The International Journal of Engineering and Science (IJES)
|| Volume || 7 || Issue || 10 Ver.III || Pages || PP 01-20 || 2018 ||
ISSN (e): 2319 – 1813 ISSN (p): 23-19 – 1805

Topic of Interest Discovery on Social Media Using Knowledge Base and Term Frequency – Inverse Document Frequency Techniques

Athman Masoud[1], Wilson Cheruiyot[1], Kennedy Ogada[1].
[1]*Institute of Computing and Information Technology,*
Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000-00200, Nairobi, Kenya.
Corresponding Author: Athman Masoud.

----------ABSTRACT----------
Online users frequently post comments in their social network profiles; these comments leave unique traces of attributes such as keywords, interests of an entity and its related connection especially in micro blogs such as twitter. The keywords updated day in day out in the user profile, build an enormous amount of data which if analyzed can define user's specific topic of interest. The dynamism of this huge, noisy and ever increasing micro-blogging data and the need to discover trending topics from this data, has led researchers to approach this problem by using several techniques such as text mining, natural language processing, statistical and other topic modeling methods. Therefore, this research presents a supervised machine learning technique with twofold approach: First, the leveraging of knowledge base for developing topic model from the training data. Second, the use of TFIDF, a feature selection method on Bag of Words (BoW) text representation model on the data corpus for discovery of user topic of interest. Experiments were carried out on 400,000 tweet texts. The text data from the tweets was pre-processed; stop-words, symbols, URLs, digits were removed and terms were lemmatized and tokenized. The corpus data was split into two sections, 80% of the data was used for training and 20% for testing. The research managed to classify twelve major topics from the training model. The topic model was tested on the held out data with K - Nearest Neighbour (KNN), Support Vector Machine (SVM) and Decision Tree (DT) algorithms. The experiment performance quality was evaluated and showed that the combination of TFIDF and Knowledge Base synsets to the BoW text representation improved the performance of SVM, which outperformed both KNN and DT.

KEYWORDS;- *topic discovery, term frequency, inverse document frequency, feature selection, machine learning*

Date of Submission: 12-10-2018 Date of acceptance: 27-10-2018

I. INTRODUCTION

Social media platforms such as twitter have been used enormously to post tweets and comments respectively by organizations or individuals from different geographical locations, religion, language and cultural background for branding, sensitization, and knowledge dissemination, message exchange etc. The real-world nature of posts is that they are noisy and complex, making text mining difficult. Tweets are intentionally short (limited to just 140-characters) which force users to be creative in how they constrain the text while preserving meaning. As with text messages sometimes users rely on common acronyms (e.g., "d/r" means "dressing room" in sports), or ("Hawks" to mean "Chicago Blackhawks," in general, this leads to noise (Dredze, McNamee, Rao, Gerber, & Finin, 2010).

Topic profiling has been another major problem in machine learning, especially the use of optimum feature selection method to discover user topic of interests, patterns and trends in social media. In practice, optimum user topic of interests depends on feature selection method, feature selection is a combinatorial optimization problem which involves identifying the best subset of features within a set (Lin, Zhang, Huang, Hung, & Yen, 2016). Mohamad and Selamat (2015), in their research presented a hybrid feature selection, which is a combination of Term Frequency Inverse Document Frequency (TFIDF) with the rough set theory in spam email classification problem that return a good result.

Kapanipathi, Jain, Venkataramani and Sheth (2014), in their approach, have presented the use knowledge bases to spot entities and create entity-based user profiles. However, exploitation of such knowledge bases to create richer user (topic) profiles is yet to be explored. Therefore, in this research, text messages were analyzed by using term frequency inverse document frequency (TFIDF) feature selection method and knowledge base Synsets merged with terms in the bag of words text representation model. In addition, various

machine learning algorithms such as K- Nearest Neighbor (KNN), Support Vector Machine (SVM) and Decision Tree (DT) were used for data training in the discovery of user topic of interest.

1.1. Research Broad objective

The broad objective of this paper is to develop a topic model by combining TFIDF feature selection method and knowledge base synsets to the document representation and evaluate accuracy of various machine learning algorithms for the discovery of user topic of interest.

1.1.2. Specific objectives

The specific objectives of this research are :

i) Analyze TFIDF feature selection method and its effect on machine learning algorithms on the discovery of user topic of interest
ii) To examine the effect of combining TFIDF feature selection method and knowledge base on topic model development for discovery of user topic of interest
iii) To develop a topic model by using knowledge base term and phrases addition to the text representation model for discovery of user topic of interests on tweets
iv) To evaluate the performance of TFIDF and knowledge base techniques on existing machine learning algorithms such as K- Nearest Neighbor, Support Vector Machine and Decision Tree on the discovery of user topic of interest.

II. LITERATURE REVIEW

This section introduces the key definitions on text mining, text classification, topic profiling, feature selection methods and machine learning algorithms. It also introduces discussion on applications of these concepts in various fields such as natural language processing, data mining (text mining) and artificial intelligence in order to assist in formulating a basis for the research, gaps identification and proposal of new methods topic profiling in social media to discover user interests, trends and patterns.

2.1. Text Mining

Text mining is a branch of Data mining. Data Mining refers to extracting informative knowledge from a large amount of data, which could be expressed in different data types, such as transaction data in Electronic Commerce applications or genetic expressions in bioinformatics research domain. The main purpose of data mining is discovering hidden data or unseen knowledge, normally in the form of patterns, from available data repository (Xu, Zhang, & Li, 2011).

2.2. Topic Profiling

According to A. K. Sehgal,(2007), a topic profile is analogous to a synopsis of a topic and consists of different types of features. Profiles are flexible to allow different combinations of features to be emphasized and are extensible to support new features to be incorporated without having to change the underlying logic. More generally, topic profiles provide an abstract framework that can be used to create different types of concrete representations for topics. Different options regarding the number of documents considered for a topic or types of features extracted can be decided based on requirements of the problem as well as the characteristics of the data. Topic profiles also provide a framework to explore relationships between topics.

2.3 Text Classification

According to Sriram et al., (2010), text classification is an area where classification algorithms are applied on documents of text. The task is to assign a document into one (or more) classes, based on its content. Typically, these classes are handpicked by humans. For example, consider the task to classify set of documents (say, each 1 page long) as good or bad. In this case, categories (or labels) "good" and "bad" represent the classes.

2.4. Term frequency – Inverse document frequency (TFIDF) Feature Selection Method

According to Erra, Senatore, Minnella and Caggianese (2015), TF–IDF is a well-known measure that is often used to construct a vector space model in information retrieval. It evaluates the importance of a word in a document. The importance increases proportionally with the number of times that a word appears in a document, compared to the inverse proportion of the same word in the whole collection of documents.

Also Lilleberg, Zhu and Zhang (2015), explained that term frequency-inverse document frequency, is used to determine what words of a corpus may be favorable to use based on each word's document frequency. TFIDF calculates a value for each word in a document through an inverse proportion of the frequencies of the

word in a certain document and to the percentage of documents to which the word appears in. The higher TFIDF values words have imply they have a stronger relationship in the document which they appear. This value is calculated from the following formula

$$(tf - idf)_{i,j} = tf_{ij} \text{ X } tf_i \quad (1.0)$$

The left side of the equation contains the two components tf and idf. The first term is the frequency of the words, expressed by formula

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2.0)$$

where $n_{i,j}$ is the number of occurrences of the term ti in the document d_j . The denominator contains the sum of occurrences all the words in the selected document d_j. The second component from (1.0) is idf_i and it is expressed by the following formula:

$$idf_i = \log \frac{|D|}{|\{ dt_i : \epsilon d \}|} \quad (3.0)$$

where |D| is the total number of documents and $\{d : t_i \in d\}$ the number of documents that contain at least one instance of document (Grycuk, Gabryel, Korytkowski, & Scherer, 2014). In text mining, the term frequency – inverse document frequency (TFIDF), is a well-known feature selection method to evaluate how important is a word in a document. TFIDF is a very interesting way to convert the textual representation of information into a Vector Space Model (VSM).

2.5. Document Representation

Documents need to be represented in a way that is suitable for a general learning process. The most widely used representation is \the bag of words" where a document is represented by a vector of features, each of which corresponds to a term or a phrase in a vocabulary collected from a particular data set. The value of each feature element represents the importance of the term in the document, according to a specific feature measurement (Ogada et al., 2015).

2.6. Document Representation Model

There are four main IR models: Boolean Model, vector space model, language model and probabilistic model. The most commonly used models in IR systems and on the Web are the first three models. Although these three models represent documents and queries differently, they use the same framework. They all treat each document or query as a "bag" of words or terms. Term sequence and position in a sentence or a document are ignored. That is, a document is described by a set of distinctive terms (Liu, 2007 as cited in Ogada et al., 2015).

