International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT20645

Comparative study of the Performance of Machine Learning Text Classifiers Applied to Afaan Oromo Text

Etana Fikadu
College of Engineering & Technology, Wollega University, Post Box No: 395, Ethiopia

Article Info
Volume 6, Issue 4
Page Number: 77-83
Publication Issue :
July-August-2020

Article History
Accepted : 10 July 2020
Published : 15 July 2020

ABSTRACT

The aim of this study is to find the optimal method that can be used to classify Afaan Oromo text among different classifier by using the same number of text document. Automatic text classification has been needed in many fields for a long time. Many methods are used to classify text. The performance of this classifier we used in this study is measured in terms of recall, precision and F-measure. Finally we compare the efficiencies of the Bayesian Network, Naïve Bayesian, IBK and SMO to classify Afaan Oromo text. Experimental results on the same set of Afaan Oromo documents used before show that SMO slightly outperforms the other methods. Comparison reported in this paper shows that the SMO classifier exceeds the other four Machine learning classifier.

Keywords : Afaan Oromo text categorization, classification algorithms, machine learning

I. INTRODUCTION

The globalization era provides a rapid growing amount of online textual information and data coming from different sources. As a result, it becomes more and more difficult for target users to select the contents. This has problems for searching the relevant documents for the entire content. Supporting the target users to access and organize the enormous and wired spread amount of information is becoming a primary issue. As a result, many services have been proposed to find and organize valuable information need by the target users [1]. But, these services are not available to fully perform the user's interest. So a solution to answer for personalized information filtering, semantic document indexing, information extraction, and automatic metadata generation for rapid growth online information. This solution is called text categorization [2].

Text categorization is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set [3]. There are various reasons for using document categorization. Automatic document categorization reduces searching time, thereby facilitating the searching process. Moreover, it facilitates access, when documents are classified based on their concept similarity; we can get hint about what the document actually contains without going through it. Document classification can be done manually or automatically.

Manual text categorization is carried out by human experts. It requires a certain level of vocabulary recognition and knowledge processing. There are some problems observed with manual classification [4]. It requires intensive human labor and affects classification results because of inconsistency due to variation in perception, comprehension, and judgment, and for the current Web based knowledge management it is almost impossible. In contrast, automatic classification is a process of classifying documents into a number of classes using machine learning methods [5].

Machine Learning is defined as "the ability of a machine to improve its performance based on previous results". In other words it is a system capable of learning from experience and analytical observation, which results in continuous self-improvement there by offering increased efficiency and effectiveness (6). Automatic text classification can significantly these days because it reduce the cost of manual classification and human intervention from manually organizing documents which is too expensive and, error prone[2].

The main stage of building a text categorization system which involves compiling and labeling text documents in data set, selecting a set of features to represent text documents in a defined set classes or categories (structuring text data),and finally choosing a suitable learning algorithm to be trained and tested using compiled corpus.

II. RELATED WORK

Different algorithms and techniques have been applied for many years in text categorization and classification. They include decision tree learning, Bayesian learning, nearest neighbor learning and artificial neural networks, early such works may be found in [11, 12].

[4] The researchers Presents the results of the Naïve Bayes and Bayes Net, classifier algorithms were implemented for training text classification model depending on different main classes of documents. The performances of these classifiers are analyzed by applying various performance factors. Among those classifications algorithms, Bayes networking algorithm shows higher performance 97.15% and hence it was utilized for constructing classification model for Afaan Oromo texts.

The researchers [5] who have been done in the area of machine learning in text categorization indicate good results. The best result obtained by Decision Tree Classifier and Support Vector Machine is on six categories data (96.58, 84.93%) respectively. This research indicated that Decision Tree Classifier is more applicable to Afaan Oromo news text than the other classifiers. Moreover, it is learnt that considering categories with equal number of news items increases the performance of the classifiers. In other words, insufficient examples in one class can affect the classifier as a whole. It was also observed that the classification of Afaan Oromo news text is possible without using the sophisticated feature reduction techniques such as information gain and odds ratio.

[7] The researcher applies different classifier on different number of classes. The best result (accuracy) obtained from both the SMO and Bayes Multi Nominal classifiers was when the number of instances are approximately equal in each class and the accuracy is 95.82% and 96.58% respectively for the category of 4 classes. Relatively lower accuracy obtained is for J48 on category of 7 classes 79.69% and on category of 11 classes 82.05% Both SMO and Bayes Multi Nominal showed high accuracy, SMO tends to have better accuracy over Bayes Multi Nominal for the Afaan Oromo news items classification.

Y. Yang, and X. Liu [8] conducted a good study comparing document categorization algorithms. Also, Hassan et al. [11] present experimental results on document clustering and classification achieved on the Arabic corpus using statistical methods.

