multiclass) and 0.764 (one vs. all). The similarities indicated that topics in *myPersonality* were more homogeneous compared to *Sentiment140* where the topics or genres of discussion were naturally more diverse or overlapping.

Table 5. Cosine similarity of SLDA topics.

	MyPersonality	***Sentiment140***
Distribution	**Intra**	**Intra**
Multiclass	0.832	0.771
One vs. All	0.827	0.764

5.2.3. Seeking Ground Truth through Trait Correlation

Based on the trait correlation criteria stated earlier, we conducted a manual inspection of the topics generated by SLDA to determine whether the generated topics correlated with the criteria stated above. Our inspection showed that SLDA can well predict the trait classes of the messages where most of the predictions matched the criteria. For example, SLDA transformed the original training instance (#Authid:c597771fab7477c2ae7b507d532130b0) "There are so many fucked up people because there are so many fucked up marriages or lack thereof" from myPersonality and modeled it as "many fuck many fuck marriages lack thereof" as well as categorized it under the Psychoticism class. In *myPersonality* self-assessment report, this instance was labeled as having the elements of Agreeableness, Conscientious, and Openness. Based on criteria stated earlier, the SLDA has correctly labeled the instance of Psychoticism as the trait correlated to Agreeableness and Conscientiousness from the Big 5 framework. Furthermore, the original instance (#Authid:172400f46880b309ca5e97d322bb8f01) "Thank you, *PROPNAME*, for still being on the ballot so that stupid people could vote for you" that correlated with Neuroticism and Openness in myPersonality was modeled as "thank still ballot stupid could vote" under the category of Neuroticism. On the other hand, the instance (#Authid:138ac63ec2b55b8f48fd19c300720cae) "Can anyone out there tell me . . . . . I need scripture reference. Thank you" corresponded to extraversion in both the myPersonality report and SLDA. Typically, the instances generated by SLDA were shorter compared to the original instance in *myPersonality* as the instances went through the pre-processing tasks.

The meaningful topics generated by SLDA showed the ground truth of topics and proof of concept of automatically categorizing the instances based on human traits. Simultaneously, the automatic labeling has also indicated the possibility of perceptional-based personality detection to be applied as an alternative to detect human traits without their psychological report. Because the SLDA displayed a promising performance, the same concept was applied in categorizing the *Sentiment140* instances. Table 6 presents the samples of the *Sentiment140* instances generated and the corresponding PEN traits. Based on the common-sense knowledge, emotional spectrum, and semantical representation of the samples [1], the sentences generated by the SLDA showed the merit and logical sense to be categorized under the corresponding PEN model traits. Furthermore, the observations on the topics generated by SLDA also revealed that our earlier assumptions to seeding the s based on one-to-one relationships insignificantly affects the inference process by SLDA. This reveals that SLDA well penetrates the semantical information encoded among the

categories rather than naively considering the presence and absence of the seed words in each of the d.

Table 6. Samples of SLDA generated *Sentiment140* instances and the corresponding traits.

Num	Instance	PEN Trait
1	"photovia fuck yeah skinny bitch people really"	Psychoticism
2	"goin kill alicia gave fucking sickness ughhh wtf"	Psychoticism
3	"really upset louisville concert cancelled scared happen wnashville"	Neuroticism
4	"well had midwife evil evil woman gave anti jab hurt like hell baby think"	Neuroticism
5	"need someone pr experience volunteer help interested helping save world"	Extraversion
6	"happy thanksgiving facebook friends family thankful wonderful"	Extraversion

Further exploration was carried out to identify text messages that may be aggregated to criminality by traversing the Psychoticism instances labeled by SLDA. Our analysis showed that none of the *myPersonality* had any elements of criminality. This may be due to the size of *myPersonality* datasets that are small and the fact that the procedures used to collect the dataset based on voluntarism were not practical, i.e., for criminals to voluntarily hand in their psychological report without prejudice [4]. Nevertheless, several instances from *Sentiment140* that had been labeled as Psychoticism by SLDA seem to semantically denote the presence of criminality aspects. Table 7 disclosed some of the messages that have the logical and semantical sense to be comprised under the aspects of bullying, harassment, sexting or humiliation.

Table 7. Samples of instances connotating the criminality aspects.

Num	Instance
1	"fucking assholes poor little girl rip khyra"
2	"wut hummm waitin cum power like bf"
3	"swearbot shit piss cunt cocksucker motherfucker tits fart turd twat blink said best"
4	"photovia fuck yeah skinny bitch people really"
5	"goin kill alicia gave fucking sickness ughhh wtf"
6	"fucking assholes poor little girl rip khyra"

5.2.4. Word Analysis

We also extended our analysis to gain insights into the significance of seed words being used by SLDA to predict the trait category of the instances albeit our earlier analysis showed that SLDA performed well on 20 seed words. Figures 3 and 4 illustrate the plotting of significant seed terms over 1×1 relationships across the traits. As the heatmap unveiled the coverage and intensities of sentiment-personality terms that exposed the psychological metrics encoded in text messages, we turned our focus to address the sensitivity of seed words frequency towards the topics generated by SLDA. During the training to identify the seed words that significantly improve the categorization, we found that higher frequency seed words are not necessarily becoming a cue for SLDA to better predict the topics although the word frequency simulated the exposure of individual differences [60]. This effect indicate that the coverage of high frequency seed words may be homogenously distributed over the topics. This effect is also clearly depicted in our earlier experiment where the coherence score of SLDA and GNMF degraded when the number of seed words increased. Regarding the optimal number of seed words required to generate the well interpretable topics, our experiments suggest that seed words in the average range of 17 to 25 boosted the predictive power of SLDA. Our literature review showed that the application of optimal size of seed words perhaps varies to each experiment as some news discourse analysis studies reported the optimal number of six to seven for each category of the topic [61].

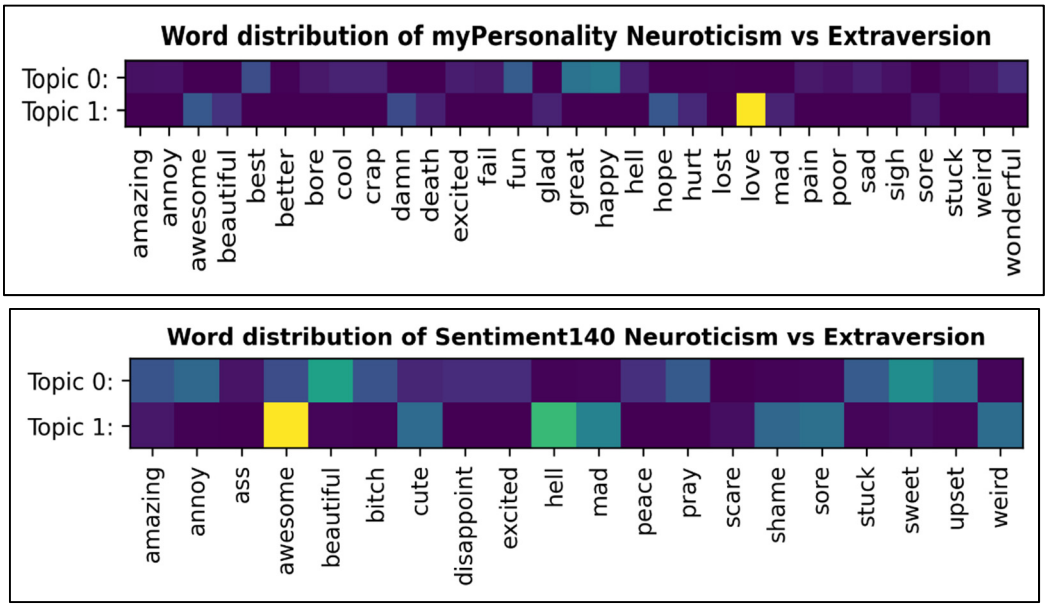

Figure 3. Word distribution of Neuroticism and Extraversion for *myPersonality* and *Sentiment140*.