2.7. The Combined Representational Model

BoW representation tends to divide text into single words and, hence, causing terms to break down into their constituent words. The model also treats synonymous words as independent features with no semantic association. These issues have been addressed by a number of researchers by representing text as concepts rather than words, using an approach known as Bag-of-Concepts (BOC). In the BOC approach, semantic knowledge bases such as WordNet, Open Directory Project (ODP) and Wikipedia are used to identify the concepts appearing within a document (Alaa, Arash, & Abdulhussain, 2014 as cited in Ogada et al., 2015).

2.8. WordNet Text Categorization

According to Elberrichi, Rahmoun, Bentaalah and Laboratory (2008), WordNet is a thesaurus for the English language based on psycholinguistics studies and developed at the University of Princeton . It was conceived as a data-processing resource which covers lexico semantic categories called synsets. The synsets are sets of synonyms which gather lexical items having similar significances, for example the words “a board” and “a plank” grouped in the synset {board, plank}. But “a board” can also indicate a group of people (e.g., a board of directors) and to disambiguate these homonymic significances “a board” will also belong to the synset {board, committee}. The definition of the synsets varies from the very specific one to the very general. The most specific synsets gather a restricted number of lexical significances whereas the most general synsets cover a very broad number of significances. The organization of WordNet through lexical significances instead of using lexemes makes it different from the traditional dictionaries and thesaurus.

The other difference which has WordNet compared to the traditional dictionaries is the separation of the data into four data bases associated with the categories of verbs, nouns, adjectives and adverbs. This choice

of organization is justified by psycholinguistics research on the association of words to the syntactic categories by humans. Each database is differently organized than the others. The names are organized in hierarchy, the verbs by relations, the adjectives and the adverbs by N-dimension hyperspaces .Elberrichi et al., (2008) explain further that some semantic relations available in WordNet are synonyms, hyponyms and hyperonyms, which are used in this paper; the examples that are given are based on words.

2.8.1. Synonymy in Wordnet

A synonym is a word, which we can substitute to another without important change of meaning. Cruse as cited in Elberrichi et al., (2008)distinguishes distinguishes further three types of synonymy:

i) Absolute synonymes.
ii) Cognitive synonymes.
iii) Plesionymes.

According to the definition of Cruse as cited in Elberrichi et al., (2008) of the cognitive synonyms, X and Y are cognitive synonyms if they have the same syntactic function and that all grammatical declaratory sentences containing X have the same conditions of truth as another identical sentence where X is replaced by Y, example: Convey /automobile. The relation of synonymy is at the base of the structure of WordNet. The lexemes are gathered in sets of synonyms ("synsets"). There are thus in a synset all the terms used to indicate the concept. The definition of synonymy used in WordNet is as follows: "Two expressions are synonymous in a linguistic context C if the substitution of for the other out of C does not modify the value of truth of the sentence in which substitution is made". Example of synset: [Person, individual, someone, somebody, mortal, human, drunk person].

2.8.2. Hyponyms and Hyperonyms in Word Net

X is a hyponym of Y (and Y is a hyperonym of X) if:

i) F(X) is the minimal indefinite expression compatible with sentence A is F(X) and
ii) A is F(X) implies A is F(Y).

In other words, the hyponymy is the relation between a narrower term and a generic term expressed by the expression "is-a". Example: It is a dog → It is an animal. A dog is a hyponym of animal and animal is a hyperonym of dog.

2.9. Data Mining

According to Han, Pei and Kamber (2011), it is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term data mining does not really present all the major components in the picture. Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is shown in Figure 2.6 as an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measure)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

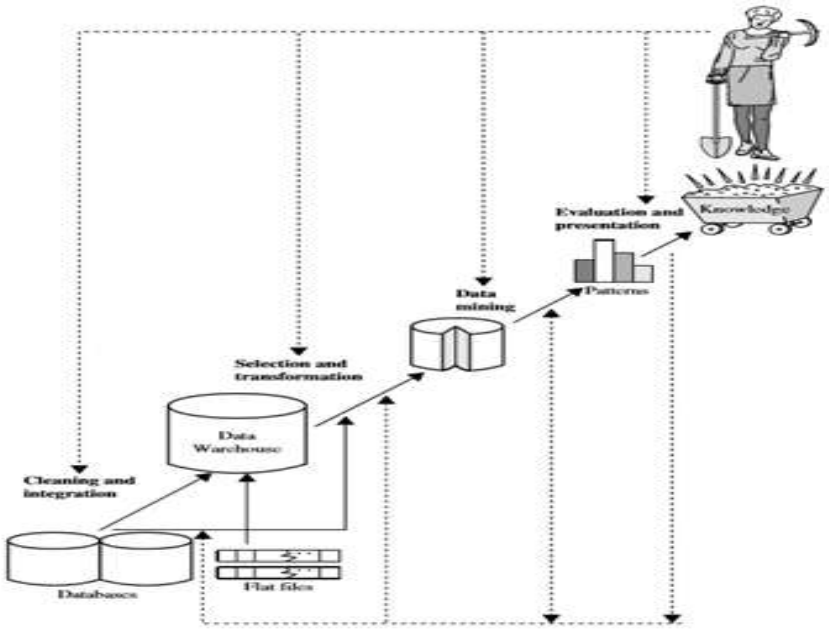

Figure 1.0 : Knowledge discovery process, adopted from Data Mining Concepts and Techniques, (Han et al., 2011).

2.10. The Artificial Intelligence

Luger & Stubblefield (1993) describe AI as the branch of computer science that is concerned with the automation of intelligent behavior. Also Kurzweil, Richter and Schneider (1990) defined AI as the art of creating machines that perform functions that require intelligence when performed by people. In summary after an extensive survey of definitions given by scientists and researchers at different times, it can be concluded that AI is the science of making a machine think and act like an intelligent person. Artificial intelligence (AI) is technology and a branch of computer science that studies and develops intelligent machines and software.

Han et al., (2011) explain further that, in order to achieve the task of imitating human behavior or acquiring human intelligence, a machine (a computer in our case) must reflect the following capabilities that are commonly inherited by an intelligent person:

(i) **Natural language processing:** Like a human, a machine should understand the spirit or the meaning of sentences spoken or written freely in natural language by humans. Human do not mind grammar as well as composition of sentences while reading or talking informally.

(ii) **Knowledge representation:** It is another great challenge how to express knowledge, which can be presented in mathematical or some logical format. Ultimate goal to get a work done by a computer will be to translate the informal sentences into formal ones, which could be well interpreted by a computer.

(iii) **Automated Reasoning:** The capability to use the stored information to answer questions and to draw new conclusions;

(iv) **Machine Learning:** Learning is an important property of humans. A machine should also be able to learn to adapt to new circumstances and to detect and extrapolate patterns.

In summary AI research areas include but not limited to Expert systems, Natural Language Processing, speech recognition, Automatic voice output, Neural Networks (Pattern recognition systems such as face recognition, character recognition, and handwriting recognition, Robotics (Industrial robots for moving, spraying, painting, precision checking, drilling, cleaning, coating, carving etc.) and Fuzzy Logic .

2.11. Machine Learning

Learning, like intelligence, covers such a broad range of processes that it is difficult to define precisely. A dictionary definition includes phrases such as to gain knowledge, or understanding of, or skill in, by study, instruction, or experience," and modification of a behavioral tendency by experience." Zoologists and psychologists study learning in animals and humans. There are several parallels between animal and machine learning. Certainly, many techniques in machine learning derive from the efforts of psychologists to make more precise their theories of animal and human learning through computational models.

It seems likely also that the concepts and techniques being explored by researchers in machine learning may illuminate certain aspects of biological learning. As regards machine learns whenever it changes its

structure, program, or data (based on its inputs or in response to external information) in such a manner that its expected future performance improves. Some of these changes, such as the addition of a record to a database, fall comfortably within the province of other disciplines and are not necessarily better understood for being called learning. But, for example, when the performance of a speech-recognition machine improves after hearing several samples of a person's speech, in that case it can be justified that the machine has learned. Generally machine learning usually refers to the changes in systems that perform tasks associated with artificial intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc.

The changes might be either enhancements to already performing systems or ab initio synthesis of new systems. To be slightly more specific, architecture of a typical AI can be used to illustrate learning by an agent. This agent perceives and models its environment and computes appropriate actions, perhaps by anticipating their effects. Changes made to any of the components shown in the figure2.6.4 might count as learning. Different learning mechanisms might be employed depending on which subsystem is being changed (Shalev-Shwartz & Ben-David, 2014).