Bawaneh et al. study [13] implemented the *K*NN and Naïve Bayes algorithms in order to make a practical comparison between them and previous studies. The basic terminology is based on the idea that a standardized text classification process passes several major phases. In the first phase (preprocessing) documents are prepared to make them adequate for further use, therefore stop words are removed. In the second phase (weighting assignment phase), it is defined as the assignment of real number that relies between 0 and 1 to each keyword and this number indicates the imperativeness of the keyword inside the document. Algorithms implementation is mainly developed for testing the effectiveness of *K*NN and Naïve Bayes algorithms when applied to Arabic text. A set of labeled text documents are supplied to the system, the labels indicate the class that the text document belongs to. All documents should be labeled in order to learn the system and then test it. The system classifies a test document comparing it to all the examples it has (i.e., the training set), the comparison is done using a two previous classifiers. Tests show that the Naïve Bayes and *K*NN achieves an accuracy of 73.6% and 84.2% respectively, these results indicated that the Naïve Bayes classifier outperform the results achieved with previous studies where the *K*NN has a poor performance when compared in these studies.

One last conclusion derived from the study is the observation that the effectiveness of the algorithms depends largely on the characteristics of the collected datasets, hence the researcher is decided to do this research on the same data set to evaluate the performance of machine leaning classifiers.

III. Waikato Environment for Knowledge Analysis (Weka 3.9.2)

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [9]. It is free software available under the GNU General Public License. It supports several standard text categorization tasks such as data preprocessing, clustering, classification etc. All of the Weka techniques are predicated on the assumption that the data is available as a single flat file or relation which each data point is described by a fixed number of attributes (numeric or nominal attributes).

The Weka main user interface is the explorer. The same functionality can be accessed through the component-based knowledge flow interface and the command line. There is also the experimenter, which allows the systematic comparison of the predictive performance of the Weka machine learning algorithms on a collection of data sets. The explorer interface has several panels that give access to the main components of text classification processes. The preprocess panel has facilities for importing data from a comma separated value (CSV) file and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria.

The classify panel enables the user to apply classification algorithms (in Weka called classifiers) to the resulting data set, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions. Weka supports different classification schemes such as bayse net, naive bayes, IBK, support vector machine, etc. In order to classify the Afaan Oromo text documents in to different categories, the researcher imports the labeled

document records in ARFF format, preprocess them using the appropriate filtering algorithms, and classify the preprocessed document text in to different categories. Finally, the detail accuracy is measured using the standard performance measurement techniques, compare them based on the same collected corpus and recommend the highest performance for future use.

A. Performance Measures

In the present study, the performance of the classification results is measured using different evaluation measures. The performance of any classification algorithm is dependent on the quality of the produced results [10]. The performance of the classification is analyzed to measure the accuracy of the classifiers in categorizing the Afaan Oromo documents in to specified categories. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications [14]. Besides, recall, precision, and F-measures of the classifier are also measured. Precision (P) measures the percentage of documents assigned to category c that are correctly assigned to category c. More formally, the precision is defined as shown in equation

$$\boldsymbol{Pi} = \frac{\boldsymbol{TPi}}{\boldsymbol{TPi+FPi}}$$

True Positive (TP): If the instance is positive and it is classified as positive

False Negative (FN): If the instance is positive but it is classified as negative

True Negative (TN): If the instance is negative and it is classified as negative

False Positive (FP): If the instance is negative but it is classified as positive

On the other hand, recall measures the percentage of total documents that are assigned to category c. It is defined as : $\boldsymbol{Ri} = \frac{\boldsymbol{TPi}}{\boldsymbol{TPi+FNi}}$

Where TPi (i.e., true positives) is the number of documents assigned correctly to category ci, FPi (i.e., false positives) is the number of documents assigned to category ci that should have been assigned to other categories, and FNi (i.e., false negatives) is the number of documents assigned to other categories that should have been assigned to category ci. The F-measure (F) is the harmonic average of precision and recall, and is defined as : $\boldsymbol{Fi} = \frac{\boldsymbol{2xPi\,xRi}}{\boldsymbol{Pi+Ri}}$
Where Pi and Ri are the precision and recall for category ci respectively.

B. Data transformation and scaling

After representative words preparation, a document is treated as a collection of words, bag of words, which are the candidate representatives of a given document. In bag of words representation the relative position of words are not used. The feature which represents a document is prepared in a format that is used by the application package; I.e. the coma separated value (CSV) or Arff (Attribute relation file format). Feature words now become the attributes of the documents, then the frequency of a word in a document (tf), in how many documents in a collection the word appears, inverse document frequency (idf) are used to calculate the weight of each word in a document and normalization to this calculation is done to find how good the word represent the document both with respect of the document itself and with respect to the document collections.

