A PREDICTIVE MODEL FOR TWEET SENTIMENT ANALYSIS AND CLASSIFICATION

Abdullah K.-K. A., Folorunso S. O., Solanke O. O., Sodimu S. M.

Department of Mathematical Sciences, Olabisi Onabanjo University, Ago Iwoye, Ogun State, Nigeria

Corresponding Author: Abdullah K.-K. A., uwaizabdullah9@gmail.com

ABSTRACT:Sentiment analysis over Twitter offers organisations and users a fast and effective way to monitor publics' feelings towards events especially during crises, hence, motivated much work on twitter data. In this study, predictions on positive, negative and neutral sentiment based on security are analysed. A polarity classification of tweet messages was done with VADER algorithm considering contextual analysis. The analysis was performed by removing stop words in the tweet along with Wordnet lemmatiser for the morphological analysis of words in the features sets. As well as subjected to word sense disambiguation to consider contextual usages of words using a path length corpus based lexicon. Term Frequency (TF) and Term Frequency Inverse Document Frequency (TFIDF) are used as feature extraction from tweets and evaluation of the features reduction was carried out by calculating the accuracy of the predictions on sentiment and tweet messages with Chi-Square to explore the possibly useful features. Finally, validations are done with machine learning models at different sequence to compare the performance between each model.

KEYWORDS: Sentiment analysis, Word sense disambiguation, Natural Language Processing, Chi-Square, VADER.

1.0 INTRODUCTION

Social networking sites also called micro blogging are the fastest emerging in recent years which become a real time communication channel. Several researches have been done in the field of social networking namely, classification of topics, sentiment analysis of users' opinion, event detection, community detection, etc. Users of social media communicate instantly about the real time happenings and express their opinion on a common topic publicly through their computers or mobile phones [M+08]**.** Twitter is one of the popular online social blog and has vast amount of information available in the form of tweets where many users post tweets related to an event. Twitter has been found to be faster than news channels but it is difficult to extract and analysed relevant information from people' opinion (sentiment) and classify users' tweet using traditional approaches. Methods of sentiment analysis can be categorised predominantly as machine-learning [MHK14], Lexicon-based [T+11] and hybrid [PT09, MM13] Mining such valuable information reflects aspects of real-world societies and helps to identify the events that occurred over space and time.

There are numerous research challenges inherent in traditional methods of analysing twitter data. The challenges faced in recent time are short messages filled with spams, slang or internet memes which lack semantics. Also, the view of the user depends on sentiment analysis of tweets not just the binary polarity (positive or negative), but the strength of the sentiment expressed in text. However, data from social media come in great volume and velocity with large number of features or parameters (dimensionality) for evaluating classifier. This makes the processing computationally infeasible and pose problems for machine learning as predictive models, therefore, run the risk of over fitting. Such large number of features may reduce the possibility of getting higher accurate results from testing datasets. Contrarily, on the context issue, word has many meanings. To extract the correct meaning of a word according to the context, the Natural Language Processing (NLP) task known as Word Sense Disambiguation (WSD) is used. A NLP embedded with Semantic Network (WordNet) utilises hard coded techniques in order to perform lexical analysis on textual data considering the number of words and function words as feature set. This can be evaluated and implemented on a semantic similarity metrics to determine the affective content of a word in different dimensions [Y+12]**.** Valence Aware Dictionary for sEntiment Reasoning (VADER) lexicon performs exceptionally well in the social media domain [HG14]in terms of polarity taking into consideration the contextual emotion such as punctuation, slang, modifiers. This mapped to intensity values and performs normalisation to strengthen the sentiment rather than positive, negative or neutral labels. For feature reduction utilises a direct change to shape streamlined information set holding the qualities of the first information set. A Chi-Square and some other feature selection solve the problem of dimensionality of features without losing

information in the original dataset. This helps to improve text classification accuracy and filtering [SCS17]. In order to add accuracy to the model and disambiguate words in twitter, a semantic Wordnet corpus is used with different supervised machine learning techniques to learn set of target words. The aim is to build a classifier that maps each occurrence of a target word in a corpus to its senses. In other words, the context of an occurrence of a target words in the corpus is represented as a feature vector, the classifier estimates the word senses on the basis of its context.

This paper focuses on security. Security is a state of being free from danger or threat. The safety of a nation or organisation must be provided against criminal activity such as terrorism, attack, kidnapping, theft or espionage. Recent research studies have demonstrated that social networks have the opportunity to strengthen national security and are used to benefit the government such as warning or trend prevention. As well as a monitoring tool, to recognise the first signs of any hostile or potential dangerous activity by collecting and analysing message. On the Twitter platform, users' interests on topics can be extracted from the published tweets to construct and then apply models to predict or make conclusion on an event.Illustration on the relation between sentiments of twitter post related to security is examined with different machine learning techniques.

The paper is organised as follows: Section 2 presents background and related works. Proposed method of sentiment analysis with classification models in section 3, while presenting the collection of data from social networks based on the vocabulary associated with security. Section 4 gives experimental results, conclusion and future works are given in Section 5.

