International Journal of Engineering & Technology, 7 (3.20) (2018) 906-909

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Categorization Arabic Text Using SVM and KNN Algorithms

Geehan Sabah Hassan

College of Education for Women
Baghdad University
**Corresponding Author Email: gee.fah@gmail.com*

Abstract

Content arrangement is a strategy for marking regular dialect writings with one or a few classifications from a predefined set. Two calculations, to be specific, bolster vector machine (SVM) and k-closest neighbor (KNN), are utilized to examine Arabic content order (TC). Distinctive Arabic datasets are utilized to analyze the two calculations. This examination has been intended to order extraordinary Arabic content. Result demonstrates that TC by means of the SVM calculation beats TC by means of KNN regarding all measures.

Index Terms: *text classification (TC), Support Vector Machine (SVM), K–Nearest Neighbor (KNN).*

1. Introduction

A case of an utilization of programmed examination of records in data recovery and information mining is content characterization subjects. TC is a condition of regulated discovering that permits classification setting and straight out instances of reports. Programmed stockpiling and recovery strategies are essential when dealing with a lot of content information, and they turn out to be very proficient and viable with the help of TC. Unique in relation to manual order, TC requires high precision.

TC plans to name normal dialect writings with one or a few classes from a predefined set. Predefined classes are normally constrained, however different applications comprise of classifications that are framed by other criteria, for example, sort and email by need. This paper centers around single-mark task. Bolster vector machine (SVM) [1], choice trees [2], k-closest neighbor (KNN) [3] and neural system [4] are a few TC procedures from information mining and machine learning. The paper concentrated on the present, and the got outcomes were contrasted and Iraqi papers and Arabic content accumulations by utilizing SVM and KNN calculations. Our tenets look at the SVM and KNN, which are the most widely recognized content assessment measures (F1, review and exactness).

This paper is sorted out as pursues: Section 2 clarifies related works. Area 3 talks about the TC issue. Area 4 exhibits the analysis results. Area 5 gives the ends and forthcoming works.

2. Literature REVIEW

Arrangement is the purpose of contact for IR and machine learning (ML); ML and IR are both used to distribute watchwords for records and characterize them into classifications. Numerous analysts have led striking works here, yet every one has related at least one issues with respect to a similar subject. Many going before works are identified with this investigation.

For instance, an assessment was led by the creators of [6] utilizing SVM and guileless Bayesian (NB) for Arabic datasets gathered from a few Saudi papers (SNP) that utilization Arabic content. Results demonstrated that the SVM calculation beats the NB as far as all measures.

Shakers measures and Manhattan separate utilizing N-gram recurrence factual procedure were analyzed against Arabic datasets made from various sites out of Arabic paper in [7]. Results exhibited that N-gram using the Dice measure surpasses Manhattan separate.

The creators of [8] examined a load balanced KNN execution that distinguishes the ideal weight vector.

[1] Employed most extreme entropy for TC on Arabic datasets and found that the normal F-measure expanded from 68.13% to 80.41% utilizing pre-handling methods (standardization, stop words evacuation and stemming).

Highlight choice, which is frequently hurtful to SVM usage, was used by [1].

Taking 1445 writings from online Arabic paper, three TC calculations, in particular, SVM, KNN and NB, were utilized by [10]. The nine arrangements of gathered writings are as per the following: Computer, Economics, Education, Engineering, Law, Medicine, Politics, Religion and Sports. Highlight choice utilized chi-square insights. [10] expressed, 'Contrasted and another arrangement ways, our framework shows a high order approval for Arabic dataset as far as F-measure (F = 88.11)'.

The creators in [12] analyzed diverse varieties
of the vector space display, for example, cosine coefficient, dice coefficient and Jaccard coefficient, utilizing KNN calculation and diverse term weighting approach. The discoveries of the normal F1 procured against six Arabic datasets showed

that Dice-and Jaccard-based TF.IDF beat cosine-based TF.IDF, cosine-based WIDF, cosine-based ITF, cosine-based log(1 + tf), Dice-based WIDF, Dice-based ITF, Dice-based log(1 + tf), Jaccard-based WIDF, Jaccard-based ITF and Jaccard-based log(1 + tf). A review of various distributed papers in TC is talked about, and the utilization of the classifiers, to be specific, remove based, KNN and NB, is broke down in [13].