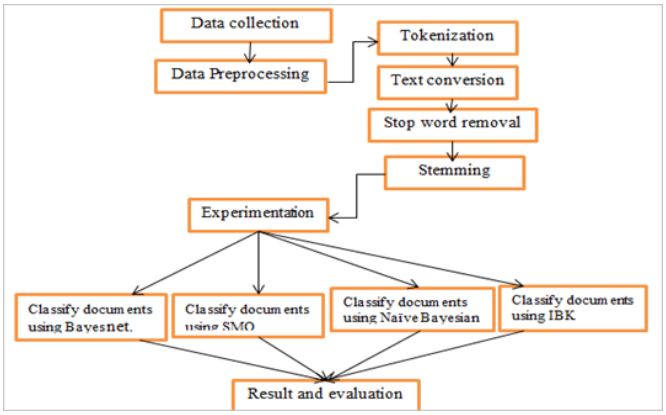
C. Testing classifier algorithms

Testing the classifier algorithms is important because it allows evaluating how reliably a given classifier will label future data, that is, data on which the classifier has not been trained.

In this research the performance of selected automatic text classifiers, for the application of Afaan Oromo text media categorization is tested. To make closer observation on the characteristics of the automatic classifiers on each and every category, Sequential Minimal Optimization (SMO) algorithm from function, IBK algorithms form lazy classifier, Bayes net and naïve Bayes algorithm from Bayesian Classifiers are used in the experiment. To ensure the application of machine learning approach to the Afaan Oromo text classification, KNN from lazy classifiers was also employed in the experiment for performance evaluation and comparison. All the selected classifiers were compared on the same data and a set of categories.

The testing with data of nine categories (Education, health, Gada system, Sport, agriculture, economy, accident, politics, religion) the all the nine categories were tested together. The whole experiment dataset has two types of attributes, numeric attributes for the weight of the feature words and nominal attributes for the class labels. Regarding the training and test sets, in all the experiments 10-fold stratified cross validation is used. A total number of documents used in this experiment are 1101 text documents.

Model frame work



IV. RESULT AND DISCUSSION

As a final step of the proposed methodology, we conduct the experiments. Four classification algorithms under test are, Sequential Minimal Optimization (SMO) algorithm, Bayes network algorithm, IBK and the Naïve Bayesian algorithm. The resulting dataset will be classified into nine classes; it will be used to assess the performance and efficiency of the Sequential Minimal Optimization (SMO) which is The WEKA version of the support vector machine algorithm (SVM). SMO implements the sequential minimal optimization algorithm for training a support vector classifier, using polynomial or Gaussian kernels. Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default. One advantage of using this implementation is that the amount of memory required by SMO is linear to the size of the data.

IBK is a k-nearest-neighbour classifier that uses the same distance metric. The number of nearest neighbours can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. IBK is a knearest- neighbour classifier. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbours. A linear search is the default but further options include KD-trees, ball trees, and so-called “cover trees”.

The final algorithm that is going to be tested using the dataset is the Naïve Bayes. A Naive Bayes classifier is a well-known and practical probabilistic classifier and has been employed in many applications. This algorithm is based on Bayes’ rule of conditional probability; the rule says that if you have a hypothesis H and evidence E that bears on the

hypothesis the conditional probability of *H* given *E* is given by: $P(H/E) = \frac{p(E/H)p(H)}{P(E)}$

Where: *P* (*H*) denotes the priori probability, the probability of the hypothesis before the presentation of any evidence.

P(*E*|*H*) denotes the conditional probability that *H* is true given evidence *E*.

P(*E*) denotes the probability of the evidence associated with the hypothesis.

Naïve Bayesian Classifier is one of the Bayesian Classifier techniques which also known as the state-of-the-art of the Bayesian Classifiers. In many works it has been proven that Naïve Bayesian classifiers are one of the most computationally efficient, effective and simple algorithms for DM applications. Naïve Bayes works very well when tested on a dataset, particularly when combined with some attribute features techniques.

The basic idea in Naïve Bayes methods is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naïve aspect of the method has to do with the fact that the dependencies between words are ignored, i.e. the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. The text documents used in this study was a total of 734 training data (documents) and 337 test data out of the total 1101 documents. The training data is approximately 2/3 of the total corpus and the rest, 1/3 hand used for testing purpose. To test the accuracy of our system, we selected 1101 documents which reside in 9 classes. Measuring the efficiency of our classifier, we used the formulas of Recall, Precision, and *F-measure*. The results are displayed in (table 1).

All classification used were compared with each other (fig.1).

