IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

Swahili Text Classification using Support Vector Machine and Feature Selection to Enhance Opinion Analysis in Kenyan Universities

Peter B. Obiria[1], Phyllis Ngigi[2], Grace Machuke[3]

*Department of Computer Science, Kiriri Women's University of Science and Technology, Nairobi, Kenya[1]

Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya[2]

Department of Mathematics, Kiriri Women's University of Science and Technology, Nairobi, Kenya[3]

Abstract: In Kenya's social media, Nairobi Swahili is the norm for communication in institutions of higher learning. Extant studies dwells on standard Swahili, affording limited text classification literature for Nairobi Swahili Natural Language Processing. The research explores how social media experience provides new ways for interactions, resulting into new challenges in managing student concerns that now require new knowledge for decision-making. Students have taken advantage of social media platforms by creating virtual discussion forums, which are quickly becoming repositories of collective knowledge. Unfortunately, institutions of higher are not able to utilize the collected knowledge through these platforms. The focus of this research is to ensure knowledge generated via social media is useful through opinion mining to enable extraction, classification and storage to support decision-making. Different algorithms were tested utilizing data from popular social media; operated by students in Kenyan universities. The results showed that SVM gives the best results when used with Linear Kernels and better performance on TF-IDF with N-grams methods. An analysis on the different SVM kernel showed linear kernel to have a better performance at 80% compared to Polynomial kernel and Radial Basis Function kernels, which both stand at 57%. To choose the best feature selection method for use along with linear SVM, TF and TF-IDF were tested. TF-IDF performed better with N-grams at 83%; rendering this research both theoretical and practical significance. The research would provide fast hand information for decision support in Kenyan higher learning institutions using text-mining tools in social media.

Keywords: Nairobi Swahili, Support Vector Machines, Feature Selection, N-grams

I. INTRODUCTION

The technological explosion in the recent past has led to an upsurge of scientific innovations; culminating to new specialties in communication [1]. With web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media, consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinions, positive or negative, regarding any product or service. These consumer voices can wield enormous influence in shaping the opinions of others and, ultimately, their brand loyalties, purchase decisions, and their own brand advocacy. Businesses can now respond to the consumer insights generated through social media monitoring and analysis by modifying their marketing messages, brand positioning, and product development [2].

Similarly, the social media experience has provided new ways for interactions amongst stakeholders in institutions of higher learning. This has resulted in new challenges in managing student issues and hence requiring new ways to gain new knowledge for decision-making. Students have taken advantage of social media platforms by creating virtual discussion forums, which are quickly becoming repositories of collective knowledge. Unfortunately, institutions of higher are not able to utilize the collected knowledge through these platforms. The focus of this research is to ensure knowledge generated via social media is useful through opinion mining, a branch of text mining to enable extraction, classification and storage of the classified comments to support decision-making.

In Kenya, Nairobi Swahili also known as Sheng or slang is the norm in social media. While extant research are based on standard Swahili (Swahili sanifu), there is limited availability of Natural Language Processing (NLP) resources for text classification of Nairobi Swahili which is the most commonly used in the social media platforms. Consequently, there is no large-scale Sheng Facebook and Twitter corpus annotated for text classification available in public domain. Sheng sentences require a clear identification. The discussion in the social media is often in informal words that are not in the current Sheng dictionary. There is use of informal words borrowed from diverse tribes in Kenya and East Africa

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

as a whole. For instance, iko shida, iko ngori, iko noma, are phrases with the same meaning, even though they appear different. Therefore, it is only by labelling and annotating of these kind of data for text classification that can make them useful.

Slang plays a significant role in international economy as it has been a language that multi-national interest groups focus on and analysts endeavours to decode on subjects matter like market analysis of consumer goods, governments' policies etc., arising from this section of the domain. The conversation being in the Sheng language, there is a great need for natural language analysis of this bulky amounts of Sheng language text and documents to back the necessary opinion analysis.

The aim of this paper is to explore the possibility of text mining for Nairobi Swahili in social media, for decision support in Kenyan higher learning institutions. This research will attempt to use text-mining tools to mine social media, a missing link that the research addresses.

II. LITERATURE REVIEW

This section describes the review of literature to lend the research both a theoretical and practical significance. It commences with a review of related work, techniques or opinion analysis, Machine Learning and Kernel Methods. It also expounds on feature selection technique, N-grams among others.