The NB calculation dependent on chi-square element choice technique was considered in [14]. A correlation of the exploratory outcomes with various Arabic TC datasets demonstrated that an expansion in grouping precision can be credited to the expulsion of uncommon terms by highlight determination. SVM for Arabic TC was at first displayed by Abdel wadood Moh'd A MESLEH [15].

The chi-square procedure was utilized by the creator for highlight choice. Stemming calculations were not used. Results recommend that SVM beat alternate calculations, to be specific, KNN and NB classifiers

3. Problem Categorization

TC is a strategy for naming regular dialect writings with one or a few classes from a predefined set. This task is identified with IR and ML. Computerized TC instruments are more appealing than manual categorisation of reports, which can be expensive or basically unconscionable on the grounds that the imperatives connected or the quantity of archives concerned [14].

TC can be utilized in numerous applications, such mechanized ordering of logical expositions based predefined thesauri of specialized terms, submit licenses in patent indexes, and disperse specific of data to buyers, computerized accumulation of various leveled lists of web assets, spam separating, assurance of record kind, attribution of initiation, review coding and even the auto-article evaluating. As indicated by [14], TC is a standout amongst the most broadly used key strategies for regulated learning in information mining. A general inductive process automatically produces a text classifier for $\mathbf{ci} \in \mathbf{C}$. This procedure is led by observing the properties of report sets before grouping as indicated by $\mathbf{c_i}$ **or** $\acute{\mathbf{c}}_\mathbf{i}$, which acquire the properties of another concealed record where $\mathbf{c_i}$ necessity fit. A document set S is necessary such that the value of $\boldsymbol{\Phi}(\mathbf{d_j}, \mathbf{c_j})$ is well known for every $(\mathbf{d_j}, \mathbf{c_j}) \in \boldsymbol{\Omega} \times \mathbf{C}$ to establish classifiers for C. $\mathbf{S}$ is generally divided into two distinct sets, namely, $\mathbf{T_r}$ (training set) and $\mathbf{T_e}$ (test set), as indicated in the experimental results for TC [15].

The record collection, which is already known in class labels, is indicated by the training dataset ($\mathbf{T_r}$). The categorisation model is established using $\mathbf{T_r}$ and is applied to the testing dataset.

The known record collection in class labels is indicated by the test dataset ($\mathbf{T_e}$).Be that as it may, exact class names ought to be come back to the records when given as a contribution to build up categorisation models.

The precision of the model can be related to the assistance of $\mathbf{T_e}$ [16].

TC for the most part experiences three key advances, to be specific, information preprocessing, content order and assessment. In the information preprocessing stage, appropriate content reports are made to prepare the classifier. A content learning approach is utilized to develop the content classifier, which is tuned against the preparation dataset. At long last, some assessment measures, for example, review and exactness, are utilized to investigate the content classifier. The two continuing subsections demonstrate the primary strides of the TC issue in relationship to the utilized information in this paper

4. Preprocessing of Arabic

The data used in our experiments are the Iraqi newspaper datasets, which consist of 5,000 Arabic documents of different lengths belonging to the following five categories: Sport "**الرياضة** "; War "**الحرب** ";General News " **اخبار عامة** "; Medicine "**الطب** "; Economic "**اقتصاد** ").The number of documents per category is offered in Table 1. Arabic is a rich dialect and its content is different; English to Arabic dialect is amazingly inflectional and derivational dialect, in this manner muddling monophonical investigation. A few vowels in the Arabic content are meant by diacritics, which commonly stays in the content and utilizes capitalisation for formal people, places or things, subsequently making equivocalness in the content [17].

Each archive record in the Arabic dataset was spared in a different document inside the coordinating index of the class. The Arabic dataset is displayed in a frame reasonable for the characterization calculation.

In this part, the information arrange [18] is pursued and the Arabic records are prepared by every content in the Arabic dataset to play out the accompanying: eliminate

Support Vector Machine

The SVM strategy in Arabic as a class of administered machine learning method is displayed in [10]. SVM is based the statute of basic limit chance. In direct grouping, a hyperplane that isolates the two sets containing information with the greatest edge is built in SVM.

