PressAcademia Procedia - (PAP), ISSN: 2459-0762

PressAcademia Procedia

2nd World Conference on Technology, Innovation and Entrepreneurship

May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

OVERVIEW AND COMPARISON OF THREE CLASSIFIERS: ARABIC DOCUMENTS AS A CASE STUDY

DOI: 10.17261/Pressacademia.2017.600
PAP-WCTIE-V.5-2017(38)-p.274-277

Essam Hanandeh [1]
[1]Zarqa University, hanandeh@zu.edu.jo

ABSTRACT

Nowadays, text classification is used in various fields of research and applications, such as information retrieval, text mining, and data mining. This study tests the Naïve Bayes, K-Nearest Neighbors, and Support Vector Machine algorithms on a relatively large dataset of Arabic documents. This dataset comprise 1,000 Arabic documents that are distributed across 10 classes. This comparison is based on recall and precision measures. The evaluation results show that the Support Vector Machine algorithms classifier outperforms the other two.

Keywords: Arabic text categorization, KNN, NB, SVM, text mining

1. INTRODUCTION

Classification constitutes a significant part of data mining, text mining, and machine Learning. In classification, a machine or human attempts to identify to which class from a set of classes a new instance belongs (Agirre et al., 2009). Machines are configured to classify different instances by referring to instances whose classes are known. Therefore, classification algorithms adopt supervised learning, whereas clustering algorithms adopt unsupervised procedures. Usually, classification algorithms use different features to determine the class of each instance under consideration. In our study, each instance represents a text document (Syiam et al., 2006). Therefore, classification of documents requires a highly dimensional feature space with scarce data. As dimensionality increases, the space of the scarce data increases. The increase in dimensionality and scarcity makes the classification problem harder to solve.

This study explores the effectiveness of three popular classification algorithms to classify Arabic text documents. The three classifiers under study are Support Vector Machine (SVM), Naïve Bayes (NB), and *K*-Nearest Neighbors (KNN).

This study also identifies the effects of reducing the advantage of these classification algorithms. Data mining and text mining are very important because they can handle the rapid growth of data that are collected and stored into large and numerous databases. These databases exceed human ability for comprehension, classification, and organization without the aid of powerful tools. Data mining is necessary for turning data that is stored in these databases into useful information which may help in decision making (Karima et al., 2005). One of the significant applications of data mining is text classification. Text classification aims to automatically assign

2. RELATED WORKS

This section presents some studies that are relevant to the present study. Therefore, only the studies that tested the effectiveness of different classifiers on a collection of Arabic documents are presented.

(Al-Kabi et al., 2007) investigated the effectiveness of six classification methods (i.e., inner product, cosine, Jaccard, Dice, NB, and Euclidean). They computed inner product, cosine, Jaccard, and Dice as associative coefficients of the vector space model (*VSM*). Their findings show that cosine is the most effective method. Furthermore, they concluded that NB is better than the other five methods tested in their study.

(Gharib et al., 2009) used SVM to classify 1,132 Arabic documents. They compared the results that they obtained from SVM with those from NB, KNN, and Rocchio. Their comparisons show that Rocchio is the best classifier for small feature sets, whereas *SVM* is the best classifier for large feature sets.

(El-halees, 2011) adopted a combined approach to extract opinions from Arabic documents. To enhance the performance of algorithms in classifying Arabic documents, he adopted a combined approach that consists of three methods. The lexicon-based method was used first. The resultant categorized documents were used as a training set for maximum entropy method, which subsequently classified other documents. Lastly, KNN was used to classify documents, which underwent both the lexicon-based method and the maximum entropy method. The results of his experiment show significant improvements in the performance of KNN.

(Wahbeh et al., 2012) conducted a comparative study on four free data mining tools (i.e., WEKA, Orange, KNIME, and Tanagra) for text classification. They concluded that WEKA is the best method.

(Khorsheed et al., 2013) conducted a study that involves a survey of Arabic text classification and a comparison of the effectiveness of different methods for Arabic text classification. They concluded that SVM is the best method, followed by the decision tree algorithm (C4.5), and then NB.

(Hanandeh and Mamoun., 2014) Conducted a study that aims at investigating different variations of vector space models (VSMs) using KNN algorithm. The Experimental results against the Saudi data sets reveal that Cosine outperformed over of the Dice and Jaccard coefficients.