2.0 BACKGROUND AND RELATED WORK

Internet users tend to express opinions, feelings and talk about lives and activities of everyday life via twitter [GME17]. Twitter allows the users to post messages called tweets which is no longer than 140 characters. Each user has several followers, who can retweet their post. Using the Twitter follower graph might improve the polarity classification [S+11]. Because of its huge users, the content they tweet also varies based on their interest and behaviour [KPA08]. Users of tweet publish tweets on real time on their opinions on any topic. Twitter can be called a dynamic source of information [A+12, LD13], mining or analysing such rich source of data can provide unknown information. It has also become a powerful tool to monitor activity such as terrorists and also predict crimes [S+11]. Mining opinion research works are done at the level of the document or sentence [WGB12]. According to [BNG11], an event is a real-world occurrence *e* with a time period $T_e$and a stream of twitter messages discussing the event during the period T_e. This definition has a twitter scope and is related to an increased amount of messages in a time window.

Sentiment analysis is treated as a task of natural language processing at several levels of granularity where classifications are generally based on the identification of opinion words. Present methodologies choose a selection of words and hashtags to follow during an event with tweets containing the selected words being deemed relevant to the event [WZL11]. In [EBG11], two approaches are suggested for sentiment analysis: the machine learning approach which converts each text into a list of words (unigram), consecutive word pairs (bi-gram) or consecutive word triplets (n-gram). Then based it upon some human coded set of texts which learn features and associate it with sentiment scores to classify the new cases. Secondly, the lexical approach uses some grammatical structure of language and some list of words with sentiment scores and polarities. Each of these methods cannot only be applied on tweet due to its volume and grammatical flaws. It this study the two methods are used to have a better accuracy.

Data collection in most existing work on twitter used Twitter4j application programming interface (API) which is is an integrated Java library with all services related to Twitter. Twitter data were extracted from Twitter by using official Twitter TweepyAPI to gather data for the prediction. The aim was to use Twitter data to understand public opinion. After the collection of data, automatic buzzer detection is used to remove unnecessary tweets/ retweets and then analysed the tweets sentimentally by breaking each tweet into several sub-tweets. Many dictionaries have been created manually such as ANEW (Affective Norms for English Words) or automatically such as SentiWordNet [BES10], TextBlob[SYR17] that allow estimating a score of the negativity, positivity and objectivity of the tweets, their polarity and subjectivity but neglect the aspect of the context especially in the area of domain. Each word may be represented in one or more sense (polysemous) [WAH14]. Word Sense Disambiguation (WSD) is the ability to disambiguate a word that can have many senses based on its usage context. This can be achieved through engaging knowledge based method, using WordNet, which is a lexical database. In Wordnet, the words are grouped into synsets representing the meaning of the words and hence semantic similarity between two synset is computed by measuring the path-length based similarity. This

leverages WordNet to score sentiment according to the English part of speech used in the text and add accuracy to the model.

A lot of work has been done to analysed tweet sentiments by applying machine learning techniques along with semantic analyses to classify product reviews or sentences. Also, different feature ranking techniques were presented which helps to improve text classification accuracy and filtering [SCS17]. However, [LD13]proposed a method which combines supervised learning that is capable of extracting, learning and classifying tweets with opinion expressions. They used mutual information and chi-square as feature selection methods to reduce the features in the dataset with Naive Bayes classifier. This improved the accuracy with better result with chi-square. This method was adopted with maximum optimisation to reduce the dimension of the features. A regression model was used in [DTF14]while [V+15, A+11] rely on other methods like Naive Bayes, Support Vector Machines (SVM), and Decision Trees (J48). Semantic user modeling has been done based on twitter posts in [MAM17]**.** They suggested a formula for user's similarity which is based on topic discussed by the users, however, all the corresponding machine learning algorithms classifier were found on the training tweets data using WEKA [The13]. In[KU16], theyemployed sentiment analysis to determine the polarity of English and Tamil tweets (bilingual) and word sense disambiguation to figure out the contextual usage and further the sentiment of the words are classified using the Support Vector Machine. Khan *et al.* [KKK13] proposed a method of word sense disambiguation (WSD) using matrix map of the semantic scores extracted from SentiWordNet of WordNet glosses terms. The correct sense of the target word is extracted and determined for which the similarity between WordNet gloss and context matrix. The method achieves an accuracy of 90.71% at sentence level sentiment classification. In [SYR17], Textblob was used for preprocessing and polarity, the polarity confidence calculation and validation are obtained by SVM and Naïve Bayes using Weka. It was reported that Naïve Bayes gives the highest accuracy. While [A+18] adopted hybrid approach that involved a sentiment analyser and machine learning, this provides comparison of techniques of sentiment analysis.

3.0 METHODOLOGY

The proposed method are categorised into two parts: the online (extraction of Twitter data) and offline (analysis of the twitter sentiment). The architecture of proposed method on security detection is based on extraction of data from Twitter in real time. VADER algorithm is used to categorise the tweets into positive, negative and neutral. This evaluates the effect on detection of events from Twitter, hence, preprocessing, feature extraction and features selection as well as WordNet semantic similarity for improving the vocabulary of the tweet.Finally, different machine learning models are used for validation.