A. Related Work

Extant studies includes works developed to monitor online dialogues, focusing into online conversations and views of people [5]. Analysis of these online reviews offers a means to better comprehend the frequent concerns and of importance to address. The Umati system is a research based on hate speech monitoring whose first stage was to scrutinize major languages in Kenya that included Kiswahili, Luo, Kalenjin, Luhya, Kikuyu and English. The system defined remarkable discourse in the Kenyan context as any speech that remarks crowds founded on tribe, religion, nationality, sexual alignment, gender, disability, influential people especially the politicians, cities/regions, and socio-economic groups. The key goal of the study was to develop a framework for opinion mining on social media to back and support decision-making. According to [10] the Umati model was inspired by the aftermath of 2007 Post Election Violence where the disputed elections made youths to fuel violence through social media.

Additionally, an experiment on data-driven part-of-speech taggers trained and evaluated on the annotated Helsinki Corpus of Swahili was been done [9]. The authors selected four of the current state of the art data driven taggers, TnT (Trigrams'n'Tags), MBT (Memory Based Learning), SVMTool (Support Vector Machines) and MXPOST (Maximum Entropy Modeling), and observed SVM as the most precise tagger for the dataset. Moreover, a method to implement sentiment classification founded on an unsupervised linguistic method was been presented [9].The authors used SentiWordNet to compute general sentiment score of each sentence.

Accordingly, [8] recommended the use of many features to advance a trained classifier of Twitter messages. the features extends the feature vector of unigram model by the theories mined from DBpedia, the verb sets and the related adjectives mined from WordNet, the Senti-features mined by SentiWordNet and some valuable domain particular features. These authors built a wordlist for sentiment icons, abbreviation and slang words in tweets, which is beneficial before outspreading the tweets with different features.

In an experimental study presented by [7] on Sentiment categorization on Chinese language documents. Four feature selection methods mutual information (MI), information gain (IG), chi-squared statistic (CHI) and document (DF) and five learning approaches which included Centroid frequency Classifier method, K-Nearest Neighbor method, Winnow Classifier method, Naive Bayes method and SVM method which were tested on a Chinese language sentiment corpus which had a size of 1021 documents. The outcomes of the experiments exhibited that IG attains the best in the selection of sentiments terms and SVM depicted to be the best in performance for the Sentiment Classification.

Accordingly, [7] used dissimilar machine learning algorithms to analyze the polarity of reviews in the Arabic criticisms, obtained from particular Websites linked to films and the movies. The two interpreted the sentiments Corpus for Arabic corpus into English, which made the English Version of Opinion Corpus in Arabic also known as EVOCA corpus. The outcome showed low accuracy due to translation.

B. Techniques for Opinion Mining

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

There are two methods for Sentiment mining or analysis. The first one being the supervised learning technique, based on machine learning classifiers. It uses training on labelled data or information before being used to the real sentiment or opinion classification chore. Naïve Bayesian, Support Vector Machines (SVM), maximum entropy and others are some of the existing supervised learning methods that can also be used to sentiment or opinion classification [7].
The other method is on lexicon, which is also known as unsupervised learning method. It does classification centered on some fixed syntactic outlines to express sentiments or opinions. It classifies the document by semantic orientation or opinion dictionary for calculating sentiment polarity of a text. Semantic Orientation - Pointwise Mutual Information-Information Retrieval algorithm (SOPMI-IR) is an example of an unsupervised method, which uses the common existence frequency of particular words to compute the opinion or sentiment Polarity [7].

C. Machine Learning

In a machine learning based classification, two sets of documents are required: training and a test set. An automatic classifier to learn the differentiating characteristics of documents uses a training set, and a test set is for validating the performance of the automatic classifier. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in text categorization. The other well-known machine learning methods in the natural language processing area are Random Forest and Logistic regression