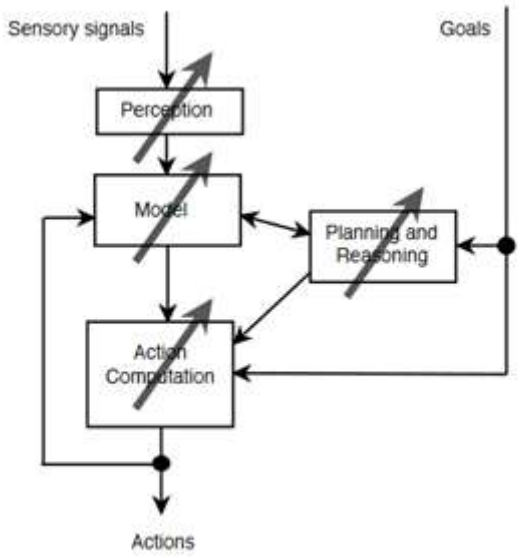

Figure 2.0 : AI system representing machine learning , adopted from the book: Understanding Machine Learning from Theory to Algorithm, (Shalev-Shwartz & Ben-David, 2014)

2.12 Supervised Machine Learning

Learning as a process of "using experience to gain expertise," supervised learning describes a scenario in which the "experience," a training example, contains significant information (say, the spam/not-spam labels) that is missing in the unseen "test examples" to which the learned expertise is to be applied. In this setting, the acquired expertise is aimed to predict that missing information for the test data. In such cases, the environment is considered as a teacher that "supervises" the learner by providing the extra information (labels) (Shalev-Shwartz & Ben-David, 2014).

2.12.1. Classification Algorithms

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown. The derived model may be represented in various forms, such as classification rules (i.e., IF-THEN rules), decision trees, mathematical formulae, or neural networks. There are many other methods for constructing classification models, such as Naıve Bayesian classification, support vector machines, and k-nearest-neighbor classification. classification model predicts categorical (discrete, unordered) labels (Han et al., 2011).

2.12.2. Support Vector Machine

Support vector machines (SVMs) are learning routines used for classification of input data received by a computing system. The input data objects may be represented by a set of one or more feature vectors, where a feature vector can include aspects of the data object that is being represented. For example an image file can be

associated with a relatively large number of feature vectors, where each feature vector of the image file represents some different aspect of the image file. The SVMs may first be trained with a number of feature vectors in order to proceed to classify other input data vectors (Eshghi & Kafai, 2016). Since SVM has the strong learning ability and is able to capture the inherent characteristics of the data, high classification efficiency results. Therefore less training samples can be used to get a trained classifier with high performance, so choice of SVM classifier for text classification is ideal (Wu & Xu, 2015 as cited in Ogada, 2015).

2.12.3. Naive Bayes

According to Bermejo, Gámez, and Puerta (2014), Naive Bayes (NB) is a probabilistic classifier based on the assumption of conditional independence among the predictive attributes given the class. The independence assumption in NB translates into a simple Bayesian network with a fixed structure having exactly n edges, which point from class to each predictive attribute. That graphical structure factorizes the joint probability distribution as follows:

$$P(C, X1, \ldots, Xn) = P(C) \prod_{i=1}^{n} P(Xi|C) \quad (4.0)$$

Therefore, only parameter estimation is needed: a marginal (multinomial) probability distribution for the class variable P(C) and a conditional probability distribution for each predictive attribute given the class P(Xi|C). Such distribution can be multinomial or Gaussian (Normal), depending on the nature of Xi (discrete or numeric, respectively), and it is estimated for each value cj of C.
The MAP principle is used for inference. That is, given an instance $< x1, \ldots, xn >$ we choose the class label c^* such that

$$c^* = \arg\max_{cj} P(C = cj|X1 = x1, \ldots, Xn = xn) \quad (5.0)$$

$$= \arg\max_{cj} P(C = cj) \prod_{i=1}^{n} P(Xi = xi|C = cj) \quad (6.0)$$

It is also well known that NB is very sensitive to the presence of redundant and/or irrelevant attributes. The presence of redundant (highly correlated) attributes can bias the decision taken by the NB classifier. Regarding irrelevant variables, although their presence should be harmless for NB classifier, in practice this is not usually the case. Thus, given an irrelevant variable Xi it would be desirable P(Xi|C = cj) to be equal for all values cj of C; however, this is not true in general due to the small sample effect and the presence of noise. Then, in the case of high-dimensional datasets, where hundreds or thousands of irrelevant variables are present, irrelevant variables can impair the precision of the NB classifier

2.12.4. K Nearest Neighbor

The main thesis in KNN is that documents which belong to the same class are likely to be close to one another based on similarity measures such as the dot product or the cosine metric. KNN assumes that the data is in a feature space, since the points are in feature space, they have a notion of distance. Also it assumes that each of the training data consists of a set of vectors and class label associated with each vector. A single number \k" is given; this number decides how many neighbors influence the classification (Al-Ghuribi & Alshomrani 2013 as cited in Ogada et el., 2015)

According to Lui (2007), KNN works as follows: Let D be the training data set. Nothing will be done on the training examples. When a test instance d is presented, the algorithm compares d with every training example in D to compute the similarity or distance between them. The k most similar (closest) examples in D are then selected. This set of examples is called the k nearest neighbors of d. d then takes the most frequent class among the k nearest neighbors. k = 1 is usually not sufficient for determining the class of d due to noise and outliers in the data. A set of nearest neighbors is needed to accurately decide the class. The general kNN algorithm is given in Figure 2.9.3.4.
Algorithm kNN(D; d; k)

1. Compute the distance between d and every example in D;
2. Choose the k examples in D that are nearest to d, denote the set by P ($\subseteq$ D);
3. Assign d the class that is the most frequent class in P (or the majority class).

In Figure 2.9.3.4, supposing the small dot in the middle need to be classified. If selecting 3 nearest neighbors, that is k = 3, there are two triangles and a square can be found in the 3 nearest shapes from the dot. So the dot should be classified in the triangles according to the definition of K-Nearest Neighbor classification. Similarly, if k = 5, the dot should be classified in the squares. So the value of k determines the category of the dot. There are other factors such as the position of the dot, the way to calculating the distance and so on. When

this classification algorithm is used for text categorization, the best parameters should be tested in the classifier to make the effect of categorization best (Yaduang et al., 2015 as cited in Ogada et al., 2015).

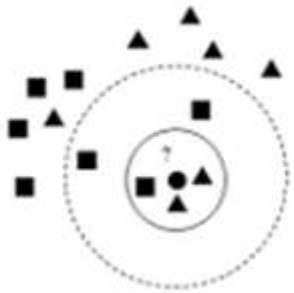

Figure 3.0: Example of KNN Model

2.12.5. Regression Algorithms

Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data. Regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels. The term prediction refers to both numeric prediction and class label prediction (Han et al., 2011). There are many other methods for constructing regression models such as linear regression, ensemble methods, decision tree, neural networks etc.

2.12.6. Linear Regression

According to Montgomery, Peck and Vining (2012), Linear regression is used to study the linear relationship between a dependent variable Y (e.g. blood pressure) and one or more independent variables X (e.g. age, weight, sex). The dependent variable Y must be continuous, while the independent variables may be either continuous (age), binary (sex), or categorical (social status). The initial judgment of a possible relationship between two continuous variables should always be made on the basis of a scatter plot (scatter graph). This type of plot shows whether the relationship is linear or nonlinear. Performing a linear regression makes sense only if the relationship is linear. Other methods must be used to study nonlinear relationships.

2.12.7. Simple linear regression

Montgomery et al.,(2012) explain further that simple or univariable linear regression studies the linear relationship between the dependent variable Y and a single independent variable X. The linear regression model describes the dependent variable with a straight line that is defined by the equation $Y = a + b \times X$, where a is the y intersect of the line, and b is its slope. First, the parameters a and b of the regression line are estimated from the values of the dependent variable Y and the independent variable X with the aid of statistical methods. The regression line enables one to predict the value of the dependent variable Y from that of the independent variable X. Thus, for example, after a linear regression has been performed, one would be able to estimate a person's weight (dependent variable) from his or her height (independent variable) The slope b
of the regression line is called the regression coefficient. It provides a measure of the contribution of the independent variable X toward explaining the dependent variable Y. If the independent variable is continuous (e.g., body height in centimeters), then the regression coefficient represents the change in the dependent variable (body weight in kilograms) per unit of change in the independent variable (body height in centimeters).

