The Impact of Transformed Features in Automating the Swahili Document Classification

Thomas Tesha
College of Informatics and Virtual Education (CIVE),
University of Dodoma (UDOM),
P.O. BOX 490, Dodoma, Tanzania

ABSTRACT
This paper describes experimental results in an attempt to identify the Transformation techniques which can be adopted to improve features for the automation of classification of Swahili documents. This means improving classification rate by enhancing separability and accuracy. The experiment involved Relative Frequency (RF), Power transformation (PT) and Relative Frequency with Power transformation (RFPT). The Term weighting with TFIDF and the absolute features (AF) were also studied. The features' dimension reduction was done by using the statistical techniques of Principal Component Analysis. In learning algorithm, the Support vector machine for classification and the *k*-NN were used, and in evaluating the effect of features' performance with the classifiers the micro averaged *f*-measure were adopted. The extensive experimental results demonstrated that the RFPT features worked better with the Support Vector Machine classifiers unlike *k*-NN in improving the classification rate by enhancing document separability and accuracy in Automation of Swahili document classification.

Keywords
Machine learning algorithm, Support vector machine, Swahili, Swahili document classification

1. INTRODUCTION
Studies on automated text classification have been one of the major and current research areas in natural language processing. It is defined as the task of automatically assigning a set of documents into appropriate categories according to their content or topic [1], [15]. The basic process is learning a classification scheme from training examples to build a model to be used in classifying unseen textual documents [1].

In fact with the status of the evolution of technology (Internet and the supporting IT infrastructure both software and the hardware) massive data in terms of text, audio, video and electronics pictures etc. (big data) are generated. However generation of these digital data is not a problem but the question that rises is how to organize and manage these piles of documents for the easy retrieval and analysis. In fact Swahili documents are among the plenty digital documents with much information not only in the Internet but also in all areas of technology including libraries as well, and if properly retrieved with easy and precise, they can solve various problems [2]. Social networks media like Facebook, Twitter, LinkedIn, Skype etc. also generate massive Swahili data all the time. So are offices, governments and organizations as well as researches. These digital documents can be harnessed and automatically organized using automated classifiers [2]. Studies on automated text classification [3- 5] have been done extensively using other languages than Swahili and the approach to improve features on Automating Text Classification in such language is done.

In this work, the focus is on the experimental study of feature transformation techniques which can improve classification rate in the automation of classification Swahili documents (text). The transformation techniques involved were RF, PT and RFPT. The experiments were also done with the Term weighting with TFIDF. In conducting this study, the selected machine learning algorithm intuitively the Support vector machine for classification (SVM) and k-nearest-neighbor (*k*-NN) were adopted.

To understand this work and the method involved, a familiarity with SVM for classification is required and a brief introduction follows. The subsequent two sections give a brief introduction on kernels and *k*-NN.

2. SUPPORT VECTOR MACHINE FOR CLASSIFICATION
The SVM [14-16] is a type of learning algorithm which was in its inception introduced largely by [16] and his colleagues and extended by the number researchers to its maturity. In fact the SVMs have proven to work successfully to many application including text classification, Optical character recognition, image processing etc. They separate a given set of binary labeled training data with a hyper-plane that is maximally distant from them (known as 'the maximal margin hyper-plane') a technique termed as optimal margin classifier. Although it involves the use of the hyperplane, It doesn't mean it handles only linear separable data, however a case of nonlinear, the nonlinearity is handled by a technique of kernels which maps the input space into high dimensional feature space which is linear separable. The kernels are discussed in brief in the following section. The hyper-plane found by the SVM in feature space corresponds to a nonlinear decision boundary in the input space.