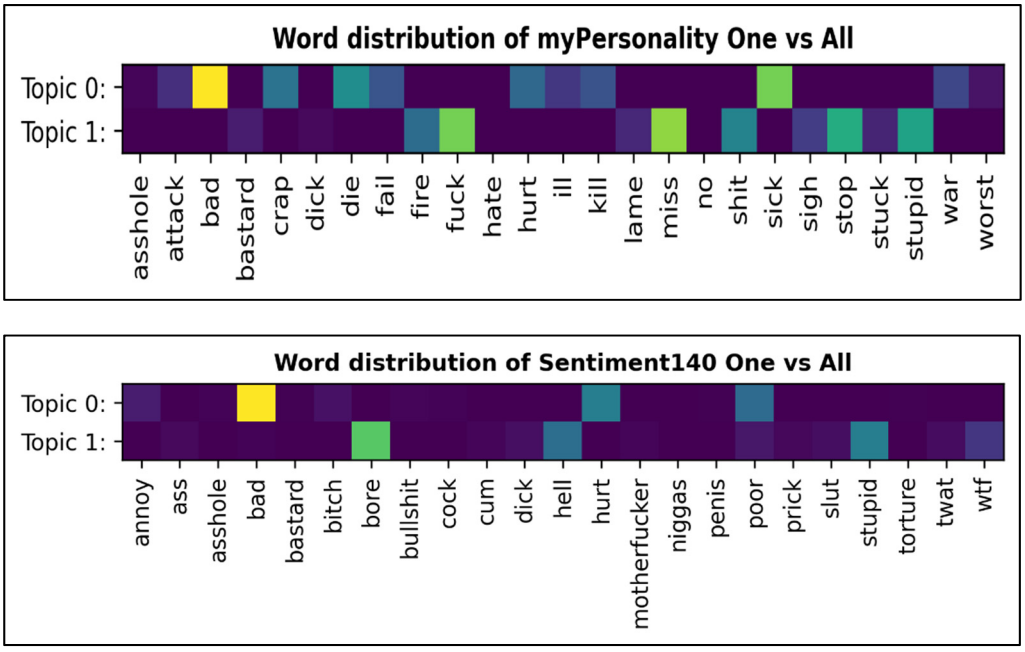

Figure 4. Word distribution of one vs. all distribution for *myPersonality* and *Sentiment140*.

Meanwhile, Table 8 presents the probability of the top three seed words that assist the modeling inference. It is noteworthy to underline that the probability difference between the top three and other seed words depicted in Figures 3 and 4 was very subtle. Apart from that, our further investigation disclosed that SLDA was insensitive for the tuning of hyperparameters such as α and β. This finding affirms the earlier suggestion [13] that SLDA is insensitive to parameter tuning, and a small number of seed words is sufficient for the model to converge and categorize the latent topics embedded in the dataset. Hence, recall that our earlier statement saying that our seed words vocabulary is sparse seems to be well synchronized with the topics in both *myPersonality* and *Sentiment140* datasets.

Table 8. Probability of top three seed words.

	myPersonality	Probability	*Sentiment140*	Probability
Multiclass	Amazing	0.040	Amazing	0.054
	Sad	0.040	Annoy	0.039
	Motherfucker	0.046	Hell	0.041
One-vs-all	Asshole	0.042	Stupid	0.052
	Fuck	0.051	Asshole	0.044
	Miss	0.062	Hurt	0.051

5.2.5. t-SNE Visualization

Figures 5 and 6 exhibit the t-distributed Stochastic Neighbor Embedding (t-SNE) visualization conducted on multiclass and one vs. all distributions of *myPersonality* and *Sentiment140*. The SLDA clustered the *myPersonality* classes with a clear boundary whereby the features were not overlapping on each other in the multiclass and one vs. all distributions. On the contrary, the plotting for *Sentiment140* revealed a strong overlapping for both distributions. In the plotting of *Sentiment140* for one vs. all distributions, the Extraversion (green) and a part of the Neuroticism (blue) instances from the multiclass distribution were merged to form the not-Psychoticism classes, whereas another part of Neuroticism fused with Psychoticism (red) to form the Psychoticism classes. As noted earlier, our pilot study has disclosed that some mutual inter-relationships are present between Extraversion and Neuroticism even though the two traits are rooted in the opposite or inverse sentiment polarities. Thus, the overlapping relationships between Extraversion and Neuroticism illustrated in *Sentiment140* one vs. all have affirmed the previous finding that the nature of human traits is mutually inclusive [21].

Furthermore, the presence of some co-existing relationships among the instances illustrates the reality of human nature even though it was not seen in *myPersonality*. Intuitively, the asynchronous association is perhaps because of the smaller size of *myPersonality* corpus. Our team is also unable to make concrete explanations regarding the formation of coalescing like adjacent soap bubbles in *myPersonality* multiclass distribution. While the exact mathematical inference for this phenomenon is not fully understood, it also may be due to the effects caused by size of the corpus towards perplexity, and such projection also can be seen in the medical domain [62]. The further experiment is required to derive a concrete interpretation and justification regarding the phenomena illustrated in Figure 5. To summarize, the t-SNE plotting illustrates that larger documents could carry more diversity in nature, and the overlapping relationships among the topics indicate that some statistically correlated words of a category could also be relevant to other categories to a certain extent.

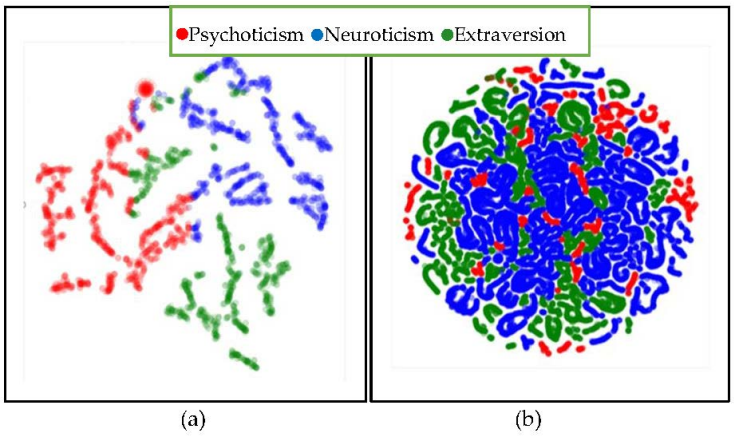

Figure 5. *t-SNE* visualization for multiclass distribution of *myPersonality* (**a**) and *Sentiment140* (**b**).

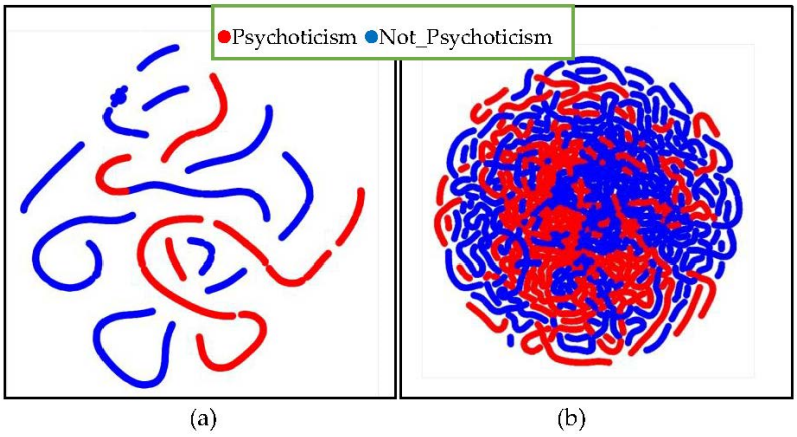

Figure 6. *t-SNE* visualization for one vs. all distribution of *myPersonality* (**a**) and *Sentiment140* (**b**).

5.3. Extrinsic Evaluation

Previously, we have demonstrated the capability of SLDA to model the unsupervised contents in the contexts of psycholinguistics. However, the motive of topic modeling is to seek the underlying topics embedded in a document in the manner of probability and without explicit training, which may produce ambiguous or less relevant topics to some extent [8,11]. In this regard, many existing studies suggested and used topic modeling to pre-model the unstructured data through an unsupervised or semi-supervised manner before feeding the output to supervised learning [9,11,42,63]. Based on this pipeline, the data science community used machine learning techniques to measure the discriminative and generalization power of the learning algorithm in order to posit how well the topic model has identified latent semantic structure that persists over time in the training document [9,64].

Typically, the research community accesses the quality of the learned topics φ by examining the predictive value of learning algorithms on unseen documents partition through a cross-validation mechanism [65]. Empirically, the considerable ambiguities in topic modeling will substantially deteriorate the discriminative power and predictive value of classifiers due to false positive and false negative caused by seed words that match unrelated texts and confound the true association between topics and words in the machine learning process [43,64]. In this sense, our team extended the experiments to measure the predictive power of machine classifiers to serve as a basis to postulate that SLDA generated meaningful topics to some extent based on PEN model traits. Because there are no adjacent personality computing works to be benchmarked with this study due to the variation in personality model, we decided to use several prominent off-the-shelf supervised classifiers namely Sequential Minimal Optimization (SMO), Naive Bayes (NB), C4.5, K-Nearest Neighbor (KNN), Random Forest (RF)., and AdaBoost [32] to predict the ground-truth topics generated by SLDA. The six classifiers were selected based on the following consideration: (1) Popularity in the area of supervised-based personality detection [1]; (2) The high dimensional unstructured data might not be linearly separable.

