National College of Ireland

# Towards an Effective Social Engineering susceptibility detection Model Using Machine Learning on the Online Social Network

MSc Internship
Cyber Security


Nelson Seyi Ayo-Akere

X18172521

School of Computing
National College of Ireland


Supervisor: Christos Grecos

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | **NELSON SEYI AYO-AKERE** |
| **Student ID:** | **X18172521** |
| **Program:** | **MSc CYBER SECURITY** **Year:** **2020** |
| **Module:** | **ACADEMIC INTERNSHIP** |
| **Lecturer:** | **CHRISTOS GRECOS** |
| **Submission Due Date:** | **12TH DECEMBER 2019** |
| **Project Title:** | **Towards an Effective Social Engineering susceptibility detection Model Using Machine Learning on the Online Social Network** |
| **Word Count:** | **9005** **Page Count:** **23** |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** ……

**Date:** **13/12/19** ……

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Towards an Effective Social Engineering susceptibility detection Model Using Machine Learning on the Online Social Network

Nelson Seyi Ayo-Akere
X18126651
MSc. Cyber Security
12TH August 2019

## Abstract

The challenge posed by social engineering has become increasing worrisome and proven over the years to be daunting to mitigate even with recent security measures in place. Humans as it is are said to be the weakest link to security which makes online social media networks (OSN) very susceptible to social engineering attacks due to its never ending increase of users who come together for the purpose of communicating and sharing information. However, OSN users have the final say as regards dissemination of information via posts, uploads and updates on OSN consequently users become susceptible to social engineering attacks via the release of personal identifiable information (PII) on OSN. This research presents a novel social engineering machine learning prediction model (SE-MLPM) to detect and extract PII in OSN user posts using natural language processing (bag-of-words) and thereafter vectorise post text into vectors utilizing the term frequency inverse document frequency (TF-IDF) vector space modelling technique and finally classify, label and predict levels of post susceptibility to social engineering attacks in addition to revealing the PII discovered to the OSN user and recommending to the user if the post should go live or not, based on PII count recovered from the post ranging from a high susceptibility level to a no susceptibility level using the logistic regression classification algorithm. This will give the OSN users the opportunity to vet their post before disseminating to the public. By so doing, SE-MLPM will minimize the enormous volume of Sensitive Personal Information (SPI) OSN users post on OSN.

**Keywords: Machine Learning, Online Social Network, Social Engineering.**

## 1. Introduction

### 1.1 Project background and motivation

Social media platforms are referred to as computer-based technologies that fosters the creation and dissemination of information, interests, social thoughts, ideas, and all other methods of human expression through networks and virtual environments. There are several kind of social media platforms around the internet that fits for anyone or everyone as well as the ones that connect people based on like hobbies, similar interests, career motivations, like demographics etc. The Social network landscape is wealthy in forums, sites, environments, and platforms where interaction is made by various people around the globe who gather together for the purpose of communicating

about events thoughts, or feelings shaping and surrounding their lives. A critical view into chats and dialogues among individuals gives a perception on the attitude, behaviour, thoughts and feelings that impact politics culture, ideology, economies, etc either from the positive or negative front. There is no lack of online goals for individuals' thoughts and emotions [1]. With the growing worldwide utilization of social media, it has become the most vulnerable platform to steal identities or execute extortion. An examination demonstrated that between April 2018 and April 2019 there are more than 3 billion clients of the OSN, a similar overview shows that Social media clients increased by 202 million [2]. Due to these evolving statistics, there is an emerging issue of computing or operating with voluminous data to deliver the desired outcome.

Social engineering is referred to as an expansive scope of dangerous activities achieved through human collaborations. It utilizes psychological manipulation to fool clients into committing security errors or parting with delicate data, thus it is understandable true that social networking platforms are the most vulnerable to abuse especially through social engineering. Attacks via social engineering takes place following some steps. An attacker initially carries out information gathering on the expected targeted client to get and accumulate vital personal identifiable information (PII, for example, weak entry points and personal information and potential feeble security protocols, then the attacker tends to gain the trust of the targeted client and provides a sensation for other acts that breaches security practices which includes allowing access to crucial data. In a study by open data security, it shows that social engineering is by far the easiest and well sorted out method of manipulating victims by hackers in which it revealed that about 1 billion dollars has been carted away with across 30 countries in 100 banks in a space of 2 years owing to social engineering attacks [3]. Furthermore, it is no hidden knowledge that individuals should avoid displaying their personal information such as date of birth, primary email address, account numbers, passwords, location etc, individuals still reveal this personal information publicly exposing them to social engineering attacks without realising it [4].

With the numerical growth of individuals who utilize social media platforms, an automated means of computing Bigdata has become paramount, coupled with combatting social engineering on social media platforms. Consequently, OSN users need to be educated on the appropriate use of social media networks in relation security and privacy protection and this is most effective if done in real time. To attain these feet, it is important to device a comprehensive solution that cuts across the whole social media platforms that can alert OSN users on how susceptible their posts are to potential social engineering attacks including its risk severity. As of late, an increasingly distinguished strategy for taking care of issues which have assumed control over the global market is machine learning.

Machine learning is a part of the field of artificial intelligence (AI), and a technology for developing problem solving models, systems or algorithms. Machine learning is thus a statistical technique utilized to teach and train computers to learn without a need for rigid set of rules. Machine learning iteration capability is paramount because new data that interacts with the model are able to adapt and be computed independently. They tend to learn from the previous dataset which was used to teach and train the model to develop consistent, repeatable outcomes and decisions. Although machine learning has been around for quite a while, the capability of computing big data faster in iteration and produce accurate results is a quite modern development [5]. There are various

algorithms available such e.g. SVM, logistic regression, naïve based etc available via this technology depending on the task, a model can be built to derive a desired result.

For this research, model has been built by combining natural language processing (NLP) model (bag of words), term frequency inverse document frequency (TF-IDF) vector space model and logistic regression machine learning classification algorithm to produce a model that would enable online social network (OSN) users to know how vulnerable their posts are to social engineering attacks before it goes online which is centred on the susceptible sensitive personal information (SPI) or personal identifiable information (PII) identified. Of cause it's still left to the user to decide whether he or she wants to go ahead with posting it online but it creates a real time awareness to the user to be able to act promptly which in turn will reduce the amount PII going online which social engineers can utilize to perpetuate their attacks. For this project, the theme for checking the post of OSN users is focused on personal identifiable information.

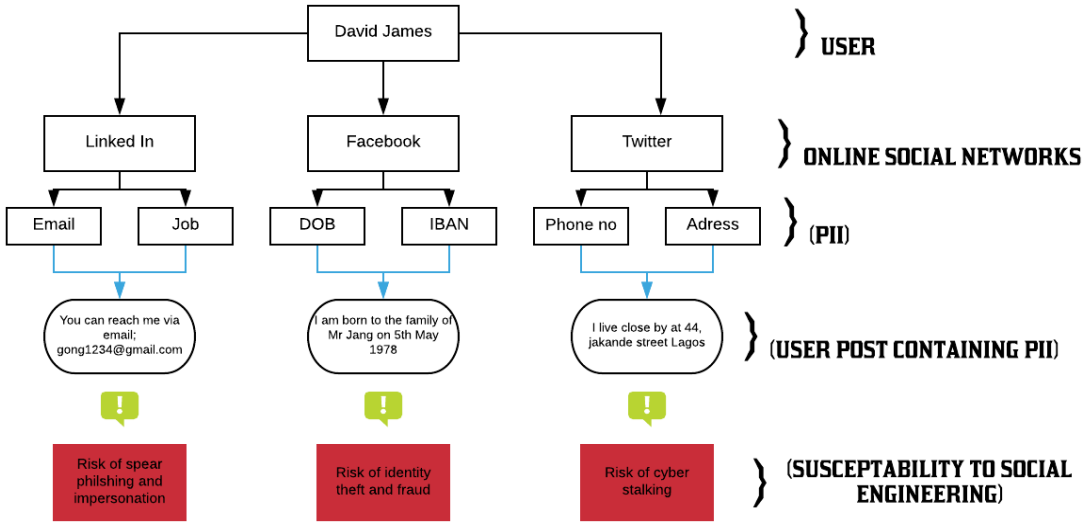


Figure 1 social engineering attack outliner.

OSN coupled with other relevant information can be used to actualize a social engineering attack. Figure 1 shows the various kinds of PII that can be gathered for social engineering via OSN and how a social engineering attack can take place.

The structure of this paper is constructed into six sections as follows; Section 1 discusses the outliners of the project which comprises of the research question, purpose of the research, research variable and project specification. Details of related work and literature review on the method and approach utilized will be discussed in section 2. A detailed breakdown on the methodology and project design will be discussed in section 3, section 4 will constitute the development and implementation of the model. Discussions, case study and evaluation of the model will be discussed in section 5 and also results produced from the running of the machine learning model, and finally in section 6, conclusions will be made, and future work as regards the project will be proposed.

## 1.2 Project requirement specification

### 1.2.1 Research Question

Based on the issues of Bigdata processing and social engineering, we ask this question - Can we reduce the amount of Personal Identifiable Information (PII) that go live on OSN platforms by predicting and notifying in real time OSN users on how susceptible their posts are to social engineering, by utilizing a machine learning model built utilizing natural language processing (NLP), TF-IDF vectorization technique and logistic regression machine learning algorithm?

### 1.2.2 Purpose

The purpose of this research is to predict the susceptibility of Online Social Network (OSN) users post to social engineering using a machine learning model and notify OSN users on how vulnerable their post could be to a social engineering attack to users.

### 1.2.3 Research Variable

Mitigating against the risk of social engineering and the accurate processing of Big data is the goal of this research and in so doing we have developed a novel solution called social engineering machine learning prediction model (SE-MLPM). SE-MLPM natural language processing (NLP) model (bag of words) and logistic regression machine learning classification algorithm in developing our model. We believe if OSN users can be notified on potential threat in posting SPI through an accurate machine learning prediction system to predict and notify users on the severity and susceptibility of the releases of such PII to the public, it will in-turn enable them to vet their post and be more cautious of posting their PIIs on the Online Social Network (OSN), consequently reducing their vulnerability to social engineering attacks. In the development of SE-MLPM, we have used software's such as Python, Python Flask, HTML, Sci-kit learn, Sublime text editor, Jupyterlab, anaconda, to build the model.