Let's consider a case of a binary classification problem. In this case we take the input vector $X^j = \{ x_1, x_2 \dots \dots \dots x_n \}$ which for a case of Swahili documents, represents a document say J^{th} in a sample S of m labeled documents say $S = \{(X^1, Y^1), (X^2, Y^2) \dots \dots \dots \dots \dots (X^m, Y^m)\}$ and this input corresponds to label $Y^j \in \{+1, -1\}$. Furthermore, by assuming all data is at least distance 1 from the separating hyperplane where for a case of support vectors the inequality becomes equality as data points touches the margin see (1) and (2) it is easy to show that the margin M can be given by $M = 2/||W||$.

$$\boldsymbol{W^T X_i + b \geq 1 \; if \; y_i = +1} \quad \text{.....(1)}$$

$$\boldsymbol{W^T X_i + b \geq 1 \; if \; y_i = -1} \quad \text{.....(2)}$$

Treating this mathematically, the SVM learning algorithm finds the hyper-plane (W, b) such that the quantity $||W||$ is minimized for (X^i, Y^i). In other words minimizing $||W||$ is equivalent to maximizing M. By taking in the consideration of the non-separable case, *see* **Fig. 1** (which illustrate this case with the data points that fall within the region of maximum margin and those which fall on the wrong side of the separating plane), this can be formulated into the primal problem (3) where $C > 0$ is a regularization constant which determines the trade-off between the flatness (misclassification) of f and the amount up to which deviations larger than ε are tolerated. Large C values favor solutions with few errors. Small values denote preference towards low-complexity. The slack variable $\xi \geq 0$ one for each data point [16] measures the deviation from the ideal situation and they are defined by $\xi_i = 0$ for data points that are on or inside the correct margin boundary while $\xi_i = |y_i - (W^T X_i + b)|$ for the other data points.

$J(W) = W^T W + C\sum_{i=1}^{m} \xi_i$ Subject to $y_i(W^T X_i + b) \geq 1 - \xi_i$ and $\xi_i > 0$……(3)

Solving the primal problem (3) can be easily tackled by converting it to its equivalently dual formulation (11) by the introduction of the Langrage multiplier (λ) for each data point (X^i, Y^i).

In fact the constrained optimization satisfies KKT conditions which require six properties (5) through (10) to hold. Thus for every data point .

$\lambda_i \geq 0$…….....(5)

$y_i(W^T X_i + b) \geq 1 - \xi_i$ ……..…..(6)

$\lambda_i[y_i(W^T X_i + b) - 1 + \xi_i] = 0$…….....(7)

$\xi_i \geq 0$ ……..…(8)

$\mu_i \geq 0$ ……....(9)

$\mu_i \mu_i = 0$ ……...(10)

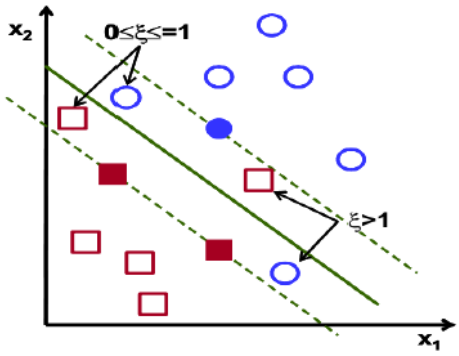

Fig. 1: Non separable case of data points

$J(W, b, \lambda, \xi) = W^T W + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \lambda_i y_i(W^T X_i + b) - 1 + \xi_i) - \sum_{i=1}^{m} \mu_i \xi_i$ …..…….. (11)

Solving (11) by taking partial derivative with respect to W, b, ξ and equate to zero it can be shown that

$W = \sum_{i=1}^{m} \lambda_i y_i x_i$, $\sum_{i=1}^{m} \lambda_i y_i = 0$ and $\lambda_i = C - \mu_i$…..(12)

And substituting (12) into (11) will give the Lagrangian dual problem (13)

Maximize $\boldsymbol{Q(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j x_i^T x_j}$

subject to $\boldsymbol{\sum_{i=1}^{m} \lambda_i y_i = 0}$ and $\boldsymbol{\lambda \in [0, C_i]}$ for $\boldsymbol{i \in [0, m]}$

……(13)