3.1 Data Collection

Twitter is structured as a directed graph, each user can choose to follow a number of other users (followees), and followed by other users (followers). Therefore, data are collected in three forms: the User type represents users' profiles; the Tweet type represents posted messages and the follow type represents relationships among users as shown in the Figure 1.

In this work, the tweets are collected using the search word defined as the query. For the extraction of Twitter, a token and access keys are obtained, hence, a Python library called Tweepy is connected to Twitter API to access the public security data. Twitter dataset is taken several times for more tweets are necessary from the Twitter page. Tweets posted by users in the form of hashtags are considered to express opinions about current security trends in Nigeria shown in figure 1. During this period, event related to security in Nigeria are considered. In order to easily determine the degree of security issues such as terrorism, insurgence, kidnapping etc. of each tweet, the recipients involved are identified and then stored. In the dataset, retweets are filtered out to avoid repetitions of tweets. The dataset contains 2,703 tweets for a period of 4weeks.

3.2 Sentiment Text Analysis

In preparing text for analysis, managing large datasets such as Twitter need preprocessing. Since the tweet are not categorized, polarity and subjectivity of the sentiment analysis was extracted using VADER algorithm to allow estimating score of the tweets for positive, negative and neutral words. VADER performs better across domains. Figure 2 categorize the sentiment into positive, negative and neutral with VADER algorithm.

Tweet	Date Created
Just In: Herdsmen Attack #Enugu Community, Machete Security Guard #Nigeria #Security https://t.co/hBPApJcTQi https://t.co/uE7aPcdCvD	2018-04-03 15:03:37
Saraki made the call at the opening of a two-day Security Summit organised by the Senate Ad-hoc Committee on Review of Current Security Infrastructure in Nigeria. #SecuritySummit Cc @bukolasaraki https://t.co/ldf8uMagBx	2018-04-03 15:01:14
Cross River vigilante group to partner security agencies in crime fighting By Joseph Kingston, Calabar The Commandant of the Vigilante Group of Nigeria (VGN), Cross https://t.co/uGUJw5UM1b	2018-04-03 14:54:32
Cross River vigilante group to partner security agencies in crime fighting By Joseph Kingston, Calabar The Commandant of the Vigilante Group of Nigeria (VGN), Cross https://t.co/rw2emlWCDj	2018-04-03 14:54:30
President Buhari Inaugurates Food Security Council-Photos President Muhammadu Buhari on Monday March 26th, inaugurated the National Food Security Council. The inauguration took place inside the... https://t.co/sHk	2018-04-03 14:47:49
@Audu Chief just the neighborhood watch here has stopped and saved lives.The security situation in Nigeria would be well if we all speak out,criminals have friends,siblings and have homes.Dole	2018-04-03 14:34:24
Our security protocol in Nigeria is almost military grade! Can you lock pick 'Jam Lock' with 4 padlock and a 'protector' behind the door? I leave that question for my opponent to answer! https://t.co/wRzP16k74c	2018-04-03 14:05:38
@sraradhana Hello greetings .am mr Joseph from nigeria. Am a paramilitary personnel.. How do i apply for security job in your firm I have 6yrs experience with proper training	2018-04-03 13:56:53
BIAFRANS CHRISTIANS ARE BEEN MURDERED DAILY BY ISLAMIC SECURITY FORCE NIGERIA GOVERNMENT UNLAWFUL? @BErNIE SANDERS CAN'T NEVER SPEAK UP? CAN'T CONDEMN IT? SHAME TO YOU @b	2018-04-03 13:41:54
River Basin boards asked to reposition to ensure food security - EnviroNews Nigeria - https://t.co/tX2ckimXqr	2018-04-03 13:19:38
If you doubt me, consider Congo DRC. Nobody gives a shit about Congo, because it doesn't fit into the "War on Terrorism". This will be our fate in Nigeria if we don't deal with our looming internal security crises - by ourselves.	2018-04-03 13:04:44
Nigeria's President has twice declared Boko Haram defeated. An attack on military barracks represents a significant escalation. https://t.co/j4w2haRALj	2018-04-03 12:56:41
Nigeria Assumes AU Security Council Chair for April https://t.co/w46a6xq2Tu via @allafrica	2018-04-03 12:55:38
@julietkego Minister of Defence. Nigeria's security is the cornerstone of any progress it makes. I would move to enhance the role of our armed services in three areas: battle readiness, the mitigation of covert and overt thre	2018-04-03 12:35:34
Well, d fight is ended. I only hope people can live with d threats they've exposed themselves too #Nigeria #security https://t.co/0JJ0MFkW1k	2018-04-03 12:29:46
Security agencies are abetting the fight against illegal Oil Bunking in Nigeria, Rivers State. @AsoRock @riversstategov @NGRSenate @RotimiAmaechi @PetersideDakuku	2018-04-03 12:22:32
@gtbank I know even Visa & MasterCard Nigeria have 3D security enabled for 1st time transactions, so why the system fail? @mypaga @interswitchGRP @Visa @VisaNigeria @mastercard @MastercardMEA @cenbank	2018-04-03 12:20:31
Retweeted The Guardian Nigeria (@GuardianNigeria): Fayose likens Buhari to a father who is protecting his children that are armed robbers but calling on security agents to arrest his neighbour's children for stealing meat from their mother's pot. https://t.co/HrhEmdf32E	2018-04-03 12:04:52

