Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 06, No. 03, September 2020.

A SCALABLE SHALLOW LEARNING APPROACH FOR TAGGING ARABIC NEWS ARTICLES[1]

Leen Al Qadi, Hozayfa El Rifai, Safa Obaid and Ashraf Elnagar

(Received: 28-Mar.-2020, Revised: 23-Jun.-2020 and 15-Jul.-2020, Accepted: 16-Jul.-2020)

ABSTRACT

Text classification is the process of automatically tagging a textual document with the most relevant set of labels. The aim of this work is to automatically tag an input document based on its vocabulary features. To achieve this goal, two large datasets have been constructed from various Arabic news portals. The first dataset consists of 90k single-labeled articles from 4 domains (Business, Middle East, Technology and Sports). The second dataset has over 290k multi-tagged articles. The datasets shall be made freely available to the research community on Arabic computational linguistics. To examine the usefulness of both datasets, we implemented an array of ten shallow learning classifiers. In addition, we implemented an ensemble model to combine best classifiers together in a majority-voting classifier. The performance of the classifiers on the first dataset ranged between 87.7% (Ada-Boost) and 97.9% (SVM). Analyzing some of the misclassified articles confirmed the need for a multi-label opposed to single-label categorization for better classification results. We used classifiers that were compatible with multi-labeling tasks, such as Logistic Regression and XGBoost. We tested the multi-label classifiers on the second larger dataset. A custom accuracy metric, designed for the multi-labeling task, has been developed for performance evaluation along with hamming loss metric. XGBoost proved to be the best multi-labeling classifier, scoring an accuracy of 91.3%, higher than the Logistic Regression score of 87.6%.

KEYWORDS

Arabic text classification, Single-label classification, Multi-label classification, Arabic datasets, Shallow learning classifiers.

1. INTRODUCTION

Large numbers of online repositories have been created continuously in the recent decades, due to the non-stop flow of information, as well as the heavy usage of the internet and Web 2.0. This increase of online documents was followed by a growing demand for automatic categorization algorithms. 80% of this information is in an unstructured form and the most common type is the "textual" data. Although it is considered to be an extremely rich source of information, it becomes harder to extract insights from or deduce trends when it's presented in enormous amounts. Machine learning techniques are often used to organize massive chunks of data and perform a number of automated tasks.

Natural Language Processing (NLP) [2], is a field of study concerned with analyzing and processing natural language data in large amounts. Machine learning algorithms, in addition to deep learning methods, are used in NLP to fulfil several tasks like the text classification task. It is the task of classifying text and assigning it appropriate tags, based on its content. The act of classification will standardize the platform, make the process of searching information easier and more feasible and simplify the overall experience of automated navigation.

Structuring data is also useful in the world of business and organizations. It will enhance the decision-making, identify current trends and predict new ones. In addition to automating regular processes, marketers can research, collect and analyze keywords by competitors.

Manual classification performed by experts is not always fruitful or efficient, due to human errors and to the long time needed to do it. Using machine learning as an alternative is proving to be more effective and, in some cases, more precise. Applications of text classification have been explored, such

[1] This paper is an extended version of a short paper [44] that was presented at the 2nd International Conference "New Trends in Information Technology (ICTCS)", 9-11 October 2019, Amman, Jordan.

[2] Natural Language Processing https://monkeylearn.com/natural-language-processing/

L. Al Qadi, H. El Rifai, S. Obaid and A. Elnagar are with Computer Science Department, University of Sharjah, UAE. Emails: u16103630@sharjah.ac.ae, u16103377@sharjah.ac.ae, u16103014@sharjah.ac.ae and ashraf@sharjah.ac.ae

as sentiment analysis [1]-[6], spam filtering [7]-[8], language identification [9], dialect identification [10] and many more.

Some languages present big challenges to many NLP applications [11] and the Arabic language is one of them. It is the mother tongue of over 300 million people and it is considered to be a significantly inflected and derived language. Compared to the English language, the scale of computational linguistic research on the Arabic language is relatively small but is now much bigger than what was available about a decade ago.

The Internet World stats reports that the Arabic language is the 4th most popular language among online users, with an estimate of 226,595,470 Arabic users. As of April 2019, that number represents 5.2% of the world's internet users. Additionally, it shows that 51.0% of all Arabic speaking people (in 2019) use the internet and that the language has the highest growth rate, of online users, among all the languages in the last 19 years, scoring a percentage of 8,917.3%.