It is further can be shown W in (1) and (2) and for the case of non-separable data can be expressed as a linear combination of the training patterns x_i which is termed as Support Vectors (SVs) expansion. In this case, the complexity of a function's representation by SVs is independent of the dimensionality of the input space $\boldsymbol{X}$, and depends only on the number of SVs

3. KERNELS

The idea of kernels is to encompass the nonlinear behavior of pattern data. It is not always that all data the SVM encounter a linear separable. A case when the algorithm attain a nonlinear input space, the kernel machine perform transformation of the input features into Hilbert feature space more often termed as feature space. This involves mapping the input say $\emptyset: I$ into feature space $\emptyset: F$ equivalently ($\emptyset: I \subseteq R^n \rightarrow F \subseteq R^n$). By introducing the kernel now (13) can be modified into (14)

Maximize $Q(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j K(x_i^T x_j)$

subject to $\sum_{i=1}^{m} \lambda_i y_i = 0$ and $\lambda \in [0, C_i]$ for $i \in [0, m]$……(14)

Various forms of kernels have been invented and these include Linear-SVM which involves inner product$x_i^T x_j$, polynomial-SVM which involves$(x_i^T x_j + 1)^P$, radial basis function (RBF-SVM) which takes the form exp $(\frac{1}{2\delta^n}||X - Y||$ etc.

4. *k*-NN

The k-nearest-neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. It involves splitting the sample into two group one being for the training and the later for the testing similar to all supervised learning algorithm. Suppose x_j is an input sample with m features $x_1, x_2 \ldots\ldots\ldots x_m$ in a set of n sample. For specificity let's use x_i^k to denote the features in the k^{th} input sample and likewise for the j^{th}, then the Euclidian distance between x_j and x_k for k in the range 1,2 ….. m is given by (15)

$d(x_j x_k) = \sqrt[m]{(x_1^k - x_1^j)^2 + \ldots\ldots\ldots + (x_m^k - x_m^j)^2}$… (15)

Using this distance metrics as one approach to measure the closeness between training sample and the test sample, the *k*-NN assign the test sample to the (label) group with which the sample appears to be closest than the rest of the other labels.

5. METHODOLOGY

This study adopted the dataset which were used in the automatic Swahili document classification [2]. The generated features were preprocessed, transformed *(not to be confused with kernel transformation*), reduced in dimension size (Dimension reduction). The resulted features were then passed into classifiers for training and testing to evaluate the effect of transformation approach adopted with the classifier involved.

5.1 The Dataset USED

The documents relevant for this study involved the Swahili dataset which had a total of 3999 random articles1 [2]. Table 2 gives the summary of the categories and the label used in

[1]In this study articles and document are used interchangeably and in any case they involve Swahili documents.

the experiment for each class [2]. The translation of the classes involved in English language is also put in bracket. Table 2 gives the summary of the dataset for each category [2].

Table 1: Categories used with their label

Article name	Label
Afya (health)	0
Elimu (academics)	1
Kilimo (agriculture)	2
Mpira (football)	3
Muziki (musics)	4
Siasa (politics)	5
Ufugaji (husbandry)	6
Ujasiriamalinabiashara (business and enterpreneurship)	7

5.2 Feature generation

In Fig. 2 is a framework adopted for the Swahili document feature generation process. The stop list which contained 300 words which were manually specified was used to filter the stopwords. The stopwords contains a list of words which appears frequently in all documents and in this case would otherwise have no impact on the contribution to the classification performance. They just increase the dimension size unnecessary. Another filtering was on special characters and the alphanumeric words. The essence was to reduce features thus including only the pertinent features that greatly would contribute to classification (separation of documents into categories).

Table 2: Articles distribution category-wise

Article category (class)	# of articles
Afya	422
Elimu	351
Kilimo	350
Mpira	848
Muziki	636
Siasa	576
Ufugaji	349
Biasharanaujasiriamali	322

Together with the filtering process for both stopwords and special characters removal, a similar approach was used to remove the duplicate words to make the wordlist2 which was then used to generate feature vector (word frequency) of the form (16). The length of the feature generated was 2993.

$$\boldsymbol{X = x_1, x_2, x_3 \ldots\ldots\ldots x_n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(16)$$

[2]In text classification the wordlist is sometimes referred to as Lexicon or vocabulary. It involves all the words in the entire dataset after duplicates, special character, alphanumeric and stop list are removed. The train wordlist is usually sorted.

