Literature Survey: Data-driven Approach for Selection of an Ensemble Model of Profane Words Detection in Social Media

Abdulrehman A. Mohamed[1], Dr. George O. Okeyo PhD[2], & Dr. Michael W. Kimwele, PhD[3]
School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, P. O. Box 62000- 00200, Nairobi, Kenya
[1]almutwafy@gmail.com, [2]gokeyo@jkuat.ac.ke and [3]mkimele@jkuat.ac.ke

Abstract - Strong language is common to most cultures, but what makes the language blasphemous; vary from region to region of the world. In social media, people chat with each other without necessary face-to-face interaction, hence promoting a sense of being liberal without breaking the cultural norms, resulting in an increased in the use of profanity in cyber space. Although most social media platforms have developed profane filters to censor profanity but the cyber bullies have continuously improve on their profanity techniques to make the existing filters less and less effective. This has led to a major problem in social media resulting in potentially serious adverse effects on young users if undetected. It is against this background that the study was motivated to explore the literature survey of data-driven approach for selection of an ensemble model for profane words detection in social media, in order to improve the accuracy and reliability in detecting profane chats. In order to achieve this goal, the study explored books, scholarly articles and all other sources published and unpublished relevant to profane detection model in social media and their relevant technological theories. By so doing, it provided a body of knowledge, using a three phase approach of: summarizing, comparing & contrasting and criticizing & synthesizing of all sources in relation to the profanity problem being investigated. The outcome of the survey was two folds: the contribution of the body of knowledge to the development of the ensemble model and the designing of the data-driven methodology.

Index Terms – Profane Words, Ensemble Detection Model, Data-driven Approach and Social Media

1 INTRODUCTION

1.1 Background Study

Profane language is common to most cultures, but what makes the language blasphemous, vary from region to region. In the physical world is not uncommon to use profane words during conversations. While on social media, people chat with each other without necessary face-to-face interaction. As a result of which, virtual world promotes a sense of being liberal without breaking the cultural norms, hence an increase of use of profane words. This view concurred with a comparison research work done by [41], where the statistics in the physical world revealed an approximation of between 0.5% and 0.7% of profane words spoken, against one in every 13 tweets containing profane words in social media.

However, verbal profane words abuses in cyber space are now a serious problem. Where some online games and chat systems have profanity filters, but these filters require high degree of detection and suppression accuracy, so as to minimize false positive results [29]. Nevertheless, several attempts have been made to developed profane filters, profane blacklist or profane libraries such as GitHub collections, but the perpetrators continuously improve their profane tweeting and chatting techniques, rendering existing filters to become less and less effective.

Therefore, it was against this background, the study was motivated to explore the literature survey of data-driven approach for selection of an ensemble model for profane words detection in social media, in order to improve the accuracy and reliability in detecting profane chats or tweets. However, according to [7], most research works have encountered several challenges in solving profanity problem in social media. The three main challenges affecting profanity detection are: category of language spoken from region to region, types of profanity filters, and techniques used for detecting profanity.

1.2 Categories of Profanity Filters

It was elaborated by [40], that categorization of commercialized profane filters are classified in to three categories: blacklist filtering (profane words), free form whitelist filtering (non-profane words), and restricted entry whitelist filtering (text prediction of non-profane words).

According to [28], blacklist filtering is a class of filers that allowed users to type any content but before the content is display to the public, is compared it with an existing blacklist library for any possible profane words. If profane phrases are found, then are remove or replace by symbols such as asterisk (*) or at symbol (@) to render them unreadable. This is evident in an online game animal jam where young teenagers are protected using this technique against profanity.

It was elaborated by [33], that free form whitelist is class of filters that uses similar procedures as blacklist filtering technique. But the only difference is that, it uses a whitelist (non-profane words) for comparison of user's content. The filter will prevent any content which is not found in the whitelist. While an enhancement of it is called the restricted entry whitelist filter. Finally, restricted entry whitelist filter class has an option for users to predictive texts of whitelist words as are they type the content. The example of such filters is found in online gaming called bubble gums.

Even though, all these filters deploy list-based techniques to prevent profane words, but current trends show that, the techniques are becoming more vulnerable to ever evolving profanity techniques. These challenges are also shared by [58], where they elaborated that, the list-based system were ineffective in detecting profane words due to their failure to capture profane words disguised as symbols (f$#%) or intentionally misspelled (shiiiiit) words.

1.3 Profanity Detection Techniques

According to [2], there are several techniques proposed for detection of profane words such as list-based detection, supervised, semi-supervised and unsupervised machine learning (ML) classification models, natural language processing (NLP) and in computational linguistics (CL). Nonetheless, the issue of single accurate technique or model to solve the detection of profane words problem remain unsolved.

This was evident in the work done by [43], where they used linguistic features of part-of-speech - POS (verbs, nouns, adverbs etc.) to detect profanity. The tweet hash tags were used for classification of sentiments as positive, negative or neutral. The n-gram model (prediction of the next word) was used to predict the next evolved profane word using ML algorithms.

Moreover, in a similar work done by [50], where a sentimental classification framework based on N-grams, bag-of-words – BoW (words presented as multiset neglecting morphology and syntax), and Skip-grams (generalization of N-gram but the prediction of the next word may be skipped by leaving a gap) for classification of sentiments in social media, is also evident.

However, all these works are linguistic-based solutions using morphology and syntax, to develop known pattern for classification and prediction. These strategies are no match to ever evolving pattern employed by the perpetrators of profane words attack on young recipients in social media.

2 METHODOLOGY OF THE STUDY

2.1 Introduction

The literature survey methodology objective was to explore books, scholarly articles, and any other sources relevant to profane detection model in social media's strength and weakness and their relevant technological theories. By so doing, the study provided a descriptive summary, and critical evaluation of these works in relation to the profanity problem being investigated. The survey was designed to provide an overview of sources explored while researching profane models and demonstrated how the survey fits within a larger field of this domain.

2.2 Approach

The approach was based on searching the existing literature and making decisions about the suitability of materials to be considered in the review [15]. There exist three main coverage strategies. First, exhaustive coverage consisting of an effort to make comprehensive exploration of all material in order to ensure that all relevant studies, published and unpublished, are included in the review and, thus, conclusions are based on all-inclusive knowledge base.

The second type of coverage consisted of presenting materials that are representative of most other works in profane detection model, by searching for relevant articles in a small number of top-tier journals in this domain [51].

