Naïve Bayes and J48 Classification Algorithms on Swahili Tweets: Perfomance Evaluation

Hassan Seif
College of Informatics and Virtual Education
University of Dodoma
Dodoma, Tanzania

***Abstract*—The use of social media has grown significantly due to evolution of web 2.0 technologies. People can share the ideas, comments and posting any events. Twitter is among of those social media sites. It contains very short message created by registered users. Twitter has played the important parts in many events by sharing message posted by registered user. This study aims on evaluating performance of Naïve Bayes and J48 Classification algorithms on Swahili tweets. Swahili is among of the African language that is growing faster and is receiving a wide attention in web usage through social networks, blogs, portals etc. To the best of the researcher's knowledge; many studies have been conducted on other language for comparing classification algorithms, but no similar studies found on Swahili language. The data of this study was collected from the top ten most popular twitter accounts in Tanzania using Nodexl. These accounts were identified according to the number of followers. The extracted data were pre-processed in order to remove noise, incomplete data, outlier, inconsistent data, symbols etc. Further, the tweets contains words which are not in Swahili language were identified and removed and filtered by removing url links and twitter user names. The pre-processed data analysed on WEKA using Naïve Bayes and J48 classification algorithms. The algorithm then evaluated based on their accuracy, precision, recall and Receiver Operator Characteristic (ROC). It has been found that; Naïve Bayes classification algorithms perform better on Swahili tweets compared to J48 classification algorithm.**

Keywords-Social media; Swahili tweets; Naïve Bayes; J48

I. INTRODUCTION

Due to the evolution of web 2.0 technologies, now days the use of social media sites has grown significantly. People communicate, posting their comments and views through social media sites depending on their interest/opinions. It is estimated that there are over 900 social media sites on the internet with more popular platforms like Facebook, Twitter, LinkedIn, Google Plus, and YouTube [1].

Twitter is a popular and massive social networking site which has a large number of very short messages created by the registered users. It is estimated that; there are about more than 140 million active users who publish over 400 million 140-character "Tweets" every day [2]. The large speed and ease of publication of Twitter have made it as an important communication medium for people. Twitter has played a prominent role in socio-political events and also has been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy [2].

The data posted on twitter can be used for various research purposes. In context of data mining, there are two fundamental tasks that can be considered in conjunction with Twitter data: (a) graph mining based on analysis of the links amongst messages, and (b) text mining based on analysis of the messages' actual text [3]. Twitter graph mining based on analysis of the links amongst message can be applied in measuring user influence and dynamics of popularity, community discovery and formation and social information diffusion. On twitter text mining based on analysis of actual message, the number of task which can be performed includes; sentiment analysis, classification of tweets into categories, clustering of tweets and trending topic detection [3], this study is based on classification of tweets into categories; where by algorithms used was to be compared.

Swahili is among of the African language that is growing faster and is receiving a wide attention in web usage through social networks, blogs, portals etc. It is spoken in several countries found in Africa such as; Tanzania, Kenya, Uganda, Burundi, DRC Congo, Rwanda, Mozambique and Somalia; and has about 50 million speakers. There are four categories of African languages namely: Khoisan, Afro-Asiatic, Nilo-Saharan and Niger–Congo Kordofanian. Swahili belongs to the Niger- Congo group of languages specifically the Sabaki subgroup of Northeastern Coast Bantu languages [4].

To the best of the researcher's knowledge several studies has been conducted for comparing classification algorithms, but many of them are based in English and other languages. There are no similar studies on Swahili language. Furthermore, there are no set of corpus of Swahili tweets which are ready made publicity available for research purpose. For this reason it can be stated that, Swahili is among of the under-resourced language. The term "under-resourced language" refers to a language with some of (if not all) the following aspects: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc [5].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

This study is intended to compare the performance of Naïve Bayes and J48 Classification algorithm on Swahili tweets.

II. NAÏVE BAYES

Naïve Bayes is a simple classifier based on the Bayes theorem. It is a statistical classifier which performs probabilistic prediction. The classifier works by assuming that; the attribute are conditionally independent.

For Naïve Bayes classification, the following equation is used [6] ;

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \quad (1)$$

From equation (1) above, the classifier, or simple Bayesian classifier, work as follows;

(1) Let D be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, X= (X1, X2....., Xn) , depicting n measurements made on the tuple from n attributes, respectively, A1, A2....., An .

(2) Suppose that there are m classes, C1, C2.....Cm . Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the Naïve Bayesian classifier predicts that tuple X belongs to the class *Ci* if and only if $P(Ci \mid X) > P(Cj \mid X)$ for $1 \leq j \leq m; j \neq 1$. Thus we maximize $P(Ci \mid X)$. The class *Ci* for which $P(Ci \mid X)$ is maximized is called the maximum posteriori hypothesis.

(3) From equation (1), as P(X) is constant for all classes, only $P(X \mid Ci)P(Ci)$ need be maximized. Then predicts data item X belongs to class *Ci* if and only if has got the highest probability compared to other class label.