2.12.8. Ensemble Methods

Al-Jarrah et al., (2015) explain that, one of the key success elements of sustainable data modeling is to maintain or improve its performance while significantly reducing its computational cost. Recent data-modeling research has shown that ensemble methods have gained much popularity as they often perform better than individual models. Ensemble method uses multiple models to obtain better performance than those that could be obtained from any of the constituent models . However, it can result in significant increase in computational cost. If the model deals with large-scale data, model complexity and computational requirements will grow exponentially. An example of such ensemble model is the Bayes classifier. In Bayes classifier, each hypothesis is given a vote proportional to the likelihood that the training dataset would be sampled from a system if that hypothesis was true. To facilitate the training data of finite size, the vote of each hypothesis is also multiplied by the prior probability of that hypothesis. The Bayes classifier is expressed as follows:

$$y = \arg\max_{c_j \in C} \sum_{h_i \in H} P(c_j \mid h_i) P(T \mid h_i) P(h_i), \tag{7.0}$$

where y is the predicted class, C is the set of all possible classes, H is the hypothesis space, P refers to a probability, and T is the training data. As an ensemble, the Bayes classifier represents a hypothesis that is not necessarily in H. The hypothesis represented by the Bayes classifier, however, is the optimal hypothesis in ensemble space (the space of all possible ensembles consisting only of hypotheses in H).

2.12.9. Decision Tree

A decision tree is a classifier which conducts recursive partition over the instance space. A typical decision tree is composed of internal nodes, edges and leaf nodes. Each internal node is called decision node representing a test on an attribute or a subset of attributes, and each edge is labeled with a specific value or range of value of the input attributes. In this way, internal nodes associated with their edges split the instance space into two or more partitions. Each leaf node is a terminal node of the tree with a class label. For example, Figure 1 provides an illustration of a basic decision tree, where circle means decision node and square means leaf node. In this example, we have three splitting attributes, i.e., age, gender and criteria 3, along with two class labels, i.e., YES and NO. Each path from the root node to leaf node forms a classification rule (Dai & Ji, 2014).

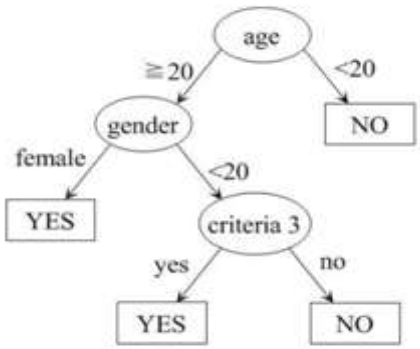

Figure 4.0 : Illustration of decision tree adopted from Dai and Ji, (2014)

2.12.10. Neural Networks

According to (Gurney, 2014) , a neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the interunit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns. Neural networks are often used for statistical analysis and data modeling, in which their role is perceived as an alternative to standard nonlinear regression or cluster analysis techniques (Cheng & Titterington 1994).

Thus, NN are typically used in problems that may be couched in terms of classification, or forecasting. Some examples include image and speech recognition, textual character recognition. This type of problem also falls within the domain of classical artificial intelligence (AI) so that engineers and computer scientists see neural nets as offering a style of parallel distributed computing, thereby providing an alternative to the conventional algorithmic techniques that have dominated in machine intelligence (Gurney, 2014)

2.13. Unsupervised Machine Learning

In unsupervised learning, the learner processes input data with the goal of coming up with some summary, or compressed version of that data, there is no distinction between training and test data. Clustering a data set into subsets of similar objects is a typical example of such a task. There is also an intermediate learning setting in which, while the training examples contain more information than the test examples, the learner is required to predict even more information for the test examples. For example, one may try to learn a value function that describes for each setting of a chessboard the degree by which White's position is better than the Black's. Yet, the only information available to the learner at training time is positions that occurred throughout actual chess games, labeled by who eventually won that game. Such learning frameworks are mainly investigated under the title of reinforcement learning (Shalev-Shwartz & Ben-David, 2014).

2.13.1. **Clustering Algorithms**

Cluster Analysis Unlike classification and regression, which analyze class-labeled (training) data sets, clustering analyzes data objects without consulting class labels. In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the inter class similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together(Han et al., 2011).

According to Jain, (2010) , Clustering algorithms can be broadly divided into two groups: hierarchical and partitional. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative mode (starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy) or in divisive (top-down) mode (starting with all the data points in one cluster and recursively dividing each cluster into smaller clusters). Compared to hierarchical clustering algorithms, partitional clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. Input to a hierarchical algorithm is an n x n similarity matrix, where n is the number of objects to be clustered.

On the other hand, a partitional algorithm can use either an n x d pattern matrix, where n objects are embedded in a d-dimensional feature space, or an n x n similarity matrix. Note that a similarity matrix can be easily derived from a pattern matrix, but ordination methods such as multi-dimensional scaling (MDS) are needed to derive a pattern matrix from a similarity matrix. The most well-known hierarchical algorithms are single-link and complete-link; the most popular and the simplest partitional algorithm is K-means.

2.13.2. K Means

According to Rajput and Patil, (2014), k-mean clustering algorithm is a special case of the generalized hard clustering algorithms. It is applied when point representatives are used and the squared Euclidean Distance is adopted to measure the dissimilarities between vectors x_i and cluster representatives Θ_j. The k-means algorithm is given below:

Algorithm:

Step1: Choose arbitrary initial estimates Θj (0) for the Θj's,
j=1,…,m.
Step2: Repeat
1. For i=1 to N
Determine the closest representative, say Θj for x_i.
Set b(i)=j;
End {for}
2. For j=1 to m
Parameter updating: Determine j as the
mean of the vectors $x_i \in X$ with b(i)=j.
End {for}
Until no change in j's occurs between two successive iterations.

2.13.3. Hidden Markov Model

A hidden Markov model (HMM) is one kind of sequence model; others are Maximum Entropy Markov Models or Conditional Random Fields. Another tool that is related to machine learning is methodological; the use of distinct training and test sets, statistical techniques like cross-validation, and careful evluation of our trained systems (Jurafsky, 2000). According to Annachhatre, Austin, and Stamp, (2015), previous research has shown that hidden Markov model (HMM) analysis is useful for detecting certain types of malware. In their research, they related the problem of malware classification based on HMMs. HMMs was trained for a variety of malware generators and a variety of compilers. More than 9000 malware samples were then scored against each model in their research and the malware samples were separated into clusters based on the resulting scores. The clusters were analyzed and showed that they correspond to certain characteristics of malware. The results indicated that HMMs are an effective tool for the challenging task of automatically classifying malware.

2.14. Topic Model Evaluation

Sathi and Ramanujapura (2016) presented Precision and Recall as two accuracy measures that are used in text classification to assess performance of the developed topic model.

- **precision** is defined as, the ratio of the number of relevant observations retrieved, to the total number of observations retrieved
- **Recall** is defined as the ratio of number of relevant observations retrieved to the total actual number of relevant observations present
- **F-measure:** It is the measure that is used to measure the overall quality performance of the model. F-measure is known by calculating the harmonic mean of precision and recall.

Precision and recall are explained with help of confusion matrix with two classes. One is labeled as positive class and other as negative as shown on table 2.9.4

	Prediction	
	Positive	**Negative**
Actual True	**TP**	**FN**
Actual False	**FP**	**TN**

Table 1.0: Description of Confusion matrix depicting true and false positives

In table 2.9.4, True Positive (TP) is the case where a particular tag (in our case) is true,and was assigned as true. Whereas a False Positive (FP) is a case where, the tag is false, but assigned a true value. Similarly, a True Negative (TN) stands for a case where a tag is false and has been assigned a false value, and a False Negative (FN) is, when a tag is true and has been assigned a false value. Then, the precision can be calculated in terms of true positives, and false positives as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{False Positives} + \text{True Positives}} \tag{8.0}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{False Negatives} + \text{True Positives}} \tag{9.0}$$

$$\text{F Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10.0}$$

2.15. Existing Approaches on Topic Modeling and Text Classification

According to Aizawa (2003), Term frequency–inverse document frequency TFIDF is one of the most commonly used term weighting schemes in today's information retrieval systems. Despite its popularity, TFIDF has often been considered an empirical method, specifically from a probabilistic point of view, with many possible variations. In the information retrieval field, term weights are mainly used to represent the usefulness of terms in the retrieval process; for example, the frequency signal-to-noise ratio.

Svore and Burges (2009) describe BM25 as one of the most important and widely used information retrieval functions. It has served as a strong baseline in the information retrieval community, in particular in the TREC Web track. Originally designed to be computed over the body and title fields of a Web document, BM25 is a nonlinear combination of three key document attributes: term frequency, document frequency, and document length

Iofciu, Fankhauser, Abel and Bischoff (2011) in their study "Identifying Users across Social Tagging Systems" combined TF-IDF and BM25 by Matching Users based on their Tags For identifying users across social systems based on their tagging behavior, they experiment with standard techniques like TF, TF-IDF and BM25 and compare it against a new symmetric variant of BM25 using site specific statistics. Their approach used two kinds of information: user ids and tags. Then introduced and compared a variety of approaches to measure the distance between user profiles for identification. With the best performing combination their method achieved accuracies of between 60% and 80%, which demonstrates that the traces of Web 2.0 users can reveal quite much about their identity. While these user identification strategies can support cross system personalization, they raise privacy concerns which the study did not handle adequately. Also investigation of network structure (such as friend links) was not considered in combination with tag-based profile features to impact user identification.