In the third strategy, was to concentrate on prior works that have been central or pivotal to this domain, which include empirical studies, conceptual papers, challenges, new methods and concepts used [15].

The final phase was to analyze and synthesize data. It consisted of summarizing, aggregating, organizing, and comparing the evidence extracted from the all the reviewed materials. The extracted data was presented in a meaningful manner suggesting a new contribution to the existing literature domain [34]. This was very important since [3] warned researchers that literature reviews should be much more than lists of papers and should provide a coherent lens to make sense of extant knowledge on a given topic.

The survey employed several methods and techniques for synthesizing quantitative (e.g., frequency analysis, meta-analysis) and qualitative (e.g., grounded theory, narrative analysis, meta-ethnography) evidence [18]. In summary the literature survey methodology employed was a 3 – Phase approach of: 1st Summarize 2nd Compare & Contrast and 3rd Criticize and Synthesize Approach as shown in the Figure 1.

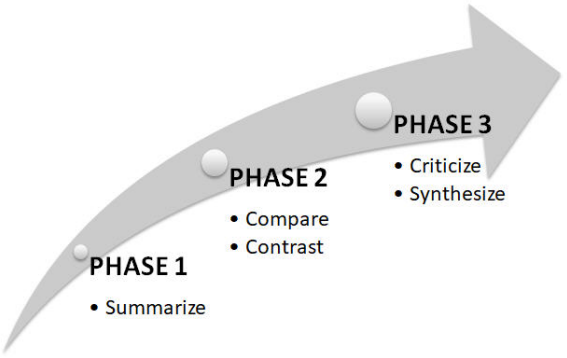

Figure 1: Summarize, Compare & Contrast and Criticize & Synthesize Approach

3 LITERATURE ANALYSIS

3.1 Introduction

The main objective of the survey was to create a body of knowledge from successful and unsuccessful works done by other researcher's experiments and theories with respect to data-driven approached for selection of an ensemble model for profane words detection in social media, in order to improve the accuracy and reliability in detecting profane chats or tweets.

To achieve these goals, the study discussed the most current relationships and gaps of profane detection models by other researches which did work and didn't work related to ten issues of profane word detection models of: one, Social Media Platforms; two, Algorithm Selection Approaches; three, Machine Learning Categories; four, Profane Detection Model Types; five, Social Media API Platforms; six, Data Preparation Techniques; seven, algorithm tuning; eight, training and test set; nine, Evaluation Model Metrics; and ten, Machine Learning Benchmarking Tools. Therefore, the literature survey was guided by following summary of the diagrammatic scheme of literature outline in Figure 2.

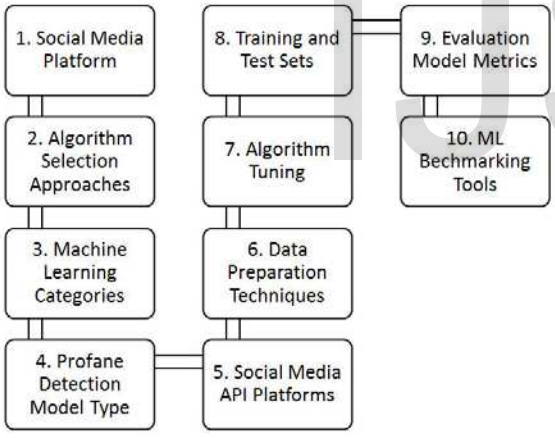

Figure 2: Survey Outline

3.2 Social Media Platforms

The main objective of this section was to review various social media platforms and select a suitable platform for proposed data-driven ensemble model development. In order to achieve this goal, the survey reviewed two popular social media platforms of: Facebook and Twitter

3.2.1 Facebook

According to [54], Facebook had an active monthly membership of 1.65 million users. They published enormous amount of data in real time, making it the single largest online publisher of content and hence becoming a world wide web on its own capacity. Likewise, it is reported by [65], that Facebook membership model collected a lot of personal data about its users, which is used for targeted marketing as its revenue model. Currently, it has 3 million active advertisers in its platform. In return, the users benefit from its collection of free service including exchange of messages, digital images, audio, video and links. Additionally, users can create and manage common interest group for private communication.

However, [3] reported that, most users of Facebook are concerned about bad sentiments including profane messages in chat rooms. Even though, according [16], that these concerns were addressed by Facebook by deploying profane filters but, these filters had failed to detect profane words disguised in form of symbols or intentionally misspelled. This was attributed by the profane model of blacklist, which was generated by the same users including the offenders. This created a need for development of alternative models to improve the accuracy and reliability of detecting profane message in Facebook.

3.2.2 Tweeter

According to [5], Tweeter was based on the concept of daylong brain storming session concept. The users discussed and contributed to a trending topic in brain storming session. The topics are classified by a phrase followed by hash sign called a hash-tag (#) for ease of following a trending topic and @ sign for identifying the creator of the message.

Even more importantly, it was reported by [23] that, Twitter had enormous impact on education, where several research works had been done by [38], [37], and [21]. However, Tweeter had also influenced political stages, where it could cause civil disobedience such as the use by Somalia's al-Shabaab rebels, who had their accounts suspended after they used the site to claim responsibility for an attack on the Westgate Shopping Mall in Nairobi in September 2013. Twitter had also been used to organize protests, sometimes referred to as "Twitter Revolutions", which include the 2011 Egyptian revolution [39], 2010–2011 Tunisian protests, 2009–2010 Iranian election protests, and 2009 Moldova civil unrest [25].

3.3 Algorithm Selection Approaches

The main objective of this section was to review various algorithm selection approaches and select a suitable approach for proposed data-driven ensemble model development. In order to achieve this goal, the survey reviewed two broad algorithm selection approaches of: Data-Driven Approach and Heuristic and Best Practices Approaches

3.3.1 Data-driven Approaches

It was asserted by [57], that data-driven approaches are based on the empirical analysis of data about a specific problem. In machine learning (ML), there are many algorithms in existence for solving various classification, detection or prediction problems. However,

the selection of an optimal algorithm to solve these particular problems remained a challenge.

Therefore, an empirical model is necessary for algorithm diversity selection of categories including rule-based algorithm, logistical regression, Bayesian, and k-Nearest Neighbor to address this gap. Hence, a data-driven approach can be used to find relationship between input and output without the explicit knowledge of the physical behavior of the model [63].