At the point when the opposite sides are equivalent, a hyperplane with the most extreme edge has the separations from the hyperplane to focuses. Numerically, the sign capacity f (x) = sign (wx + b), where w is a weighted vector in Rn, is found out by SVMs. The hyperplane y = wx + b is distinguished by SVMs by isolating the space Rn into two half-spaces with the most extreme edge. For nonlinear issues, straight SVMs can be summed up by mapping the information into another space H and applying the direct SVM calculation over this new space. TC has effectively utilized SVM [1, 21] and inferred results that are better than those of another machine learning techniques, for example, NB, choice trees and KNN, as far as exactness

KNN Classifier

Any offline learning to generate category-specific knowledge is not conducted by the KNN classifier during its learning phase, resulting in rapid training time. The cosine value distance between a test sample and specified training samples is used by the KNN classifier. The training documents are ranked when an unknown document $\vec{x}$ is evident in the similarity of

the digits, punctuation marks and non-letters (i.e. numbers and percentage signs), replace آwith double alif ا, initial أ ، إwith ا, replace all hamza forms ؤ ، ئ with ء , replacing ةby ه and removing بال،كال،فال and وال [19,20]. To enhance information retrieval and search, stop words are removed by ignoring words that generally appear in every document.

5. Classifiers

Two existing approaches to TC, namely, SVM and KNN, are presented in this section. SVM, which performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories, is an effective algorithm. KNN, which classifies objects based on the closest training examples in the feature space, is considered the easiest approach. The two subsequent subsections describe the general nature, classifier

training and document classification process, and the advantages and disadvantages of both learning methods.

Table 1: No. of text per classification.

A. Classification name	B. Number of text
C. Sport	D. 1000
E. Art	F. 1000
G. Religion	H. 1000
I. Medicine	J. 1000
K. Economic	L. 1000
M. Total	N. 5000

each one with the document $\vec{x}$. Then, the k most similar documents of $\vec{x}$ are obtained. The similarity is computed with cosine distance as follows [22]:

$$\cos(\vec{x},\vec{d}_i) = \frac{\sum_{k=1}^{n} x_k d_{ik}}{\sqrt{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} d_{ik}^2}} \ldots.1$$

All the preprocessing techniques are applied after loading the corpus. TF × IDF calculations facilitate the creation of document vectors in the feature space. This observation concludes the training phase for KNN algorithm.

The KNNs of the document are obtained by the KNN algorithm from the training documents. The document class can be predicted by utilising the class labels of these KNNs.

1. $\vec{d}_i = (\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_m)$ Where $\vec{d}_i$ indicates one training document, keeps all the training documents.

2. With the emergence of an unknown document $\vec{x}$, the training documents are ordered based on the similarity of each document with document $\vec{x}$. The k most similar documents of $\vec{x}$ are then acquired. Equation (1) computes the similarity by employing the cosine value distance, where N is the total number of feature items.

3. The class weight of $\vec{x}$ for each class C_j can be calculated as follows with the k most similar documents of $\vec{x}$:

$$p\left(\vec{x}, C_j\right) = \sum_{\vec{d}_j \in KNN(\vec{x})} sim\left(\vec{x}, \vec{d}_i\right) y\left(\vec{d}_i, C_j\right) \ldots.2$$

KNN ($\vec{x}$)represents the set of *K* most similar documents of $\vec{x}$, whereas $y(\vec{d}_i, C_j)$ indicates the classification of documents $\vec{d}_i$ for class C_j.

$$y(\vec{d}_i, C_j) = \begin{cases} 1, & \vec{d}_i \in C_j \\ 0, & \vec{d}_i \notin C_j \end{cases} \quad (3)$$

4. Lastly, the class weight of $\vec{x}$ is compared for all classes, and $\vec{x}$ is categorised to the class with the maximum class weight $p\left(\vec{x}, C_j\right)$ [22].

$$C = \arg max_{c_j}(p(\vec{x}, C_j)) \quad (4)$$

6. Experiment Results

Three calculation measures (recall, precision and F1) were utilised as the rules of our comparison, where F1 is calculated based on the following equation:

$$F1 = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Re}call + \mathrm{Pr}ecision} \ldots.5$$

Precision and recall values evaluate the performance of the categorization model. Precision computes exactness, whereas recall computes completeness [23][24]–[28]. Let TP be the number of true positives, that is, the number of documents correctly labeled and agreed upon by both the experts and the model. Let FP be the number of false positives, that is, the number of documents wrongly categorized by the model as belonging to that category. Let FN be the number of false negatives, that is, the documents number not labeled as belonging to the category but should have been labeled. Precision and recall are calculated [29] based on the following equation:

$$Precision = \frac{tp}{tp + fp} \ldots.6$$

and

$$\mathrm{Re}call = \frac{tp}{tp + fn} \ldots.7$$

Table 2 presents the F1, recall and precision results produced by the two categorisers (KNN and SVM).