In this context, we present a newly constructed Arabic news articles dataset, collected by scraping a number of websites for the purpose of our research. We implement 10 classical classifiers to predict the most suited category for a certain news article. Moreover, we design a voting classifier that classifies an article with respect to the output of selected models.

A developed system for Arabic news article single-labeling will extract text features from the text using the Tf-IDF technique. In the training phase, all the articles are turned to feature vectors, which will identify the common features that define each class separately. After the model is trained on the features, it will easily predict the class of an article after being vectorized.

To label news articles under one of 4 classes, we propose a single-class classifier, using a supervised machine learning approach. Two different vectorization methods were tested and compared to observe the effects of using each on the accuracy of the models. Lastly, the effects of using a custom-made stop words list in place of the built-in list provided by the NLTK library were also explored.

A decision to build a new multi-labeled Arabic dataset is made after analyzing the misclassified articles, classified by the single-class classifier. The need to assign multiple tags to an article instead of a single one was apparent and 2 classical classifiers were implemented for the task. The Tf-IDF technique was also used to extract the linguistic features. To decompose the multi-labeling problem into independent binary classification problems, we wrapped each of the classifiers in a OneVsRest classifier.

With an objective to assign articles to multiple labels (out of 21 labels), we propose a multi-label text classifier. We test the classifiers and evaluate them using a custom accuracy metric, along with comparing the hamming-loss scores.

The remaining of the paper is organized as follows: literature review is presented in Section 2. Section 3 demonstrates the datasets. Section 4 describes the proposed classification systems. Section 5 presents the experimental results. Finally, we conclude the work in Section 6.

2. LITERATURE REVIEW

Several papers review the various English text classification approaches and existing literature in addition to the many surveys covering the subject [12]-[14]. Some surveys that cover Arabic text categorization are also available [15]-[16].

Light has been shed on the research papers that focused on using the classical supervised machine learning classifiers, such as Decision Tree [17]-[19], NB [20]-[23], SVM [19], [22], [26]-[27] and KNN [23], [26], while other authors preferred to explore text classification using deep learning and neural networks [25] and some also witnessed an overall better performance [27]-[35].

Recently, more research works are focusing on Arabic text classification and on enriching the Arabic corpus. In [36], the authors have compared the results of using six main classifiers, using the same datasets and under the same environmental settings. The datasets were mainly collected from (www.aljazeera.net) and it was found that Naive Bayes gave the best results, with or without using feature selection methods.

Other papers focus on the feature selection method, like in [37]. Implementing the KNN classifier, the authors studied the effect of using unigrams and bigrams as representation of the documents, instead of the traditional single-term indexing (bag of words) method. Moreover, on feature selection, in [38], the authors investigated the performance of four classifiers using 2 different feature-selection methods which are Information Gain (IG) and the ($X2$) statistics (CHI squared) on a BBC Arabic dataset. The use of SVM classifier (with Chi squared feature selection) for Arabic text classification, in [39], gives the best results. In [40], a new feature selection method is presented, where it outperformed five other approaches using the SVM classifier.

Regarding the availability of Arabic datasets online, [41] suggests that some of the existing Arabic corpora are not dedicated for classification, because either there are no defined classes such as 1.5 billion words Arabic Corpus [42], or the existing classes are not well-defined. Therefore, the authors propose a new pre-processed and filtered corpus “NADA”, composed from two existing corpora: OSAC and DAA. The authors used the DDC hierarchical number system that allows for each main category to be divided into ten sub-categories… and so on. “NADA” has 10 categories in total, with 13,066 documents. We believe that the size is small with respect to the proposed number of categories.

In addition, [43] investigates text classification using the SVM classifier on two datasets that differ in languages (English and Portuguese). It was found that the Portuguese dataset needs more powerful document representations, such as the use of word order and syntactical and/or semantic information.

Overall, it is clear that the performance of classification algorithms in Arabic text classification is greatly influenced by the quality of data source, feature representation techniques as the irrelevant and redundant features of data degrade the accuracy and performance of the classifier. This work is an extension of our work [44] on single-label classification.

3. DATASET

3.1 Single-label Dataset

We propose a newly collected dataset, consisting of 89,189 Arabic articles, tagged under 4 main categories [Sports, Business, Middle East and Technology]. Using the web-scraping framework, Scrapy, we collected data from 7 popular news portals (youm7.com, cnbcarabia.com, skynewsarabic.com, Arabic.rt.com, tech-wd.com, arabic.cnn.com and beinsports.com).