3.3.2 Heuristic & Best-practices Approaches

It is reported by [58], [43] and [46] that, some of the heuristic and best practices approaches for mapping of algorithms to a given problem had been proposed by many research works to address various problems. The two main approaches used are; implementation of a general classification problem mapping of supervised, semi-supervised and unsupervised algorithms, and a specific instance problem mapping of computer vision, natural-learning processing and speech processing.

However, according to [12], these techniques are limited in the transferability of the algorithm findings from one problem to another, due to classification problem categorization. Moreover, in any given problem, it is manifested by the dataset having different attributes including data types, number of instances or the number of attributes. This manifestation will adversely affect the performance of the same algorithm given another problem of the same classification. Hence the need for data-driven approach for algorithm selection for a given problem as opposed to heuristics and best practices approaches.

3.4 Machine Learning Categories

The main objective of this section was to review various categories of machine learning algorithms, and select a suitable category for proposed data-driven ensemble model development. In order to achieve this goal, the survey reviewed two broad grouping categories of: learning model algorithms and functional similarity.

3.4.1 Learning Model Algorithms

There are different ways an algorithm can model a problem depending on its interaction with the environment or the input dataset. In order to get the best results from an algorithm, its taxonomy is essential for definition of input dataset and model preparation process. Therefore, learning model algorithms are further classified into three categories of: Supervised Learning, Unsupervised Learning, and Semi-Supervised Learning.

Supervised Learning Algorithms: It was explained by [48], that supervised Learning is learning by example, where it involves the task of deducing a function from a labeled training dataset to make prediction. The training dataset consists of input data and corresponding response values. The supervised algorithm learns from training dataset by identifying patterns and builds a model to make prediction of the response values for a new dataset. The optimal goal will allow the algorithm to correctly identify the class label for unseen dataset.

Unsupervised Learning Algorithms: It was asserted by [19], that unsupervised learning is classification of machine learning that deduces patterns from dataset consisting of input dataset without labeled response value. It explores given dataset by cluster analysis methodology to find hidden patterns. The methodology uses similarity measure metrics including Euclidean distance or probabilistic distances to deduce patterns.

Semi-Supervised Learning Algorithms: It was elaborated by [24], that semi-supervised learning is a technique that takes advantage of both supervised learning and unsupervised learning by involving the function estimation on small amount of labeled dataset and large amount of unlabeled dataset for training. The technique is motivated by the fact that labeled dataset is more expensive to generate, as oppose to unlabeled dataset.

3.4.2 Functional Similarity Algorithms

It was explained by [13], that algorithms can be grouped according to their similarities of their functions. However, there are some algorithms which may be classified to fit into multiple categories, such as Learning Vector Quantization (LVQ); can be classified in both neural networks and instanced-based. In order to address this problem, the study preferred a strong bias toward algorithms used for classification and regression for supervised ML algorithm due to the study's problem domain of profane and non-profane binary problem.

As a result of this, the study discussed the following functional similarity algorithms: ruled-based algorithms, regression algorithms, Bayesian algorithms, and instanced-based algorithms:

Rule-based Algorithms: It was elaborated by [35], that rule-based algorithms classification technique uses an algorithm to produce rules from a given dataset. The generated rules are then applied to new unseen dataset for possible prediction. Rule-based algorithm provides mechanisms that generate rules by concentrating on a specific class at a time and maximizing the probability of the desired classification. The classification rules, r=<a, c>, consists of: a (antecedent/precondition): a series of tests that are evaluated as true or false; and c (consequent/conclusion): the class or classes that are applied to instances covered by rule r.

Regression Algorithms: It was elaborated by [61], that regression is concerned with modeling the relationship between dependent (target) variables that is repeatedly refined using a measure of error in the predictions made by the model and independent variable(s) (predictor). In order to achieve its objective regression analyses data by fitting a curve/line to the data points, by minimizing the distance of data points from the curve/line. There are various types of regression techniques dependent on three metrics: number of independent variables, shape of regression line and type of dependent variable: such as linear regression; that has

continuous dependent variable, discrete /continuous independent variables(s) and shape of regression line is linear.

Bayesian Algorithms: It was described by [49] the Bayes' theorem can be used to calculate the probability of a hypothesis in a given prior knowledge. In a classification problem, hypothesis (h) may be the class to assign for a new data instance (d). Hence, the Bayes' Theorem is stated as: P(h|d) = (P(d|h) * P(h)) / P(d): where P(h|d) is the probability of hypothesis h given the data d (posterior probability); P(d|h) is the probability of data d given that the hypothesis h was true; P(h) is the probability of hypothesis h being true regardless of the data (prior probability of h); and P(d) is the probability of the data (regardless of the hypothesis). After calculating the posterior probability for a number of different hypotheses, one can select the hypothesis with the highest probability. The maximum probable hypothesis known as maximum a posteriori (MAP) hypothesis and calculated as: MAP(h) = max(P(h|d)).

Instance-based Algorithms: It was elaborated by [55], that instance-based learning also known as lazy learners is a family of learning algorithms that, compares new problem instances with instances seen in training, which have been stored in memory as oppose of performing generalization. It is known as instance-based because it constructs hypotheses directly from the training instances themselves. This means that the hypothesis complexity can increase with the data in the worst case.

3.5 Profane Detection Model Types

The main objective of this section was to investigate various types of profane detection models suitable for reliable model development. In order to qualify and quantify the model selection, the survey reviewed two categorizes of profane detection models of: single-classifier models and ensemble models in machine learning algorithms.

3.5.1 Single-Classifier Algorithm Models

It was asserted by [64], that single-classifier algorithms had been used extensively for profane word detection is social media. However, they could create an imbalance learning problem of minority classes and majority classes leading to misleading results. This was evident especially in fraud detection and medical diagnosis problems, where minority classes are more important than majority classes. As a result of this, the survey study proposed ensemble detection model for profane words detection in social media.

3.5.2 Ensemble Algorithm Models

The sub-objective of this section was to explore the optimal selection of ensemble model types. The outcome of this discussion was to inform the data-driven ensemble model development. In order to achieve this sub-goal, the survey study discussed the three most popular ensemble model types of: Bagging, boosting and stacking

Bagging Ensemble Model: The bagging ensemble model was first proposed by [10], where its main objective was to create multiple models then combined them in to a single improved model using statistical methods. Bagging is an acronym, which stands for Bootstrap Aggregation method. Where, bootstrap is a method of reducing variance and retaining bias by dividing the dataset in to multiple copies of training set using random sampling with replacement and thereafter, each of these copies are used for training different model. On the other hand, aggregation is a method that uses test dataset by combing the outputs of the different models, either by averaging (in case of regression) or by voting (in case of classification), to create a single improved model.