## 2. Literature Review

On online social networks (OSN), a great part of the implementations is made on such platforms generally to improve on the user experience and other relating activities for which they are analysed. Personal identifiable information (PII) have proven to reveal sensitive details about individual's private life, therefore paving way for social engineering attacks. Also, with the growing volume of data generated via social media, it has become paramount to implement models that would be able to compute complex Bigdata quickly and automatically and return accurate results to the user of the OSN. Our point of argument is that we believe that large volume of data generated from OSN can be automatically computed accurately and precisely to provide the desired results to notify users of OSN and get them more knowledgeable on how susceptible their posts are to social engineering attacks. We believe it is more effective to compute Bigdata through machine learning and for user data to learn from the model then subsequently display results of the analysis to OSN users to vet their intended post. In this manner, we believe it will be harder for social engineers to crawl OSN and get useful PII to enact a social engineering attack. With SEMLPS Social media managers can utilize the application model in predicting and alerting users on how vulnerable their posts can be to social engineering attacks and its degree of severity.

The review of literature pertaining to this paper will be further organised as follow:

a. Documented research based on privacy settings, access control, real time notification in relation to social engineering.
b. Appraisal of documented research on traditional programming list system and Machine Learning.
c. Literature review based on data collection.
d. Literature review based on data pre-processing.
e. Documented appraisal of vector space modelling techniques
f. Research Documentation Based on Best Suited Machine Learning Prediction and Classification Model.

## 2.1 Documented Research Based on Privacy Settings, Access Control and Real Time Notification and in Relation to Social Engineering

Over time, various access control measures have been introduced and implemented on OSN, but despite the critical implementation of these measures, users of OSN still fall prey to social engineering attacks. A research carried out by Pwint Oo described how privacy has become an enormous problem on social media platforms [6]. A typical assessment of OSN can be attributed to how social media platforms like Facebook and twitter pay little or no attention to trust whereby every follower or connection is considered a friend [7], however OSN connections cant be held responsible for the trust for such connections [8]. This aids social engineers to gain access to personal identifiable information of target victims via their profile information. For a social engineering attack to take place on a user of a social media platform, a certain level of access is needed by the social engineer to enable him or her view the profile or post of the user. Enabling the privacy setting on an OSN, can allow or disallow unwanted access to profile information shared on the platform. It is worthy to note that the measure of Implementing private security features via privacy setting is a good measure in the right direction but evidently cannot avert social engineers from accessing the target victims' profile.

Another solution was proposed by Misra in which an agent for access control on OSN decision provider that counts on the interactive bond between the information displayed and the user to determine who are authorized to have access to personal profile information depending on community network recognition [9]. In proving this, an experiment was performed by designing an application that uses Facebook Graph API and Facebook Query Language to be used by participants in making access control decisions. the acquired result of the experiment was a brilliant one but still doesn't entirely solve the problem of social engineering. Another research carried out by Bilge and his co researchers illustrated that social engineers have become cleverer in perpetuating social engineering attacks on how they can beat access control mechanisms and techniques by making a replica account or identity of an already existing friend and ask for a request to have access to the target victims' profile and information [10]. The social engineer normally studies the social relationship and replicates the identity of a friend who is inactive on the platform and completely aloof of the victims' suspicion, but whose account is still very much active on the OSN. The target victim would probably receive a friend request, requesting to be re added back to the platform using crafted messages. Although access control is a good means of preventing social engineering on social media platforms, the research carried out by Bilge [10] thus proves that the proposed "PACMAN" solution by Misra and his co researchers [9] is not entirely an effective method in preventing social engineering attacks on social media platforms.

Another research was carried out by Ololade proposed a method of real time notification of OSN user on the susceptibility of their post to social engineering by hardcoding some keywords saved in a library for the program to identify and display to the user on detection of such keywords or PII [11]. This solution has proven to be the most comprehensive and effective method of notifying users on the susceptibility of there post to social engineering and giving them an opportunity to vet their post and repost if PII are identified in the post. In this section we have reviewed access control setting, privacy and real time notification researches by various researchers and we believe that the utilization of real time notification of susceptibility to social engineering attack in the development of our machine learning model (SE-MLPM) will be the effective means to actualize the desired outcome in our research.

## 2.2 Appraisal of Documented Research on Traditional Programming List System and Machine Learning

Technology, most especially in computer science develops on a daily basis. To this end it is important that at all levels of infrastructure, technological advancement should be implemented to cater for the worlds fast growing diversities and complexities. Social media platforms in this case are no exemption. It is true that rules and instructions can be manually coded into a program to provide a desired result. Traditional programming is a manual process in-which programmers utilize to develop programs. A programmer has to manually create or code rules, without any form of logic. A research conducted made use of traditional programming skills to manually code instructions for certain set of keywords to be identified in the program and return to the users the susceptibility of their post to social engineering [11]. This method is grand, but in another research, paper proves traditional programming cannot be used to fully implement certain mostly with those regarding big data[12].

For over 10 years, an innovative kind of programming has transformed computing and data processing, especially in the field of embedded analytics. In Machine Learning**,** the input and output data are injected into an algorithm in creating a program. The algorithm automatically creates the rules from the dataset. This produces efficient insights that can be utilized to predict future results, In solving the same problem as in [11], with a full knowledge that social media platforms are responsible for the generation of millions of data within minutes and this date are largely complex, unstructured and spontaneous, which makes it tough to sort, examine and label, data engineers have thus evolved to utilizing machine learning. Instead of hardcoding instructions to create programs, machine learning simply involves collection of past data that will be utilized in the building of a automatic model. A data engineer carried out a research on the implementation of a machine learning model to detect hate speech using from a collection of data set in South Korea [13]. This research featured an automated means of detecting hate speech from a cache of dataset in South Korea, in which other historic data sets that will be introduced into the model can automatically train and learn the model to give accurate and desired results.

Another research carried out explored methods in digitally shared text entries utilizing machine learning algorithms for detecting depression, in-which the Bidirectional Encoder Representations from Transformers (BERT) model provided a result with accuracy level of 85% [14]. Thus, machine learning has proven to be more accurate and automated method for predicting outcomes. In this section we have analysed that machine learning could prove to be a more automated and accurate method for predicting results and we believe that incorporating it into our

model as the major cause for novelty an it applies to detecting of susceptible PII with OSN user post would create a semi-automated and accurate means of predicting results based on PII detected within OSN users posts. Consequently, for this to be actualised, an array of historic data has to be collected for training and learning of our model (SE-MLPM), in which future induced data can automatically learn from. This will be discussed in detail in the next section.

## 2.3 Literature Review Based on Data Collection

Following the concept of building our model it is necessary to note that historic data or a large volume of dataset (Bigdata) is required in the training and testing of our model. But the question of how we can acquire an enormous volume of data, that will be adequate and efficient enough to train and test our machine learning model arises. A proposed method for data collection was presented by innovation insights research institute where a survey was carried out for the collection of data from a selected group of individuals utilizing standardized interviews and questionnaires [15]. It later concluded that although the method of survey in collection of data is a valid method and it provides an effective way for data scientist to gather dataset, it illustrated that its time consuming and requires enormous effort from the data scientist. Also, data scientist can utilize huge volumes of data to gain insight in dealing with challenges and building models, it is easier said than done in accomplishing it. The effort is raged by privacy concerns which inherently makes it quite difficult for data to be accessed by data scientists. Additionally, various API could be used in gathering desired dataset for machine learning purposes. In a research on sentiment analysis using machine learning technique, a total of 1000 dataset was raked for twitter application programming interface (API) for its research [16].

It is important to note that in the generation of feasible dataset for extraction of keywords or personal identifiable information, the dataset should be structured and present personal identifiable information in its contents so that the desired outcome of the model will be substantially accurate. A research carried out by Neha and Roy suggested that synthetic data can be generated to overcome the challenge of production of a huge dataset and also privacy policy pertaining to collection of dataset [17]. It revealed that, there was no substantial variation between the accuracy of features generated on some form of synthesized data as against those generated on the control dataset. MIT news also validated this method by publishing it as its front runner article in 2017 [18]. Another research was carried out by Ghanem et.al on synthetic data generation for statistical testing which revealed that the results gotten suggested that within practical time, the approach can generate test data that is both logically valid an statistically representative [19]. Furthermore, another research conducted illustrates that Synthetic data is a feasible, next-step resolution to privacy problems attached to dataset collection [20]. Another research carried out by Veeramachanen discussed that according to a modern study carried out by data scientists, Kaggle tabular data is the most used platform for the acquisition of data followed closely by academia [21]. It also emphasized that data scientists are peddling synthetic data for its capability to remove a number of concerns associated with data science, including structure, removing bottle necks, clearing privacy huddles etc that are mostly associated with data access thus providing a safe data space to work with. To buttress the use of synthetic data for machine learning a research conducted shows that synthetically generated data can very much be used train and learn machine learning models [22]. It also defines Mockaroo as an application programming interface (API) that is capable of generating structured dataset with variety of cells ranging from emails, IBAN, names, passwords and regular expressions with an added

advantage of changing the values if need be with custom function. We believe that this method would enable use to be able to generate as much structured and huge dataset necessary in training and learning our machine learning model because keywords and personal identifiable information to be extracted and detected to notify users of susceptibility to social engineering will be present, which can be synthetically provided with ease and caution from privacy policies via this API. Subsequently, the appraisal of documented reviews on data processing and Personal information extraction from the dataset acquire will be discussed in the next section.

## 2.4 Literature Review Based on Data Pre-processing

Data pre-processing is an essential process in Machine Learning as it is responsible for the quality of data and presentation of useful information from the dataset, which in-turn affects the ability of the machine learning model. In machine learning data pre-processing is typically done through natural language processing (NLP). NLP is a stage in machine learning that deals with the computer's ability to analyse, understand, and manipulate text or human language [23]. In carrying out natural language processing (NLP) for Bigdata that is to be introduced into our model (SE-MLPM), the choice of natural language processing technique is very paramount since we are going to be working with OSN user posts (text) data to process the text data adequately and identify and extract personal identifiable information.