III. J48

J48 is one of the decision tree induction algorithm .It is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool [7]. This algorithm was developed by Ross Quinlan. C4.5 algorithm creates a decision tree which can be used for classification based the value which are presented on dataset. The following steps are used while the decision tree is constructed on J48 classification algorithm;

(1) In general the tree is constructed in a top-down recursive divide-and-conquer manner, at start, all the training examples are at the root, attributes are categorical (if continuous-valued, they are discretized in advance), examples are partitioned recursively based on selected attributes test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

(2) Conditions for stopping partitioning are as follows; all samples for a given node belong to the same class, there are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf, there are no samples left.

IV. RELATED WORKS

A review of literature from various scholars reveals that there are number of studies which were conducted for comparing several classification algorithms.

Goyal, A and Mehta, R [8] conduct a study on comparative evaluation of Naïve Bayes and J48 classification algorithms. The study was in the context of financial institute dataset with the aim of checking accuracy and cost analysis of these algorithms by maximizing true positive rate and minimizing false positive rate of defaulters using WEKA tool. The result showed that; the efficiency and accuracy of J48 and Naive Bayes is good [8].

Another study was conducted by Arora, R and Suman [9] which check comparative analysis on classification algorithms on different datasets using WEKA. The comparison was conducted on two algorithms; J48 and Multilayer Perceptron (MLP). The performance of these algorithms have been analysed so as to choose the better algorithm based on the conditions of the datasets. J48 is based on C4.5 decision based learning and MLP algorithm uses the multilayer feed forward neural network approach for classification of datasets. It has been found that; MLP has better performance than J48 algorithm.

Patil, Tina R and Sherekar, S S [10] did the study on comparing performance of J48 and Naïve Bayes classification algorithm based on bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool. The study found that; the efficiency and accuracy of J48 is better than that of Naïve Bayes.

Furthermore a comparative analysis of classification algorithms for students' college enrollment approval using data mining had been conducted using dataset from King Abdulaziz University database. In this study; the WEKA knowledge analysis tool is used for simulation of practical measurements. The classification technique that has the potential to significantly improve the performance is suggested for use in colleges' admission and enrollment applications. It has been found that; C4.5, PART and Random Forest algorithms give the highest performance and accuracy with lowest errors while IBK-E and IBK-M algorithms give high errors and low accuracy [11].

V. METHODOLOGY

A. Data Set Collection

The dataset of this study was collected from the top ten most popular twitter accounts in Tanzania using Nodexl. These accounts were identified according to their number of followers as presented in socialbakers sites [12]. Hot topics with their comments were identified and extracted using Nodexl

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

software. The collected data were stored in a CSV format for easy to be analysed in WEKA software.

B. Data Preprocessing

This is one of the most important steps in data mining. Since no quality data, no quality mining result. Some of data preprocessing techniques are data cleaning, data integration, data transformation, data reduction and data discretization [6]. These techniques may be combined together as a stage of data preprocessing.

The data of this study were cleaned in order to remove noise data, incomplete data, outlier, inconsistent data, symbols etc. Also, the tweets contains words which are not in Swahili language were identified. The words further are filtered by removing url links, and twitter user names. Finally the words/tweets which are not in Swahili language were removed this is because there are some tweets which was found to be in mixed language (Swahili and English) and other in English only.

C. Data Analysis

The pre-processed data were analysed by using WEKA software. "WEKA" stands for the Waikato Environment for Knowledge Analysis, which was developed by the University of Waikato in New Zealand. It is open source software issued under the GNU General Public License. WEKA has a collection of machine learning algorithms for data mining task. It has techniques for data pre-processing, classification, regression, clustering, association rules, visualization etc. It is written in Java and runs on almost every platform. It is also well-suited for developing new machine learning schemes. The tool gathers a comprehensive set of data pre-processing tools, learning algorithms and evaluation methods, graphical user interfaces (incl. data visualization) and environment for comparing learning algorithms. WEKA is easy to use and to be applied at several different levels.

WEKA has been selected because the Naïve Bayes and J48 Classification algorithm are implemented in this tool. This would results in achieving the objective of the study which is to compare the performance of Naïve Bayes and J48 classification algorithm on Swahili tweets.

D. Model Evaluation

After analyzing the data on WEKA, each algorithm was compared on their performance. Performance evaluations were based on recall, precision, accuracy and ROC curve. The formula used for evaluating these algorithms based on the following confusion matrix as described in Table 1 ;

Table 1: Confussion Matrix

		Detected	
		Positive	Negative
Actual	Positive	A:True Positive	B:False Negative
	Negative	C:False Positive	D: True Negative

Recall/Sensitivity/True positive rate it is the proportion of positive cases that were correctly identified. Recall can be calculated using the following equation:

$$\text{Recall} = \frac{A}{A + B} \quad (2)$$

Precision/Confidence denotes the proportion of Predicted Positive cases that are correctly Real Positives. Equation (3) can be used in finding precision;

$$\text{Precision} = \frac{A}{A+C} \quad (3)$$

Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The true positives, true negatives, false positives, and false negatives are also useful in assessing the costs and benefits (or risks and gains) associated with a classification model [6]. The following equation can be used to calculate the accuracy of the classifier;

$$\text{Accuracy} = \frac{A + D}{A + B + C + D} \quad (4)$$

Receiver Operator Characteristic (ROC) curve is a graphical method for displaying the tradeoff between true positive rate and false positive rate of a classifier. True positive is plotted along the Y-axis and false positive is plotted along the X-axis. The ROC has got number of properties depending on the value of its area under the curve. The following describe the nature of prediction/classification based on the value of ROC curve area (A);