Boosting Ensemble Model: The boosting ensemble model was proposed by [56], where its main objective was to convert weaker models into stronger models in order to increase the predictive accuracy. To achieve this goal, it was asserted by [62], the boosting method involves developing a model from training dataset, then create a second model that will attempt to correct any errors from the first model. Thereafter, more models are added and error corrections of preceding models are calculated by the current model. Finally, whole procedure is repeated until either, a perfect model is developed, or the maximum number of the added models is reached.

Stacking Ensemble Model: The stacking (blending) ensemble model was proposed by [67], the main objective of the model was to combine several models together to compensate their weakness and take advantage of their strength with the goal of reducing the generation error, hence also known as stacked generation. Stacked generalization is an approach of minimizing generalization error rate of one or more generalizers by deducing the biases of generalizers with respect to learning dataset. It was elaborated by [36], stacking combining mechanism is that, the output of the classifiers (Level 0 classifiers) are used as training dataset for another classifier (level 1 classifier) to estimate the same target function. This means the Level 1 classifier will try to deduce the combining mechanism. The table 1 summarizes the comparison of the three ensemble model.

Table 1: Comparison of Ensemble Model

Techniques	Bagging	Boosting	Stacking
Splitting of the data into subsets	Random sampling replacement	Giving misclassified higher preference	It uses various methods
Main Objective	Reduced variance	Increased predictive accuracy	Reduced variance and increased accuracy
Methodology used	Random Subspace	Gradient descent	Blending
Single model combine function	Weighted average or weighted voting	Weighted majority vote	Logistic regression

3.6 Social Media API Platforms

The main objective of this sub-section was to analyze various social media API platforms to be used as a data collection instrument for model development. In order to achieve the sub-goal, the study purposively discussed

the two most popular social media platforms of Twitter API platform and Facebook API platform.

3.6.1 Tweet Archivist APA

It was asserted by [20], that Tweet Archivist is a proprietary, hypermedia, and device independent API. It is built on REST architectural and supports request formats such as URI Query String/CRUD and JSON. It allows users to search Twitter for tweets by sender, recipient, object of reference, or contents. Users may then create an archive based on that search which they can analyze, export, and share tweets. The Tweet Archivist API extracts various tweets' attributes from Tweeter including hash-tags, volume over time, top users, Tweet vs. Re-tweet, top words, top URLs, and the source of Tweets.

The tweets can then be downloaded into Excel or PDF formats for further interpretation and analysis. Nonetheless, Tweet Archivist is not an open source software and hence, for research purpose where funding is limited; the study has to bear some cost. However, the non-premier service of Tweet Archivist can be access for free in a one month trial evaluation, which is sufficient time to collect the study's dataset.

3.6.2 Facebook API

It was elaborated by [11], that Facebook is a proprietary, hypermedia, and device independent API. It is built on REST architectural, RESTful protocol and responds to request formats such as URI Query String/CRUD and JSON. It is a platform for building applications that are available to the members of the social network of Facebook.

The API enable other applications to use the social connections and profile information to make applications more involving, and to broadcast events to the news feed and profile pages of Facebook, subject to specific users privacy settings. Moreover, users can add social context to their applications by exploiting profile, friend, page, group, photo, and event data.

3.7 Data Preparation Techniques

The main objective of this sub-section was to determine various data preparation techniques as a prerequisite for machine learning algorithms implementation. To achieve this goal, the survey study discussed two data preparation techniques: crowd-sourced Analytics and data preprocessing techniques.

3.7.1 Crowd-sourcing Analytics

According to [45], crowd-sourcing analytics can be defined as a platform that combines human expertise with machine learning techniques to unveil and support broader analytics use of unstructured data. The main objective of crowd-sourced analytics is to combine human knowledge and expertise with computing power to help solve problems and scientific challenges that neither machines nor humans can solve alone. According to [53] crowd-sourced analytics have been successfully applied in scientific works such as crowd tracking of hummingbirds, identification of cancer cells, classification of planet features in Mars, and more recently, classification in machine learning algorithm.

This report concurred with the views of [17], that current research in crowd-sourcing focuses on micro-tasking such as labeling dataset of images or text and designing algorithms by considering simplistic models of workers' behavior. It was elaborated by [47], that recently, there has been an increase of crowd-sourced analytics platforms including Kaggle, Amazon, CrowdAnalytix and TunedIT providing successful solution to business problems. This was evident as reviewed by [14] that successful competitions such as Kaggle's Heritage Health Prize; which generated an algorithm that could accurately predicted patients to be admitted in the hospital for proactive preventive measures

However, the selection of the crowd-sourced analytics platform will depend on various metrics such as crowd size, age, gender, nationality, and skill-knowledge-expertise. Though, the most important metrics are the crowd size and skill-knowledge-expertise. The crowd size is a measurement of the number of data scientists in a platform with respect to the extent of expertise to solve a crowd-sourced problem [9]. While skill-knowledge-expertise, is a measure of an individual skill-set or knowledge within the crowd to solve a crowd-sourced problem [47]. This measure is derived from computing the platform's ranking with respect to metrics such as postgraduate education, experience, and number of competition won by individual data scientists.

3.7.2 Data Preprocessing Techniques

It was reported by [26], that data preprocessing is a major and essential stage whose main goal is to obtain final data sets that can be considered correct and useful for further data mining algorithms. Therefore, the sub-objective of this section was to explore the machine learning data preprocessing techniques for a possible high quality dataset.

In order to achieve this sub-goal, the study discussed three data preprocessing techniques: Imperfect Data, Dimensionality Reduction and Instance Reduction.

Imperfect Data: It was elaborated by [27], the objective of imperfect data algorithms is removing the noisy data and impute the missing one. The missing values can be describe as raw data that have not been stored or gathered due to a faulty sampling process, cost restrictions or limitations in the acquisition process of data mining. While in noise treatment, data mining algorithms preferred the dataset to have a normal distribution curve. In supervised learning noise can affect the input features hence affecting the output results, leading to high biasness. Hence, noise filters are used to remove noisy instances in training sets.

Dimensionality Reduction: The objective of dimensionality reduction is to identify and remove any

irreverent and redundant information from dataset and also to reduce the number of dimensions. This is achieved by space transformation: technique that generates a set of new features by combing the original features using linear methods such as factor analysis and principle component analysis (PCA), and feature selection: a process of identifying and removing irrelevant and redundant information as much as possible by acquiring a subset of features from original data set, which is used to train the learner and reduced over-fitting.