In a recent study conducted, by Tomas and his co researchers on Distributed Representations of Words and Phrases and their Compositionality, it reveals a good method for learning high-quality distributed vector representations that covers a huge volume of precise semantic and syntactic word relationships which was the Skip-gram model [24]. Also in another Survey on Hate Speech Detection using Natural Language Processing it discusses that undeniably, larger n-grams and unigrams are mostly identified as been highly predictive, it is pertinent to note that for any task based on text classification, the most apparent method to use are surface-level feature processing methods such as bag of words [25]. It further explained that normally, a brilliant classification performance in hate speech detection is produced by utilizing bag of words features bag-of-words features as long as the target words are present in both training and test data. Although these various methods of text classification and natural language processing (NLP) processing techniques shows relationships between phrases and text, it is important to bear in mind that for the purpose of our research is projected towards the identification of personal identifiable information in a social media user's post could simply be identified by the identification of simple surface feature.

In a research paper presented in the Association for the Advancement of Artificial Intelligence (AAAI) conference give a statistical evaluation and analysis on the experiment carried out on various classes of text classification shows that, skip grams model had an accuracy of 75.4%, with a precision of 75.1%. other models evaluated were latent Dirichlet allocation (LDA), which gave a result of 72.2% accuracy and 70.8% precision. On the other hand, the bag of words model proved the most productive as it gave an accuracy of 79.7% accuracy and 79.5% precision [26].

In this light, we believe that in carrying out natural language processing and personal identifiable information (PII) identification using our dataset, the bag-of-words (BOW) model would best produce the desired result when incorporated our model SE-MLPM. On incorporation of the bag-of-word (BOW) model into our machine learning model, a vector space model (VSM) for vectorization and identified feature extraction will be discussed in the next section.

## 2.5 Documented Appraisal of Vector Space Modelling Techniques

Every natural language processing (NLP) technique e.g. bag-of-words (BOW)is incorporated with a vector space modelling technique for the vectorization of text, vector extraction and for the purpose of this project, to subsequently carry out a statistical measure of personal identifiable information or sensitive personal information (PII) pertaining to susceptibility to social engineering (SE) attacks on social media target victims, which will be fed to the machine learning model for it to understand and perform its tasks adequately. Vector space models (VSM) converts text or words having different lengths such as paragraphs, words or sentences into a numerical value that can be fed into an application e.g. a machine learning algorithm. In building our machine learning model (SE-MLPM) or any machine learning model at all, it paramount to apply a vectorization technique to convert to vectors, information retrieval as well as statistical measure and weighing of extracted PII in vector form.

Numerous researches by data scientists have been carried out to identify the best vector space modelling techniques for various tasks as it pertains to our model (SE-MLPM). A study contrasted the embedded words with the term frequency inverse document frequency (TF-IDF) vectorization technique when used with support vector machine classifier (SVM) classifier [27]. The experiment showed that the TF-IDF vector space technique gave a F1 accuracy sore of 93.1% as against 90.1% F1 accuracy of the word embedded technique when applied with the SVM. The research also revealed that less memory is been used by the TF-IDF technique and it was a brilliant technique when applied to. In another research carried out on the Comparisons and Selections of Features and Classifiers for Short Text Classification an experiment was carried out on three different vector space models namely word2vec, TF-IDF and Doc2vec methods [28]. The comparative analysis shows a significant difference in accuracy between the TF-IDF technique and the other two techniques. Statistically it revealed that when the word2vec technique was applied with three different machine learning classifiers such as naïve bayes, support vector machine and logistic regression classifiers, it gave accuracy levels of 46%, 56% and 64% respectively. The Doc2vec vector space technique for all three classifiers gave accuracy levels of 30%, 20% and 30% respectively. Finally, the TF-IDF vectorization technique for all three classifiers gave accuracy levels of 67%, 70% and 70% respectively with Bigdata.

From the basis of these reviews, we believe that in carry out vectorization and feature extraction in our model (SE-MLPM) the TF-IDF vector space technique will be most suitable in building an efficient model for the conversion of OSN user posts to machine readable vectors and extraction of personal identifiable information and for the machine learning classifier which we would be discussing about in the next section to provide the best result.

## 2.6 Research Documentation Based on Best Suited Machine Learning Prediction and Classification Model

In machine learning, various classifiers such as naïve bayes, logistic regression, support vector machine classifier, decision tree etc. are available of which each of these classifiers are brilliant in their own way depending on the kind of task, result and vectorization technique sorted. All these three factors have to be considered when choosing a machine learning algorithm. There are numerous classification algorithms available in data science but it impossible to determine which is more superior to another and this depends on the nature of the dataset and it is to be applied and the type of vector space modelling technique been applied. Classification in machine learning is the

method of predicting the class of given dataset [29]. Classes can also be called labels, categories or targets. Machine learning classifiers are also in the business of predicting outcomes depending on the dataset introduced and the vector space model applied. They break dataset into test and training data to learn the algorithm to be able to classify or make predictions towards the desired result. For the purpose of this research, A machine learning model for extraction of personal identifiable information and post classification in accordance to the degree of susceptibility of this sensitive personal information to social engineering attacks. In the building of our model (SE-MLPM) a classifier that worked efficiently with our suggested vector space modelling technique was be sorted to provide the desired result with a feasible level of accuracy, thus the reason for the review.

A research illustrated that the naïve bayes algorithm is frequently utilized in the prediction and classification of text among data scientists due to its simplicity, but its efficiency can be questioned as its not likely to classify or predict occurrences in text documents adequately [30]. In a research, an experiment was conducted on Bigdata to analyse the accuracy and effectiveness of proposed machine learning models. was carried out on test data to statistically analyse the accuracy of various models. In a Comparison of Features and Classifiers for Short Text Classification, the result of the experiment featured three machine learning classifiers namely naïve bayes, logistic regression and support vector machine (SVM) revealed that the application of TF-IDF vectorization technique on the naïve bayes classifier registered an accuracy of 67.9% accuracy, 70% accuracy on the support vector machine and 70.7% accuracy when applied with the logistic regression classification algorithm [31]. Another experimental study conducted by Arora on models for document classification and feature extraction with three distinctive vector space modelling techniques such as word2vec, Doc2vec and TF-IDF techniques on various machine learning classifiers such as naïve bayes, and logistic regression revealed that, the Doc2vec VSM applied with the naïve bayes algorithm provided an accuracy of 58.4%, while the application of Doc2vec and logistic regression prediction model produced and accuracy result of 76%. Furthermore, the application of TF-IDF vectorization technique with naïve bayes classification algorithm produced an accuracy of 73.62% while the application of TF-IDF vectorization technique with logistic regression classification algorithm produced an accuracy of 95.45% [32].

From the research reviewed we believe that in applying the TD-IDF vector space modelling technique with the logistic regression classification algorithm would best suit our model (SE-MLPM) in the efficient identification, extraction and categorization of personal identifiable information (PII) based on the number of PII to determine how susceptible online social network (OSN) users posts are to social engineering attacks and its severity classified to highly susceptible, moderately susceptible, less susceptible and not susceptible.

## 2.7 Programming Languages and Packages

**a. Python:** Python high-level programming language utilized for numerous purposes. Python can also be utilized in machine learning, building websites, GUI applications etc. it consists of several properties such as Readable and Maintainable Code, Multiple Programming Paradigms, Compatible with Major Platforms and Systems, Robust Standard Library, Many Open Source Frameworks and Tools and Simplify Complex Software Development. Python is a programming language that is flexible, stable, and offers numerous tools to software and machine learning developers, and due to all these features, Python is usually considered the first choice for machine learning**.** In building our

machine learning model (SE-MLPM) we have utilized python as the major programming language for model development. The version of python installed was python 3.7.5.

**b. Anaconda:** Anaconda is the standard platform for Python data science. The open-source Anaconda Distribution is the simplest means of performing machine learning and Python/R data science on any operating system such as window, Linux OS, MAC OS. It is a platform that can be utilized to carry out performance analysis and scalability of programs with packages such as NumPy, pandas and dask. It also contains visualization pages such as Matplotlib for visualization analysis. Anaconda navigator 2019.10 was used during the development of our model.

**c. Jupyter Notebook:** Jupyter Notebook is an open-source web application that allows developers share and create programming documents that consist of visualization, code, narrative text and visualization. It is majorly utilized for statistical modelling, coding, data cleaning and transformation, machine learning etc.in this project we have utilized Jupyter notebook 6.0.1 as our workspace to carry out coding, data cleaning and transformation as well as machine learning.

**d. HTML:** Hypertext Text Mark-up Language (HTML) is a language utilized in developing web pages. It is simply a language mostly used to build the user front end of a wed application. It can also be used to build front end interface for machine learning models.

**e. Sublime Text**: Sublime Text is a frequently utilized text editor utilized to write Python code.

**f. Python Flask:** Flask is a common Python web framework, used for integrating and developing web applications. Its major benefit is Integrating Front-End frameworks with Flask applications. It is also used in machine learning to integrate a front-end framework with machine learning models. Another benefit of utilizing flask is that provides security against flask security bugs. It avails a developer the chance to add prominent security features to flask applications such as Session based authentication, role management, basic HTTP authentication, token-based authentication, Token based account activation, user registration etc.

**g. Exploratory data Analysis (EDA) packages:** This are packages that are utilized for data analysis in summarizing heir man features or characteristics. In the data pre-processing of our model certain EDA packages were imported such as Pandas, Spacy and NumPy. Pandas makes it simpler to import, analyse and visualize dataset. Spacy on another hand helps to bring in feature to pre-process data set such as removing stop words punctuations etc. while (NumPy Array) is a multidimensional array utilized to store values of same datatype.

**Machine Learning (ML) Packages:** This are packages that are utilized for the to carry out machine learning tasks or provide various machine learning model matrixes such as logistic regression, GaussianNB, Naïve bayes etc including performance and accuracy matrixes. For this project, we use the sci-kit learn package to import these various models in carrying out our analysis.

**Visualization Packages:** Visualization packages are also important to give a pictorial view of our analysis and model prediction shape. It helped in giving a better overview and in a pictorial format of how our analysis is fairing and also our model shape. For this project the Matplotlib was utilised for that purpose.

## 3. Methodology

In this paper, a 7 steps methodology was followed In the development our Model (SE-MLPM), for the identification, extraction, prediction of the degree of severity of a social media post utilizing the degree of appearance of personal identifiable information in their posts and also notifying them of the models predictions for users of OSN to be conscious of the PII they are about to release to the public to give users the opportunity of editing their posts if they wish and by so doing reducing social engineering attacks via online social platforms. Figure 1 depicts a diagrammatic flow of the methodology followed in this study.