Abel, Herder and Krause(2011)in their study "Extraction of Professional Interests from Social Web Profiles" analyzed if professional interests of a user can be extracted from social sites (such as Facebook and Twitter activities) and be distinguished from private interests. The results indicated that performance largely depends on the size and quality of the Social Web profiles. In this study also methods for reducing noise and chatter for-high volume profiles was used to improve quality, however diversity of the profiles was deeply reduced

Michelson and Macskassy (2010) described the using of Knowledge Base (Wikipedia) to Disambiguate and Categorize Entities into two high-level steps:-

i) Categories discovery

In category discovery, the entities are identified in each Tweet, disambiguated, and the sub-tree of the folksonomy's categories that contains the disambiguated entity is retrieved. Since the output of this step is a set of categories for the Tweets, it is called the "Discover Categories" step.

ii) Generation of topic profile

Topic profile is generated for the user based on the discovered categories contained in the sub-trees. This is called the "Discover Profile" step.

The first step in discovering the categories for Tweets involves discovering the entity mentions in the Tweets themselves.

The figure 2.9.5 illustrates the steps in entities disambiguation on tweets:-

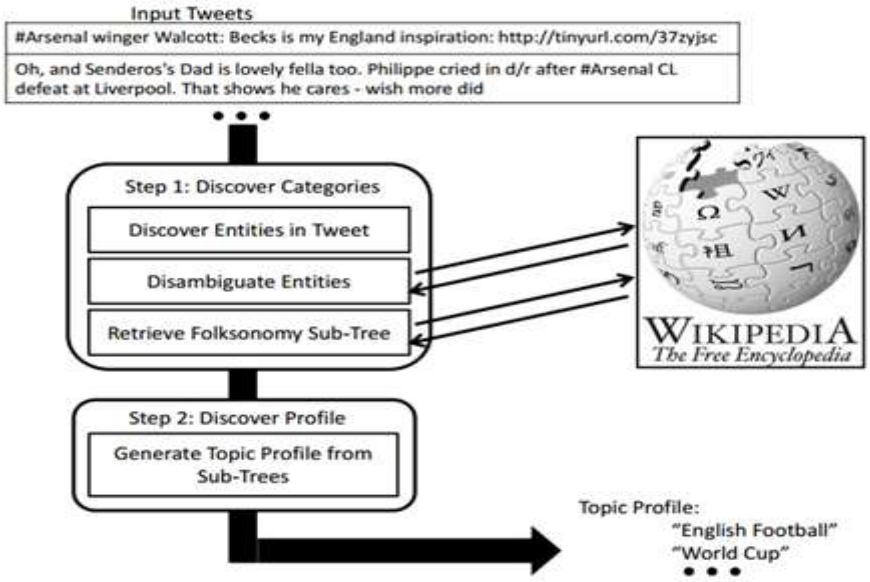

Figure 5.0 : Entity-Based Topic Profiles: Adopted from" Discovering users' topics of interest on twitter: a first look".(Michelson & Macskassy, 2010)

Michelson and Macskassy (2010) in their approach "Discovering Users' Topics of Interest on Twitter", presented early results on discovering Twitter users' topics of interest by examining the entities they mention in their Tweets. Their approach as mentioned earlier leverages a knowledge base (Wikipedia) to disambiguate and categorize the entities in the Tweets. Then a topic profile is developed which characterizes users' topics of interest, by discerning which categories appear frequently and cover the entities. The approach demonstrated that it is possible to successfully discover the main topics of interest for the users. However, this approach could not analyze Twitter to cluster and search users by their topics of interest.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei, Ng, &Jordan, 2003)

According to Xu, Ru, Xiang and Yang (2011), LDA is a generative model which represents documents as random mixtures over latent topics, and each topic is characterized by a probability distribution over words. To generate each document from a document collection, it first samples a topic from its topic distribution, and then picks a Dirichlet word from the distribution over words associated with the chosen topic.

The inference of LDA can be done by using Gibbs sampling (Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult), a relatively simple and effective algorithm. Therefore, the author-topic Model is an extension of LDA to integrate authorship information of documents into topic modeling. It assumes that each author in the document collection is represented by a distribution over topics, and each word is associated with two latent variables: an author and a topic. Differing from the generation process of LDA, the author-topic Model first chooses an author from a document's author list. Then it samples a topic from topic distribution associated with the selected author, and picks a word from the topic specific word distribution. When each document has only one author, the author-topic model is similar to the work in, which simply aggregates a big document for each user based on

his posts and runs traditional LDA on this document. The Table 2.9.8 below illustrates attributes to be extracted for a tweet in order to use LDA model which can represent an author – topic model.

Symbols	Descriptions
D	number of tweets
K	number of topics
A	number of authors
W	number of unique words
N_d	number of words in the dth tweet
w_d	content of the dth tweet
w_{dn}	nth word in the dth tweet
z_d	topic assignment of the dth tweet
z_{dn}	topic of the nth word in the dth tweet
r_d	relevance variable of the dth tweet
f_d	feature vector of the dth tweet
t_a	topic distribution of author a
α, β	Dirichlet distribution parameters
η	beta distribution parameters

Table 2.0: Illustration of attributes to be extracted from tweets for LDA Modeling

Xu et al. (2011) explain further that in this model it is assumed that each tweet is associated with a latent variable that indicates whether it is related to its author's interest, and the tweet originates either from the author's topic distribution or other distribution. Bayesian graphical model of twitter-user model is used and the description of its generative process. The model can be viewed as first generates a relevance variable and subsequently generates the tweet from either its author's topic distribution or other topic distribution. Bear in mind that this model does not analyze those interest-unrelated tweets directly. For simplicity, it is assumed that each interest-unrelated tweet is generated from a uniform topic distribution. The inference of this model can be efficiently computed using collapsed Gibbs sampling.

Xu, et al (2011) also focused on the problem of discovering users' topics of interest on Twitter. They explained that previous efforts in modeling users' topics of interest on Twitter have focused on building a "bag-of-words" profile for each user based on his tweets, they overlooked the fact that Twitter users usually publish noisy posts about their lives or create conversation with their friends, which do not relate to their topics of interest. Therefore, Xu et al, 2011, proposed a novel framework to address this problem by introducing a modified author-topic model named twitter-user model. For each single tweet, the model uses a latent variable to indicate whether it is related to its author's interest. An experiment on a large dataset was crawled using Twitter API which demonstrated that the model outperforms traditional methods in discovering user interest on Twitter.

2.16 Summary

There are several lessons learnt from the literature review on user topic of interest profiling methods especially the machine learning techniques such as text mining, text classification, feature selection, natural language processing, statistical and other topic modeling methods. The literature has explained further that the multifaceted data sets in social media, if analyzed, can give various attributes about the users, for example; Friendship, connections or relationship, locations, sentiments of the posts, gender, age and user topic of interest.

The analysis of these topic profile attributes relies on different techniques, however, these techniques have also suffered a range of gaps such as higher computational costs, lack of real time data, biasness to specific type of entities during topic identification, topic ambiguity and dependence on explicit user profile attributes which could lead to data sparsity problem and user privacy concerns

Therefore, despite several approaches that have been presented for online discovery of user topic of interest in the social media posts, it is evident that many research gaps still exist. Therefore, this research presents a supervised machine learning technique with twofold approach: First, the leveraging of knowledge base for defining, discovery of user topic of interests and developing a categorized topic model from the training data. Second, the use of TFIDF, a feature selection method on Bag of Words (BoW) document representation for the corpus data. An experiment was carried out on approximately 4.8 million words from 400,000 tweet text. The text data was pre-processed; stop-words, symbols, URLs were removed a, terms were lemmatized and tokenized. The pre-processed data was then analyzed using classifiers to discover user topic of interest in social media.

III. RESEARTCH METHODOLOGY

This paper introduces supervised machine learning techniques for modeling user topics of interest in Social web system boundaries. Knowledge base is used for text representation by merging synsets and selected keywords from which a topic model is developed. Knowledge base definitions are used to create the training data and TFIDF feature selection method is used to extract important features from the Bag of Words document (figure 7.0). The following procedures describe the steps of this paper methodology:-

(i) User Profile Selection

Active user profiles were selected from twitter, the random selection is done from the top 100 most followed profiles on the twittercounter.com website.

(ii) Data Presentation

a) All tweet texts were extracted from user profiles.
b) Stops words and non-descriptive words such as a, and, are and do, non-ascii words are removed.
c) Word stemming is carried out, i.e. words with different ending shall be mapped into single word; computerization, computer and computing are stemmed to comput.