3. EXPERIMENT RESULTS

The analysis of the results presented in the literature show that:

1. No standard Arabic corpuses are available that can be used easily. Almost all of the authors justified this finding by the lack of Arabic corpus in general (Duwairi, 2011) (Karima r al., 2005).

2. The technique of removing stop words, digits, numbers, punctuation marks, and non-Arabic words was used to prepare the text for classification. Some of the authors extracted root words (Duwairi, 2011)(Gharib et al., 2009) (Mesleh, 2008). Whereas the others preferred not to do so (Ababneh et al., 2014) (Alsaleem, 2011) (Al-Harbi et al., 2008) (Bawaneh et al., 2008). Because of the problem on the conflation of numerous terms to the same root word (Khreisat, 2009)

3. Almost all of the authors used recall, precision, and F1 (Ababneh et al., 2014) (Gharib et al., 2009) (Bawaneh et al., 2008).

4. Differences were observed among classifiers in terms of accuracy, error rate, and time taken to build the classification (Wahbeh et al., 2012).

5. SVM outperformed KNN and NB (Agirre et al., 2009) . To assess the accuracy of the proposed classifiers, Arabic text corpus was collected from online magazines and newspapers. A total of 1,000 document with varying lengths and writing styles were collected. These documents fall into 10 pre-defined categories. Every category contains 100 documents.

The set of pre-defined categories include sports, economy, Internet, art, animals, technology, plants, religion, politics, and medicine. Two authors manually categorized the collected documents. Every document was assigned to only one category. Whenever a document was found to belong to more than one category, it was assigned to the category with the maximum likelihood according to human categorizer's judgment. The accuracy of a classifier is expressed in terms of recall and precision. Figure 1 shows the recall and precision values for different values of *k*.

Figure 1: Recall & Precision for k10, k20, k50, ad k100

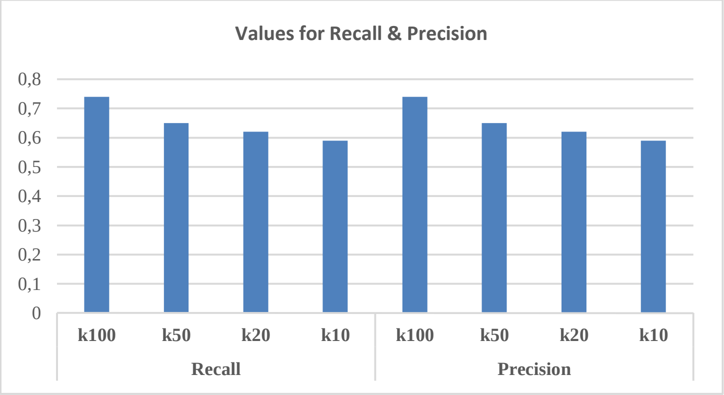

Figure 2 shows the recall and precision for all categories by all classifiers. The figure shows that the recall and precision values of SVM are better compared with NB and KNN.

Figure 2: Recall & Precision for SVM, NB, and KNN

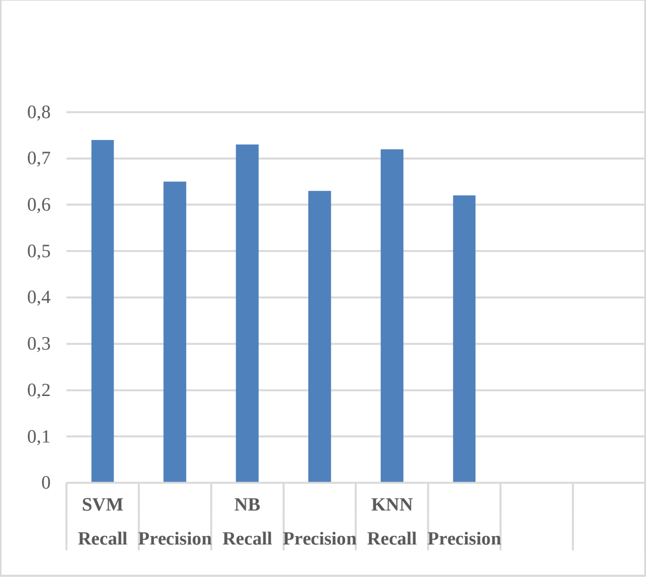

Table 1 shows the recall and precision values for all categories on the dataset for the three classifiers. The table shows that SVM has the best recall and precision values, followed by NB, and then KNN.

