Modern Applied Science; Vol. 13, No. 11; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Application of Naïve Bayes, Decision Tree, and K-Nearest Neighbors for Automated Text Classification

Jafar Ababneh[1]

[1]The world Islamic science & education university W.I.S.E, Department of Computer Networks Systems, Jordan

Correspondence: Jafar Ababneh, The world Islamic science & education university W.I.S.E, Department of Computer Networks Systems, Jordan. E-mail: jafar.ababneh@wise.edu.jo

Received: August 29, 2019 Accepted: October 5, 2019 Online Published: October 6, 2019
doi:10.5539/mas.v13n11p31 URL: https://doi.org/10.5539/mas.v13n11p31

Abstract

Nowadays, many applications that use large data have been developed due to the existence of the Internet of Things. These applications are translated into different languages and require automated text classification (ATC). The ATC process depends on the content of one or more predefined classes. However, this process is problematic for the Arabic translation of the data. This study aims to solve this issue by investigating the performances of three classification algorithms, namely, k-nearest neighbor (KNN), decision tree (DT), and naïve Bayes (NB) classifiers, on Saudi Press Agency datasets. Results showed that the NB algorithm outperformed DT and KNN algorithms in terms of precision, recall, and F1. In future works, a new algorithm that can improve the handling of the ATC problem will be developed.

Keywords: arabic data sets, data mining, decision tree, K-nearest neighbors, NB algorithms, text classification, term weighting

1. Introduction

Given the increasing global utilization of the Internet of Things, relevant data are being translated into different languages (e.g., English, French, and Arabic). For the Arabic language, the translation demonstrates high accuracy and small time delay.

Automated Text Classification has been a challenging research problem especially in information retrieval (IR) and data mining communities (Hadi, Salam, & Al-Widian, 2010). This problem was dynamic for quite a lot of years, and in recent times engrossed several scholars due to many of articles accessible on the internet, organization, and digital libraries (Thabtah, Hadi, Abdelhamid, & Issa, 2011), Researchers have conducted several studies on text classification and text mining to determine the patterns in textual data (Salloum, AlHamad, Al-Emran, & Shaalan 2018).

Automated text classification involves predicting text articles based on the content through experimental data gathering from one or more predefined classes. Unlike manual text classifications, which is time-consuming and necessitates high precision and accuracy, ATC is accurate and fast due to the automation of the procedure.

Text classification is the process in which articles are predicted into one of the predefined classes on the basis of their contents. For example, the classes of a newspaper article can be society, politics, general knowledge, technology, sport, economics, and so on. This process has several uses in many applications, such as spam detection, digital library classification, protein classification, and web mining (Cheng , Carbonell, & Klein-Seetharaman, 2005; Man Li, et al., 2018; Suprativ&Tanmay,2019). These applications are becoming increasingly significant in today's information-oriented (Hadi, Salam, Al-Widian, 2010; Mohammad et al., 2016).

ATC techniques are valuable for plain-text medical documents, in which features are extracted from free text reports (Ghulam , Liyana, Ram, Rajandram, & Khairunisa, 2018). In addition, text classification can be utilized by machine learning methods (Agarwal & Mittal, 2014).

Sebastiani (2002), automated text classification problem can be defined as follows: The articles separated in two folds, one fold for learning (training data) and the other for evaluation (testing data).

Let training data = {a_1, a_2, … , a_n}, where n articles are used as samples for the algorithm, and must contain sufficient number of positive samples for all the classes involved. The testing data {a_{n+1}, a_{n+2}, … , a_m} used to evaluate the algorithm performance. The matrix displayed in Table 1 represents samples dividing into training and testing. An article a_x is considered a positive sample to C_k if C_{kx} =1 and a negative sample if C_{kx} =0.

Table 1. Representation of text classification problem

Class	*Training data*			*Testing data*		
	a_1	….	a_n	a_{n+1}	…	a_m
C_1	C_{11}	….	C_{1n}	$C_{1(n+1)}$	….	C_{1m}
…..	…..	…..	…..	…..	…..	…..
C_y	C_{y1}	…..	C_{yn}	$C_{y(n+1)}$	…..	C_{ym}

The text classification process includes three main phases: preprocessing, which mainly involves collecting relevant documents, learning, and evaluation. Preprocessing is necessary to ensure that a text article is suitable to train the algorithm. Then, the model is created and modified using a learning approach on the training data. Finally, the model is evaluated using various evaluation measures, such as recall, precision, accuracy, and F1 (Hadi, 2015). Figure (1) presents the details of the phases.