Figure 1: Selected Tweet Datasets

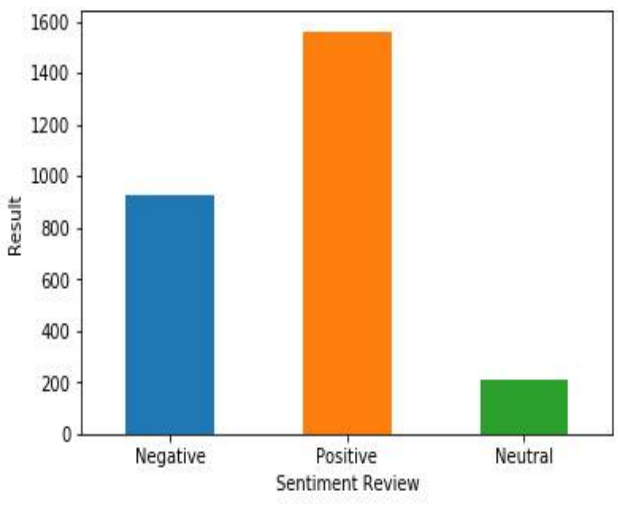

Figure 2: Categorization of Sentiment Analysis with VADER

3.2.1 Preprocessing

This is a vital part of any NLP system as characters, words, and sentences identified are crucial for preprocessing, it includes:

- *Data filtering and Data cleansing:* These removed noisy data and unwanted data such as special character (#, @, !.), URL, username, non ASCII characters from the text.
- *Normalization:* Text data are changed from upper case to lower case.
- *Tokenization:* All tweets are transformed into a well represented format known as token using Natural Language Toolkit (NLTK).
- *Uni-gram Feature:* It is assumed that words are independent from each other and it disregards the order of appearance, therefore a uni-gram (Bag of Words) is used as a baseline model. This represents text as unordered set of words.
- *Stopwords:*All stopwords whose length is greater than four (4) are removed from the stop words leaving words such as 'in', 'the', 'of' etc. with exception to words such as : ***no, not, none*** to help the sentiment negativity.
  The proposed method exploits some additional features for sentiment analysis that extend the representation of tweets.
- *Part of Speech (POS tagging):* This helps to extract contextual word in NLP, it identifies and treats different meaning of polysemy word and setting POS taggingto noun.

- *Semantic Analysis:* In order to disambiguate a word with many sense based on its usage context, Leacock and Chodrow semantic similarity WordNet is used where each word is associated with others (interconnected). That is two words are close to each other when they are semantically similar, hence, determine the pathlength similarity of the two words as synonym, then, map the words and examine their relationship. It checks the similarity of the words that the user uses in their tweets. The benchmark is set to 2.5 and above since the maximum positive score is 3.6 and maximum negative score is -3.6 for the similarity
- *Lemmatisation:* Instead of using stemming as most existing work, a WordNet lemmatiser provided by python is used. This considered the morphological analysis of words in the features sets and reduces to its dictionary based form e.g. go, goes, gone, foot, feet and improves classification performance.

3.2.2 Feature weighting

After preprocessing, the remaining data are subjected to different features methods in order to improve the accuracy of sentiment classification, for the purpose of finding strongly related words for relevant documents and dimensionality reduction of features. The extraction of each feature is transformed into feature vector in binary form. The following feature extraction and selections are used for this study and the dataset was reduced into different training set for maximum optimisation.

- ***Term Frequency (TF):*** This normalised the length to 1, no bias for short or longer words.

$$TF(t) = 1 + \log(f[t,d] \tag{1}$$

where, $f[t,d]$ is the count of term t in document d

- ***Term Frequency-Inverse Document Frequency (TFIDF):*** This determines important word in the Twitter dataset. IDF put less weight on common terms by normalising each word with the inverse in corpus frequency. Then, adjust using Inverse Document Frequency (IDF).

$$TFIDF = TF * IDF \tag{2}$$
$$idf[t] = \log(N/df[t])$$

Where, N is the total number of document in the dataset,

$df[t]$ is the number of documents containing the term t

- ***Chi-Square*** (χ^2): This is use to improve classification performance and efficiency. It normalised values by weeding out words that are independent of class, hence, irrelevant for classification. It represents the degree of relationship of features, however, use for finding spam tweets. It measures how much expected counts and observed counts deviate from each other.

$$X^2 = \sum_{i=0}^{n} \frac{observed - \exp ected)^2}{\exp ected}$$
$$\chi^2(f,t) = \frac{N(AD-CB)^2}{(A+C)(B+D)(C+D)} \tag{3}$$

Where, f is a feature (a term in the Twitter),
B is the number of times f occurs without t,
t is a target variable for prediction,
C is the number of times t occurs without f,
N is the number of observation,
D is the number of times neither t and f occurs,
A is the number of times f and t co-occur.