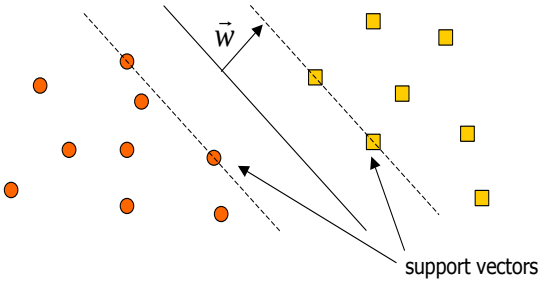

Figure 1. Illustrating Support Vector

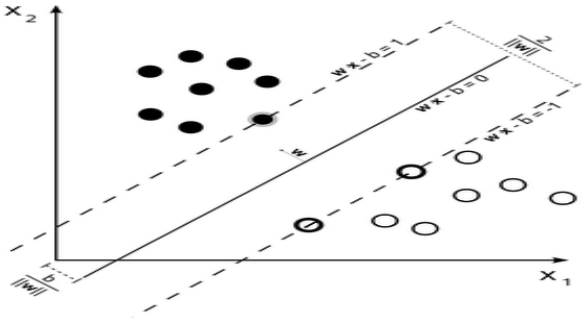

Figure 2. Linear Support Vector Machine

D. Support Vector Machine

This is a supervised machine-learning algorithm used for both classification problems and regression problems. In SVM, data element is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature as the value of a specific coordinate. SVM produces a mapping of input-output functions from an established labelled training data [8]. According to [6] a Linear SVM, as a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. The formula for which is used for a linear SVM is output is $u = \vec{w} \cdot \vec{x} - b,$ where $\vec{w}$ is the normal vector to the hyperplane, and $\vec{x}$ is the input vector.
The margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples. Maximizing the margin is as an optimization problem:
Minimize $\frac{1}{2}\|\vec{w}\|^2$ subject to $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \forall i$ where x_i is the ith training example and yi is the correct output of the SVM for the i[th] training example as shown in Figure. 1. (1)

SVM attempts to find a hyperplane , which splits the data in two classes as optimally as possible i.e. meaning that as much points as possible of label A should be separated to one side of the hyperplane and as much points of label B to the other side, while maximizing the distance of each point to this hyperplane [6]

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

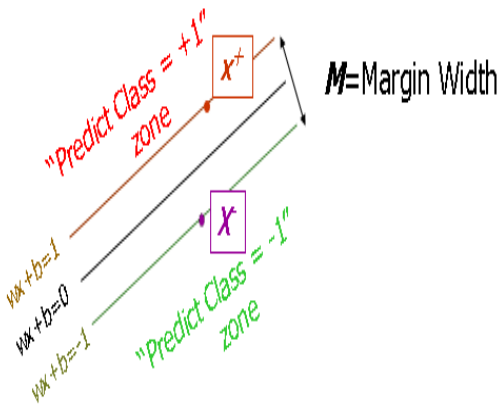

Figure 3. Illustrating the SVC classifier margin with the yellow line denoting the maximum margin

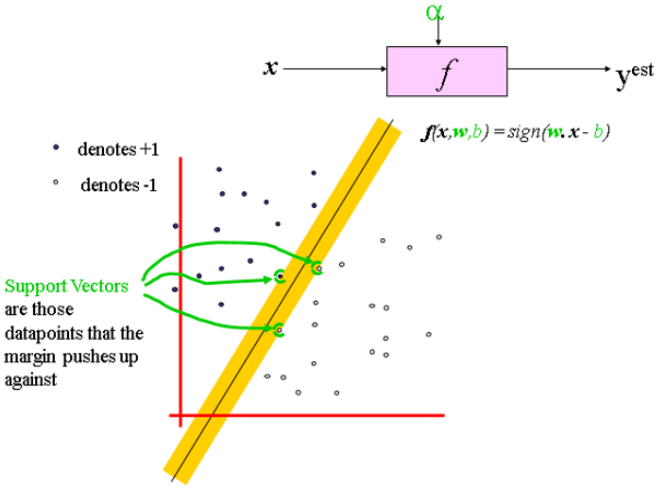

Figure 4. Specifying a line and a margin mathematically in m input dimensions

Considering that:

$\boldsymbol{w} \cdot \boldsymbol{x}^{+} + b = +1$ (2)

$\boldsymbol{w} \cdot \boldsymbol{x}^{-} + b = -1$ (3)

$\boldsymbol{w} \cdot (\boldsymbol{x}^{+} - \boldsymbol{x}^{-}) = 2$ (4)

$$M = \frac{(x^{+} - x^{-}) \cdot w}{|w|} = \frac{2}{|w|} \quad (5)$$

In addition, the first goal is, to classify all training data as shown in the formulas below