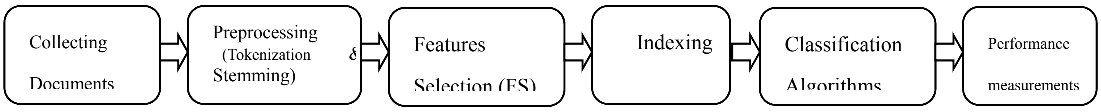

Figure 1. text classification process phases

Several text classification algorithms from machine learning and data mining communities exist such as: Support Vector Machines (SVMs) (Hadi, Salam & Al-Widian, 2010), decision trees (DTs), (Mohammad et al., 2016), multi-class classification based on association rules (MCAR) (Thabtah, Cowling, & Yonghong, 2005), and K-Nearest Neighbors (KNN) (Mohammad et al., 2016). The main goal of this study is to present and investigate results achieved against Arabic text collections using NB, KNN, and decision tree algorithms. The evaluation measures include recall, precision, and F1. In other words, this study aims to determine the algorithm that generates the best classification results.

The remainder of this paper is organized as follows: Section 2 discusses the related works. Section 3 introduces the three well-known algorithms for text classification. Section 4 presents the experiment results. Finally, Section 5 provides the conclusions and future works.

2. Related Works

(Mohammed et al., 2016) adopted three well-known algorithms, namely KNN, C4.5, and Rocchio algorithms, to classify 1,400 Arabic text datasets. These datasets were categorized into eight classes: computer (كمبيوتـــــر), economics (إقتصـــــاد), education (تعليـــــم), law (قـــانون), medicine (طبيـــة), politics (سياســـــية), religion (دينيـــة), and sports (رياضـــــية). The results obtained through Rocchio and KNN algorithms are similar. Moreover, both algorithms outperform the C4.5 algorithm in terms of recall and precision measures

(Sallam, Mousa, and Hussein, 2016) proposed automated Arabic text classification approach uses frequency ratio accumulation method (FRAM), and evaluated on three different Arabic datasets. The results achieved from the FRAM algorithm in terms of Accuracy, Recall, Precision, and F1 are generally improved and comparable to current text classification techniques.

(Ababneh et al., 2014) implemented three popular coefficient methods, namely, Cosine, Dice, and Jaccard, using the KNN algorithm based on inverse document frequency term weighting method. The experimental results from the Saudi Newspapers (SNP) datasets indicated that the Cosine coefficient outperformed Dice and Jaccard coefficients with regard to the average values of precision, recall, and F1.

(Alnababteh et al., 2014) modified the Classification Based on Association algorithm to concur the data insertion aspect. They tested this new algorithm in six UCI Repository datasets. The modified algorithm exhibited better results than the original one due to incremental learning in associate classification mining. However, the accuracy of the two algorithms is the same because of computational time delay. The new algorithm is then called Associative Classification based on Incremental Mining.

Three associative classification prediction methods, namely, full match rule, dominant class label, and average confidence per class, were tested and evaluated by (Thabtah et al., 2011) by using Reuters and Saudi Press Agency (SPA) dataset. They compared the three methods with SVMs, KNN, MCAR, NB, and C4.5 algorithms. The experimental results revealed that the proposed methods produced more competitive breakeven point measure than MCAR. Furthermore, the three generally outperformed the other classical algorithms.

(Alwedyan, Hadi, Salam, and Mansour, 2011) investigated the MCAR, NB, and SVM algorithms on SNP datasets. They reported that the MCAR algorithm produced higher precision, F1, and recall scores at 8.5%, 7.6%, and 7.3% and 4.7%, 3.8%, and 3.5% than NB and SVM, respectively.

(Hadi, Salam, and Al-Widian, 2010) used NB and SVM to classify 2,244 Arabic textual datasets called Islamic datasets (Al-Harbi, Almuhareb, & Al-Thubaity, 2008). These datasets were categorized into five classes: Feqah (الفقـــة), Tafseer (التفســـــــير), Hadeeth (الحـــديث), Lughah (اللغـــة), and Aqeedah (العقيـــدة). The comparison results indicated that the SVM classifier outperformed the NB classifier in terms of recall, precision, and F1.

3. Proposed Algorithms

Numerous classification algorithms have been developed in the recent years to achieve growth in the field of text classifications. These algorithms include KNN, SVM, neural network, Rocchio, NB classifier, and DT. The following subsections will discuss the three algorithms that are utilized in this study: DT, NB, and KNN.

3.1 Decision Tree Algorithm

Decision trees have been used for various purposes, such as pattern matching, text classification, and machine learning.

