https://doi.org/10.5281/zenodo.2544629

Arabic Text Categorization Using Three Classifiers Methods: A Comparative Study

Essam Hanandeh[1]

[1] **Computer Information Science, Zarqa University,
Zarqa, Jordan**

Abstract

Today, text categorization is usually used in various areas, such as: information retrieval, data mining and text mining. The present study aims to test the K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM) algorithms on a relatively large dataset of Arabic documents. The latter dataset includes 1,000 arabic documents that are distributed across 10 classes. The latter test is based on recall and precision measures. It was found that Supporting Vector Machine algorithms classifier outperforms the other ones.

Keywords: *Arabic text categorization, Naïve Bayes Naïve, K-Nearest Neighbors, and Support Vector Machine, text mining.*

1. Introduction

The classification considered an important part of Machine Learning, data mining, and text mining. In classification, a machine or human attempts to identify to which class from a set of classes a new instance belongs [8]. 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 the present study, each instance represents a text document [2]. Therefore, the classification of documents requires a higher 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.

The present study aimed to:

1. Exploring the effectiveness of three popular classification algorithms for classifying Arabic text documents. The classifiers that the present study aims to examine are: *K*-Nearest Neighbors (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM).
2. Exploring the effectiveness of three classification algorithms for classifying Arabic text documents. The classifiers that the present study aimed to examine are: *K*-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Naïve Bayes (NB)
3. Exploring the effects of reducing the advantage of these classification algorithms.

Data mining and text mining are considered very significant. That is because they are capable to handle the rapid growth of data which is collected and stored within large databases. Humans are not capable to organize, comprehend and classify these databases without using tools. One of the most important role in turning the data stored within those databases into useful information is data mining. Decision makers can utilize such information [1]. Also, one of the significant data mining applications is text classification.

.

2. The Relevant Literature

The present section presents some studies that are related to the problem of the present study. In other words, the present study presents the studies that deal with the effectiveness of classifiers in classifying texts.

[3] Aimed to explore the effectiveness of six classification methods (i.e. cosine, inner product, Jaccard, NB, Dice, and Euclidean). They computed cosine, inner product, Jaccard, and Dice as associative coefficients of the vector space model (*VSM*). Their findings indicate that cosine is the most effective method. Furthermore, they concluded that NB is better than the other five methods.

[5] Aimed to conduct a comparison between the effectiveness of different methods used for Arabic text classification. They concluded that SVM is ranked first in terms of effectiveness. They also concluded that the decision tree algorithm (C4.5) is ranked second in terms of effectiveness. They also found that NB is ranked third in terms of effectiveness.

[9] Employed SVM for classifying 1,132 Arabic documents. They compared the results that they obtained from SVM with the results obtained 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.

[12] Extracted opinions from Arabic documents through using a combined approach. In order to develop the performance of algorithms in classifying Arabic documents, the latter approach was adopted. It involves using 3 methods. The lexicon-based method was used first. The other methods include the lexicon-based method and the maximum entropy method. The results indicate that there is a significant improvement in the performance of KNN.

[13] Aimed to conduct a comparison between four data mining instruments (i.e. Orange, WEKA, KNIME, and Tanagra) for text classification. They concluded that WEKA is the most effective method

[15] Aimed to explore the difference between several vector space models (VSMs). He aimed to explore that through using KNN algorithm. It was found that Cosine outperforms Jaccard and Dice. [19] Aimed to shed light on the classification of Arabic texts. He conducted a comparison between five well-known algorithms used for text classification. He also aimed to explore the impacts of using various Arab logs (light stem and roots) on the effectiveness of these works. Furthermore, he conducted a comparison between several software tools used for data mining. The results indicate that a good resolution is provided by SVM workbook, especially when used along with the light swamp 10. [21] Aimed to explore the effectiveness of several methods used for text classification.She found that SEC is more effective than WBC and Al-Kabi for text classification.

3. Experiment Results

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

1. There isn’t any standard Arabic corpus that can be used easily. Most researchers attributed this results to the lack of Arabic corpus [1, 4].

2. Most of the researchers used precision, recall and F1 [9,14,16] .

3. Differences were observed among classifiers in terms of error rate, accuracy, and time required for building the classification [13].

4. The technique used for removing stop words, numbers, punctuation marks, digits, and non-Arabic words - was used for preparing the text for classification. Some of the authors extracted root words [4,6,9]. Whereas the others preferred not to do so [11,14]. Because of the problem on the conflation of numerous terms to the same root word [18].

5. SVM outperforms NB & KNN [8] . In order to evaluate the accuracy of the proposed classifiers, Arabic text corpus was collected from online newspapers and magazines. A total of 1,000 documents – which writing styles and lengths vary - were collected. Those documents fall into ten (10) pre-defined categories. Every category includes one hundred (100) documents.

The set of pre-defined categories include economy, sports, art, Internet, animals, technology, religion, plants, medicine, and politics. Two researchers categorized the collected documents manually. Each document was assigned to 1 category only. When finding a document that belongs to multiple categories, the document was assigned to the category that has the maximum possibility based on the judgment of the human categorizer. The classifier’s accuracy is represented in precision and recall. Fig 1. Presents the precision & recall values of different values of *k*.