Table 2: F1, Recall, Precision values of Arabic Text classification

Category	SVM			KNN		
Measures	PPV	Sensitivity	F1	PPV	Sensitivity	F1
General News	0.51	0.51	0.51	0.43	0.39	0.41
Medicine	0.86	0.85	0.85	0.67	0.67	0.67
War	0.63	0.61	0.62	0.50	0.54	0.52
Economics	0.96	0.94	0.95	0.96	0.91	0.94
Sport	0.90	0.86	0.88	0.86	0.83	0.84

An outstanding assessment technique in information mining is cross-approval, where the preparation information are arbitrarily isolated into n hinders, each shut is held out once and the classifier is prepared on the rest of the n – 1 squares. The mistake rate is evaluated by utilizing the holdout square. Hence, somewhat extraordinary preparing datasets are used to play out the learning methodology n times. Weka open-source programming was utilized to play out every one of the examinations [24].

The SVM categorizer beat KNN on four datasets as far as F1 results, as appeared Table 2. As demonstrated by the accuracy results, the SVM beat KNN on four datasets. The review results likewise propose that the SVM beat KNN on four datasets, and KNN was predominant on single datasets.

Another checked outcome that was likewise recorded is that all estimates exceptionally among classes. For example, the "Financial matters" classification has a clean grouping F1 of 0.94; anyway the "General News" class has a perceptibly poor Recall of 0.41 utilizing KNN. These poor outcomes demonstrate that the "General News" class is meddled with different classifications.

In conclusion, SVM classifier [30] go before unrivaled in the Iraq Newspaper informational collections [31].

7. Conclusion

The issue of programmed grouping of Arabic content records is researched in this paper. The KNN calculation, which depends on the cosine esteem separate between a test and determined preparing tests, and the SVM calculation were used to deal with such a characterization issue. Concerning F1, review and accuracy

measures, the normal of these measures was contrasted and Iraqi Newspaper Arabic datasets and showed that the SVM calculation beat the KNN calculation. Another multi-mark grouping approach dependent on affiliation rule for the TC is planned to be proposed sooner rather than later.

References

[1] Joachims T. (1999). Transductive Inference for Text Classification using Support Vector Machines. Proceedings of the International Conference on Machine Learning (ICML), (pp.200-209).1999.

[2] Quinlan, J. "C4.5: Programs for machine learning,". San Mateo, CA: Morgan Kaufmann,1993.

[3] Duwairi, R. (2007). Arabic Text Categorization. *Int. Arab J. Inf. Technol.* Retrieved from

[4] Harrag, F., Al-Salman, A. S., & BenMohammed, M. (2010). A comparative study of Neural networks architectures on Arabic text categorization using feature extraction. In *Machine and Web Intelligence (ICMWI), 2010 International Conference on* (pp. 102–107). IEEE.

[5] Shakeel PM, Manogaran G., "Prostate cancer classification from prostate biomedical data using ant rough set algorithm with radial trained extreme learning neural network", Health and Technology, 2018:1-9.https://doi.org/10.1007/s12553-018-0279-6

[6] Mohammed J. Bawaneh, M. S. A. and A. I. (2008). Arabic Text Classification using K-NN and Naive Bayes. Journal of Computer Science 4, 600–605.

[7] Laila K. "Arabic Text Classification Using N- Gram Frequency Statistics A Comparative Study,"DMIN, 2006, pp.78-82.

[8] Han, E., Karypis, G., & Kumar, V. (2001). Text categorization using weight adjusted k-nearest neighbor classification.