Instance Reduction: The objective is involving a series of techniques that must be able to choose a subset of data that can replace the original data set and also being able to fulfill the goal of a data mining application. This is achieved by either instance generation: is a randomized approach that applies instance categorization to data sampling, by generating minimum data subset used without reducing the performance [27] or instance selection: an identification approach of optimal objects subset of original training data by discarding noisy and redundant samples. They use prototype generation methods which create a representation or subset of original instances [8].

3.8 Algorithm Tuning

The objective of this sub-section was to enquire the best machine leaning benchmarking platform to implement algorithm tuning with respect to hyper-parameterization. Algorithm tuning is a process of setting different values for different models training and selecting the best values that yield the best performance. It is achieved by hyper-parameterization, defined as higher level concepts about a model including complexity or capacity to learn [4].

To achieve this objective the study discussed two machine learning benching platforms of: Python platform and WEKA benchmarking platform with respect to algorithm hyper-parameterization (tuning).

3.8.1 Algorithm Tuning in Python

It was demonstrated by [63] that Python has three basic features which can be tuned to improve the predictive power of the model: one max_features: these are the maximum number of features Random Forest is permissible to explore per tree, with three options: one, auto/none (apply to all the features in every tree or no restrictions); two, sqrt (apply square root to total number features per run); and three; 0.2 (apply 20% of variables per run or apply 0.x); two n_estimators: these are the number of trees to be developed before applying the maximum voting or averages of predictions. Higher number of trees results in performance slower code.

The choice of high value is processor dependent to make predictions more stable and reliable, and three; min_sample_leaf: smaller leaf generates models which are more likely to capture noise in the training data. However, one should explore multiple leaf sizes to realize the most optimum results for a specific domain.

3.8.2 Algorithm Tuning in WEKA

Waikato Environment for Knowledge Analysis (WEKA) machine learning tool provides a graphical user interface for exploring and experimenting with machine learning algorithms on datasets using Java, unlike Python platform; where the researcher is forced to write code. Therefore, WEKA can be used by researchers to implement machine learning algorithm with no prior knowledge of any programming language. Furthermore, WEKA can be used as an experimental tool to tune algorithms such as random forest or k-nearest neighbor also known as IBk to increase predicting power of the model.

The WEKA GUI is user friendly where to run the experiment, simply click run in WEKA environment to configure the experiment and start to begin the experiment. To review the results, simply click analyze to open experimental results panel. To rank the tuning, choose ranking on the test base then apply perform test. To check performance, in the test base choose IBk algorithm with Manhanttan Distance then apply perform test, similarly repeat the process with Euclidean distance and Chebyshev distance to compare the performance.

3.9 Training & Test Sets

The objective of this sub-section was to query the best machine leaning strategy of training and test sets for running experiments for model building.

To achieve this objective the survey study reviewed two machine learning strategies of training and test sets: cross validation and percentage split

3.9.1 Cross Validation

According to [30], cross-validation is a model validation technique for evaluating how the outcomes of a statistical analysis can be generalized into an independent data set. It is basically applied in settings where the objective is prediction, and one needs to approximate how accurately a predictive model will perform. In prediction problem, a model is given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).

The objective of cross validation is to define a dataset to "test" the model in the training stage (i.e., the validation dataset), in order to reduce problems like over-fitting, give an insight on how the model will generalize to an independent dataset (unseen dataset).

3.9.2 Percentage Split - Partition

It was asserted by [60], the percentage split or partition is used for model evaluation and for model selection but the role of the test data is at best indirect. The simplest partition possible for cross-sectional data is a two-way random partition to generate a learning (or training) set and a test set (or validation set). The division of the data into learning dataset and testing dataset must be created

carefully to avoid introducing any systematic differences between learning and testing. In order to avoid systematic difference between the partitions the use of random assignment is recommended.

3.10 Model Evaluation Metrics

The objective of this sub-section was to review the measure of quality for evaluating the performance of machine learning and classier algorithms. To achieve this objective the survey study reviewed Confusion Matrix and three performance metrics of: accuracy, ROC area and F-Measure.

3.10.1 Confusion Matrix

It was elaborated by [52], that machine learning have several measures of evaluating the performance of learning algorithms and the classifiers generated. The measures of the quality of classification are built from a confusion matrix which records correctly and incorrectly recognized examples for each class. The Table 2 below presents a confusion matrix for binary classification, where tp are true positive, fp – false positive, fn – false negative, and tn – true negative counts.

Table 2: Confusion matrix for binary classification [42]

Class/Recognized	As Positive	As Negative
Positive	tp	fn
Negative	fp	tn

The following are some of the equations for calculating various evaluation measures:

1. Accuracy = {tp+tn}/{tp+fp+fn+tn} Equation 1
2. Sensitivity = {tp}/{tp+fn} Equation 2
3. Specificity = {tn}/{fp+tn} Equation 3
4. Precision (P) = {tp}/{tp+fp} Equation 4
5. Recall (r) = {tp}/{tp +fn}- Sensitivity Equation 5
6. F-Measure = 2[{P*r}/{P+r}] Equation 6

Therefore, in summary, confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix [42].

3.10.2 Performance Matrix

The most commonly used performance matrix in machine learning algorithms based on confusion matrix are: Accuracy, ROC and F-Measure:

Accuracy: It can be defined as the ratio of total of true positive and true negative to a total of all prediction. Accuracy = (1 – error rate) is a standard method used to evaluate learning algorithms. It is a single-number summary of performance [6]. It can also be defined as degree of closeness of measurement of a quantity to that quantity's true value [31].

ROC: Receiver Operating Characteristic can be defined as the graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

F-Measure: F-Measure, or F1 score, or F-Score, can be defined as the measures of accuracy using precision p and recall r. While, Precision is the ratio of true positives (tp) to all predicted positives (tp + fp), and Recall is the ratio of true positives to all actual positives (tp + fn). The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously.

3.11 Machine Learning Benchmarking Tools

The objective of this sub-section was to realize the best machine leaning benchmarking platform to conduct domain's experiments.