3.3 Classification Model

The model was tested on the training dataset to obtain the accuracy result of each classifier method using 10-fold cross-validation. The transformed dataset with *tfidf* and *TF* features using prediction such as Gaussian Naïve Bayes, Logistic regression, Support Vector Machine, Decision Tree and Random Forests models are used in prediction tasks. The description of each model and reasons for selecting these models are as follows:
Given a polarity label y where, y = {positive, negative and neutral}, and features vector x, target function f and Probability P.

1. Logistic Regression

Logistic regression uses a Logistic function to estimate probabilities between positive or negative label y and features vector x_i for the purpose of classification.

$$f_\theta(y/x) = \sum_{i}^{n} -\log(1+e^{-x\theta}) + \sum_{y_i=0}^{n} -x_i\theta - \lambda\|\theta\|_2^2 \tag{4}$$

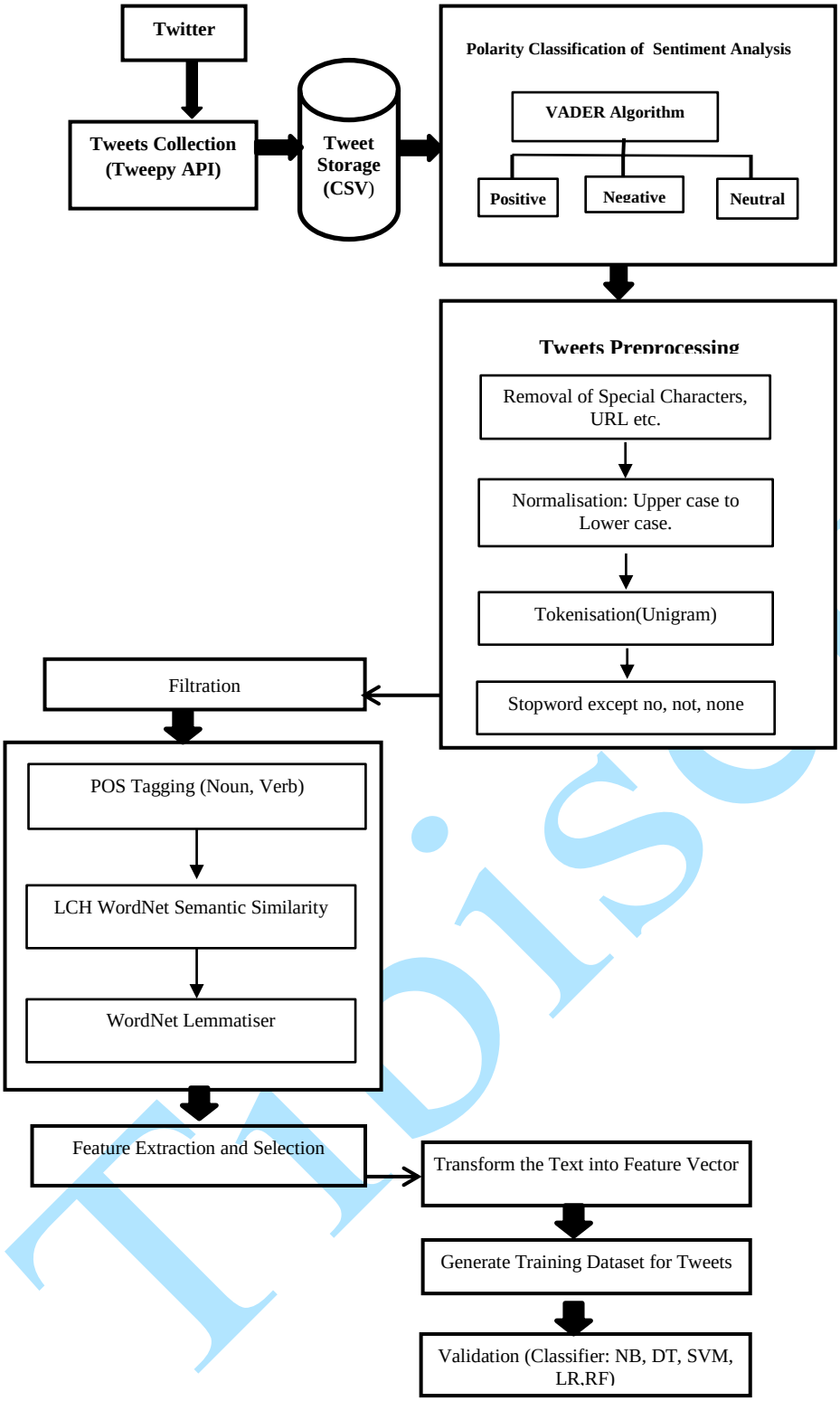

Figure 3: Proposed Architecture of Tweet Sentiment Analysis and Classification

2. Nave Bayesian

When assumption of independence hold

$$p(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (5)$$

where y is the label and x is a dependent feature vector of size n and $X = (x_1, x_2, \ldots x_n)$.

3. Support Vector Machine

It is used to minimize the misclassification error (Optimisation problem)

$$y = \arg\max_{y \in Y} f_y(x_i) \quad (6)$$

where $f_y(x) = w_y x$ is a classifier associated with label y .

