Arabic Text Categorization

Jazya Moftah[1] * and Mabrouka Amhamed[2]
e-mail: jazamoftah@su.edu.ly

Sirte University/Faculty of sciences (SU), Libya

Abstract

In this paper, the researcher compared the performance of two classifiers for Arabic text classification. Naïve Bayes and Key Nearest Neighbor (KNN) were used to classify the documents. These documents which were not classified were preprocessed by removing stop words and punctuation marks from them. The word in each document was presented as a vector . These vectors were used in WEKA tool to give the results. The accuracy of two algorithms was compared using precision, recall, f-measure. The results showed that the accuracy Naïve Bayes algorithm was better than Key Nearest Neighbor(KNN) algorithm .

Keywords: *text classification, categorization, naïve Bayes , Key Nearest Neighbor (KNN), Arabic language.*

1. Introduction

With emergence of internet, huge amount of electronic information is available, so the process of retrieving becomes more difficult because there's no good indexing and summarization of documents contents [1]. Besides that, anyone can post documents to the internet without constraint; as well the amount of document increase, the retrieval for information becomes difficult problem. Text categorization is taken to deal with sorting documents by content, while text classification is used to classify the documents by any kind of assignment of documents to classes [2] . A text categorization system can be used in indexing documents to assist information retrieval tasks as well as in classifying e-mails, memos or web pages [3].

Text categorization (classification) is a solution for the problem. Text Classification (TC) is the task using to classify a specific dataset into different classes [4], Categorization process can be done by manual or automatic, but manual categorization is impossible to apply because the documents are very huge and the time to complete this mission will be too much. so, the optimal approach is automatic categorization . Achieving high accuracy in Arabic text categorization depends on the preprocessing techniques used to prepare the data set[5] .Many methods have been applied in this field .some of these methods are related to statistical like regression models, nearest neighbor Classifiers, decision trees, rule learning algorithms, neural networks. Categorization problem is viewed as supervised learning which means to assign unlabeled tuples to predefined categories [6]. Text categorization refers to the process of classifying category. Using any approach like machine learning done by learning the system how to classify using training data. Text Categorization or Classification (TC) is concerned with placing text documents in their proper category according to their contents[20]

Reuter's data collection is used to evolution English text categorization according to [1] the Reuters has many version one of them which is commonly used Reuters 215782 collection. this standard can be used to help researches to evaluate and compare their results with other results, but in Arabic there is no public data set of collection like English .so the Arabic document was one of challenge that had been shown when started this paper, for that we developed our small collection which consisting of 200 document belong to 4 different categories.

Another difficulty was the Arabic text and its morphology, the process of Automatic text categorization depends on the contents of documents, a huge number keywords, can lead to a poor of both accuracy and time, in Arabic text you can find a huge number of morphemes generated out from one root, if treat all of morphemes independently with very large number of keywords that will reduce efficiency and the accuracy of classification, one solution to reduce problem is the use of roots, which leads to a smaller and more efficient feature space

2. Methodology

In testing the Naive Bayes and the KNN algorithms to classify unlabeled document .The documents are collected from online websites. The corpus contain of 200 documents divided into 4 categories (biology, mathematics, physics, chemistry). These documents distributed 50 documents to every category. I got program to help me doing preprocessing for document. The output of this phase will be the input for the next step. The next step it will be using WEKA. In WEKA The data was used into two ways, in the first one data will be divided into two partitions, where the first partition is 66% from the dataset, and it will be used for training phase, the second partition is 33% in size and will be used for testing phase.

The second form for using the dataset for training and testing is using cross validation, in k-fold cross-validation technique; the dataset are randomly partitioned into k mutually exclusive subsets or “folds”, each of approximately equal size. Training and testing is performed k times. In iteration i, partition is reserved as the test set, and the remaining partitions are collectively used to train the model. So the process runs as follow, in the first iteration, subsets D2.....Dk collectively serve as the training set in order to obtain the first model, which is tested on D1; the second iteration is trained on subsets D1, D3, ...Dk and tested on D2 ; and so on.

3. Preprocessing

there is no publicly Arabic corpus to test the proposed classifiers and compare between them I have to use my own document which I collected it from internet. In this Arabic dataset saved each document in separate file within category's directory. Document preprocessing includes much process such as remove punctuation marks, prepositions, conjunction, and pronouns. The rest of words are refered as key words The experiment is performed after normalizing the key words. In normalization, some Arabic litter are normalized such as (hamzah) أ, إ and ا Converted to ا and ى replaced by ي and ة to ه.

4. WEKA

The Waikato Environment for Knowledge Analysis inception in1992, Collecting the learning schemes together was a daunting at best. Weak is not just a toolbox of learning algorithms, but also its consider a frame work which the research can implement new algorithm without conserved about the infrastructure . Weka is a data mining tools. It is contain the many machine leaning algorithms [7]. WEKA contains algorithms for regression, classification, clustering, and attribute selection and association rule. WEKA can be used for classification, and gave me option to run the cross validation for a selected learning algorithm on the data set which prepared in preprocessing phase. Classification is considered as a supervised algorithm. Also weka supports application of unsupervised like clustering algorithm. Weka can provide visualization for some task, because many of application, the data visualization give important insight.