(iii) Topic Model Development

a) Keyword definition from the training data using wordNet
b) Generation of text representation model, which was done by merging keywords with their associated Synsets.
c) Generation of Knowledge Base defined topics using hyperonymy of the hyponymy through WordNet: Exmple [person, individual, someone] **(hyponymy)** implies[Human] **(hyperonymy)**
d) The merged Synsets and Keywords were categorized according to the discovered topics.
e) Data labels (Topics) are pre-determined according to the identified topics
f) Topic model was developed based on combination of Synsets and selected keywords with topic categories as follows: **[Movies, Music, Media, Crime, Politics, Events, Economy, Science and Technology, Legal, Fashion, Sports, Humanity , Social]** (Figure 6.0)

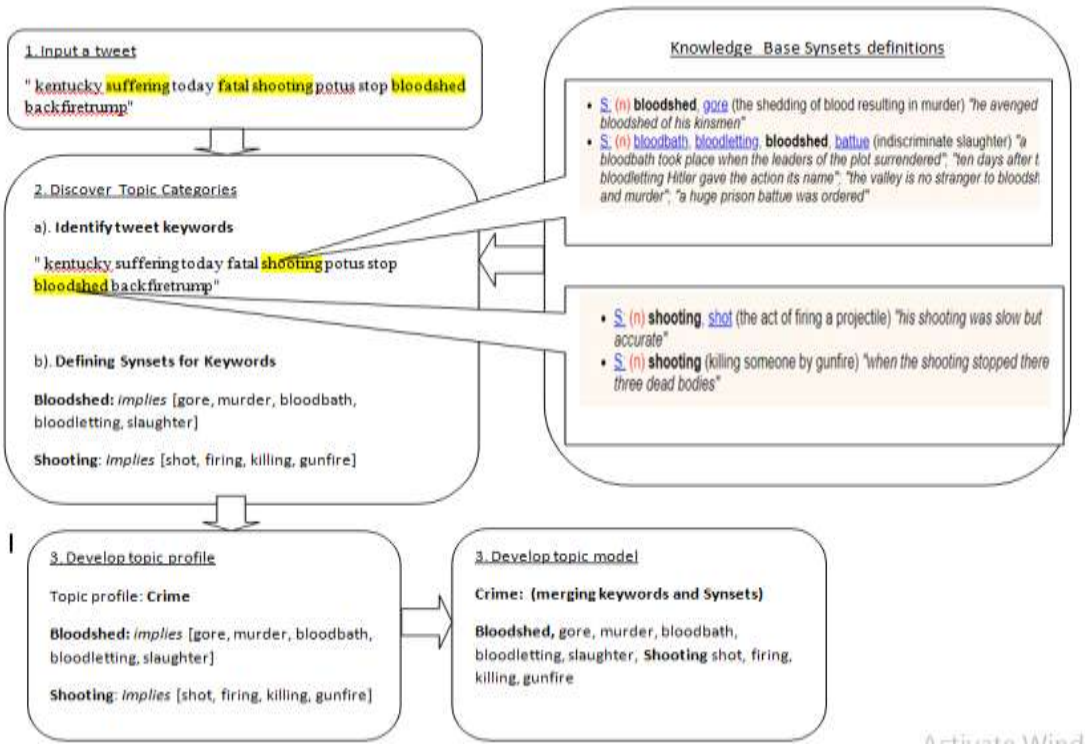

Figure 6.0: Topic Model Development

(iv) Feature Selection

a) All tokenized features from the topic model were fitted to TFIDF (Term frequency - inverse document frequency).
b) Important terms were selected from the topic model based on the TFIDF weighting.

(v) Topic Model Classification

The developed topic model was used to train data with different classifiers such as Support Vector Machine, Naive Bayesian, K-nearest Neighbor and Decision Tree.

vi) Topic model Evaluation

The performance of the topic model was measured using precision, recall and F- measure. Results were compared among the different machine learning algorithms for analysis and discussion.

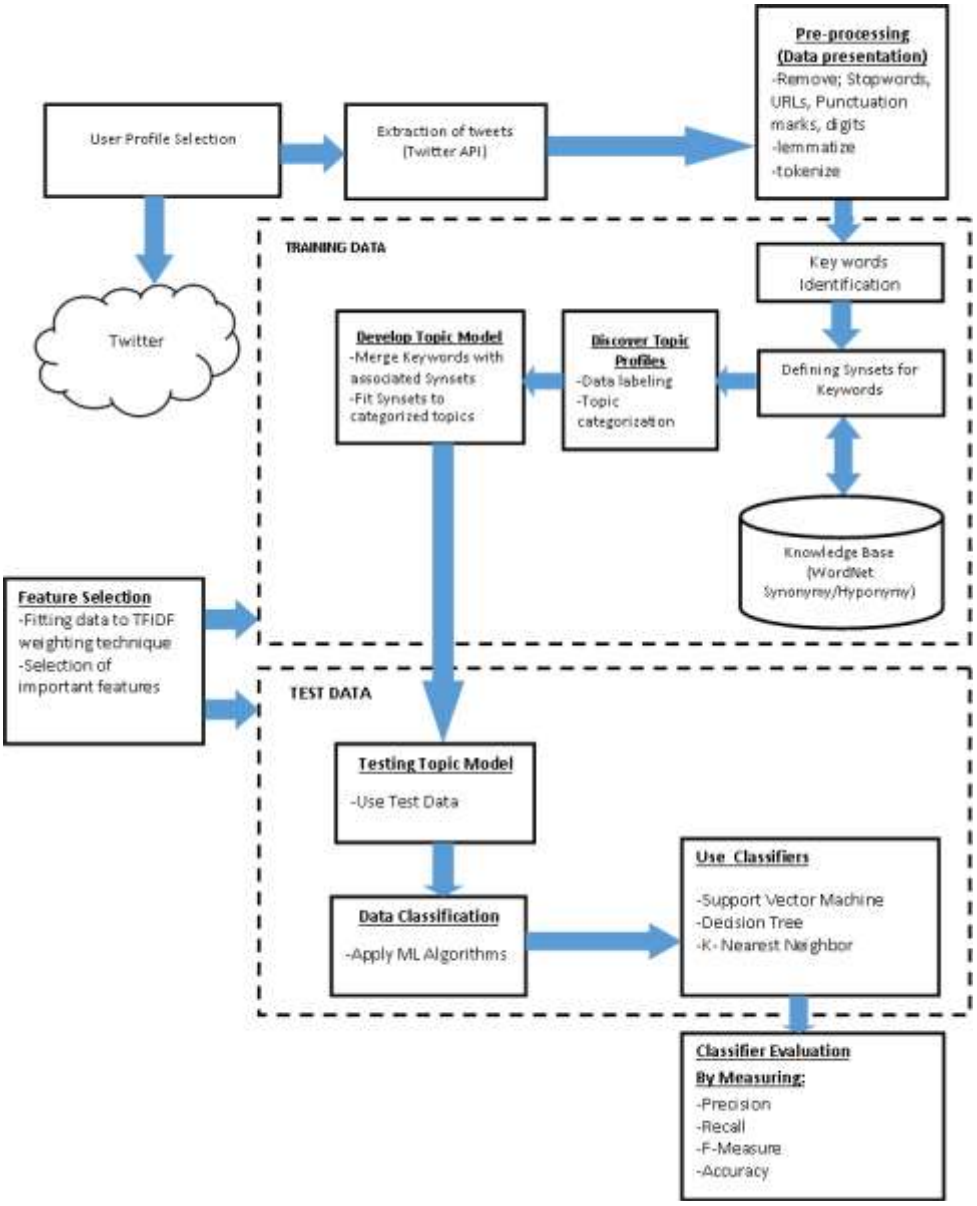

Figure 7.0: Research Methodology Work Flow Diagram

3.1. Population and Sample

The Population of the study consists of twitter user profiles from English speaking social media platform. 400,000 tweets of 100 user profiles were randomly selected from Twittercounter.com. Tweet texts were extracted and pre-processed, Topics of interest were categorized based on the knowledge base definitions. Topic model was developed from 80% of the training data with12 topics as discovered by the knowledge base. The topic model was tested on the remaining 20% of the tweet texts using different classifiers such as Support Vector Machine, K-nearest Neighbor and Decision Tree for analysis and results discussion.

3.2. Data Collection

The research data (tweets) was collected from Twitter.com social network site user profiles to represent a domain of discourse; Selection of user profiles was done manually through twittercounter.com website where the top 100, most followed twitter users were selected . R-studio extraction tool was used to collect 500 fresh tweets from each of the 100 user profiles, which makes 400,000 tweets.