Table 1. Shows the comparison between four classifier

Name of classifier	Precision	Recall	F-Measure
Bayes network	97.5	95.9	95.9
Naïve Bayes net.	95.8	95.7	95.7
IBK	96.1	96.1	96.1
SMO	97.7	97.5	97.5

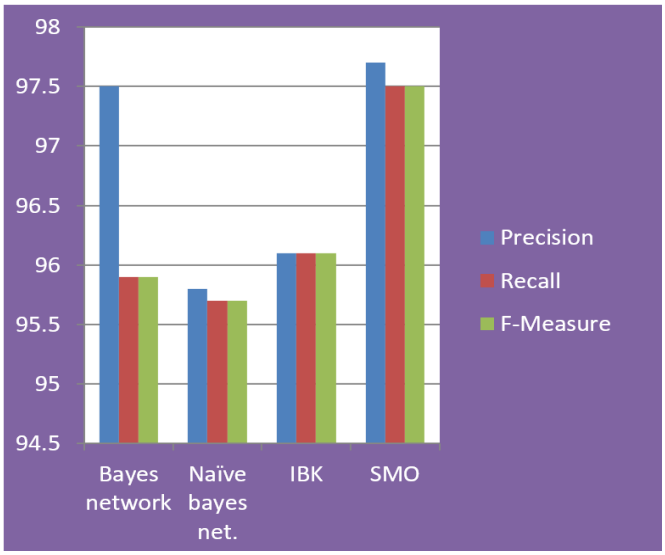


Figure 1. Comparison between classification algorithms.

As we observed from the experiment the values of Recall, Precision, and *F-measure* are calculated for each of the classifier. So, from the four classifier SMO shows the best performance than the other.

V. CONCLUSION

Automatic text categorization is the task of automatically assigning one or several predefined category labels (or classes or topics) to a given text written in a natural language, according to its similarity with respect to a previously labeled corpus used as a reference set [4]. Many of researcher done research on this area and they propose different

classifier on different Collected Afaan Oromo corpus. But, there is no common developed classifier model for Afaan Oromo texts. This paper comes up with comparison of different machine learning classifier on collected Afaan Oromo text documents. In this paper we have evaluate the performance of different machine learning classifier on the same corpus to conclude the highest performance for the four classifier algorithms. Based on the experiment SMO shows the best performance and the researcher recommend this classifier for future use for Afaan Oromo text Classification.

Author Profile

Etana Fikadu Dinsa (MSc in computer science) working as senior lecturer in Department of Computer Science, Wollega University, Nekemte, Ethiopia. His research interest is Machine learning, Natural Language Processing and Data mining.

VI. REFERENCES

[1]. A. Addis,"Study and Development of Novel Techniques for Hierarchical Text Categorization," PhD Thesis, University of Cagliari, 2010.

[2]. F. SEBASTIANI, “Machine Learning in Automated Text Categorization”, vol. 34, 2002.

[3]. R.S. Feldman, J., the Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press,USA, New York, 2006.

[4]. Etana Fikadu and Ramesh Babu P, “Application of Data Mining Classification Algorithms for Afaan Oromo Media Text News Categorization”, published 7 - july 2019

[5]. Kamal, et al “Afaan Oromo News Text Categorization using Decision Tree Classifier and Support Vector Machine: A Machine Learning Approach”, published May, 2017

[6]. Nigam K, American Association for Artificial Intelligence (AAAI),. Inc., A Nonprofit California Corporation, Al Topics / Machine Learning, 2008.

[7]. Abera Diriba, Automatic classification of Afaan Oromo news text: The case of radio fana, master’s thesis, (2009)

[8]. Y. Yang, and X. Liu, (1999). A re-examination of text categorization methods.

[9]. I.and Frank, "Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, USA: ," Morgan Kaufmann, 2005.

[10]. Y. Zhao, "Comparison of Agglomerative and Partitioning Document Clustering Algorithms.," Washington DC: ACM Press., (2002).

[11]. S. Hassan, H. Ney, and J. Zaplo, (France, Friday 6 July 2001), Statistical Classification Methods for Arabic News Articles, ACL/EACL 2001, Germany.

[12]. A. Bensaid, et al (2004). Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm.

[13]. Sakhr, Categorization, (2007). http://www.sakhr.com/

[14]. R. and Ribeiro-Neto, B. Baeza-Yates, "Modern Information Retrieval. New," 1999.

[15]. http://www.cs.waikato.ac.nz/ml/weka/ (Visited January 8th, 20120).

Cite this article as :

Etana Fikadu, "Comparative study of the Performance of Machine Learning Text Classifiers Applied to Afaan Oromo Text", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 4, pp.77-83, July-August-2020. Available at doi : https://doi.org/10.32628/CSEIT20645

Journal URL : http://ijsrcseit.com/CSEIT20645