This paragraph will briefly introduce all the classifiers stated above. SMO is a variant of SVM that optimizes a problem iteratively by splitting it into a series of subproblems using two Lagrange Multipliers [66]. Because SMO uses the kernel function to compute the dot product of the vectors in feature space, this study has applied the popular kernel function termed Polynomial [27,28] to transform the nonlinear space into linearly separable space. NB is a probabilistic classifier based on the Bayes rules with strong independence assumptions, which determine whether the presence or absence of a certain feature of a class is related to the presence or absence of other features. The C4.5 classifier is a variant of the Decision Tree Algorithm based on the Iterative Dichotomizer learn to determine the target values of new samples based on the various attribute values in the data [67]. Generally, KNN processes the classification by predicting new data points from known data

points based on the nearest k points where the k is evaluated using the distance function. In this study, the size of k was fixed to l, and the Euclidean Distance Function was embraced to evaluate the performance of KNN. RF is a supervised text classification algorithm dealing with the high dimensional text data and philosophically operate by building multitude binary-based decision trees $\{RF(x,\ \theta j),\ j = 1 \ldots, \}$, where RF is a meta estimator, x is an input vector, and θj are lists of the independent identically distributed random vectors [68]. On the other hand, the AdaBoost (Ada) is a meta-learning classifier that makes predictions based on constructing and adjusting the weights from multiple weak classifiers [69].

During preprocessing, the attributes in the training set were transformed into unigram, bigram, and trigram using n-gram mechanism and limited the minimum number of attribute frequency to 3 as well as vocabulary size to 3000, 1500 for unigram features of *Sentiment140* and *myPersonality*, respectively. In the case of bigram and trigram, we fixed the vocabulary size to 1500 and 1000 for respective *Sentiment140* and *myPersonality* corpuses. The values were fixed based on trial and error to eliminate issues caused by sparsity. The probability of the attribute a in the given instances was approximated based on (8), and the conditional probabilities of the n-gram language models were computed using (9), where the size of bigram and trigram attributes represented as $n = 2$, and $n = 3$ respectively.

$$p(a_1, \ldots, a_z) \approx \prod_{i=1}^{z} p(a_i | a_{i-(n-1)},\ \ldots,\ a_{i-1}) \tag{8}$$

$$p(a_i | a_{i-(n-1)},\ \ldots,\ a_{i-1}) = \frac{Count(a_i \left| a_{i-(n-1)},\ \ldots,\ a_{i-1},\ a_i \right.)}{Count(a_i \left| a_{i-(n-1)},\ \ldots,\ a_{i-1} \right.)} \tag{9}$$

Similar to the previous topic modeling process, this supervised learning also treated the problem as multiclass and one vs. all classification. In this investigation, the training set was comprised of n pairs of numbers, i.e., (x_1, y_1), (x_2,y_2), ... , (x_n,y_n) for multiclass and one vs. all distributions, where each x_n was a measurement referring to a single data point, whereas y_n was the label for that point. Then, the Synthetic Minority Oversampling (SMOTE) class balancing technique with a synthetical seed number of 5 was applied on both distributions of *myPersonality* and *Sentiment140* to improve the machine learner's performance due to the high distribution variation nature of the classes. In order to measure the consistency of the topics predicted by SLDA, the training documents were partitioned according to 10-fold cross-validation. In this experiment, we used a personal computer with the specification of 8 GB RAM and Intel(R) Core (TM) i7-3610QM CPU@ 2.30 GHz processor.

5.3.1. Evaluation Metrics

We chose sensitivity/recall (10), precision (11), F1 score (12), Area Under Curve (AUC) and Geometric Mean (GM) [70–72] (14) as evaluation metrics because of the popularity in personality detection experiments [17,32,34,36], which were quantifiable to binary and multiclass experiments [70,71,73] and were suitable for our asymmetric distribution datasets [74,75]. AUC and GM metrics are also suitable for the class imbalance problem [71] as the SLDA generated datasets seems to be imbalanced. AUC is computed by plotting the recall (10) against the False Positive Rate (FPR) (13) at various threshold settings. Literally, the better generalization by the heuristic models illustrated through the high score of the three metrics will provide additional merit to the topics modeled by SLDA. Machine learning studies apply the statistical metric known as the confusion matrix to measure the strength of an algorithm in solving the given problem automatically. The confusion matrix is comprised of four elements, namely true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In order to measure the time taken by the classifiers to build

the model, we followed the time complexity notion stated in [66] for SMO, ref. [69] for NB and KNN, ref. [75] for DT, ref. [76] for RF, and ref. [77] for AdaBoost classifier.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{12}$$

$$FPR = \frac{FP}{TN + FP} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

$$GM = \sqrt{Recall \cdot Specificity} \tag{15}$$

5.3.2. Machine Learning Classification

Table 9 lists the machine learning classification results for multiclass distribution where the training document was comprised of all the instances categorized under Psychoticism, Extraversion and Neuroticism. In the viewpoint of language models, our experiments showed that the unigram feature channeled better discriminative information to the classifiers to attain up to 0.995 of recall, precision and F1, respectively, compared to bigram and trigram features from *Sentiment140*. In the same manner, the unigram features also well discriminated the classes on *myPersonality* where the SMO achieved up to 0.979 of F1 and 0.968 of GM scores. The SMO classifier also achieved the highest AUC score (0.989) for the classification of *Sentiment140* unigram features. The highest AUC score obtained by SMO seems to be in a similar range with other Big 5 based personality-based recognition studies [78,79] and other nontextual affecting computing [80,81]. The size of the training documents extensively influenced the vocabulary and efficiency of the algorithms. Based on this factor, the SMO seems to be preferable because the classifiers registered insignificantly better predictions with reasonable time processing compared to the other three classifiers. The steady prediction of SMO agrees with the findings of previous personality recognition studies that have examined the representation of English messages [82,83]. As SMO aims to forward the Lagrange multipliers or alphas that satisfy the actual inherent learning process by identifying the support vectors [66], the transformation of inputs by the kernel function and optimization of subproblems minimized the computational cost for a large volume of the matrix.

Apart from SMO, the KNN also registered steady performances using bigram and trigram features extracted from both *myPersonality* and *Sentiment140* datasets. The steady performances of KNN to predict the topics without actual training by determining the closeness amongst the d using Euclidean Distance also asserted the presence of reliable latent structures that are meaningful to the corresponded topics generated by SLDA. Furthermore, NB recorded negligible recall (0.284), AUC (0.376) and GM (0.264) for the classification using *Sentiment140* bigram features and the similar effects also minorly featured on the experiment using *myPersonality* unigram features. The low GM value indicates that the NB poorly predicts the positive instances, and it even seems to correctly classify the negative instances. The further investigation on the NB prediction showed that a large rate of misclassification occurred on Extraversion and Neuroticism instances because it cannot penalize the false negative from the majority classes and possibly because of the unfathomable reason of independent assumption set by the classifier. In other words, when NB assumed that features in the vocabulary are not independent, then each attribute exclusively contributes to the discriminative power towards the learning process. The assumption caused the magnitude of the weights for classes with strong feature dependencies to be larger than

for classes with weak word dependencies. Thus, the magnitude of the weights for weak attributes will indirectly channel ambiguous information to the prediction process [84]. The AdaBoost classifier also comparatively achieved better predictions using bigram and trigram attributes. The good performance of AdaBoost on bigram features was also reported in detecting offensive language [82].

Table 9. Multiclass classification performance.