3.3. Data Analysis

The research data was randomly collected from tweet texts, and then recorded in CSV and text files using Twitter API and R. Studio tools. 400,000 tweets were pre-processed, Wordnet Knowledge base was leveraged to add new terms to document representation, topic categorization and labeling, normalized TFIDF was used for feature selection. The data was split into 80% of training data from which a topic model was developed and 20% test data, Machine Learning algorithms such as KNN, SVM and DT were used as classifiers for user topics of interest discovery. The KNN, SVM, and DT were evaluated for performance measuring. The combination of knowledge base and TFIDF was able to discover user topic of interest as expected. The results of the developed user topic of interest discovery model were presented in tables.

IV. RESULTS AND DISCUSSION

The performance of this research approach was assessed by evaluating bag of words (BoW) document modeling technique on machine learning algorithms such as KNN, SVM and DT. The aim of these experiments was to assess the effect of classifier performance on the TFIDF selection method on text representation, evaluation of performance of combining knowledge base synonyms and TFIDF feature selection method to represent text for classification with machine learning algorithms, compare the performance with the baseline results; unweighted bag of words text representation.

The experiment setup of this research used classifiers provided by R Studio. The data sets which were split into 80% training data and 20% testing data . The data sets were fitted to machine learning algorithms; K-Nearest Neighbour (KNN), Support Vector Machines (SVM), and Decision Trees (DT). The experiments started with BoW text representation by introducing neither TFIDF nor Knowledge base Synsets (Terms and Phrases), the results were used for baseline performance, and they were named as BoW_UNWEIGHTED. Then followed by introducing TFIDF to a bag of words, which was named BoW_TFIDF, then TFIDF was introduced to the Bag of Words combined with Synsets (KB Terms and phrases) which was named BoW_TFIDF_KB. These algorithms were used to predict the testing data user topic of interest. Experiment results were then presented in tables, analyzed and interpreted as detailed in section 4.1 to 4.5

4.1. Experiment 1; Training using Unweighted Bag of Words (BoW) with KNN, SVM and DT

A classifier describing labeled set of topics was developed. This was done using training data with only Bag of words document representation and the performance of classifiers was recorded in tables. The accuracy and precision performance was evaluated and results were recorded. BoW_UNWEIGHTED which was a bag of words consisting of keywords only without using TFIDF Feature selection method on text representation. The three machine learning algorithms; KNN, SVM and DT were used and the results are show in the following table 3.0.

Table 3.0: BoW_Unweighted Using KNN, SVM and DT

BoW_UNWEIGHTED	Precision	Recall	F-Measure	Accuracy
KNN	0.348	0.186	0.436	0.494
SVM	0.441	0.254	0.681	0.718
DT	0.011	0.097	0.136	0.270

The unweighted bag of words (BoW_ Unweighted) text representation model in this experiment contains features (keywords only) which were not fitted to any feature selection method (e.g TFIDF). Various machine learning algorithms such as KNN, SVM and DT were used to train the testing data and their performance were evaluated. BoW_Unweighted produced accuracy of 49.4%, 71.8% and 27% for KNN, SVM and DT respectively. The performance of BoW_TFIDF gives this research a baseline for assessing the effect of TFIDF feature selection method, the effect of adding knowledge base synsets to the text representation model and the effect of combining both the TFIDF with Synsets from KB to the text representation. When TFIDF feature selection method is not used and Knowledge base synsets are not introduced to the BoW text representation, SVM classifier outperforms KNN and DT in accuracy.

4.2. Experiment 2; Training using TFIDF on Bag of Words (BoW) with KNN, SVM and DT

A classifier describing labeled set of topics was developed. This was done using training data with only Bag of words representation model and the performance of classifiers was recorded in tables. The performance evaluation was carried out and recorded. BoW_TFIDF which was a bag of words of keywords (without synsets) only with TFIDF method being used for feature selection. The three machine learning algorithms (KNN,SVM and DT) were used and the results are show in the following table 4.0.

Table 4.0: BoW_TFIDF Using KNN, SVM and DT

BoW_TFIDF	Precision	Recall	F-Measure	Accuracy
KNN	0.248	0.096	0.252	0.318
SVM	0.526	0.328	0.689	0.726
DT	0.111	0.971	0.148	0.280

The weighted bag of words (BoW_TFIDF) text representation model in this experiment contains features (keywords only) which were fitted to TFIDF feature selection method. Various machine-learning algorithms such as KNN, SVM and DT were used to train the testing data and their performance was evaluated. BoW_TFIDF recorded accuracy of 31.8%, 72.6% and 28% for KNN, SVM and DT respectively. When TFIDF feature selection method alone is introduced to the BoW text representation, it does not improve the performance of KNN since its accuracy was reduced from 49.4% to 31.8%, but it improves the performance of SVM since its accuracy increased from 71.8% to 72.6% and that of DT increased from 27% to 28% compared to the baseline results.

4.3. Experiment 3; Training using Unweighted Bag of Words (BoW) and KB with KNN, SVM and DT

A classifier describing labeled set of topics was developed. This was done using training data with introduction of knowledge base synsets and the performance of classifiers recorded in tables. The performance evaluation was carried out and recorded. BoW_KB_Unweighted which was a combination of keywords and Synsets without TFIDF Feature selection method being used. The three machine learning algorithms (KNN,SVM and DT) were used and the results are show in table 4.4.

Table 5.0: BoW_KB_UNWEIGHTED Using KNN, SVM and DT

BoW_KB_UNWEIGHTED	Precision	Recall	F-Measure	Accuracy
KNN	0.479	0.188	0.463	0.516
SVM	0.549	0.322	0.677	0.713
DT	0.112	0.097	0.132	0.261

The unweighted bag of words with Knowldge base Synsets (BoW_KB_UNWEIGHTED) text representation model in this research contains both keyword and synsets (features) which were not fitted to TFIDF feature selection method. Various machine learning algorithms such as KNN, SVM and DT were used to train the testing data and their performance were evaluated. BoW_KB_UNWEIGHTED recorded accuracy of 51.6%, 71.3% and 26.1% for KNN, SVM and DT respectively. When Knowledge base synsets alone are introduced to the BoW text representation, they improve the performance of KNN since its accuracy increased from 49.4% to 51.6, while that of SVM slightly reduced from 71.8% to 71.3% and that of DT reduced from 27% to 26.1% compared to the baseline results.

4.4. Experiment 4; Training using Weighted Bag of Words (BoW) and KB with KNN, SVM and DT

A classifier describing labeled set of topics was developed. This was done using training data, which combined knowledge base synsets and TFIDF feature selection method in a BoW. The performance evaluation was carried out and recorded. BoW_KB_TFIDF means a combination of keywords and Synsets with TFIDF Feature selection method in a BoW. The three machine learning algorithms (KNN,SVM and DT) were used and the results are show in table 6.0.

Table 6.0: BoW_KB_TFIDF Using KNN, SVM and DT

BoW_KB_TFIDF	Precision	Recall	F-Measure	Accuracy
KNN	0.418	0.127	0.384	0.438
SVM	0.513	0.297	0.687	0.723
DT	0.112	0.097	0.130	0.259

The weighted bag of words with Knowledge base Synsets (BoW_KB_TFIDF) text representation model in this experiment contains both keyword and synsets (features) which were fitted to TFIDF feature selection method. Various machine learning algorithms such as KNN, SVM and DT were used to train the testing data and their performance were evaluated. BoW_KB_TFIDF recorded accuracy of 43.8%, 72.3% and 25.9% for KNN, SVM and DT respectively. Combining TFIDF feature selection method and Knowledge Base synsets on the BoW text representation does not improve the performance of KNN and DT since their accuracy reduced compared to the baseline results. However, the combination improves the performance of SVM since its accuracy increased from 71.8% to 72.3%.

4.5. Comparison of Performance of Various Machine Learning Algorithms

In accuracy performance, SVM classifier performed the highest with 72.6 % on BoW_TFIDF text representation, 71. 3 % on BoW_KB_UNWEIGHTED and 72.3% on BoW_KBsv_TFIDF. In the other hand KNN performed its best accuracy with BoW_KB_UNWEIGHTED at 51.6% , SVM performed its best with BoW_TFIDF at 72.6% . From table 4.6.1 it is clear that DT accuracy performance is below average even from the results given by its respective baseline.