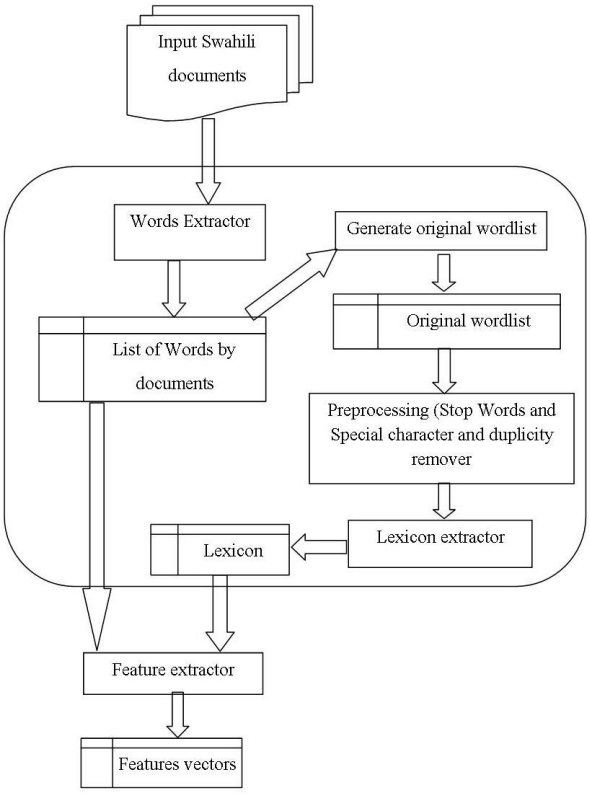

Fig. 2: Swahili document feature generation process

5.3 Term selection and the feature transformation

In this study both the term selection and feature transformation were studied for the sensitive power to represent feature in Swahili document classification. We examine the usage of these tools in this section

5.3.1 Term Selection

In this study feature selection involved term weighting with the Term frequency–inverse document frequency (TFIDF). The idea was to provide term weighting based on the frequency of occurrence of words, thus the more a word appears in a document, its term frequency (TF) became high and the more it is estimated to be significant in this document [7]. Each word in the text document is weighted according to its uniqueness [5] to capture its relevancy among words, text documents and particular categories.

5.3.2 Feature transformation using Relative frequency

The absolute word frequency in generating the feature vectors has one major drawback of depending on text length leading into lower classification rates [8], [9] and it is because text length may differ within the same class hence lower separability [8]. Then performing relative frequency, an improvement or increase in the separability of documents in their respective categories is expected [2]. This is because the relative frequency does not depend on the text length and therefore the within-class variance of the relative frequency is smaller than the absolute frequency [9]. The relative frequency equation takes the form (17)

$$Y_i = \frac{x_i}{\sum_{j=1}^{n} x_j} \qquad (17)$$

5.3.3 Feature transformation using Power Transformation

The power transformation was also studied and it is a function which varies continuously with respect to the power parameter V [9], [10] which takes values in the range [0,1] for the data vector Xi's [10] ie. that is for the data vector of the form (18) as shown in (19)

$$X = x_{1,}, x_2 \dots \dots \dots, x_n \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (18)$$

$$Z_i = X_i^V \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (19)$$

This was employed to improve the classification accuracy. It improves the symmetry of the distribution of the frequency which is noticeably asymmetric near the origin [5]. In the experiment V was experimentally chosen to be 0.5.