***myPersonality* (Multiclass)**							
Language Model	**ML Classifier**	**Recall**	**Precision**	**F1**	**AUC**	**GM**	**Time Complexity**
Unigram	SMO	0.979	0.979	0.979	0.970	0.968	4.45 s
	NB	0.739	0.897	0.810	0.764	0.819	2.05 s
	C4.5	0.967	0.965	0.966	0.960	0.965	92.54 s
	KNN	0.939	0.932	0.935	0.930	0.930	0.01 s
	RF	0.979	0.961	0.970	0.965	0.966	71.40 s
	Ada	0.968	0.962	0.965	0.954	0.964	9.91 s
Bigram	SMO	0.899	0.883	0.888	0.889	0.890	0.28 s
	NB	0.891	0.895	0.893	0.885	0.890	0.28 s
	C4.5	0.888	0.887	0.857	0.884	0.886	0.53 s
	KNN	0.897	0.882	0.887	0.888	0.889	0.28 s
	RF	0.895	0.875	0.885	0.887	0.887	99.33 s
	Ada	0.893	0.895	0.894	0.888	0.892	0.24 s
Trigram	SMO	0.930	0.920	0.925	0.921	0.924	0.05 s
	NB	0.929	0.918	0.923	0.920	0.921	0.05 s
	C4.5	0.918	0.914	0.916	0.911	0.915	0.05 s
	KNN	0.930	0.920	0.916	0.924	0.923	0.05 s
	RF	0.930	0.920	0.916	0.924	0.922	8.46 s
	Ada	0.920	0.920	0.920	0.916	0.919	0.04 s
***Sentiment140* (Multiclass)**							
Language Model	**ML Classifier**	**Recall**	**Precision**	**F1**	**AUC**	**GM**	**Time Complexity**
Unigram	SMO	0.995	0.995	0.995	0.989	0.989	172.43 s
	NB	0.841	0.939	0.887	0.814	0.863	24.68 s
	C4.5	0.991	0.991	0.991	0.981	0.984	3379.83 s
	KNN	0.967	0.966	0.966	0.979	0.956	15.04 s
	RF	0.986	0.986	0.986	0.983	0.978	2589.97 s
	Ada	0.990	0.990	0.990	0.984	0.982	194.9 s
Bigram	SMO	0.959	0.958	0.958	0.954	0.987	654.59 s
	NB	0.284	0.815	0.421	0.376	0.264	64.33 s
	C4.5	0.952	0.951	0.951	0.952	0.944	12,409.02 s
	KNN	0.954	0.946	0.946	0.953	0.941	27.45 s
	RF	0.948	0.946	0.947	0.948	0.940	5800.23 s
	Ada	0.954	0.951	0.952	0.944	0.949	157.20 s
Trigram	SMO	0.947	0.947	0.947	0.946	0.941	2.34 s
	NB	0.947	0.935	0.929	0.937	0.939	26.01 s
	C4.5	0.947	0.947	0.947	0.939	0.944	690.05 s
	KNN	0.947	0.929	0.938	0.934	0.940	0.05 s
	RF	0.942	0.942	0.942	0.921	0.941	795.85 s
	Ada	0.945	0.945	0.945	0.922	0.942	126.08 s

In the comparison of tree-based classifiers, although the C4.5 and RF technically gave similar performances for both datasets, the time complexities to predict the training classes by the two classifiers were very complex for *Sentiment140* as the dataset was comprised of high dimensional attributes. For instance, the use of unigram and bigram features from *Sentiment140* caused the C4.5 model to take 3379.83 s and 12,409.02 s, respectively, to accomplish the prediction process. Regarding the processing time inefficiency, C4.5 built the network with 195 nodes and 98 leaves for the unigram feature, while it used 105 nodes with 53 leaves using bigram attributes. Meanwhile, for *myPersonality*, the model only generated a tree with 31 nodes and 16 leaves for unigram and 13 nodes with 7 leaves for bigram. The longer time complexity manifested by those tree-based classifiers for *Sentiment140* indicates the inefficiency of the models for the training on big datasets, although the inductive network can handle the high dimensional data besides yielding a promising prediction score. Therefore, it is practical to use other algorithms such as SMO, KNN or AdaBoost instead of C4.5 and RF because the tree-based networks may encounter the bloating issue for high dimensional data [85].

Overall, the identical and good performance of the classifiers on most of the experiments using multiclass topics indicates that the models can capture and well generalize the underlying structure encoded in the document resampled through cross-validation. This can also be affirmed by observing the GM metric where most of the classifiers obtained very high GM values. In this regard, we can make the inference that topics generated by SLDA are consistent and structurally correlated with the traits because those classifiers can gain better insights about the nature of the data regardless of significant biases (i.e., overfitting). Consequently, this experiment was directed to examine the one vs. all distribution because most of the off-the-shelf classifiers performed well using multiclass distributions.

Table 10 lists the performance capabilities of the six machine learners for one vs. all distribution. In this experiment, all the classifiers registered more than 0.9 of recall, F1, AUC and GM except for the prediction by NB using *myPersonality* unigram and *Sentiment140* bigram features. As seen previously, the poor performance of showed by NB on the *Sentiment140* bigram feature also persisted on one vs. all classification where the model merely achieved an AUC of 0.461, 0.105 of recall and 0.126 of GM. This indicates that NB performance using *Sentiment140* bigram features is worse than the previous multiclass classification. The unsatisfactory performance may be due to the concept of independent assumption of terms that diverted the magnitude of the weights among the features in the training classes. We believed that the underperformance of NB could be improved by embracing the ensemble technique in future. The SMO and AdaBoost presented a good, steady and consistent performance in all the n-gram features from *myPersonality* and *Sentiment140*. Similar to multiclass classification, SMO also achieved highest AUC (0.997) and GM (0.998) on *myPersonality* unigram features. Our literature review showed that the training using support vector machine-based classifiers and a SMOTE class balancing technique seems to yield better GM scores in certain research areas [86,87].

In contrast with multiclass classification, C4.5 and KNN performed insignificantly better than SMO using bigram attributes of *myPersonality* and trigram features of *Sentiment140*. Nevertheless, C4.5 and RF were less efficient based on the time complexity in comparison with KNN. The analysis of *Sentiment140* using bigram attributes showed that the C4.5 classifier took 15,467.56 s to build the network. Upon inspecting the size of the tree generated by C4.5, the model was comprised of 367 nodes with 184 leaves. Therefore, the bloated size of the C4.5 tree caused the model to take additional time to execute pruning and make the decisions based on ratio of information gain. Besides the pitfall, C4.5 and RF also required more resources to make the decisions.

Table 10. One vs. all classification performance.

***myPersonality* (One vs. All)**							
Language Model	**ML Classifier**	**Recall**	**Precision**	**F1**	**AUC**	**GM**	**Time Complexity**
Unigram	SMO	0.999	0.999	0.995	0.997	0.998	1.07 s
	NB	0.822	0.958	0.885	0.794	0.781	1.47 s
	C4.5	0.992	0.999	0.999	0.991	0.995	11.41 s
	KNN	0.992	0.993	0.992	0.992	0.995	0.05 s
	RF	0.996	0.996	0.996	0.985	0.994	35.66 s
	Ada	0.945	0.945	0.945	0.945	0.939	0.26 s
Bigram	SMO	0.969	0.964	0.966	0.966	0.955	4.01 s
	NB	0.942	0.939	0.940	0.938	0.922	6.16 s
	C4.5	0.965	0.963	0.964	0.961	0.919	36.94 s
	KNN	0.946	0.947	0.946	0.947	0.931	0.09 s
	RF	0.940	0.964	0.952	0.931	0.936	141.61 s
	Ada	0.964	0.959	0.961	0.953	0.961	6.67 s
Trigram	SMO	0.998	0.998	0.998	0.996	0.997	0.04 s
	NB	0.998	0.998	0.998	0.995	0.995	0.19 s
	C4.5	0.989	0.989	0.988	0.983	0.985	1.04 s
	KNN	0.998	0.998	0.998	0.996	0.997	0.04 s
	RF	0.998	0.998	0.998	0.996	0.997	42.95 s
	Ada	0.998	0.998	0.998	0.995	0.997	0.46 s
***Sentiment140* (One-vs.-All)**							
Language Model	**ML Classifier**	**Recall**	**Precision**	**F1**	**AUC**	**GM**	**Time Complexity**
Unigram	SMO	0.996	0.996	0.996	0.989	0.996	295.83 s
	NB	0.950	0.960	0.955	0.948	0.948	48.56 s
	C4.5	0.992	0.992	0.992	0.973	0.991	3345.64 s
	KNN	0.989	0.990	0.989	0.968	0.987	0.05 s
	RF	0.989	0.989	0.989	0.967	0.987	2970.03 s
	Ada	0.994	0.994	0.993	0.990	0.993	1865.02 s
Bigram	SMO	0.958	0.955	0.956	0.826	0.958	546.02 s
	NB	0.105	0.850	0.187	0.461	0.126	120.02 s
	C4.5	0.941	0.938	0.939	0.874	0.925	15,467.56 s
	KNN	0.942	0.937	0.939	0.816	0.921	965.27 s
	RF	0.940	0.938	0.939	0.856	0.918	5634.67 s
	Ada	0.944	0.944	0.944	0.861	0.922	1259.64 s
Trigram	SMO	0.949	0.921	0.935	0.933	0.941	24 s
	NB	0.950	0.952	0.951	0.940	0.939	29.02 s
	C4.5	0.950	0.952	0.951	0.940	0.942	128.65 s
	KNN	0.950	0.946	0.948	0.939	0.942	347.28 s
	RF	0.905	0.903	0.904	0.941	0.912	504.24 s
	Ada	0.947	0.947	0.947	0.939	0.941	630.32 s