$wx_i + b \geq 1 \quad if\ y_i = +1$

$wx_i + b \leq 1 \quad if\ y_i = -1$

$y_i(wx_i + b) \geq 1 \quad$ for all i (8)

$$M = \frac{2}{|w|}$$

Then the second goal is to maximize the Margin

Which is the same as to minimize $\frac{1}{2} w^t w$

We now can formulate a Quadratic Optimization Problem and solve for w and b

Minimize $\Phi(w) = \frac{1}{2} w^t w$ (9)

Subject to $y_i(wx_i + b) \geq 1 \quad \forall i$ (10)

Now solving the optimization problem

Find w and b such that

$\boldsymbol{\Phi(w)} = \frac{1}{2}\, \mathbf{w}^{T}\mathbf{w}$ is minimized;

And for all $\{(\mathbf{x_i}, \mathbf{yi})\}$: $y_i\,(\mathbf{w}^{T}\mathbf{x_i} + b) \geq 1$ (11)

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

Need to optimize a quadratic function subject to linear constraints. Quadratic optimization problems are a well-known class of mathematical programming problems, and many (rather intricate) algorithms exist for solving them. The solution involves constructing a dual problem where a Lagrange multiplier αi is associated with every constraint in the primary problem:

Find α1…αN such that
$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}$ is maximized and

(1) $\Sigma\alpha_i y_i = 0$

(2) $\alpha i \geq 0$ for all αi (12)

The Optimization Problem Solution:
$\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x_i}$ $b = y_k - \mathbf{w}^T\mathbf{x_k}$ for any $\mathbf{x_k}$ such that $\alpha_k \neq 0$

Each non-zero αi indicates that corresponding xi is a support vector. Then the classifying function will have the form:
$f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x_i}^T\mathbf{x} + b$

Notice that it relies on an inner product between the test point x and the support vectors xi – we will return to this later. Also, keep in mind that solving the optimization problem involved computing the inner products xiTxj between all pairs of training points.

According to [11] SVM is an incomplete case of kernel-based approaches. It maps feature vectors into a higher-dimensional space by using a kernel function and constructs a best and optimal linear discriminating function in an optimal hyper-plane fitting into the training data. The solution is the best or optimal, as the margin amongst the separating hyper-plane and the nearest feature vectors from both classes is highest in a two classes' problem. The feature vectors close to the hyper-plane are the support vectors, and the location of the other vectors does not affect the hyperspace or the hyper-plane, which is the decision function. By finding the hyper-plane that differentiates the two classes very well, the classification is performed [13]. Used Support Vector Machine (SVM), Naive Bayes and Maximum Entropy Classification to implement classification of sentiments. This was on movie review records. In the research, SVM did the best while combine with unigram.

The experiments by [12] on movie review data from Epinions.com showed that crossbreed SVMs which combine with unigram features with those built on real-values to be favored by measures acquired higher performance in the results.

The researcher converted each training section to an actual vector, xi that comprises of a set of substantial features signifying the related document, di. Therefore, Tr+ = Pn i=1(xi, +1) for the positive sample set and Tr− = Pn i=1(xi, −1) and the negative sample set, [13].

SVM has been successful used in much text classification and this is due to their many advantages that includes robustness especially in the high dimensional spaces and when samples sets are scarce; and the fact that most text categorizations problems are linearly separable, SVM have attained worthy results [11].

E. Kernel Methods

The researcher then plotted the training vectors (xi) into a higher dimensional space by the function φ in the Kernel functions. SVM finds a linear splitting hyperplane with the utmost margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. Furthermore, K (xi, xj) ≡ φ (xi) T φ (xj), [11] .The following are the four basic kernels use in SVM:

Linear: $K(xi, xj) = x\ T\ i\ xj$

Polynomial Kernel: $K(x_i, x_j) = [(x_i . x_j + 1)]^q$

Radial Basis Kernel: $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2\alpha^2}\right)$ (13)

Sigmoid Kernel: $K(x_i, x_j) = \tanh(v(x_i, x_j) + C)$

F. Feature Selection

The objective of feature selection method is to discover the best important features for the classification job and to eliminate unrelated, blaring and redundant information data, [17] Feature selection has as another goal of the reduction

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

of the dimensionality of feature space and time processing. Precise feature selection is quite crucial for the performance of any classifier.