4. Decision Tree

A Decision tree impelling chooses significant features. Choice tree actuation is the learning of choice tree classifiers, building tree structure where every inside hub (no leaf hub) signifies quality test.

$$D = \sum_{i=1}^{n} P_i \log_2(p) \quad (7)$$

Where p_i is the probability that arbitrary vector in D belongs to label i.

5. Random Forest

This builds multiple decision (ensemble) trees and merges them together to get better accurate prediction.

$$P(x|y) = \sum_{1}^{n} P_n(x|y) \quad (8)$$

Algorithm *: The sentiment Analysis algorithm for a Twitter Text*

Input: *A Twitter Text*

Output: *Sentiment Polarity and Classification*

1. // use vader sentimiment analyzer for Twitter data

```
sid = SentimentIntensityAnalyzer()
#0 - Negative, #1 - Positive, #2 - Neutral
for i  in range(0,len(train_dataset_filter)):
    ss = sid.polarity_scores(original_tweet)
    if ss["compound"] == 0.0:
        vader_dataset['vader'][i] = 2
    elif ss["compound"] > 0.0:
        vader_dataset['vader'][i] = 1
    else:
        vader_dataset['vader'][i] = 0
```

2. custom stop word removal(limit to 4 letters to avoid negativity filtering)

```
tweet = customStopwordRemoval(tweets)
```

3. // wordSenseDisambiguation using leacock for all the tweets

```
tweet = lch_similarity(word1_syn, word2_syn)
```

4. //turn all tweets into matrix/vector

```
bow = TfidfVectorizer(tweet)
```

5.

```
kbest_feature = [{}];
kbest_num = 5000;
counter = 0
for (k = 0; k< kbest_num; k++){
    kbest_feature = getChi2Features(bow)
    counter = counter + 1
    if len(kbest_feature) == 1000 :
        break
```

6. Classifier algorithm

4.0 RESULT AND DISCUSSION

In order to evaluate the effectiveness of the proposed method, an experiment were performed on 2703 tweets dataset collected and transformed into 5283 bags of words after preprocessing and feature weighting.In calculating the average classification accuracy for the test data, then equation 9 is used.

$$\text{Accuracy} = \sum_{i=1}^{y} = \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i} \quad (9)$$

Where, tp_i are true positive

tn_i are true negative

fp_i are false true positive

fn_i are false true negative

y is the number of labels

Table 1 : TFIDF with Chi-Square on MLT

KBest	NB	LR	DT	RF	SVM
1000	0.7468	0.7523	0.7135	0.7597	0.5638
1500	0.7671	0.7486	0.7246	0.7542	0.5638
2000	0.7394	0.756	0.7079	0.7431	0.5638
2500	0.7283	0.756	0.7079	0.7579	0.5638
3000	0.732	0.7579	0.7098	0.7412	0.5638
3500	0.7652	0.756	0.7153	0.7357	0.5638
4000	0.7079	0.7505	0.7043	0.7375	0.5638
4500	0.6969	0.7486	0.6876	0.7449	0.5638
5000	0.671	0.7412	0.7024	0.7468	0.5638

Table 2: Term Frequency with Chi-Square on MLT

KBest	NB	LR	DT	RF	SVM
1000	0.7079	0.7985	0.7227	0.7505	0.5638
1500	0.7837	0.7967	0.7043	0.7338	0.5638
2000	0.7116	0.7967	0.7098	0.7375	0.5638
2500	0.7190	0.7856	0.7172	0.7542	0.5638
3000	0.7116	0.7893	0.7116	0.7449	0.5638
3500	0.7338	0.7819	0.7061	0.7431	0.5638
4000	0.6987	0.7837	0.7098	0.7375	0.5638
4500	0.6987	0.7837	0.7024	0.7634	0.5638
5000	0.6691	0.7726	0.6950	0.7616	0.5638