5.3.4 Feature transformation using Relative frequency with Power transformation

Another transformation technique deployed was the hybrid that combined the relative frequency which increases separability [8] and power transformation to improve the classification accuracy [9]. When the RF transformed features is re-transformed using power transformed technique; see in (19), the new features combine the strength of power transformation with that of relative frequency and can be abbreviated as RFPT. For this case RFPT features have superior properties to simplify the process of learning by a classification system [10]. The actual formulation for this technique then is as shown in (20).

$$Z_i = \left(Y_i = \frac{x_i}{\sum_{i=1}^{n} x_j}\right)^v \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (20)$$

5.4 Dimensionality reduction

Though the features were reduced by stopwords, duplicate word and alphanumeric character removal, the dimension reduction technique was also adopted to reduce the features further. This is because high dimension data increase the search space. They are also problematic in terms of computational time and storage resources [8], [11]. The other problem is over-fitting as a phenomenon by which the classifier is turned to the contingent characteristics of the training data rather than just the constitutive characteristics of the categories [1], [11]. In this study the statistical technique Principal Component Analysis (PCA) was used to reduce features thus retain only the principal components as the ones to be passed to the classifier/discriminant functions.

5.5 Training/ learning and testing and evaluation

Three sets of features were created. The first set consisted of the features which were not transformed at all but only reduced in dimensionality using the stoplist also known as stopwords and alphanumeric filtering process and the dimension reduction. These features will be referred to as absolute features (AF). The second group had the transformed features which included RF features, PT features and RFPT features. The last group had the TFIDF features. The classifiers used were the linear-SVM, polynomial-SVM, RBF-SVM and *k*-NN. The training phase was to build the model in which the test documents were later fed for prediction. Each of the classifier used in the training phase contributed to build the model that was used for classifying the test documents and the result for each classifier was recorded.

6. EXPERIMENTAL SETUP AND PERFORMANCE EVALUATION

The dataset were split into two groups. The first group was for the training and the other for the testing purpose. The training group had a total of 2993 articles and the testing had a total of 1006 articles. Table 2 presents the summary of the distribution of the articles involved in the training and testing group based on the available data in each category. On extraction *see fig. 2* the train documents generated *list of words by documents*.

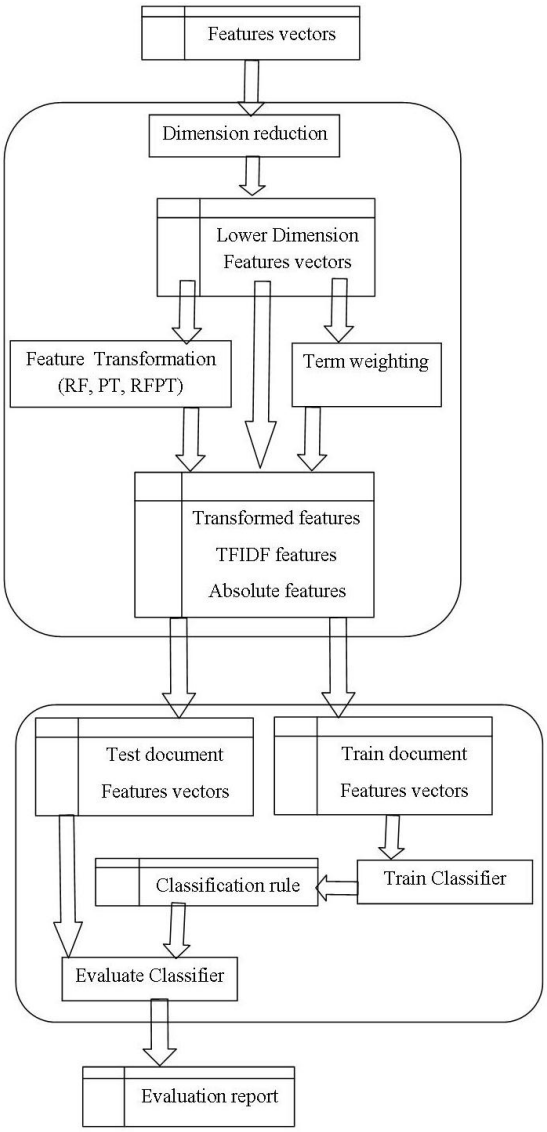

Fig. 3: Applying transformation, and term weighting with TFIDF, Training and evaluating the classifiers