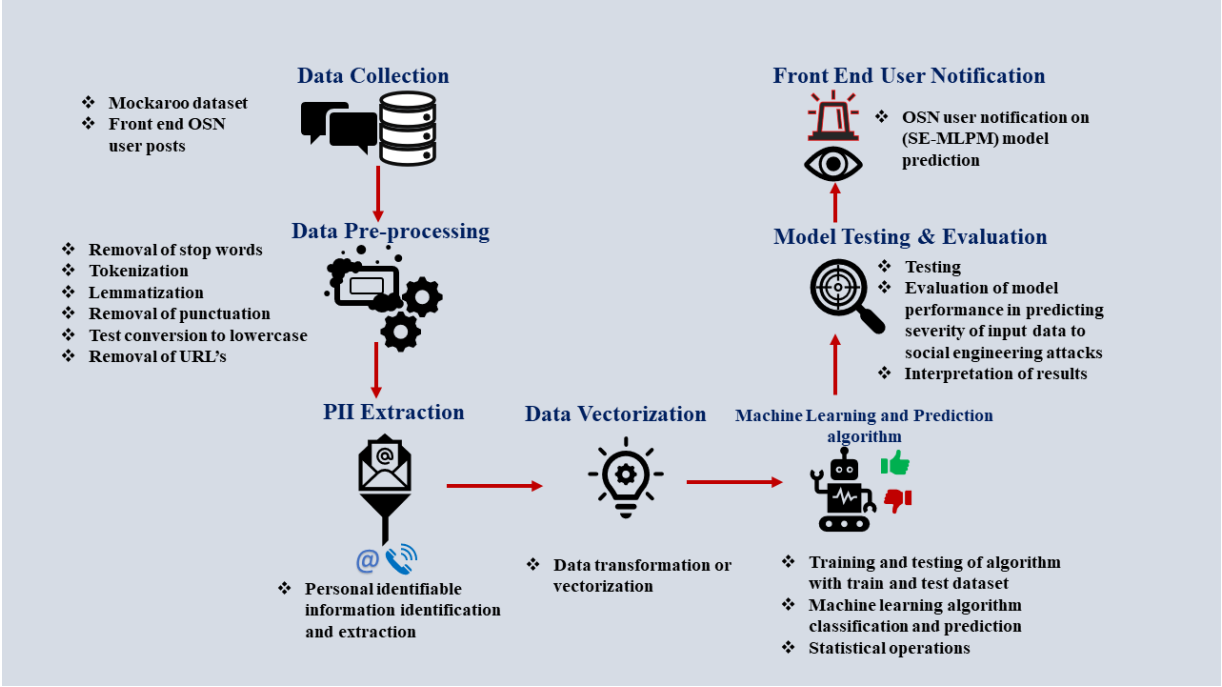


Figure 2. Methodology Model

### 3.1 Data Collection

In the collection of data that will be utilized in building our model we decided to use to utilize the method of synthetic data generation. Synthetic data is artificially generated rather than being produced real individuals or real dataset collection API's. It is frequently utilized for a broad range of events, such as test data for model validation, new tools and products and in machine learning requirements, produced with the aid of algorithms. Synthetic data is beneficial because it can be generated to meet very precise requirements or conditions that cannot be gotten in real existing data that is readily available. This has proven to be an alternative as it can be utilized when there are privacy limitations or when the dataset required for an experiment or research simple isn't available [33]. Self-driving car simulators initiated the use of synthetic data in the field of machine learning. Other sectors that utilize synthetic data include research, financial sectors and health sectors among others.

As earlier reviewed, data collection generation and collection face a lot of bottle necks such as its availability for the particular task intended, ethical issues and volume. To this effect we have justified the use of synthetic data in the training and testing of our model (SE-MLPM). Synthetic

data can be generated from various online platforms but for the purpose of this research we have chosen to generate our dataset from Mockaroo data generator. Thus, a total of 4000 rows of dataset was generated via the platform which included columns that constituted of personal identifiable information (PII) such as email addresses, phone numbers, IBAN etc. Other columns included we regular phrases and expression, to make up the complete data set. After the automatic generation of dataset via Mokaroo data generator, it was downloaded into a csv file for proper visualization.

## 3.2 Data Pre-processing

Generally, the whole process was carried out using the anaconda platform. Jupyter notebook is by default installed in anaconda and was used in writing our code and building the model. The selection of an accurate representation of text in a dataset is important in obtaining an accurate classification outcome. Likewise, the transformation of the input data to vectors and feature extraction is also important. However, prior to the extraction of features pre-processing of data is needed. In carrying out pre-processing of Bigdata, backed with our research outcome, we have used the bag of words model in data pre-processing for each text document or post. The bag-of-words model is a simple and popular technique utilized in the processing of dataset to label the existence of every word within a dataset or document. In the bag-of words (BOW) model, certain procedures for pre-processing of data are carried out depending on the outcome required. Here, there are also numerous tools contained in the bag of words model that can be used in data pre-processing, but we have used Spacy and NLTK for cleaning of our data. Other packages were also imported to aid in the data pre-processing stage such as Pandas and NumPy. The following processes were carried out in cleaning our dataset:

**a. Tokenization:** Tokenization is the way toward breaking the sentences just as the document into word delimited by white space, new line or tab and so forth. Result of this tokenization stage is a group of words delimited by white space. Its typically a method of transforming a document to a list of words that can be used to prepare a Metrix via vectorization.

**b. Punctuation Removal:** Special characters and punctuation(, :,{,},[,],^,&,*,(,) , | etc.) were removed from the dataset provided, also tabs, excess spaces shifts were also removed.

**c. Lemmatization:** Is described as the morphological analysis of data with the proper utilization of vocabulary aimed at extracting inflectional endings and to output just the recognised dictionary word called the lemma.

**d. Stop words Removal:** Stop words are a set of words which frequently occur and add no relevance or meaning to the dataset or does not provide any relatable information to the impended task. Stop word are frequently utilized word such as “an”, “the”, “in”, “a”. Thus, these words will need to be removed before additional processing can be carried out.

## 3.3 Feature Extraction

For the identification and presentation of personal identifiable information in our dataset to the machine learning model and likewise the predictive values to the OSN user, it is paramount to carry out important feature extraction or this personal identifiable information from the injected dataset and in the future OSN users’ posts.

Some personal identifiable information or PII that our model is expected to extract are email addresses, phone number, IBAN etc. The technique we have chosen to be part of our model is the bag-of-words technique. There are several power tools that can be used in the bag of words model to identify patterns of personal identifiable information for its extraction but for this project we have utilized regular expression (REGEX) for important information retrieval. This tool can be used for information retrieval from the dataset and then we can represent each word as a vector of 0 and 1 for feature extraction by a vector space modelling technique. This method helps to grade PII count to various degrees of post susceptibility to social engineering attacks such as highly susceptible, moderately susceptible, less susceptible and non-susceptible.

## 3.4 Data Vectorization

For a machine learning algorithm to understand dataset or text document, the corpus has to be converted to vectors. The process of converting this corpus to vectors is known as vectorization which is carried out by vector space techniques. Also, there are various vectorizing techniques around but choosing the right vectorization technique depends on the particular task needed to be carried out. From our review and for the specific task of feature extraction and vectorization, we have identified the TF-IDF vectorizer to be the most efficient in building our model (SE-MLPM). TF-IDF is a statistical measure utilized to assess the significance of a word to a document in a dataset.

TF-IDF is utilized for vectorizing text or data with TF–IDF scores. In context, the TF-IDF Vectorizer utilizes the Count Vectorizer estimating technique in counting the bag of words vectors and how many times the extracted tokens occur, then further utilises the inverse document frequency (IDF) Transformer in normalizing the occurrence of the count. Raw dataset and text document inform of strings works and integers are supposed to be the input of the TF-IDF vectorizer, thus a standardize vectorization is applied. The vectorizer returns a vector matrix representation where it is statistically calculated and given a TF-IDF score which can now be fed into the machine learning algorithm.

## 3.5 Machine Learning and prediction algorithm

There are numerous methods for developing machine learning models for different text-based applications which critically depends on the dataset and definition of problem to be solved. After clean-up or pre-processing of the dataset, we can now build our classifier/prediction machine learning algorithm to identify the statistical vectorized score and predict an outcome based an the label of levels of susceptibility of the post to social engineering attacks placed on the personal identifiable information PII which has been extracted during the processing and extraction phase. From our detailed review on the type of machine learning algorithm that will best suit our chosen vector space model and produce the best accuracy for our model to was the logistic regression classification model which proved to work more efficiently with the TF-IDF vectorizer as compared with other models incorporated with various vector space techniques. The skylearn (0.20.3) machine learning package was imported to avail us the chosen algorithm as well as other algorithms to be used for evaluation and testing. In carry out classification and prediction of susceptibility of post to social engineering we have utilized the logistic regression machine learning algorithm. we first split the data set into training and testing (X,Y) dataset in the ratio of 70% to 30% respectively. By so doing, the algorithm will be able to automatically learn and predict results via the test dataset been used. And for subsequent posts from OSN users, the algorithm will also be able to predict its susceptibility to social engineering

attacks based on the count of PII and recommend if the post should go live or not. Of cause it's still left to the user to decide whether he or she wants to go ahead with posting it online but it creates a real time awareness to the user to be able to act promptly which in turn will reduce the amount PII going online which social engineers can utilize to perpetuate their attacks.

## 3.6 Front End User Notification

To finally achieve our aim and purpose, the results on prediction of our model (SE-MPLM) has to be displayed to the user to give the OSN user a real time awareness on the susceptibility of his or her post to social engineering attacks and pointing out the personal identifiable information(s) contained in the post. To append results of our model to the frontend user we made use of the python sublime text editor to import flask and integrate our model to a front end designed from HTML. The level of severity was labelled non-susceptible, less susceptible, moderately susceptible and highly susceptible depending on the score of the number of PII contained in the best. We believe that by doing so, the social media platform user will be able to be given real time awareness and also be able to vet his post if need be. There are various tools and software used in defining and building this model for every step of the methodology. We shall be defining what they mean before analysing how they are been implemented in our model.