A DT essentially classifies the training data into sets. These sets are formed by branching on the attribute values of the examples in the training data. A perfect DT is constructed when all training examples at a node in the tree are under the same classification. However, this phenomenon rarely happens because of the presence of a few outliers due to noise (Wahiba & Ahmed, 2016). The prediction at the node is represented by the majority of the classification of various examples with incorporated bias. When the formation of the tree is completed, the prediction can be performed. Given a piece of data, a path from the root to the leaf of the tree is traversed by taking the edges corresponding to the value that the data depicts for the attribute at the node. The prediction of the leaf node is the returned prediction (Mohammad et al., 2016; Kim, 2016; Khedr, Idrees, & El Seddawy, 2016).

However, DT is inapplicable for large numbers of attributes because the nodes will explode if every attribute is branched. For example, for 30 binary attributes, the total number of leaf nodes is 230. The chi square value is then used to test for statistical significance (Russell & Norvig, 2009). The attribute that will maximize the amount of information gain will be selected (Thabtah et al., 2011).

3.2 Naïve Bayes Algorithm (NB)

NB classification is a popular algorithm that is widely used due to its excellent performance in many applications, including text classification, phishing detection, medical diagnosis (Sonia & Maheshwar, 2019), job search engine (Slamet, Andrian, Maylawati, Suhendar, Darmalaksana & Ramdhani, 2018)and spam detection. This algorithm is also characterized by simplicity and acceptable efficiency. NB estimates the probability, as shown in Equations 1 and 2, that an instance x belongs to a class y and predicts the class with the highest value of P(y|x).

$$\mathrm{P(y\backslash x)} = \frac{\mathrm{P(y)}}{\mathrm{P(x)}}\mathrm{P(x\backslash y)} \tag{1}$$

$$\mathrm{P(y\backslash x)} = \frac{\mathrm{P(y)}}{\mathrm{P(x)}}\prod_{i} \quad \mathrm{P(x_i\backslash y)} \tag{2}$$

Equations (1) and (2) represent the Bayes theorem and the NB assumption, respectively. The latter assumption is made because learning P(x|y) for all x is difficult, unlike learning $P(x_i|y)$. For a binomial attribute x_i, learning is simply counting the number of occurrences in the training set S of x_i in each class y, as well as the number of instances in each class.

Even in basic form, the NB classifier performs satisfactorily. However, the greatest weakness of this algorithm lies in its simplicity, in which each attribute is assumed to be autonomous. Another drawback of NB is that when the training data set is skewed (e.g., no training instance is obtained for a particular class), the performance of the algorithm is not that satisfactory. For such cases, as suggested in (Sallam, Mousa & Hussein, 2016), we assign $P(x_i|y) = \varepsilon > 0$.

3.3 K-Nearest Neighbors Algorithm (KNN)

Due to the feature space and depending on closest training samples, this algorithm can predict and classify objects. A prediction depends on the nearest neighbor is given with percentage of confidently, this prediction result obtained firstly by checking the feature space, this is really how KNN algorithm method work (Deng, Zhu, Cheng, Zong, & Zhang, 2016).

The KNN algorithm is considered lazy learning because it rely on predictions from just only specific selection of instances most similar to the test set instance, (Gongde, Hui, David, & Yaxin, 2004; Sonia & Maheshwar 2019). beside that it is the most simple method or algorithm for data mining and machine language, (Mohammad et al., 2016).

The training instances are described by n-dimensional numeric features, and each instance represents a point in an n-dimensional space; all training instances are stored in an n-dimensional pattern space. For an unseen instance (test), the KNN algorithm searches the pattern space for the K-training instances that are closest to the unseen instance, and these instances are the “nearest neighbors” (Hadi, 2015).

4. Experiments Results

In this study, three well-known data mining algorithms, namely Decision tree, KNN, and NB algorithms are used to classify 1562 Arabic articles collected from Saudi Press Agency (SPA) (Al-Harbi, Almuhareb & Al-Thubaity,2008), SPA datasets are categorized into six classes: Culture news ,"اخبار ثقافية" Sport news "اخبار رياضية",Social news ,"اخبار إجتماعية" Economics news ,"اخبار إقتصادية" Political news ,"اخبار سياسية" and General news."اخبار عامة"

The investigation is conducted using Weak software (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) , and the classification results are evaluated through a 10-fold cross-validation. As previously mentioned, three different assessment measures are used to assess the performance of the three algorithms: recall, precision, and F1. The recall for a class is defined as the percentage of correctly classified articles among all articles belonging to that category equation (3), the precision is the percentage of correctly classified articles among all articles that were assigned to the class by the algorithm equation (4), and the F1 score is the harmonic mean of the two measures equation (5).