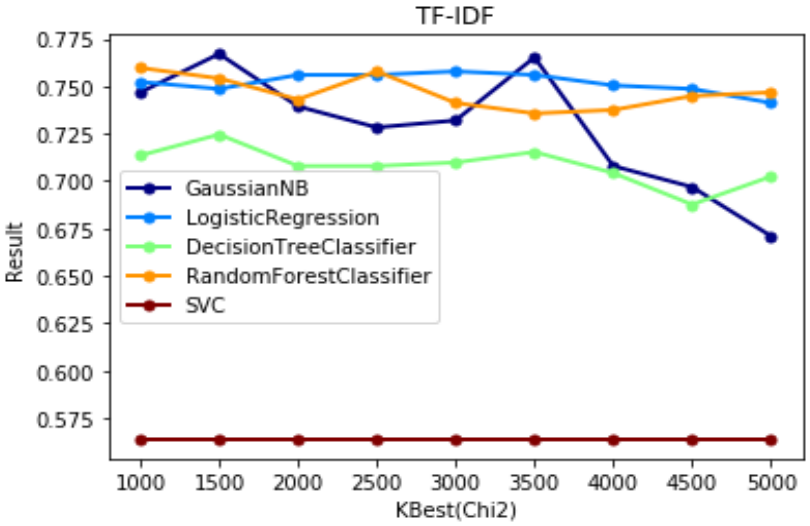

Figure 4: TFIDF with Chi-Square on MLT

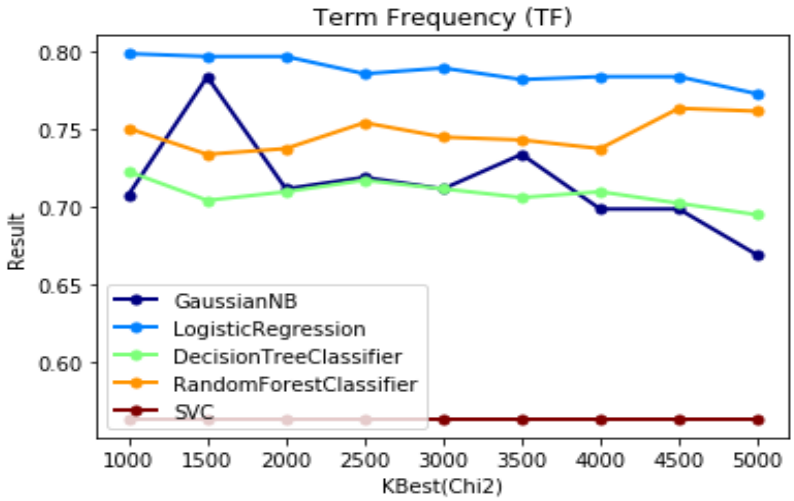

Figure 5: TF with Chi-Square on MLT

Table 1 and 2 with figure 4 and 5 illustrate the comparison of classification accuracy of Gaussian Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), Logistic Regression (LR) with the Chi-Square optimization algorithm in relation to the feature extraction (TFIDF and TF)to the classification accuracy and the number of tweet inUni-grams.The findings indicate that the Logistic Regression classification method with the given tweet data is the best (min 7412 andmax 0.7579) for *tfidf* and (min 7726 and max 0.7985) for *tf* withclassification accuracy in comparison to the analyzed classifiers. Support Vector Machine sentiment classification method performed poorly and is stable as the values of average classification accuracy are constantin comparison to other methods while Random Forest performed better than Decision Tree as an ensemble method. However, Random Forest achieved 0.01-0.02 higher average classification accuracy results in comparison to Naïve Bayes and Support Vector Machine, but the difference is not statistically significant. Also, tfidfis used to generate the most frequent terms in the bag of words as in figure 6 such as herdsmen, attack, community, machete, security, guard, Nigeria etc.

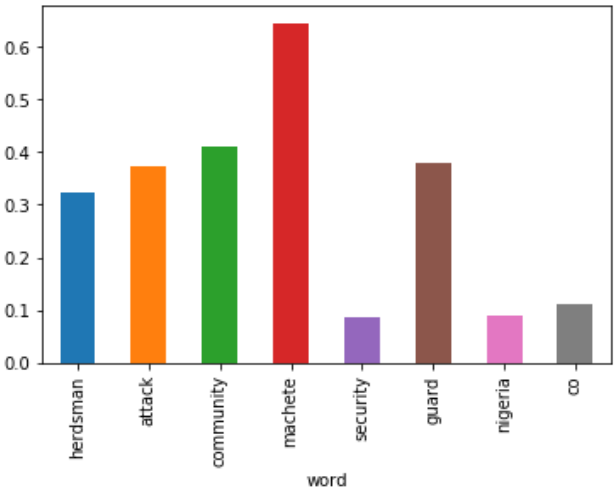

Figure 6: Tweet Most Frequent Words

CONCLUSION

The comparison of Gaussian Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), Logistic Regression (LR methods for sentiment analysis classification is presented in this paper. The experimental results have shown that the Logistic Regression classification method for sentiment analysis performed higher average of classification accuracy than other method, but the difference is not statistically significant. However, it shows that the most frequent terms in the tweet is Machete, this shows that machete is mostly used in the attack in the Nigeria security for the attacks.
The investigation indicates that increasing the size of the training data set from 1000 to 5000 leads to insignificant growth of the classification for the classifiers. These results show that a training set size of 2500 per sentiment class is sufficient for all analysed classification accuracy related to the *uni*-gram features. More work can be done including the stopwords and using bi-gram and n-gram words with the classification.

REFERENCES

[A+11] **F. Abel, Q. Gao, G. J. Houben, K. Tao** -*Semantic enrichment of Twitter posts for user profile construction on the social web*. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, Springer, Heidelberg,vol. 6644, pp. 375–389, 2011.

[A+12] **S. Amer-Yahia, S. Anjum, A. Ghenai, A. Siddique, S. Abbar, S. Madden, A. Marcus, M. El-Haddad** - *Maqsa: A system for social analytics on news.* In Proceedings of the 2012 ACM SIGMOD International Conference, 2012.

[A+18] **H. Ali, M.I. Sana, K. Ahmad, S. Shahaboddin**-*Machine Learning-Based Sentiment Analysis for Twitter Accounts*. Mathematical and Computational Applications. 23(11). 1-15, 2018.

[BES10] **S. Baccianella,A. Esuli, F. Sebastiani** -*SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA). 2010.

[BNG11] **L.Becker, M. Naaman, L. Gravano**-*Beyond trending topics: Real-world event identification on twitter*. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), pp. 1–17, 2011.