## 3.7 Model Testing and Evaluation

In carrying out evaluation of our model, the process will be broken into three parts; Model interpretation, evaluation of performance and accuracy and Testing.

### a. Model interpretation

The interpretation of the model was carried out to ascertain why our model made such predictions. This is helpful to further understand our model and the kind of results it gives. In the interpretation of our model, a function called ELI5 will be used to know why our model made such prediction.

### b. Evaluation of performance and accuracy

For our Model to be certified reliable in detecting personal identifiable information in a post to ascertain its level of risk to social engineering attacks, a performance and accuracy scale of measurement has to be introduced. For this project, the confusion Metrix was utilized. A confusion Metrix is used to measure the performance of a machine learning model in this case (SE-MLPM). It is very useful in measuring recall, precision, accuracy and specificity. It measures in four scopes such as positive prediction when it is actually true (true positive), negative prediction and its true (true negative), positive prediction and its false (false positive) and negative prediction and its false (false negative).

### c. Testing

Testing of the model was carried out in a controlled situation where inputted posts in the front end of the application were controlled to contain personal identifiable information PII which could be detected and analysed by our model.

## 4. Implementation

In the implementation of SE-MLPM, the process was carried out in two phases. First is the development of SE-MLPM at the back end and its integration to the front end which was carried out using an application called Flask. The front end was designed to actualize the purpose of our model to OSN users which is to notify them of how susceptible their post is to social engineering attacks based on the number of PII extracted from their post. The figure below gives a comprehensive workflow of the implementation process.

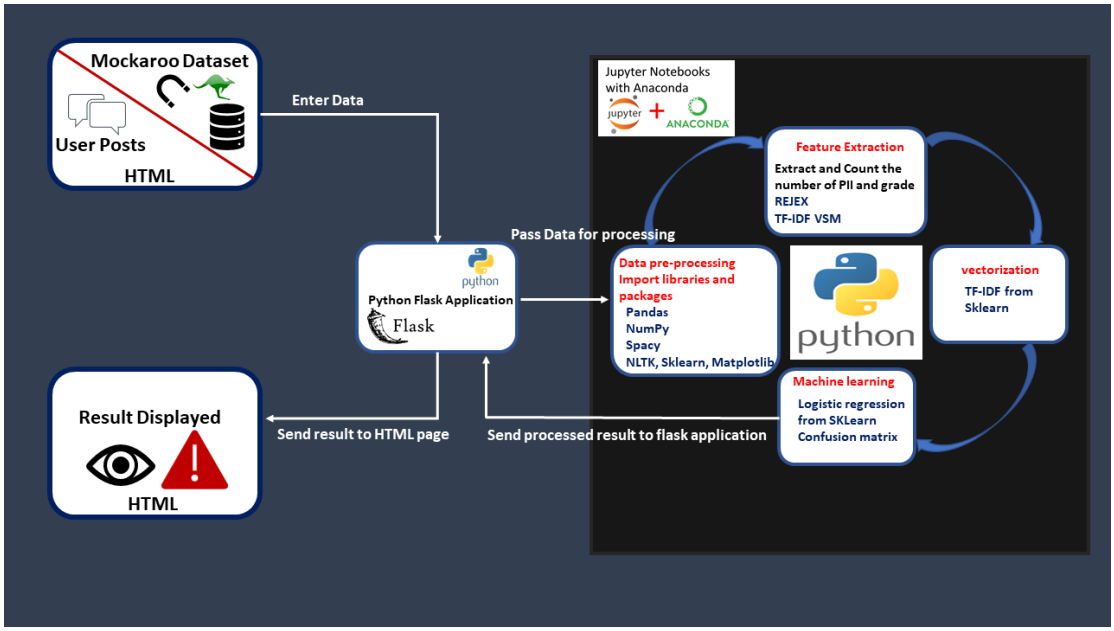


Figure 3. implementation flow process

**a. Stage 1**. Altogether, the data was collected from Mockaroo database generator. For this experiment, 5 attributes were selected which we believed were most useful in providing PII and useful in predicting susceptibility of posts to social engineering attacks. In addition, the dataset contained a total of 4000 rows and 7 columns. The attributes that were selected to be generated by mockaroo dataset generator were email address, phone number IBAN, state, names, time and regular phrases and expressions. The next step was to install the necessary programming language and packages locally and from the command line that would aid in the processing of the dataset, which included python 3.7.5, Anaconda navigator 3, Pandas 0.25.3, flask 1.1.1, Sklearn 0.21.3, Joblib 0.14.0 etc. followed closely by launching Jupyter notebook via anaconda command prompt and loading the data set first in the environment. the next step was data pre-processing which is an integral step to clean data and prepare it to be useful for any experiment related to machine learning.

**b. Stage 2:** In this study some pre-processing Steps were employed on the generated database. To begin with the size and shape of the dataset was analysed and found to be good enough for the machine learning implementation approaches. As described earlier, the shape and volume of our dataset could affect the result provided by the machine learning model. Thus, the application of synthetic data set could help us maximize the capacity of the dataset as earlier discussed. Secondly using Spacy and natural language tokenizer (NLTK), we were able to remove stopwords, punctuations, and tokenization was performed to clean the dataset. In order to address other discrepancies in our dataset, data lemmatization was performed as well using spacy.

**c. Stage 3:** The next step was important information retrieval using regular expression (REGEX). This tool defines patterns to which personal identifiable information could attribute to and thereby extracting them from the data. Then the entity count function is used to count the number of PII identified per roll in the dataset, which was further arranged into arrays and mapped to dictionaries of different susceptibly levels ranging from non-susceptible to moderately susceptible to less susceptible and highly susceptible using the map function. Furthermore, matplotlib was utilized to give a graphical representation of the extracted PII count, found within the dataset introduced.

**d. Stage 4:** The next step was to convert our data values into appropriate machine-readable formats. In this study two vectorizers were implemented to further help in the evaluation of our selected choice of vectorization technique and saved using the Joblib package. The term frequency – inverse document frequency was implemented as our choice of vectorization technique and the count vectoring technique was also implemented and saved separately for comparison on performance purpose. The vectorizers were imported using the sklearrn package. It is paramount to convert data type to binary attributes that machine learning algorithms can understand since we are going to be integrating the vectorized dataset with multiple machine learning algorithms for comparison and evaluation sake.

**e. Stage 5:** Here the data set was split into train and test data of the ratio of 70% (3200) and 30% (800) respectively, to learn the classification/prediction algorithm. In the building of our proposed model SE-MLPM, the logistic regression classification algorithm was utilized to classify and predict outcomes of PII detection based on their predefined classes of severity of risk to social engineering attacks using the Sklearn package. Furthermore, other models such as the naïve Bayes algorithm was separately implemented and saved to further compare its accuracy and precision with our suggested model. After the back-end implementation, it shows that our model had an accuracy of 56% which was higher than that of the Naïve Bayes classifier model on any of the VSM techniques of 52% (Naïve Bayes with TF-IDF) and 47% (Naïve Bayes with count vectorizer).

**f. Stage 6:** Lastly the sublime text editor was utilized to import the flask application to integrate our already developed model with the front end designed with HTML for the results of the classifiers to be displayed on front end using the flask application which helps to integrate machine learning models and front-end frameworks together. The results showed the prediction of how susceptible a post can be to social engineering attacks, also presenting to the user the various PII that makes up the post. The PII found and levels of risk to social engineering is displayed for the user to make a conscious effort on vetting his post and thus reducing the amount of PII that goes online thereby reducing SE attacks.

## 5. Experiment and Evaluation

### 5.1 SE-MLPM Performance and Accuracy

Due to our study and model exhibition, the novel solution proposed was able to detect and notify users of all the learned personal identifiable information such as phone numbers, credit card details and email addresses. An experiment was carried out to evaluate the accuracy and performance of our model. For SE-MLPM we have prioritised certain PII based on how to determine risk levels. In SE-MLPM we have prioritised emails, phone numbers, IBAN and bank details because of its

importance in social engineering attacks. These keywords hold high, moderate, less or no severity depending on the count.

The figure below is a typical representation of the use case scenario of our model.

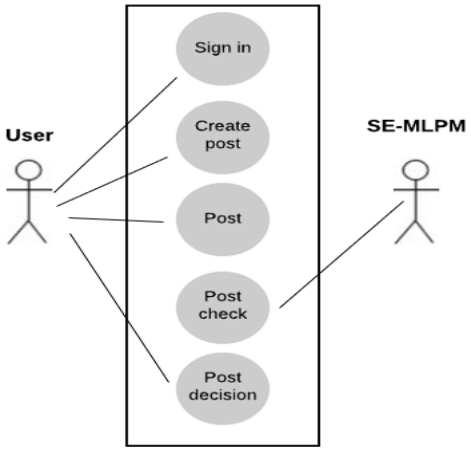


Figure 4. Use case for SE-MLPM

## 5.2 Use Case Testing

The use case testing of our model was carried out among 17 volunteered participants, this was done by deploying SE-MLPM on four computer systems. This was carried out in a controlled situation where participants were informed of the objective of SE-MLPM, and a guide that contained the pattern of the Personal identifiable information was handed over to them. In testing for the various susceptibility levels, participants were made to launch the application and print random controlled post in them. The figures below show the various results obtained by virtue of severity to social engineering attacks using the logistic regression prediction algorithm

**a. Non susceptibility:** First a non-susceptible post was inserted into the post section by a participant and the logistic regression model was selected. As shown in the figure below SE-MLPM predicted correctly. No PII was found in the post.