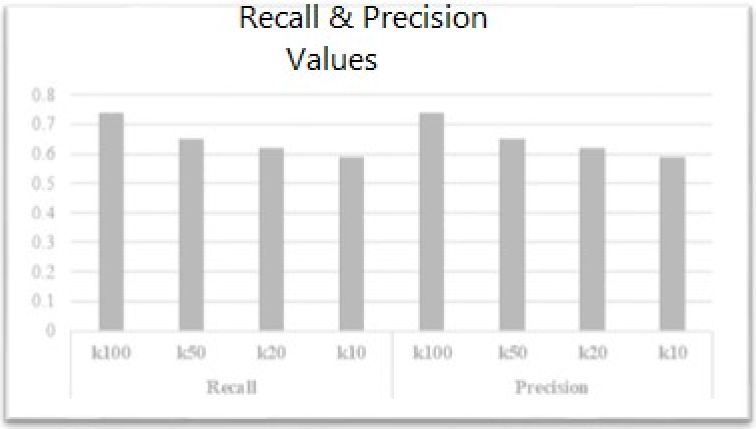

Fig 1. Recall & Precision for k10, k20, k50, ad k100

Fig 2. Presents the precision and recall values of all the three classifiers. It also indicates that the precision and recall values of SVM are better than the counterpart values of KNN and NB.

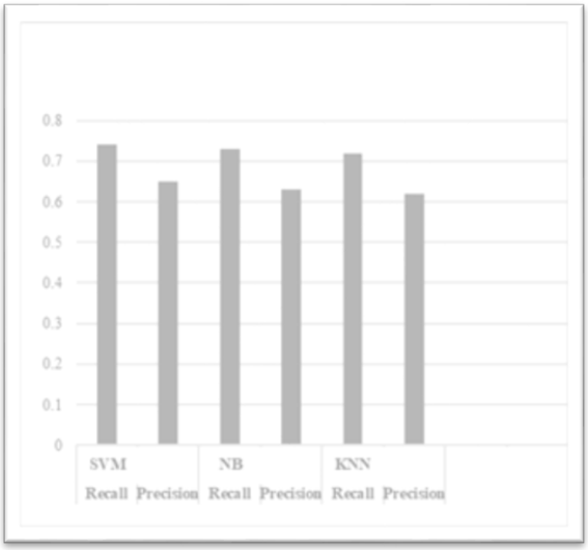

Fig 2. Recall & Precision for SVM, NB, and KNN

Table 1 presents the precision and recall values of the three classifiers. The table shows that SVM has the best recall and precision values, followed by NB, and then KNN.

TABLE I: 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

The comparison was conducted between the three classification techniques in terms of the time required for building the models. These models are used for testing the classifiers' accuracy. It was found that NB requires the least time for building the model, followed by KNN, and SVM respectively

4. Conclusion

The present study aims to conduct a comparison between 3 classification techniques through using Arabic text documents which fall under 4 classes. The latter comparison was conducted in term of the classifiers' time and accuracy. It was found that NB requires the least time for building the model, followed by KNN, and SVM respectively. It was found that SVM shows the highest accuracy, followed by NB, and KNN respectively.

References

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

[2] M. Syiam., Z. T. Fayed and M. B. Habib, "An intelligent system for Arabic text categorization." IJICIS, Vol.6, No. 1 JANUARY 2006.

[3] M. Al-Kabi, & S. I. Al-Sinjilawi, "a Comparative Study of the Efficiency of Different Measures To Classify Arabic Text." University of Sharjah Journal of Pure & Applied Sciences, vol 4, pp. 13–26, 2007.

[4] R. Duwairi, "Arabic Text Categorization," The International Arab Journal of Information Technology, Vol. 4, 2007.

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

[6] A. Mesleh, "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.

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

[8] E. Agirre, O. Lacalle, and A. Soroa, "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.

[9] T. Gharib, M. Habib, & Z. Fayed, "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/75679,2009.

[10] L. Khreisat, "A machine learning approach for Arabic text classification using N-gram frequency statistics," Journal of

Informatics, Volume 3, 2009.

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

[12] A. El-halees, "Arabic Opinion Mining Using Combined Classification Approach. *Proceeding The International Arab Conference On Information Technology, Azrqa, Jordan.2011.*

[13] A. Wahbeh, Q. Al-Radaideh, M. Al-Kabi, & E. Al-Shawakfa, "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), pp.19–26.2012.

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

[15] E. Hanandeh, & M. Rababah, "The Automated VSMs to Categorize Arabic Text Data Sets", INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY,VOL 13, NO 1 (2014)MARCH-2014.PP.4047-4081

[16] S.Charanyaa and K.Sangeetha, "Term Frequency Based Sequence Generation Algorithm for Graph Based Data Anonymization, International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 2, February 2014, ISSN(Online): 2320-9801

[17] N. A. Abdulla, M. Al-Ayyoub, and M. N. Al-Kabi. "An extended analytical study of arabic sentiments. International Journal of Big Data Intelligence, 1(1):103–113, 2014

[18] A. Alwajeeh, M. Al-Ayyoub, and I. Hmeidi. "On authorship authentication of arabic articles. In Information and Communication Systems (ICICS), 2014 5th International Conference on, pages 1–6. IEEE, 2014

[19] H. Ismail, "Automatic Arabic text categorization: A comprehensive comparative study " *Journal of information Science, vol 41,*pp. 114-124, Feb 2015

[20] B. Al Shboul, M. Al-Ayyoub, and Y. Jararweh. "Multi-way sentiment classification of arabic reviews. In Information and Communication Systems (ICICS), 2015 6th International Conference on, pages 206–211. IEEE, 2015

[21] A. Nabil, M. Mohammed*, Q. Said, E. NourEddine,* "Impact of stemming on Arabic text summarization", *Information Science and Technology(CiSt) 2016 4th IEEE International Colloquium* , pp. 338-343, 2016

[22] Fawaz S. Al-Anzi, Dia AbuZeina, "A Micro-Word Based Approach for Arabic Sentiment Analysis", *Computer Systems and Applications (AICCSA) 2017 IEEE/ACS 14th International Conference on*, pp. 910-914, 2017