[9] Preeth, S.K.S.L., Dhanalakshmi, R., Kumar, R.,Shakeel PM.An adaptive fuzzy rule based energy efficient clustering and immune-inspired routing protocol for WSN-assisted IoT system.Journal of Ambient Intelligence and Humanized Computing.2018:1–13. https://doi.org/10.1007/s12652-018-1154-z

[10] Mesleh, A. A. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System," Journal of Computer Science (3:6), 2007, pp. 430-435.

[11] Thabtah F., Hadi W., Al-shammare G. (2008) VSMs with K-Nearest Neighbour to Categorise Arabic Text Data. In The World Congress on Engineering and Computer Science 2008. (pp.778-781), 22-44 October 2008.

[12] Thabtah F., Eljinini M., Zamzeer M., Hadi W. (2009) Naïve Bayesian based on Chi Square to Categorize Arabic Data. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt 4 - 6 January. (pp. 930-935).

[13] Abdelwadood Moh'd A MESLEH. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System". Journal of Computer Science 3(6). Pages 430-435.2007.

[14] Shakeel, P.M., Tolba, A., Al-Makhadmeh, Zafer Al-Makhadmeh, Mustafa Musa Jaber, "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks", Neural Computing and Applications,2019,pp1-14.https://doi.org/10.1007/s00521-018-03972-2

[15] Sebastiani, .F "A Tutorial on Automated Text Categorization," In Proceedings of the ASAI-99,1st Argentinian Symposium on Artificial Intelligence, 1999. pp. 7-35.

[16] Yang, Y. (2001). A study of thresholding strategies for text categorization. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 137–145). ACM.

[17] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern Recognition and Machine Learning (Vol. 1). springer New York.

[18] Hammo, B., Abu-Salem, H., Lytinen, S., and Evens, M. 2002. "QARAB: A Question Answering System to Support the Arabic Language". Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. pp. 55-65.

[19] El-Kourdi, M., Bensaid, A., and Rachidi, T. "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," 20th International Conference on Computational Linguistics, 2004, Geneva.

[20] Samir, A., W. Ata, and N. Darwish. "A New Technique for Automatic Text Categorization for Arabic Documents," 5th IBIMA Conference (The internet & information technology in modern organizations), 2005, Cairo, Egypt.

[21] Geehan S. hassan, S.K. Mohammad and F.M. Alwan, 2015. Categorization of 'Holy Quran-Tafseer' using Knearest neighbor algorithm. Int. J. Comput. Appl., 129(12).

[22] Manogaran G, Shakeel PM, Hassanein AS, Priyan MK, Gokulnath C. Machine-Learning Approach Based Gamma Distribution for Brain Abnormalities Detection and Data Sample Imbalance Analysis. IEEE Access. 2018 Nov 9.DOI 10.1109/ACCESS.2018.2878276

[23] Joachims T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In Proceedings of the European Conference on Machine Learning (ECML), 1998, pp.173-142, Berlin.

[24] Lu, F., & Bai, Q. (2010). A refined weighted K-Nearest Neighbors algorithm for text categorization. Intelligent Systems and Knowledge Engineering (…, 326–330. doi:10.1109/ISKE.2010.5680854

[25] Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, (December). Retrieved from WEKA. Data Mining Software in Java: http://www.cs.waikato.ac.nz/ml/weka

[26] L. Haoyu, L. Jianxing, N. Arunkumar, A. F. Hussein, and M. M. Jaber, "An IoMT cloud-based real time sleep apnea detection scheme by using the SpO2 estimation supported by heart rate variability," *Futur. Gener. Comput. Syst.*, 2018.

[27] P. M. Shakeel, S. Baskar, V. R. S. Dhulipala, and M. M. Jaber, "Cloud based framework for diagnosis of diabetes mellitus using K-means clustering," *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, p. 16, 2018.

[28] M. A. Mohammed *et al.*, "Genetic case-based reasoning for improved mobile phone faults diagnosis," *Comput. Electr. Eng.*, 2018.

[29] S. K. Abd, S. A. R. Al-Haddad, F. Hashim, A. B. H. J. Abdullah, and S. Yussof, "Energy-Aware Fault Tolerant Task offloading of Mobile Cloud Computing," in *Proceedings - 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2017*, 2017.

[30] Saleh Alsaleem (2010). Automated Arabic Text Categorization Using SVM and NB. In International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011

[31] Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 60(5), 493–502.