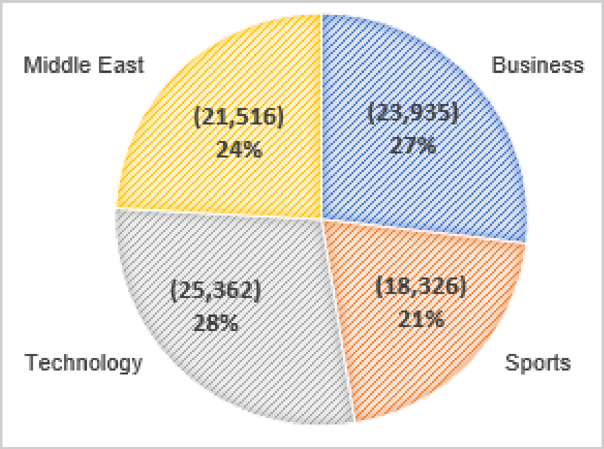


Figure 1. Single-labeled dataset distribution percentages.

With more than 32.5M words, all the articles in the dataset have no dialects and are written in Modern Standard Arabic (MSA). To avoid bias, it was essential to build a balanced corpus. On average, each category almost has 22k articles. Table 1 and Figure 1 show the exact distribution of the articles, with the count.

3.2 Multi-label Dataset

The same web scraping technique was used, in addition to others like BeautifulSoup and Selenium, in a hunt for websites that publish Arabic articles with several tags. Numerous amounts of multi-labeled

Table 1. Articles count for each scraped news portal.

Websites	Classes	Articles Count
Sky News Arabia	Sports	7923
CNN Arabia	Sports	3800
	Tech.	1680
	Middle East	21516
	Business	3908
Bein Sports	Sports	6603
Tech-wd	Tech.	23682
Arabic RT	Business	896
Youm7	Business	14478
CNBC Arabia	Business	4653

articles, written in (MSA), were collected from 10 websites (cnbcarabia.com, beinsports.com, arabic.rt.com, tech-wd.com, youm7.com, aitnews.com, masrway.com, alarabiya.net, skynewsarabic.com and arabic.cnn.com).

As it is shown in Figure 2, the first collected dataset "Dataset_1" has 284,860 articles in total and was used to train and test the multi-labeling classifiers, giving the results apparent in Table 3 and Table 5. After studying the precision, recall and F1-score of each label, we decided to refine the dataset by removing redundant articles found in categories like: "Business" and by enriching the rest of them. The resulting dataset is called "Dataset_2" and contains 293,363 articles.

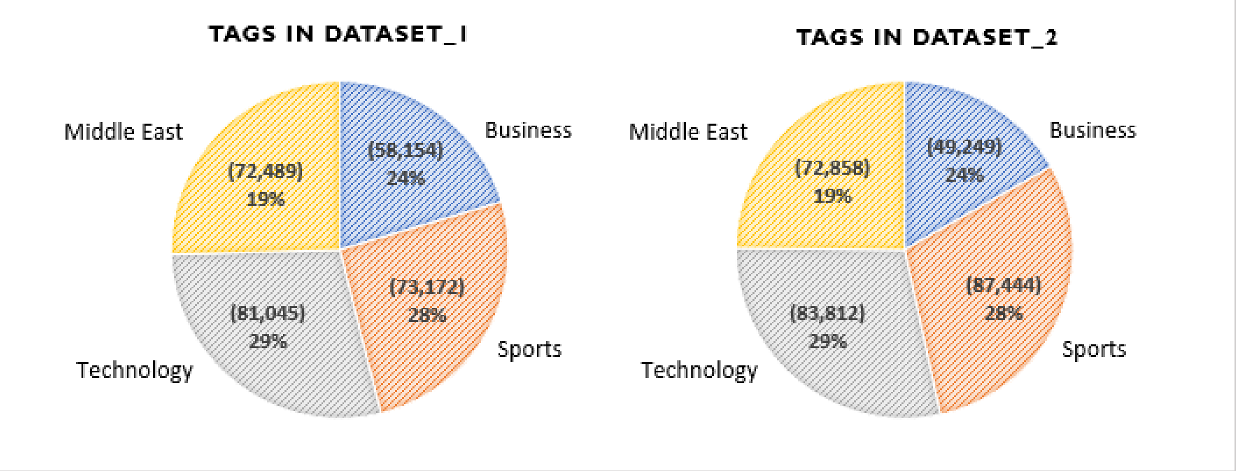


Figure 2. Article distribution in multi-labeled dataset.