To achieve this objective the study survey reviewed two machine learning benching platforms of: WEKA (Waikato Environment for Knowledge Analysis) benchmarking platform Rapid-Miner

3.11.1 Rapid-Miner Environment

Rapid-Miner provides a graphical user interface (GUI) and a Java API for developing customized applications. It provides data treatment, visualization and modeling with a suite of machine learning algorithms. Moreover, according to [66], Rapid-Miner is being used in various industries including automotive, banking, insurance, life Sciences, manufacturing, oil and gas, retail, telecommunication, and utilities.

Furthermore, the tool have predefined blocks which act as plug and play devices, and more importantly, they allow custom R and Python scripts to be integrated into the system. The GUI is based on a block-diagram approach, similar to WEKA which is fully open source. However, the Rapid-Miner is open-source for only the old version (below v6) but the latest versions come in a 14-day trial period.

Even more importantly, it was reported by [1], that Rapid-Miner was originally started in 2006 as an open-source stand-alone software named Rapid-I. The current products of Rapid-miner have premier services including: Rapid-Miner Studio: stand-alone software which can be used for data preparation, visualization and statistical modeling; Rapid-Miner Server: an enterprise-grade environment with central repositories which allow easy team work, project management and model deployment; Rapid-Miner Radoop: implements big-data analytics capabilities centered around Hadoop; and Rapid-Miner Cloud: cloud-based repository which allows easy sharing of information among various devices

3.11.2 WEKA Environment

It was elaborated by [32], that WEKA (Waikato Environment for Knowledge Analysis) is a machine learning platform developed by the University of Waikato, New Zealand. It is programmed in Java and provides three interfaces: a graphical user interface (GUI), command line interface and Java API. It is perhaps the most popular Java machine learning library and is recommended for both leaners and experts for practicing and experimenting research works in machine learning. This popularity is attributed due to its support by machine learning community for help and its full availability as open source software unlike Rapid-Miner.

Moreover, it was reported by [44], that WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code through an API. More importantly, WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, visualization, and integrates R and Python.

Even more importantly, according to [22], that WEKA could be used by beginners in data science and the best part is that it is open-source. One can learn about it using the Massive Open Online Course (MOOC) offered by University of Waikato. This makes WEKA to be the preferred choice in the academic community and even in some industries.

4 FINDINGS OF THE STUDY

4.1 Introduction

The objective of this section was to explore body of knowledge contribution from the literature survey study. In order to achieve this goal the survey evaluated the literature survey evaluation and findings

4.2 Survey Evaluation

The critical evaluation consisted of synthesizing relationships and gaps of related works in Social Media Platforms, Algorithm Selection Approaches, Machine Learning Categories, Profane Detection Model Types, Social Media API Platforms, Data Preparation Techniques, algorithm tuning, training & test set, Evaluation Model Metrics and Machine Learning Benchmarking Tools as given below:

Social Media Platforms: The survey preferred Twitter over Facebook due to one, the preprocessing of data to remove noise is more effective with structured sentences than larger unstructured messages; two, Twitter provides an open source API for ease of data mining, as oppose to Facebook; three, there is a higher probability of detecting profane words is topic dependent, where the hash-tags of trending topics in Twitter are more structured than the groups of common interest in Facebook; and finally four, the daylong brain storming session concept of Twitter, make it more suitable for research works than friends networking concept of Facebook

Algorithm Selection Approaches: The survey preferred a data-driven approach over heuristic and best-practice approach for algorithm selection since it leads to limitation of transferability of the algorithm findings from one problem to another, given the same algorithm. While data-driven approach implements an empirical evaluation of a diversity of algorithms on profane dataset.

Machine Learning Categories: The results of the review preferred an empirical selection of a suite of supervised algorithm from rule-based algorithms (ZeroR as baseline Performance), Regression algorithm (Logistic Regression algorithm), Bayesian (Naïve Bayes algorithm), and Instanced-based algorithms (k-Nearest Neighbor algorithm) for data-driven ensemble model development over functional similarly algorithms since they introduce over fitting of the dataset

Profane Detection Model Types: The survey preferred ensemble model over single-classifier model, since model choice will depend on the domain's yielded results and prior selection of the best would difficult to apply in practice and hence preferred an empirical selection of ensemble model type of bagging, boosting and stacking.

Social Media API Platforms: The survey preferred Tweet Archivist API platform over Facebook API, although both APIs shared similar technologies in design and architecture model, but the choice of the API platform will depend on the domain specific to yield the desired results

Data Preparation Techniques: The survey preferred both techniques of crowd-sourcing analytics and data preprocessing as to argument each other, since each technique had uniquely addresses a specific problem, although they share similar goals. However, there is no single technique superior to another technique, since the choice of a particular technique will depend on specific algorithm area of application.

Algorithm tuning: The survey preferred WEKA platform over Python platform, where each technique had biasness towards certain parameter, although they share similar goals. However, WEKA was friendlier and did not need any prior knowledge of programming to be used, unlike Python where the researcher needed to learn the language before using the tool.

Training & Test Set: The survey preferred cross-validation test over percentage split test since each implemented different strategy to achieve similar goal. However, cross validation employed a more structure process of sampling as oppose to percent split which uses some randomness to achieve similar goal.

Evaluation Model Metrics: The survey preferred evaluation metrics of accuracy, ROC and F-Measures with respect to confusion matrix, since there is no single metric superior to another metric

Machine Learning Benchmarking Tools: The survey preferred WEKA benchmarking tool over Rapid-Miner, since WEKA was more favorable due to its

popularity among research community, ease of learn-ability and resource support, and more importantly, free download, as open source software.

4.3 Literature Survey Findings

The objective of this sub-section was to explore the results of the literature survey. In order to achieve this sub-goal the survey reviewed ten outcomes contributing knowledge to ensemble model development and data-driven methodology.

4.3.1 Contribution to Ensemble Model Development

It was evident from the discussion of the literature survey that the outcome contributions to the facilitation of an ensemble model development were achieved through the following five knowledge deliverables:

1. Tweets from Tweeter platform as suitable dataset
2. WEKA benchmarking platform to implement algorithm tuning.
3. Tweet Archivist API platform as a suitable data collection instrument
4. Cross validation technique for training and test dataset to run experiments
5. Evaluation metrics of accuracy, ROC and F-Measure for model evaluation

4.3.2 Contribution to Data-driven Methodology

It could be deducted from the literature survey that the outcome contributions to the facilitation of data-driven methodology were achieved through the following five knowledge deliverables:

1. Empirical test harness process for proposed data-driven methodology
2. Empirical selection of a suite of supervised algorithm from four categories
3. Empirical ensemble model type selection from ensemble models
4. Empirical data preparation of crowd-sourcing and ML preprocessing filters
5. WEKA platform to conduct experiments

5 SUMMARY & CONCLUSION

The main objective of the literature survey study was to review all materials of profane detection models to inform data-driven methodology and ensemble model development.