G. Term Frequency and Term Frequency Inverse Document

This helps in eliminating the words that happen repeatedly in a corpus, and remove terms that appear infrequently. These terms are not distinctive between documents. The aim of term frequency is to alter the text to a vector representation. The weight of the word is then determined with the reverence to the document having that word. There are numerous weighting structures namely Term Frequency (TF) weighting, Inverse Document Frequency (IDF) weighting, Boolean weighting and Term Frequency Inverse Document Frequency (TFIDF), [16].
The Boolean weighting (presence) is simple and the weight of the word is one (1) if a word it occurs in the document and (0) if otherwise. TF (term frequency) calculates the raw frequency of a term in a document – this is the number of times that a term t appears in the document. The Inverse Document Frequency (IDF) shows how much information the word provides; this is whether the word is common or infrequent across all documents. To calculate IDF, the number of documents containing the term; and then taking the logarithm of that fraction [16] divides the total number of documents.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document)

IDF (t) = log_e (Total number of documents / Number of documents with term t in it).

Value = TF * IDF (14)

Example of TF-IDF

A document has 100 words and the word Nokia appears 3 times. The term frequency (i.e., tf) for Nokia is then (3 / 100) = 0.03. We assume we have 10 million documents and the word Nokia appears in one thousands of these. The researcher then calculated the inverse document frequency (idf) using this formula:

Log (10,000,000 / 1,000) = 4

Thus, the Tf-idf weight is the product of these quantities:

0.03 * 4 = 0.12. (15)

H. N-grams

According to [14]. N-grams perform better than the use of light stemming. This is because they easily match stems of words and that they are free from a pre-set vocabulary like in morphological analysis. N-grams that include prefixes and suffixes appear more frequently than n-grams that contain stems, and hence the use of inverse document, the IDF would automatically demote the weight of n-grams that have prefixes and suffixes and promote the weight of n-grams that include stems.

I. Bag of words

To achieve machine learning on text documents, there is need of turning the text information into numerical feature vectors. The Bag of Words Counts the frequency of each word or pair of consecutive words in each document. It helps change the raw data in text into numerical feature vectors with a permanent size [15].

III. METHODOLOGY

In this section, the methodology used for collecting Facebook updates and tweets from twitter are described in forming a relevant subset selection appropriate for machine learning.

A. Collecting comments from social media sites

The first step in this methodology was to collect a large set of comments from social media sites. For twitter we used a library called Twython, a python package API for Twitter. For Facebook we used FacePy, a python package API for

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

Facebook. A total number of 2537 documents were collected. The data were then loaded from the disk into memory using Pandas; this helped in data exploration, performing summary statistics and creation of the vector from text messages.

B. Testing

The model accuracy has been used to measures the percentage of inputs in the test set that the model classifier has correctly labelled. The precision is the number of true positives in the class (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class). Recall is defined as the number of true positives divided by the total number of elements that belong to the positive class (i.e. the sum of true positives and false negatives, which are items, not labelled as belonging to the positive class but should have been). The analytic formulas for these measures are:

Precision = TP TP+FP
Recall = TP TP+FN
Where TP stands for True Positives, FP as False Positives and FN as False Negatives
F- Score is a Harmonic mean of Precision and Recall
F-Score =2 (P*R) / (P + R) (16)

C. Corpus

The comments were collected using ptywthon and facepy from the social media, and 2587 posts from the higher learning institutions domain, 1811 were Negative, 338 were Neutral while 386 were positive as shown in table 1.

TABLE 1
DESCRIPTION OF CORPUS USED IN THE EXPERIMENTS

	Number of words	**% of words**
Inciting	386	15.2
Non-Inciting	1811	71.4
Neutral	338	13.3
Total	2535	

o = negative/Not inciting; 1=Neutral; 2=Positive / inciting. The opinion classifier needed a comments list, initially the data had three classes which were Neutral, Negative and positive comments. Since the main objective of the model was to assist, the management of higher learning institutions avert crisis. The negative class of 1811 comments sampled to 500 to avoid biasness. (Table 2).