4. Proposed Classification Systems

4.1 Text Features

Vectorization is the process of feature building, by turning text into numerical vectors. In text processing, words of the articles represent categorical features. This is a crucial step, as machine learning algorithms are unable to understand plain text. The most commonly used methods for this task are the Count Vectorizer and the Tf-IDF Vectorizer. Using a Count Vectorizer creates a Bag of words that counts the frequency of each word. However, using a Tf-IDF Vectorizer increases the value of the word proportionally to its count in a document, but is inversely proportional to its frequency in the corpus. The Tf-IDF Vectorizer is composed by two terms:

- Term Frequency (TF): measures how frequently a word occurs in an article. Since every article is different in length, it is possible that a term would appear much more times in long articles than in shorter ones.
- Inverse Document Frequency (IDF): measures how important a word is by weighing down the frequent terms and scaling up the rare ones.

It should be noted that unigram features are used in this work, as they reported better results when compared to bigram features. To show the effects of the using each of vectorizers on the model's accuracies, we made a comparison to rule out the best method. We used a portion of the dataset, containing approximately 40k articles belonging to 3 categories: (Middle East, Business and Sports). Figure 3 shows the results of the comparison, where the higher accuracy percentages were scored by the models using the Tf-IDF Vectorizer, which we decide to adopt in our work.

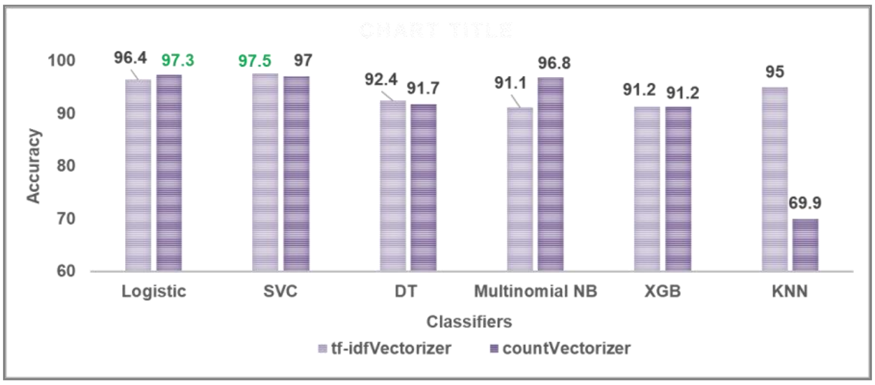


Figure 3. Accuracy comparison between TF-IDF Vectorizer and Count Vectorizer.

In addition to that, we put together our own custom-made stop words list and tested it against the NLTK built-in list. Higher accuracies were achieved using our list and we used it in further experiments.

Lastly, Figure 4 describes the overall workflow of our system. As this is a supervised machine learning approach, labels were one-hot encoded using sk-learn's Label Encoder and fed to the algorithms along side the vectors, during the training of the models.

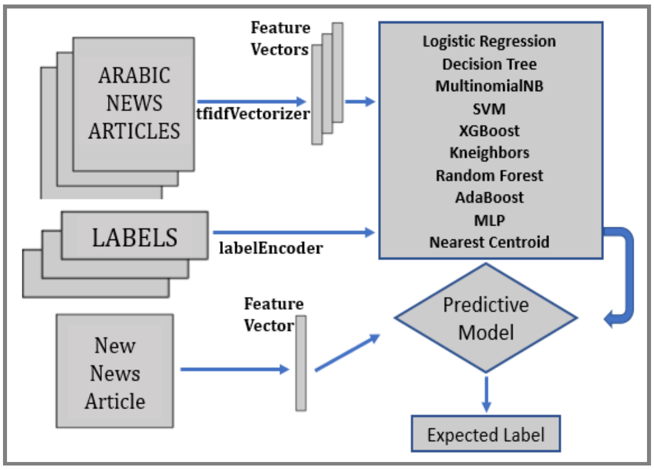


Figure 4. Summary of the work-flow of the classifiers.

4.2 Selected Classifiers

There exists an array of supervised shallow learning classifiers that are suitable to perform the text classification task. These classifiers have to map input data to a specific predicted category. We compared the results of 10 classifiers, in addition to a majority voting classifier. The classifiers are:

- Logistic Regression: Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- Multinomial Naïve Bayes: Using Bayes Theorem, this classifier calculates the probability of each label for a given data, then outputs the label with the highest probability. The classifier

assumes that the attributes are independent of each other. In other words, the presence of one feature does not affect the presence of another; therefore, all the attributes contribute equally in producing the output.