The results of the study were realized in two folds: facilitation of an ensemble model development and facilitation of data-driven methodology. The result of the survey study had five constructs contributing to ensemble model development and another five constructs to data-driven methodology. The Table 3 below summarizes the ten techniques, objectives, and the results of the study's literature survey.

Table 3: Summary of Literature Survey

TECHNIQUES	OBJECTIVES	RESULTS	TECHNIQUES	OBJECTIVES	RESULTS
1. Social Media Platforms	To explore categories of social media platform for a suitable dataset for reliable profane word detection model	Preferred tweets from Tweeter platform as suitable dataset for profane word detection	6. Data Preparation Techniques	To determine various data preparation techniques as a prerequisite for machine learning algorithms implementation.	Preferred an empirical data preparation of combination of crowdsourcing and machine learning preprocessing filters
2. Algorithm Selection Approaches	To discover algorithm selection approaches for reliable profane word detection model	Preferred and empirical test harness process for proposed data-driven methodology	7. Algorithm Tuning	To enquire the best machine leaning benchmarking platform to implement algorithm tuning with respect to hyper-parametrization	Preferred WEKA benchmarking platform to implement algorithm tuning.
3. Machine Learning Algorithm Categories	To review various categories of machine learning algorithms, to select a suitable category for ensemble model development.	Preferred an empirical selection of a suite of supervised algorithm from four categories	8. Training and Test Sets	To query the best technique of training and test dataset for running experiments for ensemble model development	Preferred cross validation technique for training and test dataset for running the experiments
4. Profane Detection Model Types	To investigate various categories of profane detection model suitable for reliable model development	Preferred an empirical ensemble model selection from bagging, boosting and stacking models	9. Model Evaluation Metrics	To study the best model evaluation metrics for model development	Preferred metrics of accuracy, ROC and F-Measure for model evaluation
5. Social Media API platforms	To analyze various social media API platforms suitable for data collection instrument for model development	Preferred Tweet Archivist API platform as a suitable data collection instrument	10. Machine Learning Benchmarking Tools	To realize the best machine leaning benchmarking platform to conduct experiments	Preferred WEKA platform to conduct experiments

BIBLIOGRAPHY

[1]. Aarshay, J. (2016). 19 Data Science Tools for people who aren't so good at Programming. Retrieved from https://www.analyticsvidhya.com/blog/2016/05/19-data-science-tools-for-people-dont-understand-coding/

[2]. Aggarwal, A. (2012). Scunthorpe Problem: Software Filters Misunderstand Words. Retrieved from http://www.labnol.org/tech/scunthorpe-clbuttic-mistake/14186/

[3]. Agrawal, H. (2015). Facebook Page Feature : Block Words and Profanity Blocklist. Retrieved April 16, 2016, from http://www.shoutmeloud.com/facebook-page-profanity-blocklist-spam-words.html

[4]. Amatriain, X. (2016). What are hyperparameters in machine learning? Retrieved from https://www.quora.com/What-are-hyperparameters-in-machine-learning

[5]. Avery, H., Kang, B., Mark, C., Yaschur, M., & Carolyn, M. (2014). Seeking and Sharing: Motivations for Linking on Twitter. Communication Research Reports, 31(1), 33–40.

[6]. Badgerati, A. (2010). Machine Learning – Confusion Matrix. Retrieved from https://computersciencesource.wordpress.com/2010/01/07/year-2-machine-learning-confusion-matrix/

[7]. Benedictus, L. (2014). What are the most popular swearwords on Twitter? Retrieved May 31, 2015, from http://www.theguardian.com/technology/shortcuts/2014/feb/23/most-popular-swearwords-on-twitter

[8]. Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowlege-Based Systems, 8(6), 33–45.

[9]. Brabham, D. C. (2010). Moving the Crowd at Threadless. Information Communication Society, 13, 1122–1145.

[10]. Breiman, L. (1996). Bagging Predictors. Kluwer Academic Publishers, 24, 123–140.

[11]. Brennan, M. W. (2015). Most Popular APIs used at Hackathons. Retrieved from http://www.programmableweb.com/api/faceboo k

[12]. Brownlee, J. (2014). A Data-Driven Approach to Choosing Machine Learning Algorithms. Retrieved from http://machinelearningmastery.com/

[13]. Brownlee, J. (2015). Supervised and Unsupervised Machine Learning Algorithms. Retrieved from http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[14]. Ching Ho, C., & Ting, C.-Y. (2014). Measuring Crowd Sourced Analytics: A Review. Multimedia Tools and Applications.

[15]. Copper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in Society, 1(1), 104–126.

[16]. Daniel, N. (2012). Cookies and Web tracking, battle for the internet. The Guardian.

[17]. de Vreede, T., Nguyen, C., de Vreede, G. J., Boughzala, I., Oh, O., & Reiter-Palmon, R. (2013). A Theoretical Model of User Engagement in Crowdsourcing in Collaboration and Technology. Springer Berlin Heidelberg, 94–109.

[18]. Dixon-Woods, M., Agarwal, N., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of Health Services Research and Policy, 10(1), 45–53.

[19]. Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Unsupervised Learning and Clustering: Pattern classification (Second Edition). Wiley.

[20]. DuVander, A. (2012). Today in APIs: foursquare in TomTom, Facebook Mentions and 15 New APIs. Retrieved from http://www.programmableweb.com/category/all /news?keyword=Tweet%20Archivist%20API

[21]. Ebner, A., Lienhardt, M., Rohs, S., & Meyer, K. (2010). Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning? Computers & Education, 55, 92–100.

[22]. Eibe, F., Mark A., H., & Ian H., W. (2016). WEKA Workbench Online Appendix for Data Mining. WEKA.

[23]. Elavsky, C. M., & Mislan, C. (2011). When talking less is more: exploring outcomes of Twitter usage in the large- lecture hall. Media Technology, 36(3).

[24]. Fazakis, N., Karlos, S., Kotsiantis, S., & Sgarbas, K. (2015). Self-Trained LMT for Semisupervised Learning. Computational Intelligence and Neuroscience, (1–13). https://doi.org/10.1155/2016/3057481

[25]. Fox, A., & Zoe, K. (2012). How the Arab World Uses Facebook and Twitter. Mashable.