TABLE 2
DESCRIPTION OF COMMENTS USED IN PROPOSED LEXICON-BASED CLASSIFIER

	Number of words	**% of words**
Inciting	386	43.6
Non-Inciting	500	56.4
Total	886	100

D. Pre-processing

This stage involved data collection, data cleaning and filtering using in-built Natural Language Toolkit stop word removal.The following is the list of custom words removed.

'babu','owino','owino','sonu','mbithi','mbeche',
'jkuat','uon','university of nairobi','nairobi',
'ku','kenyatta university','raila',,''professor',
'the university of nairobi','odm','jap','raila',
'odinga','waiguru','likes','university','kiriri'
'magoha' 'jacobs','jacob','mike' ,'kwust',

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

After that, TFIDF (Term Frequency–Inverse Document Frequency) was done on vector representation for the terms from their textual representations was obtained by performing weight and TF (term frequency) which is a well-known weight presentation of terms often used in text mining [16].

E. Calculating a relevancy score for each bi-gram

The researcher used n-grams range (1, 3) in addition to TF and TFIDF, calculated for each of n-gram ranges. Moreover, a k-fold cross-validation, the dataset is partitioned into k subsets, performing the classification on one subset (the training set), and validating the model on the rest (k-1) subsets (called the validation set or testing set). This operation is repeated k times for every subset. Then validation results averaged over the k iterations. In addition, in order to evaluate the classification techniques, the researcher calculated the most widely used performance measures in the classification task; precision, recall, F measure, and accuracy. The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative (Godbole & Sarawagi, 2004).

The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.The F-beta score weights recall more than precision by a factor of beta. Beta == 1.0 means recall and precision are equally important.

	message	label	clean_message
1568	Kuna mathe hapa ananiangalia kwa macho ya huru...	0	kuna mathe hapa ananiangalia kwa macho ya huru...
751	@WKabeo Congratulations comrade	0	congratulations comrade
1142	RT @Tsharz: SMH INFERTILITY.RT @Otuambala: SMH...	0	smh infertility.rt smh need birth control meas...
768	RT @PewaAbagenge: Hahahaha "#Someonetellsoutha...	0	hahahaha police vs police
866	RT @bocholah: @UoN_Comrades conngratulation al...	0	conngratulation elected.parklands conducted fr...
1681	Things are not good at hall 9!!! Mike Jacobs h...	2	things good hall 9!!! jacobs refused part stri...
1388	Comrades must cook since there is no roasted u...	2	comrades must cook since roasted uji cafeteria...
1566	this issue is simple, they either release babu...	2	issue simple, either release comrades release
424	RT @arabellakiki: UON students now outside Haz...	2	students outside hazina towers.more tear gas
1827	We will not rest. We are tired of dictatorship...	2	rest. tired dictatorship mediocrity

Figure 5. A sample of uncleansed and cleaned comments

IV. RESULT AND DISCUSSION

Table3 shows the performance of different algorithms on the classification of text. From the analysis, based on precision measure, Random forest performs better at 78%. Based on recall measure, Naives Bayes performs better at 97%. Based on F-score, Linear SVM is the best at 74%. Lastly based on accuracy measure, the best algorithm is the Linear SVM at 83%.

F Score or the F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0. Hence a better measurement metrics than precision and recall. For this research accuracy and F score are the two performance metrics measurements are used measure.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

TABLE 3:
THE PERFORMANCE OF DIFFERENT ALGORITHMS ON THE CLASSIFICATION OF TEXT

Algorithms	Precision (%)	Recall (%)	F-Score (%)	Accuracy (%)
Linear SVM	77	71	74	83
Random forest	78	56	65	79
Naïve Bayes	41	97	58	79
Logistic regression	50	92	65	81

TABLE 4:
ACCURACY OF KERNEL METHODS WITHOUT N-GRAMS

Kernel	Min Range (*100%)	Max Range (*100%)
Linear	0.78	0.80
Poly	**0.57**	**0.57**
RBF	0.57	0.57

The table 4 above indicates the different SVM kernels that used to classify text without N-grams. For the analysis using python, Linear SVM performed the best at 80%.