Our observation also showed that the unigram attributes channeled a better discriminative power to the classifiers to predict the training instances compared to bigram model for both *myPersonality* and *Sentiment140*. The better informative representation value channeled by unigram attributes has been reported in other text-based personality investigations [88]. The salient unigram attributes that have enhanced the prediction process are also similar to approaches taken by psychologists. In general, psychologists would observe the words (e.g., adjectives) used by people to identify their personality characteristics based on the assumption that individual differences are encoded in the natural language, where the noticeable markers are more likely to be expressed in single words. This representation is known as the "Lexical Hypothesis" [89]. In the viewpoint of GM, the very good results obtained in the *Sentiment140* trigram feature indicate the central tendency of the classifiers and strengthen the prediction performance of the classifiers. Indirectly, it indicates that most of the classifiers can well distinguish the traits classes labeled by SLDA.

To summarize, the good performances illustrated by the classifiers in most of the one vs. all experiments indicate that SLDA well distinguished the latent features embedded in Psychoticism and not_Psychoticism texts. Instead of merely modeling the topics using prior knowledge, SLDA also reduced the dimensionality and eliminates the noises so the internal structural complexities embedded in the documents can be minimized to some extent. In this sense, the SLDA transformed the high dimensional unstructured data to lower-dimensional so that supervised classifiers can well predict the classes using the high-quality discriminative features. Therefore, we can deduce that promising prediction by the classifiers in both multiclass and one vs. all classifications indicates that topics discovered by SLDA according to predefined psychological themes were meaningful and consistent.

5.3.3. Confusion Matrix

The confusion matrix tabular (Figures 7 and 8) presented the probability of true and false classification of the classifiers that attained reasonable time processing on each language model. We randomly picked several confusion matrixes of the classifiers to discuss in this section. As shown in Figure 8, the prediction probability of 0.0000 depicted in several classifications in fact implied that the number of false cases is relatively too small. For instance, the classification by SMO using trigram attributes from *Sentiment140* (multiclass) only misclassified two Psychoticism instances as Extraversion.

As disclosed in the previous section, we can stress that unigram features that represent the word contexts channeled better statistical properties to the prediction compared to bigram and trigram attributes. Again, this illustration is also in the line with our earlier finding that claimed there are overlapping structural relationships between the instances of Extraversion and Neuroticism in our dataset despite the two traits being rooted in opposite sentiment polarities. However, further inspection on minority class distribution shows that the prediction probability of false positive caused by the weights of bigram and trigram is significantly larger than true positive prediction. For instance, the prediction by C4.5 using bigram features in *myPersonality* (one vs. all) yielded a false positive probability of 0.0336 that is equivalent to the misclassification of 71 instances where the figure is substantially larger than the probability of true positive (0.0118) or 202 instances. In this regard, we can infer that respective machine learning predictions substantially contributed by the features and weights derived from the not_psychoticism class. This type 1 error occurred due to the asymmetric distribution between the classes that lead to the data distortion. This can be seen obviously were the instances from not_Psychoticism dominated the majority class region.

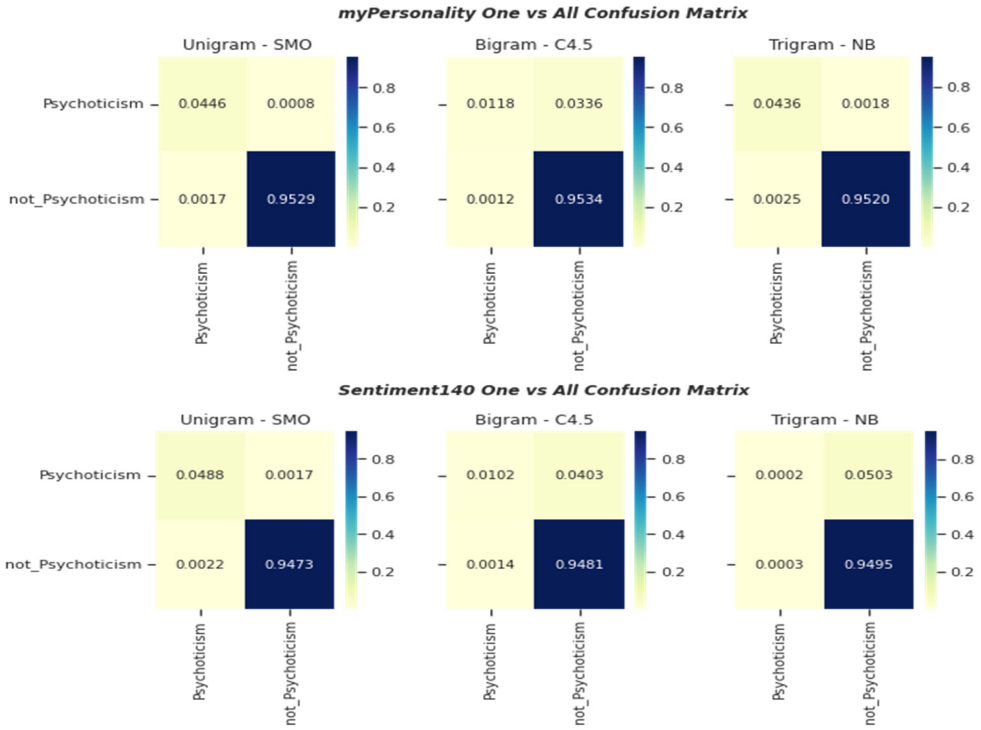

Figure 7. Confusion Matrix of one vs. all Classification.

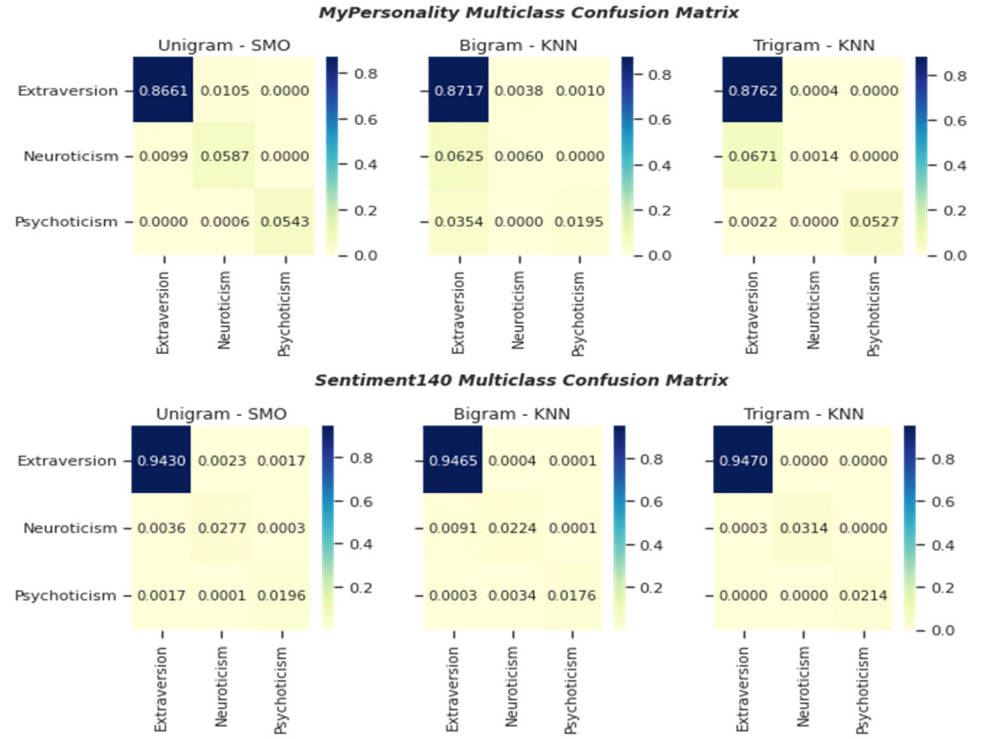

Figure 8. Confusion Matrix of Multiclass Classification.