A= 1.0: perfect prediction
A= 0.9: excellent prediction
A= 0.8: good prediction
A= 0.7: mediocre prediction
A= 0.6: poor prediction
A= 0.5: random prediction
A= <0.5: something wrong

VI. RESULT AND DISCUSSION

Experiments were performed on Swahili tweets data set which was extracted by using Nodexl. The total number tweets on a data set were 276 with 5 attributes. These data analysed on WEKA tool by using Naïve Bayes and J48 Classification algorithm. Before the data set tested on classification algorithm, attribute subset selection measure were used in order to select the best attribute and removing all weak irrelevant attribute. Since the high dimension data will make testing and training of general classification methods to be difficult [13]. The Heuristic method used was stepwise forward selection (Best First) whereby the best of the original attribute is determined by added to the reduced set of attribute.

The accuracy of the selected algorithms (J48 and Naïve Bayes) was tested by cross validation method. In this method, 10-fold cross validation was used where by a data set were

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

randomly partitioned into 10 mutually exclusive folds each of approximately equal size. Training and testing is performed 10 times. For each iteration, one fold is selected as the test set and the remaining data in another nine folds used as a training set. The testing is repeated 10 times. The final accuracy of an algorithm will be the average of the 10 trials.

Table 2 shows the result of the experiment on Swahili tweets using Naïve Bayes and J48 classification algorithms on WEKA

Table 2: Experiment Results

Evaluation / Algorithm	Accuracy	Precision	Recall	ROC
Naïve Bayes	36.96%	0.294	0.37	0.525
J48	34.78%	0.121	0.348	0.461

It has been found that; Naïve Bayes classification algorithms perform better on Swahili tweets compared to J48 classification algorithm. The model were evaluated by using accuracy, precision, recall, and ROC; and it has been found that, Naïve Bayes has the highest accuracy (36.96%) compared to J48 classification algorithm (34.78%). This implies that; the total number of instances that are correctly classified by Naïve Bayes is larger than the total number of instances that are correctly classified by J48. Furthermore; Naïve Bayes has been found to be the best in terms Precision (0.294), Recall (0.37) and ROC (0.525) compared to J48 in terms of Precision (0.121), Recall (0.348) and ROC (0.461).

VII. CONCLUSION

The Naïve Bayes has been found to be the best classification algorithm on Swahili tweets data set compared to J48 classification algorithm in terms of accuracy, precision, recall and ROC. In general the performance of Naïve Bayes and J48 algorithm on Swahili tweets was very poor. This is because; the values of their evaluation measure (accuracy, precision, recall, and ROC) are very small.

More research should be conducted in order to identify the best algorithm which will give highest performance in terms of accuracy, precision, recall, ROC and other evaluation measures. Further research also should be conducted in order to find the way on how to increase the performance of both algorithm (Naïve Bayes and J48) in terms of accuracy, precision, recall, ROC and other evaluation methods.

REFERENCES

[1] R. C. M. Jr and F. M, "Social Media Analytics : Data Mining Applied to Insurance Twitter Posts," *Casualty Actuar. Soc. E-Forum*, vol. 2, 2012.

[2] S. Kumar, F. Morstatter, and H. Liu, *Twitter Data Analytics*. Springer, 2013.

[3] A. Bifet and E. Frank, "Sentiment knowledge discovery in Twitter streaming data," in *Discovery Science*, 2010, pp. 1–15.

[4] S. Marjie-okyere, "Borrowings in Texts : A Case of Tanzanian Newspapers," *New Media Mass Commun.*, vol. 16, no. Marjie 2010, pp. 1–9, 2013.

[5] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic Speech Recognition for Under-Resourced Languages : A Survey."

[6] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second. San Francisco, CA: Morgan Kaufmann, 2006.

[7] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, 2014.

[8] A. Goyal and R. Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms," *Int. J. Appl. Eng. Res.*, vol. 7, no. 11, 2012.

[9] R. Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," *Int. J. Comput. Appl.*, vol. 54, no. 13, pp. 21–25, 2012.

[10] T. R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *Int. J. Comput. Sci. Appl. ISSN 0974-1011*, vol. 6, no. 2, pp. 256–261, 2013.

[11] A. H. M. Ragab, A. Y. Noaman, A. S. AL-Ghamd, and A. I. Madbouly, "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining," in *Workshop on Interaction Design in Educational Environments*, 2014, p. 106.

[12] "Most popular Twitter accounts in Tanzania _ Socialbakers." [Online]. Available: http://www.socialbakers.com/statistics/twitter/profiles/tanzania/. [Accessed: 13-Dec-2015].

[13] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500