5. Classifier

- **KNN Classifier**

The KNN is one of the basic classification and regression methods. The KNN initially used to solve the problem of text categorization and later widely used in various fields of pattern recognition . KNN is known of the top methods, many studies applied KNN on their classification. The idea of KNN can be demonstrated as follows:-given asset of best data (document) to be classified, the algorithm use the pre-classified from the training document to search for the K-nearest neighbors based on similarity measure and ranks those K neighbors based on their similarity. The KNN classifier scans its documents for every test. The drawback of KNN is the difficulty of determination the value of K has to traverse all the training document many times to determine the best value of K. see Appendix A

- **naïve Bayes classifier**

Naïve Bayesian is a machine learning approaches used in text classification[8]. At the first the probabilities are calculated for each category in the naïve bayes classifier. This can be done by first calculating the prior probability P(v) for each category; which was 1/4 for each category. Next I calculated the probability given a word from the Vocabulary (wk) that the category will be v. This probability will be referred to as P(wk|v) .

$$P\ (wk|v) = (nk + 1) / (n + |Vocabulary|)$$

To classify the documents .we just chooses the highest probability.

6. Experiment Results and Discussion

Experiment our methods, Arabic text corpus was collected from online sitses.200 document were collected. Every category contain 50 document the predefined category include: physics, biology ,chemistry ,and mathematical. I applied two techniques, KNN and naïve bayes. when I determined the train set and test set document ,I used two different approach : 1-cross validation , 2-percentage split .In our experiment I filtered my documents from stop word and normalize it them I used WEKA tool to give me the result. The accuracy of these experimental shows. that the nave bayes is better that KNN in both technique split of data set .10 fold cross validation and percentage split as table 1

Table 1":the accuracy for both classifier (Naïve Bayes, KNN)

	Accuracy	
	10 Fold Cross Validation	Percentage Split
Naïve Bayes	70%	63.24%
KNN, K=5	62%	48.50%

Table[2] shows recall, precision and f-measure when naïve bayes classifier used in two cases (10 fold cross validation and percentage split).It can be seen from table2; precision and recall reach to 0.7 for 10fold cross validation and precision, recall is 0.66, 0.63 respectively foe percentage . When classifier used where K=5 the precision and recall showed as in table [4] The precision for each category is varying from others but in general the precision for mathematical was the best and it's reach to 0.98 but the physics was the worst category where the precision reach to 0.54.

Also from table 4,5,6,7 showed that the cross validation was better than percentage spilt when we determine the test document.

Table 2:the recall, precision and f-measure for classifier (Naïve Bayes)

	Naïve Bayes	
	10 Fold Cross Validation	Percentage Split
Precision	0.7	0.664
Recall	0.7	0.632
F-Measure	0.697	0.64

Table 3:the recall, precision and f-measure for classifier(KNN)

	KNN with K=5	
	10 Fold Cross Validation	Percentage Split
Precision	0.775	0.58
Recall	0.62	0.485
F-Measure	0.603	0.453

Table 4: Naïve Bayes Confusion Matrix Percentage Split.

	Naïve Bayes Confusion Matrix Percentage Split				
Precision	a	b	c	d	**< Classified as**
0.69	16	3	2	2	a = Biology
0.72	1	8	1	1	b = Chemistry
0.58	1	2	10	4	c = Mathematics
0.53	1	5	2	9	d = Physics

Table 5: Naïve Bayes Confusion Matrix Cross Validation

	Naïve Bayes Confusion Matrix Cross Validation				
precision	a	b	c	d	< **Classified as**
0.68	34	3	9	4	a = Biology
0.72	2	36	1	11	b = Chemistry
0.86	2	3	43	2	c = Mathematics
0.54	5	14	4	27	d = Physics

Table 6: KNN Confusion Matrix Percentage Split

	KNN Confusion Matrix Percentage Split				
Precision	a	b	c	d	< **Classified as**
0.47	11	4	7	1	a = Biology
0.55	0	6	5	0	b = Chemistry
0.88	0	1	15	1	c = Mathematics
0.05	0	8	8	1	d = Physics

Table7: KNN Confusion Matrix Cross Validation

	KNN Confusion Matrix Cross Validation				
precision	a	b	c	d	< **Classified as**
0.68	34	2	14	0	a = Biology
0.6	0	30	20	0	b = Chemistry
0.98	0	1	49	0	c = Mathematics
0.22	1	11	27	11	d = Physics

7. Conclusion

This paper has been carried to automatically classify Arabic document using Naive Bayes and KNN algorithm. Unclassified documents were preprocessed and I used WAKA tool to apply naïve bayes and KNN classifier. The results showed that the performance of the naïve bayes classifier was better than KNN classifier.

References

[1] Kanaan ,G.G. 2006. Arabic Text Categorization Using kNN Algorithm, Proceedings of The 4th International Multiconference on Computer Science and Information Technology, Amman, Jordan, Vol 4.

[2] Elhassan ,R,M. 2015. Arabic Text Classification Process, International Journal of Computer Science and Software Engineering, pp. 258-265

[3] Errub ,A.A**. 2014.** Arabic text categorization algorithm using vector evaluation method, international journal of computer science & information technology , Vol 6.

[4] Shalabi.R. 2015. Different classification algorithms based on Arabic text classification, Study International Journal of Advanced Computer Science and Applications, Vol. 6, No 2.

[5] Alshammari.R. 2018. Arabic Text Categorization using Machine Learning Approaches, International Journal of Advanced Computer Science and Applications, Vol 9,No 3 .

[6] Zhu,S.O. 2008. Text categorization via generalized discriminate analysis, Information Processing and Management ,an International Journal**, Vol**. 44, PP. 1684-1697.

[7] Litoriya,R. 2012. Comparison the various clustering algorithms of weka tools, International Journal of Emerging Technology and Advanced Engineering .