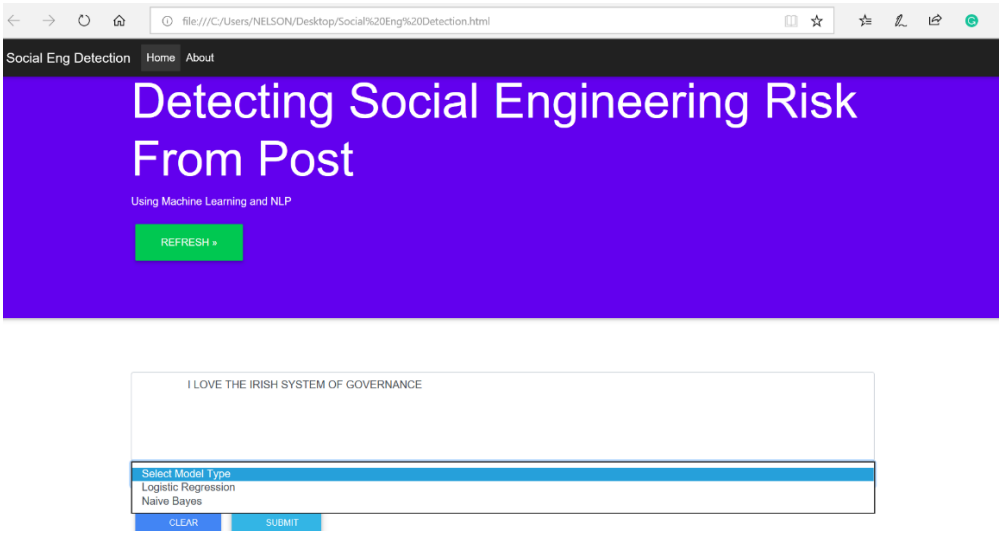


Figure 5. Selection of classification algorithm.

After click on the submit button the following result was obtained.

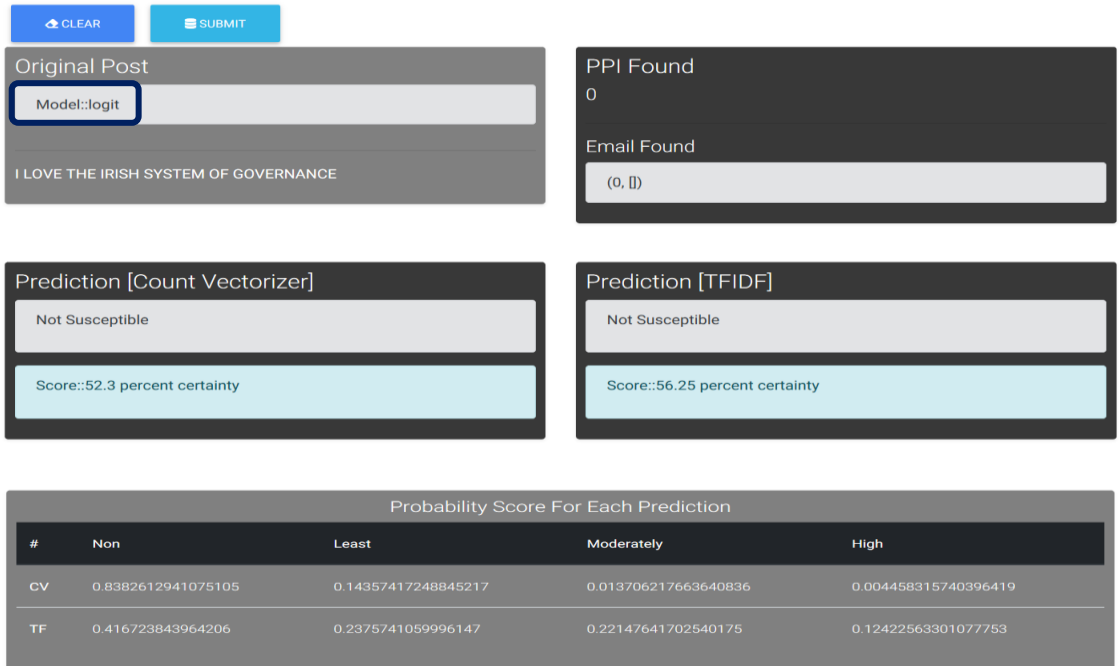


Figure 6. Non susceptibility result.

**b. Less susceptibility:** In the same vein, a less susceptible post was inserted by a participant in the post column and SE-MLPM using logistic regression predicted correctly. For PII count of 1, SE-MLPM predicts correctly that the post is less susceptible.

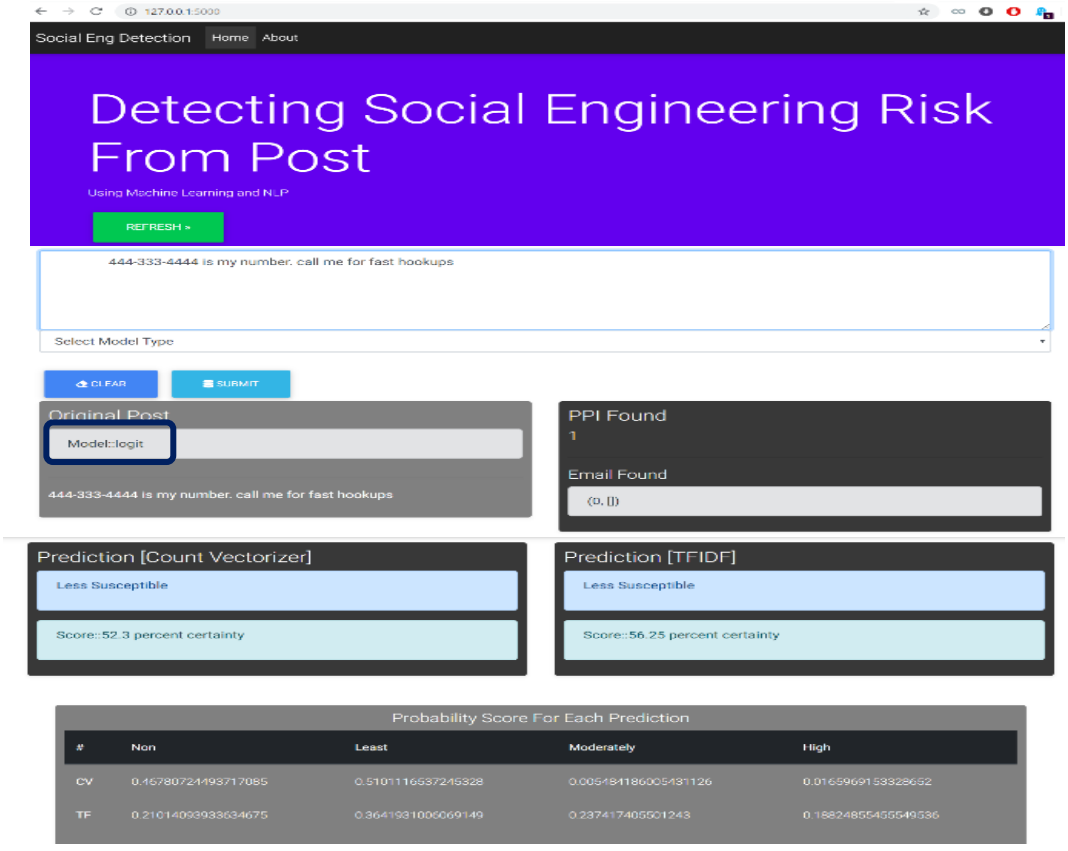


Figure7. Less susceptibility results

**c. Highly susceptibility:** Also, for count of PII of 3 and above SE-MLPM predicted it as highly susceptible and recommends the user not to post the original post.

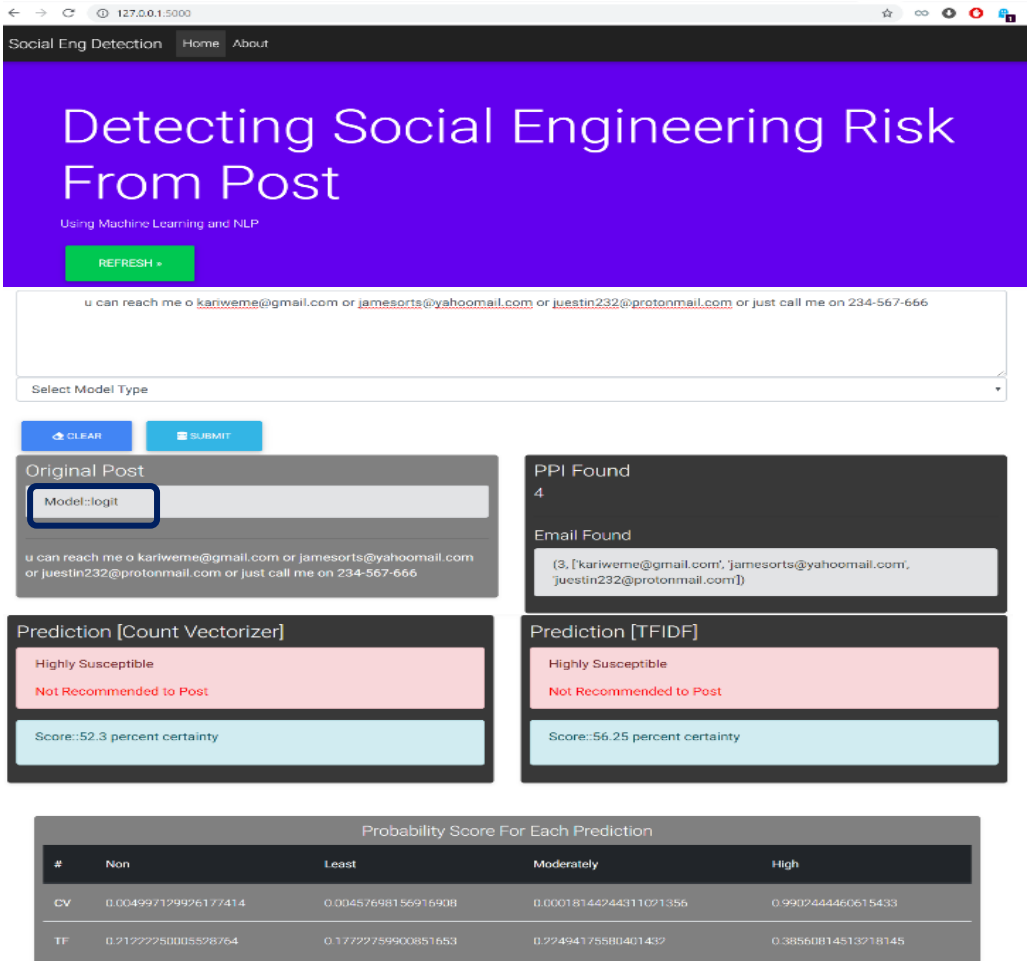


Figure 8. High susceptibility results

## d. Model interpretation

Utilizing the Elif5 function and confusion matrix, a graphical and tubular representation of the evaluation and interpretation of our model was produced. The figures below show the Python elif5 interpretation and confusion matrix evaluation of our model.