Using this *list of words by documents*, a list containing 79644 words were generated which contained stoplist, special characters and duplicates. This list was termed as original wordlist. In fact the original wordlist was reduced to 9230 different words following preprocessing which involved the removal of the stoplist, special character and duplicated words. In this stage the list was passed through the lexicon extractor which sorted to prepare the lexicon which is also known as wordlist. The lexicon contained the words which allowed as making the feature vector using the *list of words by documents*. The complete process is given in fig. 2

The generated feature vectors as described in *Fig. 2* were then further reduced into lower dimension using the statistical principal component analysis (PCA) and the originated features were passed into the transformation (RF, PT and RFPT), TFIDF to generated respective features as well as retain the original copy which was termed as absolute feature (AF) meaning the copy was neither transformed nor passed into the TFIDF. These processes are well described in Fig. 3. The RF generated RF features, PT generated PT features, RFPT generated RFPT features and the TFIDF generated TFIDF features. These features along with the AFfeatures were then used to train the classifier to generate classification rule which were used to classify unseen document (test document) for the evaluation process. The results for each experiment with the features were recorded as can be seen in Table 3.

In the experiment, SVMlight package [12] and the *k*-NN were adopted for testing and evaluation. The value of *k* in *k*-NN was experimentally chosen and the SVMlight involved the classifiers linear-SVM, polynomial-SVM and RBF-SVM. To evaluate the performance of the classifiers, we adopted F-measure metric [13].

7. RESULTS AND DISCUSSIONS

The results of all the classifiers used from the SVMlight along with the *k*-NN on the features applied are as shown in Table 3. The Fig. 4 and Fig. 5 show the comparison between classifiers and features that were adopted for the classification of Swahili documents. The aim was to identify the best transformation techniques for automating Swahili Document Classification. A closer stare in Fig. 4 and Fig. 5 show, RFPT features with the Linear SVM, Polynomial-SVM kernels and SVM-RBF outperformed the other features. On the other hand with the k-NN, the TFIDF worked better in comparison with the RFPT.

Table 3. Classifiers performance evaluation

Classifier	AF	RF	RFPT	TFIDF	PT
k-NN	91.63	91.66	93.17	93.29	93.17
Linear-SVM	93.06	93.73	93.94	93.78	93.78
Polynomial-SVM	90.72	93.75	93.89	93.76	93.13
RBF-SVM	92.45	90.88	93.50	93.29	92.80

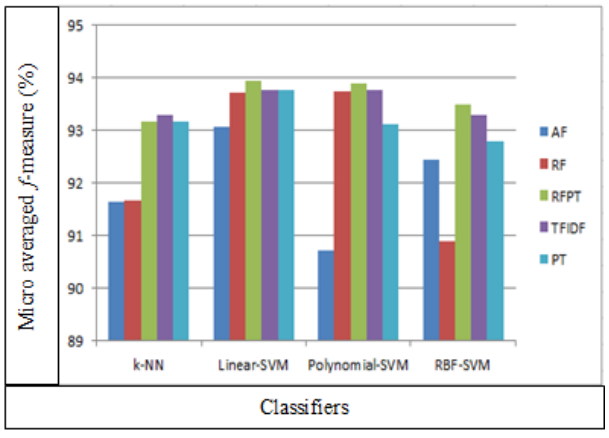

Fig. 4: Classifiers performance evaluation

The general observation regardless of the performance, the RFPT works well with Swahili under the SVM environment. Another observation, Fig. 5 shows that, the Linear-SVM had the best performance with the RFPT than are the counterparts' classifiers. This indicates RFPT is better at representing Swahili feature in comparison with the other feature transformation technique experimented