In precision performance (Table:8.0), SVM classifier also performed the highest with 52.6 % on BoW_TFIDF text representation, 54.9 % on BoW_KB_UNWEIGHTED and 51.3% on BoW_KB_TFIDF. in the other hand KNN, SVM and DT performed their best precisions with BoW_KB_UNWEIGHTED at 47.9% , 54.9% and 11.2% respectively. However DT is still under performing with the below average precision results

Table 7.0: A comparison of Accuracy Performance between various text representations

Document Representations	Accuracy		
	KNN	SVM	DT
BoW_UNWEIGHTED	0.494	0.718	0.270
BoW_TFIDF	0.318	0.726	0.280
BoW_KB_UNWEIGHTED	0.516	0.713	0.261
BoW_KB_TFIDF	0.438	0.723	0.259

Table 8.0: A comparison of Precision Performance between various text representations

Document Representations	Precision		
	KNN	SVM	DT
BoW_UNWEIGHTED	0.348	0.441	0.011
BoW_TFIDF	0.248	0.526	0.111
BoW_KB_UNWEIGHTED	0.479	0.549	0.112
BoW_KB_TFIDF	0.418	0.513	0.112

V. DISCUSSION

The results of various experiments are discussed in this section. First, the effect of TFIDF feature selection method with respect to user topic of interest discovery on BoW text representation and data training is described. Secondly, the effect of adding Synsets (terms and phrases) from KB definitions with respect to user topic of interest discovery on training data is also described. The effect of combining both the Synsets and TFIDF technique to the text representation model is presented. The performance evaluation of machine learning algorithms for test data classification in discovering user topic of interest is also discussed. Finally, the comparison in classifier performance for discovering the user topic of interest in test data is describe.

TFIDF feature selection method affect the quality of training data, when BoW text representation is fitted to the TFIDF selection methods it affects the Performance of machine learning algorithms in various ways. Accuracy and precision of some classifier increases and others reduces, this is shown in table 4.0 and 6.0. Also In some instances when TFIDF is introduced it records highest values and in some it reduces the values compared to the baseline results. This can be concluded that it is not always that when TFIDF technique is used gives the best results in performance; the performance also depends on the selection of an appropriate machine learning algorithm.

The addition of terms and phrases (Synsets) from KB also affect the quality and size of training data, when Synsets are added to BoW text representation model they affect the Performance of machine learning algorithms in various ways. Accuracy of some classifier increases and others reduces, this is shown in table 5.0 to 6.0 Also In some instances when Synsets are introduced the classifiers it record highest values and in some it reduces the values compared to the baseline results. This can be concluded that it is not always that when Synsets are added to text representation model gives the best results in performance; the performance of classifier also depends on the selection of an appropriate machine learning algorithm.

Combining TFIDF feature selection method and addition of Synsets to the text representation model to discover user topic of interest affects the of machine learning algorithms in various ways. Accuracy of some classifier increases and others reduces; this is shown in table 6.0. Also In some instances when Synsets combined with TFIDF technique are introduced to classifiers they record highest values and in some it reduces the values compared to the baseline results. This can be concluded that it is not always that when the combined approach is used on text representation model gives the best results in performance; the performance of classifier also depends on the selection of a appropriate machine learning algorithm.

VI. CONCLUSION AND RECOMMENDATION

The performance comparison of various machine learning algorithms in this experiment shows that SVM gives a higher value when used with combined (TFIDF and KB) approach on training data. However, the same classifier, SVM, gives even a highest value in performance compared to all other classifiers in the

experiment when used on BoW and KB without TFIDF. Also the experiments show that KNN can be considered for combined (BoW and KB without TFIDF text representation) approach as shown in table 7.0. In conclusion, the combination of TFIDF feature selection method and Knowledge Base synsets to the BoW text representation improves the performance of SVM, which outperformed both KNN and DT. Therefore, SVM machine learning algorithm is best suited for this approach.

VII. FUTURE WORK

There is need to explore more feature selection methods in combination with Knowledge Base Synsets for the text representation models; Information gain, Knowledge gain and Chi-square that could be combined with knowledge base terms, phrases and concepts for text representation. This research also needs to be enhanced by experimenting with more classifiers such as Naive Bayes (NB), Radial Support Vector Machine (RSVM) and Maximum Entropy (ME) for the discovery of user topic of interest in social media. Furthermore, there is need for more research by considering other tweet attributes such as location and connection between users for the discovery of user topic of interest on social media.

REFERENCE

[1]. Abel, F., Herder, E., & Krause, D. (2011). Extraction of professional interests from social web profiles. Proc. UMAP, 34.

[2]. Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. Big Data Research, 2(3), 87–93.

[3]. Annachhatre, C., Austin, T. H., & Stamp, M. (2015). Hidden Markov models for malware classification. Journal of Computer Virology and Hacking Techniques, 11(2), 59–73.

[4]. Bermejo, P., Gámez, J. A., & Puerta, J. M. (2014). Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. Knowledge-Based Systems, 55, 140–147.

[5]. Castells, M. (2007). Communication, power and counter-power in the network society. International Journal of Communication, 1(1), 29.

[6]. Certeau, M. de. (1988). The practice of everyday life. Trans. Steven Rendall. Berkeley: University of California Press. Retrieved from http://library.wur.nl/WebQuery/clc/1846341

[7]. Dai, W., & Ji, W. (2014). A mapreduce implementation of C4. 5 decision tree algorithm. International Journal of Database Theory and Application, 7(1), 49–60.

[8]. Dredze, M., McNamee, P., Rao, D., Gerber, A., & Finin, T. (2010). Entity disambiguation for knowledge base population. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 277–285). Association for Computational Linguistics.

[9]. Elberrichi, Z., Rahmoun, A., Bentaalah, M. A., & Laboratory, E. (2008). Using WordNet for Text Categorization, 5(1), 9.

[10]. Escalante, H. J., García-Limón, M. A., Morales-Reyes, A., Graff, M., Montes-y-Gómez, M., & Morales, E. F. (2014). Term-Weighting Learning via Genetic Programming for Text Classification.

[11]. Eshghi, K., & Kafai, M. (2016). Support vector machines with sparse binary high-dimensional feature vectors. Hewlett Packard Labs, Palo Alto, CA, USA, Tech. Rep. HPE-2016-30, 1–10.

[12]. Feenberg, A. (2009). Technology, Power, and Freedom. Readings in the Philosophy of Technology, 139.

[13]. Grycuk, R., Gabryel, M., Korytkowski, M., & Scherer, R. (2014). Content-based image indexing by data clustering and inverse document frequency. In International conference: beyond databases, architectures and structures (pp. 374–383). Springer.

[14]. Gurney, K. (2014). An introduction to neural networks. CRC press.

[15]. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier. Iofciu, T., Fankhauser, P., Abel, F., & Bischoff, K. (2011). Identifying Users Across Social Tagging Systems. In ICWSM.

[16]. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[17]. Jurafsky, D. (2000). Speech & language processing. Pearson Education India.

[18]. Kapanipathi, P., Jain, P., Venkataramani, C., & Sheth, A. (2014). User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In The Semantic Web: Trends and Challenges (pp. 99–113). Springer, Cham. https://doi.org/10.1007/978-3-319-07443-6_8

[19]. Kurzweil, R., Richter, R., & Schneider, M. L. (1990). The age of intelligent machines (Vol. 579). MIT press Cambridge.

[20]. Lin, K.-C., Zhang, K.-Y., Huang, Y.-H., Hung, J. C., & Yen, N. (2016). Feature selection based on an improved cat swarm optimization algorithm for big data classification. The Journal of Supercomputing, 72(8), 3210–3221.

[21]. Luger, G. F., & Stubblefield, W. A. (1993). Artificial intelligence: its roots and scope. Artificial Intelligence: Structures and Strategies For, 1–34.

[22]. Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. In Proceedings of the fourth workshop on Analytics for noisy unstructured text data (pp. 73–80). ACM.

[23]. Mohamad, M., & Selamat, A. (2015). An evaluation on the efficiency of hybrid feature selection in spam email classification. In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on (pp. 227–231). IEEE.

[24]. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[25]. Ogada, K., Mwangi, W., & Cheruiyot, W. (2015). N-gram Based Text Categorization Method for Improved Data Mining. Journal of Information Engineering and Applications, 5(8), 35–43.

[26]. Raad, E., Chbeir, R., & Dipanda, A. (2010). User profile matching in social networks. In Network-Based Information Systems (NBiS), 2010 13th International Conference on (pp. 297–304). IEEE.

[27]. Sathi, V. R., & Ramanujapura, J. S. (2016). A Quality Criteria Based Evaluation of Topic Models.

[28]. Scott, P. R., & Jacka, J. M. (2011). Auditing social media: A governance and risk guide. John Wiley & Sons.

[29]. Sehgal, A. K. (n.d.). Profiling Topics on the Web for Knowledge Discovery (A report submitted in partial fulfillment of the requirements of the Ph. D Proposal Examination in the Department of Computer Science).

[30]. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.

[31]. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 841–842). ACM.
[32]. Van Dijck, J. (2013). The culture of connectivity: A critical history of social media. Oxford University Press.

Athman Masoud. "Topic of Interest Discovery on Social Media Using Knowledge Base and Term Frequency – Inverse Document Frequency Techniques"The International Journal of Engineering and Science (IJES),), 7.10 (2018): 01-20