| **y=not_susceptible** top features | | **y=less_susceptible** top features | | **y=moderately_susceptible** top features | | **y=highly_susceptible** top features | |
|---|---|---|---|---|---|---|---|
| **Weight²** | **Feature** | **Weight²** | **Feature** | **Weight²** | **Feature** | **Weight²** | **Feature** |
| +1.575 | <BIAS> | +1.097 | x9540 | +3.137 | x7980 | +2.853 | x7530 |
| +1.081 | x13845 | +1.010 | x11071 | +1.450 | x6536 | +2.436 | x11958 |
| *… 1570 more positive …* | | +0.949 | x3593 | +0.922 | x3368 | +2.099 | x9291 |
| *… 12074 more negative …* | | +0.948 | x12668 | +0.881 | x943 | +2.002 | x8309 |
| -0.986 | x4312 | +0.946 | x7262 | +0.740 | x2446 | +1.950 | x11614 |
| -0.987 | x1260 | +0.941 | x4822 | +0.740 | x792 | +1.758 | x10279 |
| -0.993 | x1183 | +0.910 | x5068 | +0.716 | x13105 | +1.598 | x13012 |
| -1.029 | x4442 | +0.908 | x14634 | +0.705 | x4610 | +1.549 | x5603 |
| -1.046 | x3064 | +0.906 | x3173 | +0.684 | x3399 | +1.416 | x7895 |
| -1.066 | x4296 | +0.904 | x5433 | +0.680 | x3756 | +1.386 | x3108 |
| -1.067 | x1479 | +0.888 | x13635 | +0.677 | x4954 | +1.345 | x7509 |
| -1.069 | x1241 | +0.863 | x14796 | *… 5385 more positive …* | | +1.292 | x2430 |
| -1.132 | x1873 | *… 3812 more positive …* | | *… 8259 more negative …* | | +1.273 | x5609 |
| -1.185 | x5236 | *… 9832 more negative …* | | -0.674 | x10626 | +1.236 | x2492 |
| -1.228 | x2435 | -0.862 | x1855 | -0.714 | x9167 | +1.216 | x2134 |
| -1.337 | x5609 | -0.910 | x7509 | -0.736 | x9554 | +1.144 | x1175 |
| -1.341 | x4231 | -0.949 | x10279 | -0.742 | x11614 | +1.140 | x3097 |
| -1.397 | x1175 | -1.004 | x3399 | -0.786 | x11958 | +1.131 | x4583 |
| -1.399 | x1253 | -1.101 | x11958 | -0.798 | x3108 | +1.030 | x4940 |
| -1.880 | x6536 | -1.124 | x7530 | -0.801 | x1183 | *… 6014 more positive …* | |
| -2.204 | x7530 | -1.338 | x2476 | -0.886 | x7530 | *… 7630 more negative …* | |
| -2.578 | x7980 | -1.455 | <BIAS> | -4.357 | <BIAS> | -5.241 | <BIAS> |

Figure 9. SE-MLPM model interpretation

The best possible tuning was applied on the model to provide a suitable performance with the following parameters:
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='warn', n_jobs=None, penalty='l2', random_state=None, solver='warn', tol=0.0001, verbose=0, warm_start=False)

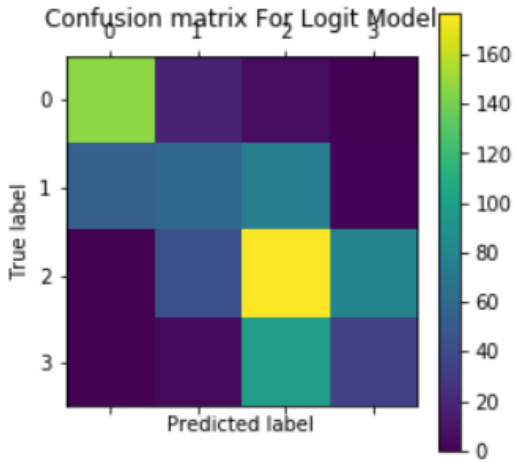


Figure 10. Confusion matrix for SE-MLPM

## Evaluation of performance

Three models were evaluated and compared with SE-MLPM, namely the TF-IDF Naïve Byes model, the countVec logistic regression model, and the countVec Naïve Bayes Model by 20 participants. Also, the performance confusion matrix for both classifiers were evaluated. . The same input data was used for the 3 models and the results were evaluated.

**a. TF-IDF Naïve Bayes Model:** From figure 9 below highlighted in green, we can see from our front end that the TF-IDF Naïve Bayes Model did not provide the desired result on evaluation. From the objective set, a PII count of 4 is supposed to reveal a highly susceptible label, instead it reveals a moderately susceptible label, and it didn't recommend to the user not to post.

**b. CountVec Naïve Bayes Model:** Also, from figure 9 below highlighted in red we can see from our front end that the CountVec Naïve Bayes did not provide the desired result on evaluation. From the objective set, a PII count of 4 is supposed to reveal a highly susceptible label, instead it reveals a moderately susceptible label, and it didn't recommend to the user not to post.

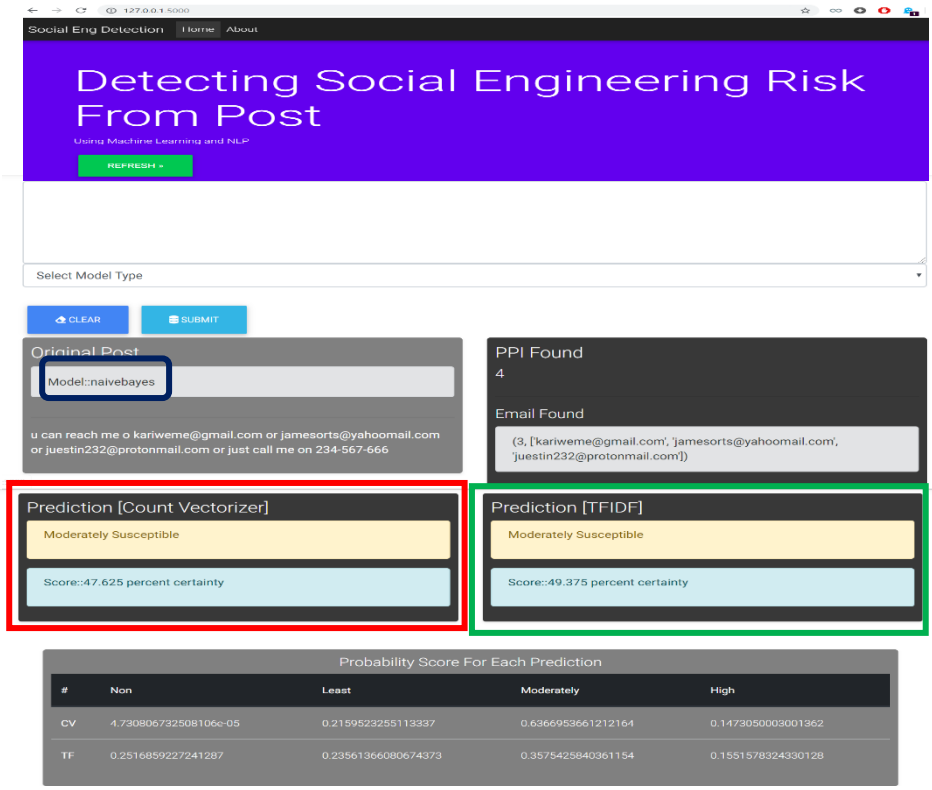


Figure 9. CountVec Naïve Bayes Model and TF-IDF Naïve Bayes Model prediction

**c. CountVec logistic Regression Model:** From figure 10 below, it shows that the accuracy on prediction was 52.3% (In green) which was less than that of our model of 56%. This means that our model has a greater probability of predicting the right result even though at this time both models predicted correctly.

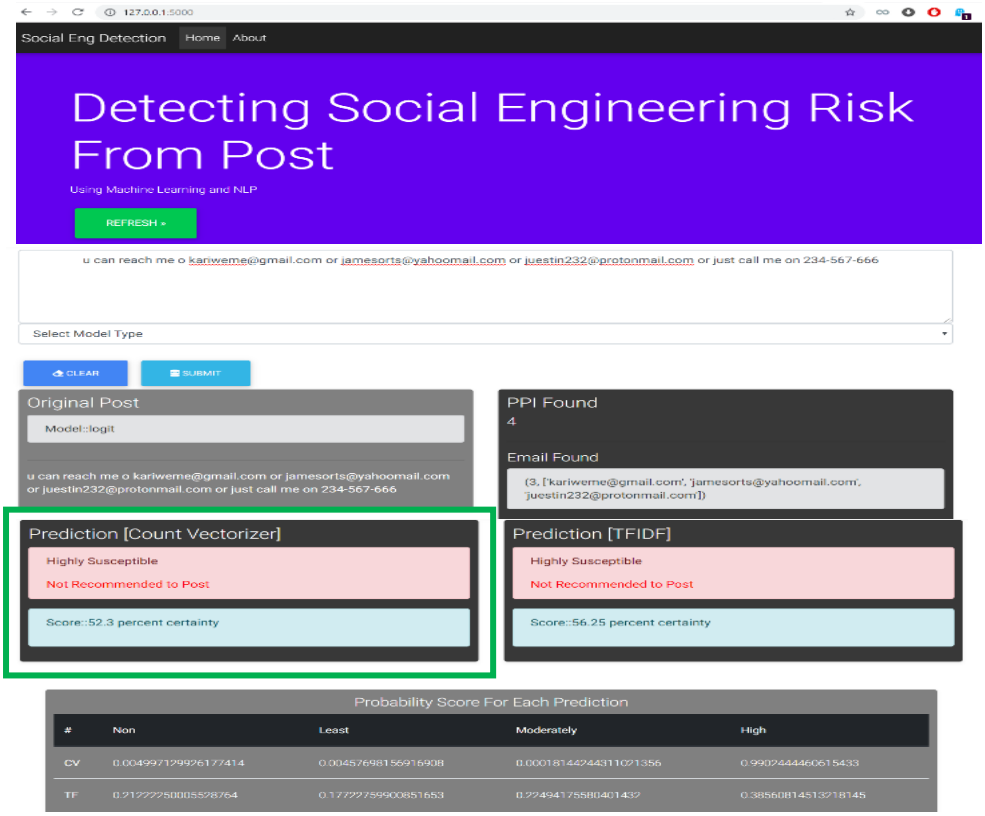


Figure 10. CountVec logistic Regression Model comparism

## 6. Conclusion

Although Online social networks are considered to be highly beneficial to OSN users, it can also prove be a two-edged sword. From our research, it is evident that the major prerequisite for a social engineer to carry out social engineering attacks is are basically personal identifiable information which could be released via OSN, thus the more users share PII on OSN, the more susceptible they are to social engineering attacks such as spear phishing attack, identity theft etc. we have understood that if we prevent the dissemination of personal identifiable information from the users, it will be tough for social engineers to gain SPI, hence it will not be possible for them to spring and attack. From our research we also believe that in recent times there is an automated technology that can be implemented to cater for this solution, called machine learning. Consequently, we have been able to build a social engineering risk detection model (SE-MLPM) which incorporated the TF-IDF VSM technique and the logistic regression classification algorithm, which have proved its potential of been able to predict and notify users on the susceptibility of their intended post to social engineering attacks with an added advantage of revealing the PII detected in the post, and categorizing posts based on PII count found. The major limitation encountered while developing SE-MLPM is that we could not get real and accurate data from individuals even with the acquisition of ethics form due to trust issues from the part of the individuals as regards the conscious release of their personal identifiable information.