Figure (2) shows that the NB algorithm produces better classification results than KNN and DT algorithms in terms of all measures. NB obtained a 3.8% and 6.2% higher recall than KNN and DT, respectively. In addition, NB outperformed KNN and DT by 2.9% and 5.8%, respectively, in terms of precision and 3.35% and 6%, respectively, in terms of F1 score. However, the general news (اخبـــار عامه) class obtained unacceptable results, which can be attributed to the extreme association of the terms in the general news

class to the terms in other classes. In conclusion, data mining and machine learning algorithms perform well in classifying Arabic articles.

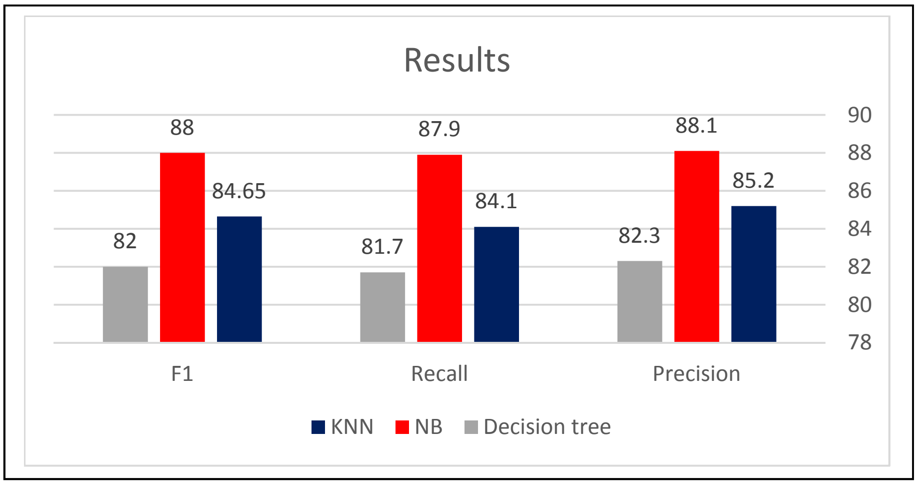

Figure 2. Classification results of KNN, NB, and Decision tree Algorithms

.

$$F1 = \frac{2*Precision*Recall}{Recall+precision} \tag{3}$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

5. Conclusions

Building an ATC to classify text articles on the basis of the appropriate classes is considered a serious problem, especially for Arabic texts. In this study, we utilized KNN, DT, and NB algorithms to address this problem by using the SPA dataset. The results showed that the NB algorithm outperformed KNN and DT algorithms in terms of all evaluation measures (precision, recall, and F1 score). In future works, a new algorithm that can further improve ATC will be developed.

References

Ababneh, J., Almomani O., Wael H., El-Omari N. K. & Al-Ibrahim, A. (2014). Vector Space Models to Classify Arabic Text, *International Journal of Computer Trends and Technology (IJCTT), 7*(4), 219-223, Published by Seventh Sense Research Group, http://doi.org/ 10.14445/22312803/IJCTT-V7P109.

Agarwal B. & Mittal N. (2014). Text Classification Using Machine Learning Methods-A Survey. In: Babu B. et al. (eds) Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. *Advances in Intelligent Systems and Computing, 236*, Springer, New Delhi, https://doi.org/10.1007/978-81-322-1602-5_75.

Al-Harbi, S, Almuhareb, A, Al-Thubaity, A, Khorsheed, M. S. & Al-Rajeh, A. (2008). *Automatic Arabic Text Classification.* Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data, France, https://eprints.soton.ac.uk/id/eprint/272254.

Alwedyan, J., Hadi, W. M., Salam, M. & Mansour, H. Y. (2011). *Categorize arabic data sets using multi-class classification based on association rule approach*. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications - ISWSA '11, 1-8, New York, New York, USA: ACM Press. http://doi.org/10.1145/1980822.1980840.

Betty Yee Man Cheng, Jaime Gual Carbonell & Judith Klein-Seetharaman (2005). Protein classification based on text document classification techniques. *Proteins Structure Function and Bioinformatics, 58*(4), 955-70 , https://doi.org/10.1002/prot.20373 ·

Deng, Z., Zhu, X., Cheng, D., Zong, M. & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing, 195*, 143-148. http://doi.org/10.1016/j.neucom.2015.08.112.

Ghulam Mujtab, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram & Khairunisa Shaikhde (2018). Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study, *Journal of Forensic and Legal Medicine, 57*, 41-50, https://doi.org/10.1016/j.jflm.2017.07.001.

Gongde Guo, Hui Wang, David A. Bell & Yaxin Bi. (2004). KNN Model-Based Approach in Classification, https://doi.org/10.1007/978-981-13-2354-6_12.