However, the imbalance ratio between Psychoticism and not_Psychoticism distribution also approximately implied the real-world statistics where the number of people with psychosis disorders is relatively rare and estimated to be 1% of the world population [90]. The domination of Extraversion instances (multiclass) in the majority class region also promoted the universal positivity biases of human language that simulated the access of Extraversion characteristics and promoted prosocial behaviors in SN sites [54]. On analysis

of true positive of majority class, the probability of unigram features was insignificantly lower than the probability of bigram features. For example, SMO correctly predicted the Extraversion instances up to 0.9465 or 94.65% using bigram features from *Sentiment140* (multiclass) compared to 0.9430 or 94.30% using unigram features from the same dataset. In addition, the probability of false positive unigram features from majority class is quite larger than the probability of false positive of bigram and trigram. This effect can be seen absolutely in multiclass classification. As such, the mixed sensitivity shown by the language models can implicitly support or undermine the classifier's performance, although the distribution of the classes is simplified. We believed that additional experiments using other evaluation metrics may provide additional insights regarding the structure of data and capability of these machine learning classifiers.

6. Threat to Validity

Several factors may affect the validity of this experiment. First and foremost, there may be an argument raised regarding the scale set to bridge the traits and sentiment through valence and criteria fixed for trait correlations between the Big 5 and PEN model. As noted earlier, we set the scale and criteria based on the statistical findings from the psychological studies [14,21] and general perceptional towards the sentiments of emotional words [23]. Although it is intractable to define the exact relationships between the seed words and each characteristic stated in Table 1, we made the assumptions that there are strong relationships existing between emotional states of human behavior and polarity as well subjectivity encoded in natural language. Based on the common understanding, we decided to manually inspect the reliability of topics generated by SLDA using the information stated in Psychological indicators provided along with *myPersonality* dataset and applied the same settings on *Sentiment140* to conduct for cross-domain analysis. Next, we validated the topics generated by SLDA on *Sentiment140* based on our understanding of the semantics of the instances, and it should not necessarily be correlated with the actual traits of the writer because the human personalities are not mutually exclusive [21]. Furthermore, our assumptions to link the seed words with the trait classes based on one-to-one relationships might not represent the actual contextual meaning of the training instances. Then, the validity of the topics generated by SLDA could be ambiguous or inaccurate to a small fraction of extents due to the dynamicity of SN messages and effects of negation words.

7. Limitation and Future Direction

Although SLDA shows promising performance in modeling the SN texts based on personality theme, this probabilistic model also has several limitations that can be improved in the future. Because SLDA required auxiliary knowledge, it is vulnerable to identifying contents written using figurative languages such as irony and sarcasm. Our team is also susceptible to the capability of SLDA in classifying the long contents because the degree of information encoded in such instances is complex and dynamic. Therefore, we are planning to improve and train the SLDA with other hard real-world datasets especially in the contexts of long texts and personality.

Despite the fact that most of the machine learning classifiers performed well on the datasets, we are also interested in extending this work to investigate the predictive performance of other machine learning models such as Extreme Gradient Boosting [91] and fusion deep learning model [92] to gain additional insights about the data and the nature of the algorithms. Apart from cosine similarity, we are also interested to explore the nature of the dataset produced by SLDA from the perspective of other metrics such as Euclidean Distance.

On the other hand, the promising performance of the machine learners using bigram and trigram attributes also sparked the curiosity to investigate the effects caused by contiguous sequences of tokens during the modeling process. As another limitation of SLDA to merely exploit the degree of co-occurrence of word probability between the regular

attributes and seed words to identify the relevant topics, we believe that adjacent words embedded within an instance could offer other useful information encapsulated in the text. Furthermore, the proposed SLDA architecture could be improved further by integrating computational components such as Attention to exploit the contextual information embedded in the texts. In this regard, we are recommending the future study to explore the effects of adjacent attributes along with seed words using biterm or contextual approaches such as window size to better discover the abstracts embedded in the document.

Our proposed framework is also practical to be applied in author profiling tasks because the goal of the task is to predict or infer as much knowledge as possible about an unknown author through analyzing a specific text written by him/her. Furthermore, the framework also could be applied in the end-to-end framework as our t-SNE visualization illustrated the clustering of similar data points in certain regions. Eventually, we hope our experiment shed light on the tendency of perceptional-based psycholinguistic seed words in guiding the modeling of texts based on personality traits through the dataless technique. We also encourage future studies to devise new modeling strategies as well as further explore the applicability of human traits such as Psychoticism in the context of digital forensics so that it can help to reduce the circle of suspects.

8. Conclusions and Discussion

This paper proposed a mechanism to incorporate a small set of seed words collected through public perception into a dataless unsupervised model called SLDA to automatically generate topics aggregated to PEN model traits. The proposed framework exposed the functions of sentiments to bridge the gap between linguistics and personality and the practicality of SLDA to modeled SN messages based on human traits. The satisfactory performance of SLDA against GNMF and other nonseeded models exposed the capability of our proposed method to penetrate the personality–emotional structures embedded in SN texts. The intrinsic evaluations presented various information and insights such as the nature of topics modeled by SLDA that could be used for further experiments. Furthermore, the good performances shown by the machine learning classifiers also explicitly strengthen the reliability and consistency of the topics predicted by the SLDA. The supportive expressions of the unigram in the extrinsic evaluations also strengthen our earlier premise that hypothesized that a small set of seed words is enough to supervise the modeling process and that emphasized the notion that individual differences encoded in language are more likely to be articulated by single words. On other hand, the overlapping characteristics between Extraversion and Neuroticism showed in t-SNE analysis and confusion matrix emphasized the reliability of intrinsic and extrinsic evaluations, where both evaluations yielded identical results. Furthermore, our experiments indicate that the properties of Psychoticism traits have the potential to be applied in detecting the trustworthiness of social networks users. This is practical because the effects of trust strongly correlated to the behavior of human beings and extreme levels of negativity and cunningness articulated through linguistics could expose the real intentions of social networks users. Therefore, we believe that further investigation on the exposure of Psychoticism in natural language may reveal additional insights that could be used as a source of information to detect the level of trust between a pair of users.

Comprehensively, the entire experiments not only demonstrated the applicability of perceptional seed words using the unsupervised technique to model the topics as well as to measure the extent of machine learners to predict the classes, but it also contributed to the traits-related wordlists collected from our preliminary study and a dataset labeled according to PEN model. Because there is no domain knowledge and training corpus based on the PEN model, the wordlists and datasets could be used in future to explore and exploit the psycholinguistics elements encoded in natural texts. The *Sentiment140* datasets modeled by SLDA can be downloaded from (https://studentusm-my.sharepoint.com/:f:/g/personal/saravanan_18_student_usm_my/EsXxLhTqWFpGnTv1mCS4RbYBAVTJOVGGtS9HpTqhZ3f1A?e=5TfiPh)

(last accessed on 2 March 2022) for empirical purposes, whereas the PEN model wordlists were already published in our preliminary study.

Author Contributions: Conceptualization, S.S., N.H.A.H.M. and M.H.H.; Formal analysis, S.S.; Investigation, S.S.; Methodology, S.S., N.H.A.H.M. and M.H.H.; Supervision, N.H.A.H.M. and M.H.H.; Visualization, S.S.; Writing—original draft, S.S.; Writing—review & editing, S.S., N.H.A.H.M. and M.H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Research University Grant (RUI) 2017, Division of Research and Innovation, Universiti Sains Malaysia (1001/PKOMP/8011035).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Two publicly available dataset namely *my Personality* (the my Personality corpus provider had decided to stop providing the dataset since April 2018) and *Sentiment140* (https://www.kaggle.com/kazanova/sentiment140) (accessed on 21 March 2014) were used for this study. However, permission was acquired from *myPersonality* data provider to publish this work since the experiment was conducted before 2018 and the thesis was already published. We also shared the corpus generated by SLDA for further empirical studies.