TABLE 5:
ACCURACY OF TF AND TF-IDF WITHOUT N-GRAMS

Feature selection method	Accuracy
TF	0.69
TF-IDF	0.79

The researcher tested Linear SVM on TF and TF-IDF feature selection methods without n-grams. From the analysis, TF-IDF showed better results with an accuracy of 79% in comparison to TF (69%) as shown on table 5.

TABLE 6:
ACCURACY OF TF AND TF-IDF WITH N-GRAMS

Feature selection method	Accuracy
TF	0.68
TF-IDF	0.83

The findings also exhibited that TF-IDF had better results with an accuracy of 83%% in comparison to TF (68%) as shown table 6. In addition, TF-IDF had better performance with or without N-grams. However, TF-IDF performed better when combined with N-gram method. TF and TF-IDF features, Precision, Recall and F score were methods used to measure the quality of the classification. Table7 and table 8 below indicate the output of the analysis.

TABLE 7:
TF-IDF REPORT

label	Precision	Recall	F score
0	0.81	0.88	0.85
2	0.84	0.75	0.82

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 5, Issue 1, January 2018

TABLE 8:
TF REPORT

label	Precision	Recall	F1-score
0	0.62	0.99	0.77
2	0.97	0.29	0.44

The results show that TF-IDF with N-grams for non-inciting comment (0) performs better with F score of 85% while that of TF is at 77%. For inciting text (2), TF-IDF performs better with an F score of 82% while that of TF is 44%.

V. CONCLUSION

This research presented a model for unconventional language text classification. The research found through the analysis of Random Forest, Logistic Regression, Naives Bayes and SVM, that SVM performance was highest with 83% compared to Logistic Regression at 81%, Random Forest at 79% and Naives Bayes at 79%. An analysis on the best SVM kernel showed linear kernel to have a better performance at 80% compared to Poly kernel and RBF kernels, which both stand at 57%. To choose the best feature selection method for use along with linear SVM, TF and TF-IDF were tested. TF-IDF performed better with N-grams at 83%.

REFERENCES

[1] P. B. Obiria and M. W. Kimwele, "A location-based privacy-preserving m-learning model to enhance distance education in Kenya," *Journal of Computers in Education*, pp. 1-23, 2017.

[2] J. Zabin and . A. Jefferies, "Social media monitoring and nalysis: Generating consumer insights from online conversation.," *Aberdeen Group Benchmark Report*, vol. 39, no. 9, 2008.

[3] Rushdi-Saleh and Mohamed, Bilingual experiments with Atrabic English corpus for opinion mining, 2011.

[4] E. Amberber, "Umati: Kenyan platform to fight online hate speech with NLP and Machine Learning in Africa," yourstory.com, 2014. [Online]. Available: https://yourstory.com/2014/09/umati-hate-speech/. [Accessed 2017].

[5] N. Sambuli,, A. Crandall, P. Costello and C. Orwa, "Viability, Verification, Validity:3Vs of Crowdsourcing TESTED IN ELECTION-BASED CROWDSOURCING," iHub Research , Nairobi, 2013.

[6] V. Vapnik, "The Nature of Statistical Learning Theory," *Data mining and knowledge discovery*, 1995.

[7] S. Tan and . J. Zhang, "An empirical study of sentiment analysis for chinese documents.," *Expert Systems with Applications, 34(4),* p. 2622–2629., 2008 .

[8] Q. Béchet, , . A. Shilton and B. Guieysse, "Modeling the effects of light and temperature on algae growth: state of the art and critical assessment for productivity prediction during outdoor cultivation.," *Biotechnology advances, 31(8), ,* pp. 1648-1663., 2013.

[9] P. C. A. Odinga, "Use of new media during the kenya elections.," 2013.

[10] . G. De Pauw,, P. W. Wagacha and G. M. de Schryver, "The SAWA corpus: a parallel corpus English-Swahili.," in *In Proceedings of the First Workshop on Language Technologies for African Languages*, 2009.

[11] S. Kotsiantis, D. Kanellopoulos and . P. Pintelas, "(2006). Handling imbalanced datasets," *A review. GESTS International Transactions on Computer Science and Engineering,*, pp. 25-36., 30(1).

[12] T. Mullen and N. Collier, " Sentiment Analysis using Support Vector Machines with Diverse Information Sources.," *EMNLP, Vol. 4,*, pp. 412-418, 2004.

[13] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques.," in *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* , 2002.