### 6.1 Future Research

1. Spotting further improvements to the SE-MLPM model is expected to provide a better productive future work for this research. The introduction of real individual datasets can be sorted to implement the model. Although due to certain factors such as data ethics and collection volume required, it has proved to be a dauting task while carrying out this research. This can be sorted to be merged with or compared to the synthetic dataset to see if accuracy levels can be improved.

2. It is important to note that we have based our research on certain important and most prominent social engineering susceptible keyword patterns which has given us accurate results. For future works more of these PII patterns could be included during feature extraction such as location pet names, hobbies etc. to further give a robust field of PII detection.

3. As result of previous issue, other Natural Language Techniques analysis may lead to improvements in SE-MLPM. Natural language techniques evolve as improvement in technology is sorted, to this light a deeper natural language processing technique may improve on the model's efficiency.

4. New Machine Learning algorithms evolve continuously as technology improves. Some of the current algorithms have been reviewed and from our research our model had topped the chart in terms of accuracy and precision. But these algorithms are inexhaustive as technology advances. So future algorithms could be tested to see if the model can be improved upon.

5. Sometimes web browser users throw caution to the wind and input SPI in websites maybe due to free offers or naivety. Research can be carried out on how to integrate SE-MLPM with various web browser extensions. This would thus, be able to detect and notify web browser uses in real time on the on potential threats of information released online on browsers and the risk involved.

6. Finally, we envisage that SE-MLPM can be offered to business organisations as a service to be incorporated in their business applications or models which can detect and notify the business

enterprise on the threat of disseminating certain SPI that could be related to their business enterprise. In doing so we believe that fraud and identity theft can be mitigated via this process by reducing the amount of business SPI that can be revealed to unwanted personalities within or outside the organisation.

## 7. References


[1] D. Gonimah, "The Complexity of the Social Media Landscape," 13 Aug 2018. [Online]. Available: https://storyful.com/thought-leadership/the-complexity-of-the-social-media-landscape/. [Accessed 23 Aug 2019].

[2] K. Smith, "126 Amazing Social Media Statistics and Facts," 13 June 2019. [Online]. Available: https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/. [Accessed 24 Nov 2019].

[3] ODS, "The Most Famous Cases of Social Engineering," 2019. [Online]. Available: https://opendatasecurity.io/the-most-famous-cases-of-social-engineering/. [Accessed 12 Oct 2019].

[4] K. Lewis, "How Social Media Networks Facilitate Identity Theft and Fraud," 2019. [Online]. Available: https://www.eonetwork.org/octane-magazine/special-features/social-media-networks-facilitate-identity-theft-fraud. [Accessed 11 Nov 2019].

[5] S. institute, "Evolution of machine learning," 2019. [Online]. Available: https://www.sas.com/en_ie/insights/analytics/machine-learning.html. [Accessed 2 Nov 2019].

[6] S. H. P. Oo, "Intelligent access control policies for social network site," June 2013. [Online]. Available: http://www.airccse.org/journal/jcsit/5313ijcsit15.pdf. [Accessed 2 Aug 2019].

[7] M. S. A. A. G.-J. a. K. I. Shehab, "Access control for online social networks third party applications, Computers & Security 31(8): 897–911.," 2012. [Online]. Available: URL: http://linkinghub.elsevier.com/retrieve/pii/S0167404812001186. [Accessed 5 Aug 2019].

[8] D. M. a. E. N. B. Boyd, "Social Network Sites: Definition, History, and Scholarship, Journal of Computer-Mediated Communication 13(1): 210–230.," 2007. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/abstract. [Accessed 8 Aug 2019].

[9] G. a. S. J. M. Misra, "PACMAN: Personal Agent for Access Control in Social Media, IEEE Internet Computing 21(6): 18–26," 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8114620/. [Accessed 9 Nov 2019].

[10] L. S. T. B. D. a. K. E. Bilge, "All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks, Web Security p. 10," 2009. [Online]. Available: https://www.researchgate.net/publication/221023904_All_your_contacts_are_belong_to_us_Automated_identity_theft_attacks_on_social_networks. [Accessed 3 Aug 2019].

[11] O. Ololade, "Towards a Conceptual Model for Mitigating against Social Engineering on the Online Social Network," 2018. [Online]. Available: http://trap.ncirl.ie/3559/1/olabodeololade.pdf. [Accessed 3 June 2019].

[12] O. Kharkovyna, "Machine Learning vs Traditional Programming," 22 Apr 2019. [Online]. Available: https://towardsdatascience.com/machine-learning-vs-traditional-programming-c066e39b5b17. [Accessed 2 June 2019].

[13] Y. B. Woo, "My First Machine Learning Project: Designing a Hate Speech Detecting Algorithm," 3 May 2018. [Online]. Available: https://towardsdatascience.com/my-first-machine-learning- project-designing-a-hate-speech-detecting-algorithm-56ab32f10833. [Accessed 14 Nov 2019].

[14] K. Cornn, "Identifying Depression on Social Media," 2006. [Online]. Available: https://web.stanford.edu/class/cs224n/reports/custom/15712307.pdf. [Accessed 5 Nov 2019].

[15] i. insight, "Using Surveys for Data Collection in Continuous Improvement page 7," Aug 2006. [Online]. Available: https://www2.virginia.edu/processsimplification/resources/PennState%20Surveys.pdf. [Accessed 4 Aug 2019].

[16] R. R. M S Neethu, "Sentiment analysis in twitter using machine learning techniques," 2013. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6726818. [Accessed 9 Aug 2019].

[17] R. W. Neha Patki, "The Synthetic data vault," 2014. [Online]. Available: https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf. [Accessed 9 Aug 2019].

[18] M. news, "Artificial data give the same results as real data — without compromising privacy," 2017. [Online]. Available: https://lids.mit.edu/news-and-events/news/artificial-data-give-same-results-real-data-%E2%80%94-without-compromising-privacy. [Accessed 4 Nov 2019].

[19] G. S. et.al, "Synthetic Data Generation for Statistical Testing," 2017. [Online]. Available: https://people.svv.lu/soltana/papers/ASE17.pdf. [Accessed 2 Aug 2019].

[20] S. M. B. e. al, "Privacy and synthetic datasets," 19 Sept 2017. [Online]. Available: file:///C:/Users/NELSON/Downloads/Privacy%20and%20Synthetic%20Datasets.pdf. [Accessed 8 Nov 2019].

[21] V. e. al, "Synthesizing Tabular Data using Generative Adversarial Networks," 27 Nov 2018. [Online]. Available: https://arxiv.org/pdf/1811.11264.pdf. [Accessed 12 Nov 2019].

[22] A. Gonfalonieri, "Do You Need Synthetic Data For Your AI Project?," 21 Oct 2019. [Online]. Available: https://towardsdatascience.com/do-you-need-synthetic-data-for-your-ai-project-e7ecc2072d6b. [Accessed 23 Aug 2109].

[23] B. Shetty, "Natural Language Processing(NLP) for Machine Learning," 24 Nov 2018. [Online]. Available: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b. [Accessed 21 Nov 2019].

[24] M. e. al, "Distributed Representations of Words and Phrases and their Compositionality," 2013. [Online]. Available: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf. [Accessed 18 Nov 2019].

[25] A. e. al, "A Survey on Hate Speech Detection using Natural Language Processing page 2," Apr 2017. [Online]. Available: https://www.aclweb.org/anthology/W17-1101.pdf. [Accessed 8 Sept 2019].

[26] Y. Liu, "Topical Word Embeddings," 2015. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9314/9535. [Accessed 3 Aug 2019].

[27] D. Cam-Stei, "Word Embedding Explained, a comparison and code tutorial," 2019. [Online]. Available: https://medium.com/@dcameronsteinke/tf-idf-vs-word-embedding-a-comparison-and-code-tutorial-5ba341379ab0. [Accessed 4 Aug 2019].

[28] Y. W. e. al, "Comparisons and Selections of Features and Classifiers for Short Text Classification," 2017. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/261/1/012018/pdf. [Accessed 3 Aug 2019].

[29] S. Asiri, "Machine Learning Classifiers," 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623. [Accessed 3 Nov 2019].

[30] M. I. e. al, "Text Classification Using Machine Learning Techniques," Aug 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9153&rep=rep1&type=pdf. [Accessed 2 Aug 2019].

[31] W. e. al, "Comparisons and Selections of Features and," 2017. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/261/1/012018/pdf. [Accessed 4 Aug 2019].

[32] I. Arora, "Document feature extraction and classification," 19 Mar 2017. [Online]. Available: https://towardsdatascience.com/document-feature-extraction-and-classification-53f0e813d2d3. [Accessed 4 Aug 2019].

[33] A. multiple, "Synthetic Data: An Introduction & 10 Tools [2019 Update]," 2019. [Online]. Available: https://blog.aimultiple.com/synthetic-data/. [Accessed 3 Aug 2019].