Acknowledgments: The research team want to thank Universiti Sains Malaysia for providing fund and technical supports for this experiment.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mehta, Y.; Majumder, N.; Gelbukh, A.; Cambria, E. Recent trends in deep learning based personality detection. *Artif. Intell. Rev.* **2020**, *53*, 2313–2339. [CrossRef]
2. Boduszek, D.; McLaughlin, C.; Hyland, P. Criminal attitudes of ex-prisoners: The role of personality, anti-social friends and recidivism. *Int. J. Crim.* **2011**, *9*, 1–10.
3. Kamaluddin, M.R.; Shariff, N.S.M.; Othman, A.; Ismail, K.H.; Saat, G.A.M. Linking psychological traits with criminal behaviour: A review. *ASEAN J. Psychiatry* **2015**, *16*, 13–25.
4. Wang, Z.; Wu, C.; Zhe, W.; Niu, X.; Wang, X. SMOTETomek-based resampling for personality recognition. *IEEE Access* **2019**, *7*, 129678–129689. [CrossRef]
5. Zha, D.; Li, C. Multi-label dataless text classification with topic modeling. *Knowl. Inf. Syst.* **2019**, *61*, 137–160. [CrossRef]
6. Wang, D.; Thint, M.; Al-Rubaie, A. Semi-supervised latent dirichlet allocation and its application for document classification. In Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China, 4–7 December 2012.
7. Ferner, C.; Havas, C.; Birnbacher, E.; Wegenkittl, S.; Resch, B. Automated seeded latent dirichlet allocation for social media based event detection and mapping. *Information* **2020**, *11*, 376. [CrossRef]
8. Jin, Y.; Bhatia, A.; Wanvarie, D. Seed word selection for weakly-supervised text classification with unsupervised error estimation. *arXiv* **2021**, arXiv:2104.09765.
9. Kherwa, P.; Bansal, P. Topic Modeling: A Comprehensive Review. *EAI Endorsed Trans. Scalable Inf. Syst.* **2020**, *7*, e2. [CrossRef]
10. Toubia, O.; Iyengar, G.; Bunnell, R.; Lemaire, A. Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *J. Mark. Res.* **2018**, *56*, 18–36. [CrossRef]
11. Li, C.; Xing, J.; Sun, A.; Ma, Z. Effective document labeling with very few seed words: A topic model approach. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management—CIKM'16, Indianapolis, IN, USA, 24–28 October 2016.
12. Li, X.; Li, C.; Chi, J.; Ouyang, J.; Li, C. Dataless text classification: A topic modelling approach with document manifold. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management—CIKM'18, Torino, Italy, 22–26 October 2018.
13. Li, C.; Chen, S.; Xing, J.; Sun, A.; Ma, Z. Seed-guided topic model for document filtering and classification. *ACM Trans. Inf. Syst.* **2019**, *37*, 1–37. [CrossRef]
14. Lynam, D.R.; Miller, J.D. On the ubiquity and importance of antagonism. In *Handbook of Antagonism*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 1–24.
15. Ghafari, S.M.; Beheshti, A.; Joshi, A.; Paris, C.; Yakhchi, S.; Jolfaei, A.; Orgun, M.A. A dynamic deep trust prediction approach for online social networks. In Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia, Chiang Mai, Thailand, 30 November–2 December 2020; pp. 11–19.

16. De Meo, P.; Musial-Gabrys, K.; Rosaci, D.; Sarnè, G.M.L.; Aroyo, L. Using centrality measures to predict helpfulness-based reputation in trust networks. *ACM Trans. Internet Technol.* **2017**, *17*, 8. [CrossRef]
17. Alkhamees, M.; Alsaleem, S.; Al-Qurishi, M.; Al-Rubaian, M.; Hussain, A. User trustworthiness in online social networks: A systematic review. *Appl. Soft Comput.* **2021**, *103*, 107159. [CrossRef]
18. Argamon, S.; Dhawle, S.; Koppel, M.; Pennebaker, J.W. Lexical predictors of personality type. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, St. Louis, MI, USA, 8–12 June 2005; pp. 1–16.
19. Park, G.; Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.L.; Kosinski, M.; Stillwell, D.J.; Ungar, L.H.; Seligman, M.E. Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* **2015**, *108*, 934–952. [CrossRef] [PubMed]
20. Ruch, W.; Wagner, L.; Heintz, S. Humor, the PEN model of personality, and subjective well-being: Support for differential relationships with eight comic styles. *Riv. Ital. di Studi sull'Umorismo* **2018**, *1*, 31–44.
21. Sáez, Y.; Navarro, C.; Mochón, M.A.; Isasi, P. A system for personality and happiness detection. *Int. J. Interact. Multimed. Artif. Intell.* **2014**, *2*, 7. [CrossRef]
22. Sagadevan, S.; Malim, N.H.A.H.; Husin, M.H. Sentiment valences for automatic personality detection of online social networks users using three factor model. *Procedia Comput. Sci.* **2015**, *72*, 201–208. [CrossRef]
23. Mohammadi, G.; Vinciarelli, A. Automatic personality perception: Prediction of trait attribution based on prosodic features extended abstract. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 484–490.
24. Finn, E. Swearing: The good, the bad & the ugly. *ORTESOL J.* **2017**, *34*, 17–26.
25. Nielsen, F.A. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv* **2011**, arXiv:1103.2903.
26. Hoekstra, R.; Vugteveen, J.; Warrens, M.J.; Kruyen, P.M. An empirical analysis of alleged misunderstandings of coefficient alpha. *Int. J. Soc. Res. Methodol.* **2019**, *22*, 351–364. [CrossRef]
27. Oberlander, J.; Nowson, S. Whose thumb is it anyway? Classifying author personality from weblog text. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, 17–18 July 2006; pp. 627–634.
28. Celli, F.; Pianesi, F.; Stillwell, D.; Kosinski, M. Workshop on computational personality recognition: Shared task. In Proceedings of the International AAAI Conference on Web and Social Media, Cambridge, MA, USA, 8–11 June 2013; Volume 7.
29. Iacobelli, F.; Gill, A.J.; Nowson, S.; Oberlander, J. Large scale personality classification of bloggers. In *Affective Computing and Intelligent Interaction*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 568–577.
30. Junior, R.A.P.; Inkpen, D. Using cognitive computing to get insights on personality traits from twitter messages. In *Advances in Artificial Intelligence*; Mouhoub, M., Langlais, P., Eds.; Canadian AI 2017. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10233. [CrossRef]
31. Sharma, S. Predicting Employability from User Personality Using Ensemble Modelling. Master's Thesis, Thapar University, Patiala, India, 2015.
32. Kunte, A.V.; Panicker, S. Using textual data for personality prediction:a machine learning approach. In Proceedings of the 2019 4th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 21–22 November 2019; pp. 529–533.
33. Saini, M.; Sharan, A. Ensemble learning to find deceptive reviews using personality traits and reviews specific features. *J. Digit. Inf. Manag.* **2017**, *12*, 84–94.
34. Levitan, S.I.; Levitan, Y.; An, G.; Levine, M.; Levitan, R.; Rosenberg, A.; Hirschberg, J. Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, San Diego, CA, USA, 12–17 June 2016.
35. Agarwal, B. Personality detection from text: A review. *Int. J. Comput. Syst.* **2014**, *1*, 1–4.
36. Mulay, P.; Joshi, R.R.; Misra, A.; Raje, R.R. Detection of personality traits of sarcastic people (PTSP): A social-IoT based approach. In *Intelligent Systems Reference Library*; Springer International Publishing: Cham, Switzerland, 2019; pp. 237–261.
37. Liu, Y.; Wang, J.; Jiang, Y. PT-LDA: A latent variable model to predict personality traits of social network users. *Neurocomputing* **2016**, *210*, 155–163. [CrossRef]
38. Moreno, D.R.J.; Gomez, J.C.; Almanza-Ojeda, D.-L.; Ibarra-Manzano, M.-A. Prediction of personality traits in twitter users with latent features. In Proceedings of the 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), Cholula, Mexico, 27 February–1 March 2019; pp. 176–181.
39. Kwantes, P.J.; Derbentseva, N.; Lam, Q.; Vartanian, O.; Marmurek, H.H. Assessing the Big Five personality traits with latent semantic analysis. *Pers. Individ. Differ.* **2016**, *102*, 229–233. [CrossRef]
40. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
41. Chen, X.; Xia, Y.; Jin, P.; Carroll, J. Dataless text classification with descriptive LDA. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
42. Vendrow, J.; Haddock, J.; Rebrova, E.; Needell, D. On a guided nonnegative matrix factorization. *arXiv* **2021**, arXiv:2010.11365v2.
43. Jagarlamudi, J.; Daume, H.; Udupa, R. Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012; pp. 204–213.
44. Fard, M.M.; Thonet, T.; Gaussier, E. Seed-guided deep document clustering. In *Lecture Notes in Computer Science*; Springer Science and Business: Cham, Switzerland, 2020; pp. 3–16.