TABLE 1: VALUES OF RECALL AND PRECISION FOR THREE CLASSIFIER

Category Name	**SVM**		**NB**		**KNN**	
	Precision	Recall	Precision	Recall	Precision	Recall
Sports	0.97	0.891	0.962	0.884	0.95	0.871
Internet	0.662	0.241	0. 643	0.231	0.634	0.21
Art	0.412	0.862	0.402	0.842	0.386	0.811
Economy	0.401	0.978	0.392	0.964	0.378	0.943
Animals	0.921	0.674	0.903	0.654	0.896	0.642
Plants	0.941	0.593	0.933	0.587	0.912	0.567
Technology	0.492	0.398	0.478	0.379	0.463	0.362
Politics	0.994	0.454	0.983	0.441	0.976	0.434
Religion	0.873	0.601	0.865	0.599	0.851	0.583
Medicine	0.795	0.697	0.787	0.682	0.772	0.671

Another measure that was obtained from the experiments is the amount of time taken to build the models, which are used for testing the accuracy of the classifiers. This measure illustrates that NB takes the shortest amount of time to build the model, followed by KNN, and then SVM.

4. CONCLUSIONS

This study aims to compare three classification techniques using Arabic text documents which fall under four classes. The comparison is based on two main aspects of the classifiers: accuracy and time. In terms of time, results show that NB takes the shortest time to build the model, followed by KNN, and then SVM. On the other hand, SVM achieves the highest accuracy, followed by NB, and then KNN.

REFERENCES

Ababneh, J., Almomani, O., Hadi, W., El-Omari, N.K.T., and Al-Ibrahim, A., "Vector Space Models to Classify Arabic Text," International Journal of Computer Trends and Technology (IJCTT), vol 7, 2014

Agirre E., Lacalle O., and Soroa A., "Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD," *in Proceedings of the 21st International Joint Conference on Artificial Intelligence*, San Francisco, USA, pp. 1501-1506, 2009.

Alsaleem, S., " Automated Arabic Text Categorization Using SVM and NB," International Arab Journal of e-Technology,
Vol. 2, 2011

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S. and Al-Rajeh, A. "Automatic Arabic Text Classification," Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, Lyon-France, 2008

Al-Kabi, M. N., & Al-Sinjilawi, S. I. (2007). a Comparative Study of the Efficiency of Different Measures To Classify Arabic Text. *University of Sharjah Journal of Pure & Applied Sciences*, *4*(2), 13–26.

Bawaneh, M.J., Alkoffash, M.S., and Al Rabea A.I."ArabicText Classification using K-NN and Naive Bayes". Journal of Computer Science, vol. 4, 2008.

Duwairi, R. "Arabic Text Categorization," The International Arab Journal of Information Technology, Vol. 4, 2007.
El-halees, A. (2011). Arabic Opinion Mining Using Combined Classification Approach. *Proceeding The International Arab Conference On Information Technology, Azrqa, Jordan.*

Gharib, T. F., Habib, M. B., & Fayed, Z. T. (2009). Arabic Text Classification Using Support Vector Machines. *International Journal of Computers and Their Applications*, *16*(4), 192–199. Retrieved from http://purl.utwente.nl/publications/75679

Hanandeh E., Mamoun S.The Automated VSMs to Categorize Arabic Text Data Sets,INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY,VOL 13, NO 1 (2014)MARCH-2014.PP.4047-4081

Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources and Evaluation*, *47*(2), 513–538. http://doi.org/10.1007/s10579-013-9221-8
Khreisat, L. "A machine learning approach for Arabic text classification using N-gram frequency statistics," Journal of Informatics, Volume 3, 2009.

Karima, A, Zakaria, E and Yamina, T.G. "Arabic Text Categorization: A Comparative Study of different Representation Model, " Journal of Theoretical and Applied Information Technology, Vol. 38, 2005.

Mesleh, A.M.A. Support Vector Machine text Classifier for Arabic Articles: Ant Colony Optimization-based Feature Subset Selection., The Arab Academy for banking and financial Science, PHD. Thesis, 2008.

Syiam. M. M., Z. T. Fayed & M. B. Habib. An intelligent system for Arabic text categorization. IJICIS, Vol.6, No. 1 JANUARY 2006.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, , M. N., & Al-Shawakfa, E. M. (2012). A Comparison Study between Data Mining Tools over some Classification Methods. *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, , 2*(8), 19–26.