Hadi, W. (2015). ECAR : A New Enhanced Class Association Rule. *Advances in Computational Sciences and Technology, 8*(1), 43-52, Research India Publication http://www.ripublication.com

Hadi, W. M., Salam, M. & Al-Widian, J. A. (2010). *Performance of NB and SVM Classifiers in Islamic Arabic Data.* Proceedings of the 1st International Conference on Intelligent Semantic Web-Services and Applications, 14, http://doi.org/10.1145/1874590.1874604.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter, 11*(1), 10. http://doi.org/10.1145/1656274.1656278.

Khedr, A. E., Idrees, A. M. & El Seddawy, A. I. (2016). Enhancing Iterative Dichotomiser 3 algorithm for classification decision tree. *Data Mining and Knowledge Discovery, 6*(2), 70-79. http://doi.org/10.1002/widm.1177

Kim, K. (2016). A Hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recognition*. http://doi.org/10.1016/j.patcog.2016.04.016

Man Li (2018). Classification of G-protein coupled receptors based on a rich generation of convolutional neural network, N-gram transformation and multiple sequence alignments, *Amino Acids, 50*(2), 255- 266, https://doi.org/10.1007/s00726-017-2512-4.

Mohammad, A. H., Al-momani, O. & Alwada'n, T. (2016). Arabic Text Categorization using k-nearest neighbour , Decision Trees (C4 . 5) and Rocchio Classifier: A Comparative Study. *International Journal of Current Engineering and Technology, 6*(2), 477-482.

Mohammed H. Alnababteh , M. Alfyoumi, A. Aljumah & J. Ababneh,(2014). Associative Classification Based on Incremental Mining (ACIM), International Journal of Computer Theory and Engineering (IJCTE (International Association of Computer Science and Information Technology Press (IACSIT Press), *Singapore, 6*(2), 135-140, https://doi.org/10.7763/IJCTE.2014.V6.851.

Russell, S. & Norvig, P. (2009). Artificial Intelligence: A Modern Approach. Pearson; 3 edition. Retrieved from http://aima.cs.berkeley.edu/

Sallam, R., Mousa, H. & Hussein, M. (2016). Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computer Applications, 135*(2), 38-43. http://doi.org/10.5120/ijca2016908328.

Sallam, R., Mousa, H. & Hussein, M. (2016). Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computer Applications, 135*(2), 38-43. http://doi.org/10.5120/ijca2016908328

Salloum S.A., AlHamad A.Q., Al-Emran M., Shaalan K. (2018). A Survey of Arabic Text Mining. "Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, *Springer, 740*, Cham., https://doi.org/10.1007/978-3-319-67056-0_20

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47. http://doi.org/10.1145/505282.505283.

Slamet C., R Andrian, D S Maylawati, Suhendar, W Darmalaksana & M A Ramdhani (2018). Web Scraping and Naïve Bayes Classification for Job Search Engine IOP Conf. *Series: Materials Science and Engineering, 288*(2018), 012038.

Sonia Goyal & Maheshwar (2019). Naïve Bayes Model Based Improved K-Nearest Neighbor Classifier for Breast Cancer Prediction, In book: Advanced Informatics for Computing Research https://doi.org/10.1007/978-981-15-0108-1_1.

Suprativ S., Tanmay B.(2019). *A Novel Approach to Find the Saturation Point of n-Gram Encoding Method for Protein Sequence Classification Involving Data Mining: Proceedings of ICICC 2018*, 2, In book: International Conference on Innovative Computing and Communication.

Thabtah, F., Cowling, P. & Yonghong Peng. (n.d.) (2005). *MCAR: multi-class classification based on association rule.* In The 3rd ACS/IEEE International Conference onComputer Systems and Applications, 2005, 130-136. IEEE. http://doi.org/10.1109/AICCSA.2005.1387030.

Thabtah, F., Hadi, W., Abdelhamid, N. & Issa, A. (2011). Prediction Phase in Associative Classification Mining. *International Journal of Software Engineering and Knowledge Engineering, 21*, 855-876. Retrieved from <Go to ISI>://INSPEC:12687537.

Vandana Korde, and C Namrata Mahender, (2012). EXT CLASSIFICATION AND CLASSIFIERS:A SURVEY, *International Journal of Artificial Intelligence & Applications (IJAIA), 3*(2), https://doi.org/10.5121/ijaia.2012.3208.

Wahiba, B. A. & Ahmed, B. E. F. (2016). *New Fuzzy Decision Tree Model for Text Classification*, 309-320. http://doi.org/10.1007/978-3-319-26690-9_28

Copyrights