45. Li, C.; Chen, S.; Qi, Y. Filtering and classifying relevant short text with a few seed words. *Data Inf. Manag.* **2019**, *3*, 165–186. [CrossRef]
46. Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5802–5805. [CrossRef]
47. Alec, G.; Richa, B.; Lei, H. *Twitter Sentiment Classification Using Distant Supervision*; CS224N Project Report; Stanford University: Stanford, CA, USA, 2009; pp. 1–12.
48. Sagadevan, S. Comparison of Machine Learning Algorithms for Personality Detection in Online Social Networking. Ph.D. Thesis, Universiti Sains Malaysia, Penang, Malaysia, 2017.
49. Porter, M.F. An algorithm for suffix stripping. *Program* **1980**, *14*, 130–137. [CrossRef]
50. Li, N.; Chow, C.-Y.; Zhang, J.-D. Seeded-BTM: Enabling biterm topic model with seeds for product aspect mining. In Proceedings of the 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Zhangjiajie, China, 10–12 August 2019; pp. 2751–2758.
51. Anoop, V.; Asharaf, S. A topic modeling guided approach for semantic knowledge discovery in e-commerce. *Int. J. Interact. Multimed. Artif. Intell.* **2017**, *4*, 40. [CrossRef]
52. Scholte, R.H.; De Bruyn, E.E. Comparison of the Giant Three and the Big Five in early adolescents. *Pers. Individ. Differ.* **2003**, *36*, 1353–1371. [CrossRef]
53. Dodds, P.S.; Clark, E.M.; Desu, S.; Frank, M.R.; Reagan, A.; Williams, J.R.; Mitchell, L.; Harris, K.D.; Kloumann, I.M.; Bagrow, J.; et al. Human language reveals a universal positivity bias. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2389–2394. [CrossRef] [PubMed]
54. Rocha, A.; Goldenstein, S.K. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 289–302. [CrossRef] [PubMed]
55. Tijare, P.; Rani, P.J. Exploring popular topic models. *J. Phys. Conf. Ser.* **2020**, *1706*, 012171. [CrossRef]
56. Ray, S.K.; Ahmad, A.; Kumar, C.A. Review and implementation of topic modeling in Hindi. *Appl. Artif. Intell.* **2019**, *33*, 979–1007. [CrossRef]
57. Albalawi, R.; Yeap, T.H.; Benyoucef, M. Using topic modeling methods for short-text data: A comparative analysis. *Front. Artif. Intell.* **2020**, *3*, 42. [CrossRef]
58. Towne, W.B.; Rose, C.P.; Herbsleb, J.D. Measuring similarity similarly: LDA and human perception. *ACM Trans. Intell. Syst. Technol.* **2016**, *8*, 7. [CrossRef]
59. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; pp. 399–408.
60. Monaghan, P.; Chang, Y.-N.; Welbourne, S.; Brysbaert, M. Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *J. Mem. Lang.* **2017**, *93*, 1–21. [CrossRef]
61. Watanabe, K.; Zhou, Y. Theory-driven analysis of large corpora: Semi supervised topic classification of the UN speeches. *Soc. Sci. Comput. Rev.* **2020**. [CrossRef]
62. Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10*, 5416. [CrossRef]
63. Phan, X.-H.; Nguyen, L.; Horiguchi, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th Information Conference on World Wide Web (WWW'08), Beijing, China, 21–25 April 2008.
64. Resch, B.; Usländer, F.; Havas, C. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartogr. Geogr. Inf. Sci.* **2018**, *45*, 362–376. [CrossRef]
65. Andrzejewski, D.; Zhu, D.; Craven, M.; Recht, B. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
66. Platt, J.C. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*; Technical Report MST-TR-98-14; Microsoft: Redmond, WA, USA, 1998.
67. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
68. Van, T.P.; Thanh, T.M. Vietnamese news classification based on BoW with keywords extraction and neural network. In Proceedings of the 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), Hanoi, Vietnam, 15–17 November 2017; pp. 43–48.
69. Chen, S.; Shen, B.; Wang, X.; Yoo, S.-J. A strong machine learning classifier and decision stumps based hybrid adaboost classification algorithm for cognitive radios. *Sensors* **2019**, *19*, 5077. [CrossRef] [PubMed]
70. Zadeh, P.; Hosseini, R.; Sra, S. Geometric mean metric learning. In Proceedings of the 33nd International Conference on Machine Learning (ICML), New York, NY, USA, 20–22 June 2016; pp. 2464–2471.
71. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
72. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
73. Livieris, I.; Kiriakidou, N.; Stavroyiannis, S.; Pintelas, P. An Advanced CNN-LSTM model for cryptocurrency forecasting. *Electronics* **2021**, *10*, 287. [CrossRef]

74. Mustafa, M.; Zeng, F.; Ghulam, H.; Arslan, H.M. Urdu documents clustering with unsupervised and semi-supervised probabilistic topic modeling. *Information* **2020**, *11*, 518. [CrossRef]
75. Salem, H.; Shams, M.Y.; Elzeki, O.M.; Elfattah, M.A.; Al-Amri, J.F.; Elnazer, S. Fine-tuning fuzzy KNN classifier based on uncertainty membership for the medical diagnosis of diabetes. *Appl. Sci.* **2022**, *12*, 950. [CrossRef]
76. Shaukat, K.; Luo, S.; Chen, S.; Liu, D. Cyber threat detection using machine learning techniques: A performance evaluation perspective. In Proceedings of the 2020 International Conference on Cyber Warfare and Security (ICCWS), Islamabad, Pakistan, 20–21 October 2020; pp. 1–6.
77. Freund, Y.; Schapire, R.E. A Decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
78. Adi, G.Y.N.N.; Harley, M.; Ong, V.; Suhartono, D.; Andangsari, W. Automatic personality recognition in bahasa indonesia: A semi-supervised approach. *ICIC Express Lett.* **2019**, *13*, 797–805. [CrossRef]
79. Markovikj, D.; Gievska, S.; Kosinski, M.; Stillwell, D. Mining facebook data for predictive personality modeling. In Proceedings of the International AAAI Conference on Web and Social Media, Cambridge, MA, USA, 8–11 July 2013.
80. Kamble, K.S.; Sengupta, J. Ensemble machine learning-based affective computing for emotion recognition using dual-decomposed EEG signals. *IEEE Sens. J.* **2021**, *22*, 2496–2507. [CrossRef]
81. Dupré, D.; Krumhuber, E.G.; Küster, D.; McKeown, G.J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE* **2020**, *15*, e0231968. [CrossRef] [PubMed]
82. Abro, S.; Shaikh, S.; Hussain, Z.; Ali, Z.; Khan, S.; Mujtaba, G. Automatic hate speech detection using machine learning: A comparative study. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 8. [CrossRef]
83. Alam, F.; Riccardi, G. Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In Proceedings of the International Conference of Inter Speech, Lyon, France, 25–29 August 2013.
84. Rennie, J.D.M.; Shih, L.; Teevan, L.; Karger, D.R. Tackling the poor assumptions of naive Bayes text classifiers. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington, DC, USA, 21–24 August 2003.
85. Brownlee, J. *Deep Learning for Natural Language Processing: Develop Deep Learning Models for Your Natural Language Problems*; Machine Learning Mastery: San Francisco, CA, USA, 2017; Available online: https://www.technocourses.com/wp-ontent/uploads/2020/09/nlp.pdf (accessed on 28 October 2017).
86. Cao, H.; Li, X.-L.; Woon, Y.-K.; Ng, S.-K. SPO: Structure preserving oversampling for imbalanced time series classification. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining, Vancouver, BC, Canada, 11–14 December 2011; pp. 1008–1013.
87. Tang, Y.; Zhang, Y.-Q.; Chawla, N.V.; Krasser, S. SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man Cybern. Part B* **2009**, *39*, 281–288. [CrossRef] [PubMed]
88. Solé, X.; Ramisa, A.; Torras, C. Evaluation of random forests on large-scale classification problems using a bag-of-visual-words representation. In *Proceedings of the Catalan Conference on Artificial Intelligence, Frontiers in Artificial Intelligence and Applications*; IOS Press: Barcelona, Spain, 2014; pp. 273–276.
89. Mairesse, F.; Walker, M. Words mark the nerds: Computational models of personality recognition through language. In Proceedings of the Annual Meeting of the Cognitive Science Society, Vancouver, BC, Canada, 26–29 July 2006; Volume 28.
90. McGrath, J.; Saha, S.; Chant, D.; Welham, J. Schizophrenia: A concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **2008**, *30*, 67–76. [CrossRef]
91. Amirhosseini, M.H.; Kazemian, H. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technol. Interact.* **2020**, *4*, 9. [CrossRef]
92. Madisetty, S.; Desarkar, M.S. A neural network-based ensemble approach for spam detection in twitter. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 973–984. [CrossRef]