- Decision Tree: This classifier resembles a tree, with each node representing a feature/attribute and each corresponding leaf representing a result. Each branch represents a condition and whenever a condition is answered, a new condition will be distributed recursively until a conclusion is reached. Recursion is used to partition the tree into a number of conditions with their outcomes.
- Support Vector Machines (SVM): This is a supervised non-probabilistic binary linear classifier that is extremely popular and robust. It constructs a model and outputs a line, known as the hyperplane, between classes. This hyperplane separates the data into classes. Both linear and nonlinear classification can be performed by the SVM classifier. The hyperplane can be written as the vector of input articles xx satisfying $w.x - b = 0w.x - b = 0$, where ww is the normal vector to the hyperplane and bb is the bias.
- Random Forest: This is a supervised ensemble learning-based classifier. It uses an array of decision trees. The outcome class is determined as an aggregate of such trees. Technically, given a set of articles $x_1, x_2, \dots, x_n x_1, x_2, \dots, x_n$ and their corresponding classes $y_1, y_2, \dots, y_n y_1, y_2, \dots, y_n$, each classification tree fb is trained using a random sample $(X_b, Y_b)(X_b, Y_b)$, where b ranges from 1 to the total number of trees. The predicted class shall be produced using a majority vote of all used trees.
- XGBoost Classifier: This is a supervised classifier, which has gained popularity because of winning a good number of Kaggle challenges. Like Random Forest, it is an ensemble technique of decision trees and a variant of gradient boosting algorithm.
- Multi-layer Perceptron (MLP): This is a supervised classifier. It consists of three (or more) layers of neuron nodes (an input and an output layer with one or more hidden layers). Each node of one layer is connected to the nodes of the next layer and uses a non-linear activation function to produce output.
- KNeighbors Classifier: This is a supervised classifier. In order to classify a given data point, we take into consideration the number of nearest neighbors of this point. Each neighbor votes for a class and the class with the highest vote is taken as the prediction. In other words, the major vote of the point’s neighbors will determine the class of this point.
- Nearest Centroid Classifier: This is a supervised classifier. It’s a no parameter algorithm, where each class is represented by the centroid of its members. It assigns to tested articles the label of the class of training samples with mean (centroid) closest to the article.
- AdaBoost Classifier: This is a supervised classifier. It is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset, but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.[3]
- Voting Classifier: It is a very interesting ensemble solution. It is not an actual classifier but a wrapper for a set of different classifiers. The final decision on a prediction is taken by majority vote.

It should be noted that only a number of supervised classical classifiers are suitable for the multi-labeling tasks and they are implemented using specific methods. For example, Logistic Regression and XGBoost both can classify text into multiple labels if each of them is wrapped in a OneVSRest Classifier. This will divide the bigger classification problem into many sub-problems.

We used the TF-IDF technique to vectorize the articles and used the default hyperparameters for each classifier. To encode the labels, we used MultiLabelBinarizer () that returns the string labels assigned to each article in a one-hot encoded format.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

5. Experimental Results and Discussion

5.1 Setup and Pre-processing

5.1.1 Single-label Text Classification

The main objective is to conduct a comparative study on 11 classification models, testing them in classifying Arabic news categories. All classifiers are implemented using the popular scikit-learn (machine learning in Python). The experiment begins with classifying the articles, then analyzing the results and determining the best classifier. After that, the same algorithms will be trained and tested on another newly reported dataset called ‘Akhbarona’ [34]-[44], that divides the articles into 7 classes.

We used the 80/20 ratio to split our dataset, where 80% of the data is in the training set, consisting of 71,707 articles and 20% in the testing set, containing 17,432 articles. More than 344k features were extracted from the training set. It should be noted that 10% of the training dataset is used for validation purposes in order to fine-tune parameters.

We report the accuracy score for each classifier, to evaluate their performance, which is calculated as the ratio of the number of correctly classified articles.

Moreover, cleaning and pre-processing the text are mandatory and recommended specially for text collected from the web. All non-Arabic content is removed and the articles are processed by eliminating punctuation, isolated characters, qur’anic symbols, elongation and other marks, along with the stop words.

Because the dataset is large enough to assign enough samples for each Arabic character, we believed that the normalization step is not necessary. In contrast to most research works on Arabic computational linguistics that apply normalization on the collected corpus, we skipped it as it can affect the meaning of some Arabic words. The stemming step is also skipped, as it results in a less deep view of the semantic relationships of the words, as it is concluded in [45].