[26]. Garcia, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowlege-Based Systems, 9(8), 1–29. https://doi.org/10.1016/j.knosys.2015.12.006

[27]. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 9. https://doi.org/10.1186/s41044-016-0014-0

[28]. Geographic, N. (2016). Animal Jam. Retrieved from http://www.animaljam.com/

[29]. Gibbs, S. (2014). UK porn filter blocks game update that contained "sex." The Guardian. Retrieved from http://www.theguardian.com/technology/2014/jan/21/uk-porn-filter-blocks-game-update-that-contained-sex?CMP=twt_gu

[30]. Grossman, R., Giovanni, S., Elder, J., Agarwal, N., & Huan, L. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan and Claypooi. https://doi.org/10.2200/S00240ED1V01Y200912DMK002

[31]. Hernandez-Orallo, J., Flach, P. ., & Ferri, C. (2012). A unified view of performance metrics: translating threshold choice into expected classification loss. Journal of Machine Learning Research, 13(1), 2813–2869.

[32]. Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. Computational Statistics, 24(2), 225–232. https://doi.org/10.1007/s00180-008-0119-7

[33]. Interactive, B. G. (2013). Bubble Games. Retrieved from http://www.bubbleguminteractive.com/games/

[34]. Jesson, J., Matheson, L., & Lacey, F. M. (2011). Doing your literature review: traditional and systematic techniques. SAGE Publication.

[35]. Jiawei, H., Micheline, K., & Jian, P. (2012). Data Mining Concepts and Techniques (Third Edition). Morgan Kaufmann.

[36]. Joshi, A. (2016). What is stacking in machine learning? Retrieved from https://www.quora.com/What-is-stacking-in-machine-learning

[37]. Junco, R., Elavsky, C. M., & Heiberger, G. (2012). Putting Twitter to the test: assessing outcomes for student collaboration, engagement,

and success. British Journal of Educational Technology. https://doi.org/10.1111/J

[38]. Junco, R., Heiberger, G., & Loken, E. (2011). The effect of Twitter on college student engagement and grades. Journal of Computer Assisted Learning, 27(2), 119–132.

[39]. Kareem, F. (2011). Protesters in Egypt Defy Ban as Government Cracks Down. The New York Times.

[40]. King, M. (2013). Types of Profanity Filters for online Safety. Retrieved from https://www.inversoft.com/blog/2013/03/28/types-of-profanity-filters-for-online-safety/

[41]. Knibbs, K. (2014). Curses! People swear a lot on Twitter, and here are the most popular words. Retrieved May 31, 2015, from http://www.digitaltrends.com/social-media/popular-curse-words-twitter/

[42]. Kohavi, R., & Provost, F. (1998). Applied Research in Machine Learning. In Application of Machine Learning and the Knowledge Discovery Process (Vol. 30). New York USA: Columbia University.

[43]. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Association for Avdancement of Artificial Intelligence.

[44]. Kunal, K. (2015). Weka – GUI way to learn Machine Learning. Retrieved from https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/

[45]. Leopoid, G. (2016). Crowdsourcing Used to Augment Machine Learning. Retrieved from https://www.datanami.com/2016/04/13/crowdsourcing-used-augment-machine-learning/

[46]. Lim, J. D., Choi, B. C., Han, S. W., & Lee, C. H. (2011). Automatic Classification of x-rated videos using obscene sound analysis based on a repeated curve-like spectrum feature. The International Journal of Multimedia & Its Applications (IJMA), 3(4).

[47]. Marr, B. (2014). Amazing Big Data World: Kaggle and Crowdsourced Data Scientist. Retrieved from http://www.smartdatacollective.com/bernardmarr/209091/amazing-big-data-world- kaggle-and-crowd-sourced-data-scientist

[48]. Mohr, M., Rostamizadeh, A., & Talwalker, A. (2012). Foundations of Machine Learning. MIT Press.

[49]. Narasimha, M., M., & Susheela, D., V. (2011). Pattern Recognition: An Algorithmic Approach.

[50]. Ogada, K. O., Mwangi, W., & Cheruiyot, W. (2016). N-grams for Text Classification Using Supervised Machine Learning Algorithms (PhD Thesis). Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.

[51]. Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. Information & Management, 52(2), 183–199.

[52]. Powers, D. (2011). Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63. Journal of Machine Learning Technologies, 2(1), 37–63.

[53]. Prpić, J., Taeihagh, A., & Melton, J. (2015). The Fundamentals of Policy Crowdsourcing. Policy & Internet, 7(3), 340–361. https://doi.org/10.1002/poi3.102.

[54]. Ruche, D. (2012). Facebook IPO sees Winklevoss twins heading for $300 million fortune. The Guardian.

[55]. Russell, S., & Norvig, P. (2003). Artificial Intelligence: A Modern Approach (Second Edition). Prentice-Hall. Retrieved from 0-13-080302-2

[56]. Schapire, R. (1990). The Strength of Weak Learnability. Kluwer Academic Publishers, 5(2), 197–227.

[57]. Solomatine, D., See, L. M., & Abrahart, R. J. (2008). Data-Driven Modelling: Concepts, Approaches and Experiences.

[58]. Sood, S. O., Antin, J., & Churchill, E. (2012). Using Crowdsourcing to Improve Profanity Detection. Association for the Advancement of Artificial Intelligence.

[59]. Srivastava, T. (2015). Tuning the parameters of your Random Forest model. Retrieved from https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

[60]. Steinberg, D. (2014). Why Data Scientists Split Data into Train and Test. Retrieved from http://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data-into-Train-and-Test

[61]. Sunil, R. (2015a). 7 Types of Regression Techniques you should know! Retrieved from https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[62]. Sunil, R. (2015b). Qick Introduction to Boosting Algorithms in Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/

[63]. Ticlavilca, A. M., & Torres, A. (2010). Data Driven Models and Machine Learning (ML)

Approach in Water Resources Systems. University of Rochester.

[64]. Wang, S., & Yao, X. (2013). Relationships between Diversity of Classification Ensembles and Single- Class Performance Measures (Vol. 25). IEE Transaction on knowledge and data engineering.

[65]. Weigel, M. (2012). Why Facebook user get more than they give. Journalist Resource - Harvard Kennedy School.

[66]. Witten, I., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). San Francisco: Morgan Kaufmann.

[67]. Wolpert, D. H. (1992). Stacked Generation. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1