[**DTF14]** **M. Dadvar, D. Trieschnigg,F. Jong**-*Experts and machines against bullies: A hybrid approach to detect cyberbullies*. In Canadian AI, 2014.

[EBG11] **P. Earle, D. Bowden, M. Guy**-*Twitter earthquake detection: earthquake monitoring in a social world*. Annals of Geophysics 55: pp 708–713, 2011.

[GME17] **D. Gonzalez-Marron, D. Mejia-Guzman, A. Enciso-Gonzalez**-*Exploiting Data of the Twitter Social Network Using Sentiment Analysis*. In: Sucar E., Mayora O., Munoz de Cote E. (eds) Applications for Future Internet. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 179. Springer, Cham, 2017.

[HG14] **C. J. Hutto, E. E. Gilbert** -*VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Eighth International Conference on Welogs and Social Media (ICWSM-14) Ann Arbor, MI, June 2014.

[KU16] **N. Kausikaa, V. Uma** - *Sentiment Analysis of English and Tamil Tweets using Path Length Similarity based Word Sense Disambiguation.* Journal of Computer Engineering (IOSR-JCE) Vol 18 (3). I. pp 82-89, 2016.

[KKK13] **M. F. Khan, A. Khan,K. Khan**-*Sentence Level Sentiment Classification of Online Reviews*. Science International (Lahore),25(4) pp 937-943, 2013.

[KPA08] **B. Krishnamurthy, P. Gill, M. Arlitt**-*A few chirps about Twitter*. In Proceedings of the First Workshop on Online Social Networks, WOSN '08, ACM, New York, NY, pp. 19–24, 2008.

[LD13] **P. W Liang, B. R. Dai**- *Opinion Mining on Social Media Data. Mobile Data Management (MDM)*, 2013 IEEE 14th International Conference on. Vol. 2. IEEE, 2013.

[MM13] **R. McCreadie, C. Macdonald** - *Scalable distributed event detection for Twitter*. In: 2013 IEEE International Conference on Big Data, IEEE pp. 6–9, 2013.

[MAM17] **B. Marouane, B. Abderrahim, E. Mohammed**-*Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks*. The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017), Procedia Computer Science 113 (2017) pp 65–72., 2017.

[MHK14] **W. Medhat, A. Hassan H. Korashy**- *Sentiment analysis algorithms and applications: A survey*. Ain Shams Eng. J. 5, 1093–1113, 2014.

[M+08] **S. Milstein,A. Chowdhury, G. Hochmuth, B. Lorica,R. Magoulas**- *Twitter and the micro-messaging revolution:Communication, Connections and immediacy140 characters at a time*. O'Reilly Media, 2008.

[PP10] **A. Pak,P. Paroubek**- *Twitter as a corpus for sentiment analysis and opinion mining*. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[PT09] **R. Prabowo, M. Thelwall** -*Sentiment analysis: A combined approach*. J. Informetr. 3. pp143–157., 2009.

[SCS17] **J. Shashank, M. Calvin, S. Sumathy** - *Sentiment analysis of feature ranking methods for classification accuracy*. IOP Conf. Series: Materials Science and Engineering 263 (2017), pp 2-6, 2017.

[SYR17] **S. Saha, J. Yadav, P. Ranjan**- *Proposed approach for sarcasm detection in twitter*. Indian Journal of Science Technology, 10. 2017.

[S+11] **M. Speriosu, N. Sudan, S. Upadhyay, J. Baldridge**-*Twitter polarity classification with label propagation over lexical links and the follower graph*. Proceedings of the First Workshop on Unsupervised Learning in NLP. Edinburgh, Scotland, Association for Computational Linguistics: pp 53-63, 2011.

[The13] **M. Thelwall**- *Heart and soul: sentiment strength detection in the social web with SentiStrength*. In: Proceedings of the CyberEmotions, pp. 1–14, 2013.

[T+11] **M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede** - *Lexicon-based methods for sentiment analysis*. Comput. Linguist 37. pp 267–307. 2011.

[V+15] **C. Van-Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, V. Hoste** - *Automatic detection and prevention of cyberbullying*. In Human and Social Analytics, 2015.

[WAH14] **M. Walaa, H. Ahmed,K. Hoda** - *Sentiment analysis algorithms and applications: A survey*,Ain Shanz Engineering Journal,5(4), pp 1093-1113, 2014.

[WGB12] **X. Wang, M. S. Gerber,D. E. Brown**- *Automatic crime prediction using events extracted from Twitter posts*. In Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction, SBP'12. Springer Verlag: Berlin, Heidelberg, pp. 231–238, 2012.

[WZL11] **Y. Q. Wu, X. H. Zhang,W. Lide** - *Structural opinion mining for graph-based sentiment representation*. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-2011), 2011.

[Y+12] **D. Yang, H. Zheng, J. Yan, Y. Jin** - *Semantic Social Network Analysis with Text Corpora*. In: Tan PN., Chawla S., Ho C.K., Bailey J. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2012. Lecture Notes in Computer Science, vol 7301. Springer, Berlin, Heidelberg, 2012.