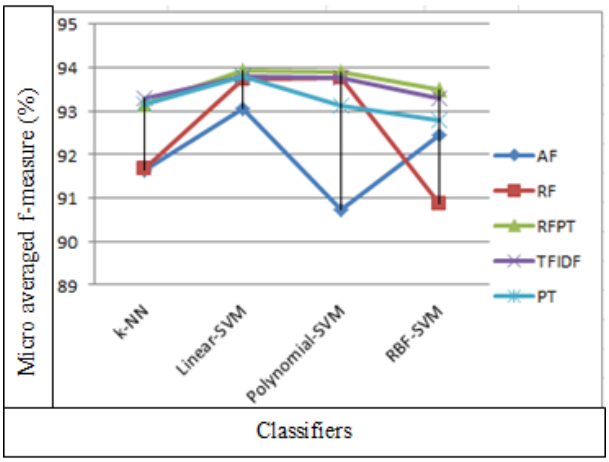

Fig. 5: Features and classifiers performance

8. CONCLUSION

This paper has discussed a comparative study of the transformed, the term weighting with TFIDF and the absolute features in automating Swahili document classification. The transformation techniques involved were the RF, PT and the hybrid of these two which involves RFPT. In evaluating the effect of features' performance with the classifiers the micro averaged *f*-measure was adopted and the results demonstrate that, the RFPT features worked better in representing Swahili features with the SVMs features than *k*-NN features. It was also demonstrated that the linear-SVM had the highest performance in comparison with the polynomial-SVM and RBF-SVM in the family of SVMs kernels involved. In all cases the extensive experiments reveal that, RFPT improves the classification rate by enhancing Swahili document separability and accuracy.

9. REFERENCES

[1] Sebastiani F. (2002), "Machine learning in automated text categorization", ACM Computing Surveys, Vol. 34, pp. 1-47.

[2] T. Tesha, L. S. P. Busagala, "Automatic Swahili documents classification", unpublished.

[3] Al-Harbi, S. et. Al (2008), "Automatic Arabic Text Classification", Journées internationales d'Analyse statistique des Données Textuelles

[4] A. Kokawa et al (2011), "Feature Selection and Integration in Automatic Classification of Japanese Texts", Graduate School of Engineering, Mie University, Tsu-shi, Japan, Sokoine University of Agriculture, Morogoro, Tanzania

[5] Z. Shuigeng and G. Jihong (2002), "Chinese Documents Classification Based on N-Grams", Proceeding CICLing '02 Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing pp 405-414

[6] Y. Zhang, L.Gong, Y. Wang (2005), "An improved TF-IDF approach for text classification", Journal of Zhejiang University SCIENCE ABC, Vol.6 No.1 Pp.49~55

[7] P.Soucy, G. Mineau (2005), Beyond TFIDF Weighting for Text Categorization inthe VectorSpace Model, In Proceedings of the Proceedings of the 19th International Joint Conference on Artificial Intelligence

[8] L. S. P. Busagala et al (2005), "Machine learning with transformed features in automatic text classification", Mie University, Kurimachiya-cho, Tsu, Mie, Japan

[9] Z. Guoweiet. Al (2003), "Accuracy improvement of automatic text classification based on feature transformation", [Online]. Available: http://www.hi.info.mie-u.ac.jp/publication/archive/Guowei_Proc_2003_11.pdf

[10] A. Malero., L. S. P. Busagala (2011), "Transformed features in automatic spam filtering", Journal of Informatics And Virtual Education, Tanzania

[11] M.Ikonomakis, S. Kotsiantis, and V.Tampakas (2005), "Text Classification Using Machine Learning Techniques", WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, pp. 966-974

[12] T. Joachims (1999), Making large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, [PDF] [Postscript (gz)] [BibTeX]

[13] O. Arzucan, O.Levent and G. Tunga (2005), "Text Categorization with Class-Based and Corpus-Based Keyword Selection", Springer-Verlag Berlin Heidelberg

[14] N Cristianini, J Shawe-Taylor (2000), "An introduction to support vector machines and other kernel-based learning methods" Cambridge university press

[15] V. Vapnik (1998), Statistical learning theory. Vol. 1. New York: Wiley

[16] C. Cortes and V. N. Vapnik (1995), "Support vector networks" Machine Learning