5.1.2 Multi-label Text Classification

For this classifying approach, the dataset is split into 80% training set consisting of 118,700 labeled articles and 20% testing set consisting of 29,676 articles. The same text pre-processing steps used on the single-labeled dataset are used on the multi-labeling dataset.

A portion of the multi-labeled dataset proposed earlier, “Dataset_2” has been used to train and test the models. We chose to train on the labels with the highest frequency, because the performance of supervised deep learning classifiers is highly dependent on the number of instances for each label. Figure 5 shows the count of the 21 labels chosen from the dataset.

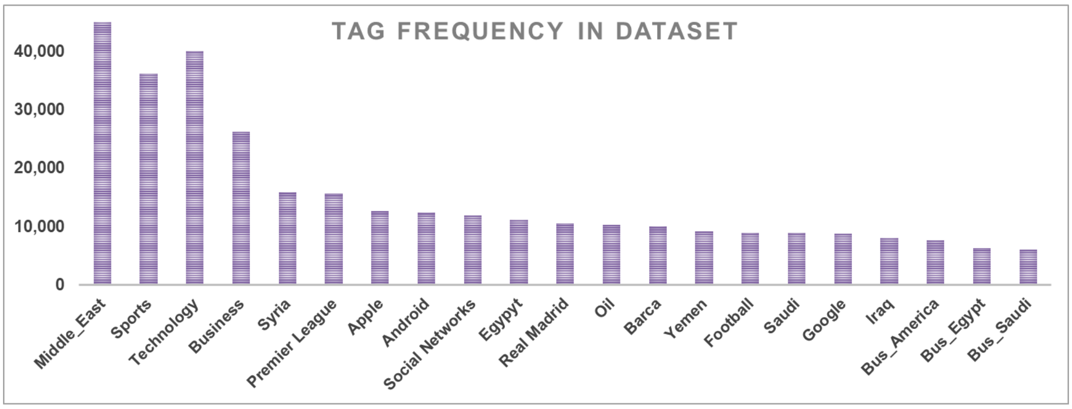


Figure 5. Tag count used in multi-labeling experiment using “Dataset-2”.

One evaluating metric we used is a custom accuracy metric, to evaluate the accuracy of the predictions. It calculates the ratio of correctly predicted tags (output as 1) over total expected tags (originally 1 in dataset). The more correct labels the model predicts, the more accurate it is. We choose a threshold of 50%, meaning that if the probability percentage of the tag is equal to or higher than 50%, then its value will be set as 1.

The second metric is the hamming loss, which is a commonly used metric for multi-labeling tasks. It is the fraction of wrongly predicted labels to the total number of labels. The smaller the value, the better results the model is achieving.

Figure 6 shows the relative distribution of the tags in “Dataset_2”, where the highest number of tags for an article is 6 and the lowest is 2.

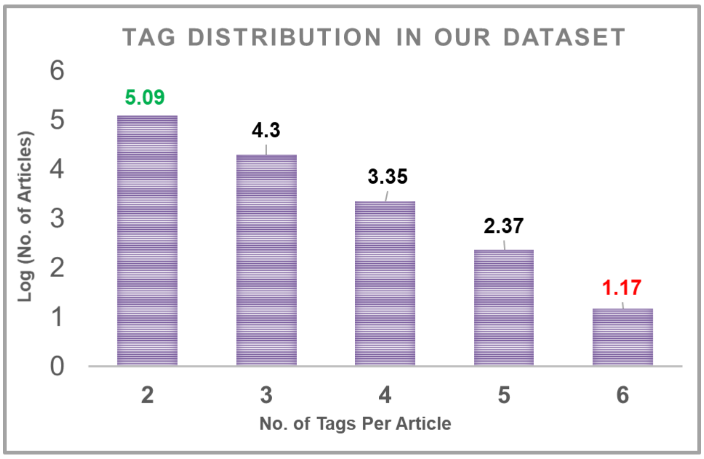


Figure 6. Tag count used in multi-labeling experiments.

5.2 Performance Evaluation

5.2.1 Single-label Text Classification

All the classifier models were implemented using Scikit-learn and by using the default hyper-parameters as a black-box, with the addition of L1 penalty for some of the classifiers. The proposed classifiers were tested using the testing set. Figure 7 shows the accuracy scores of each classifier. The percentages are exceptionally high and prove the strength of the default hyper-parameters used in the system.

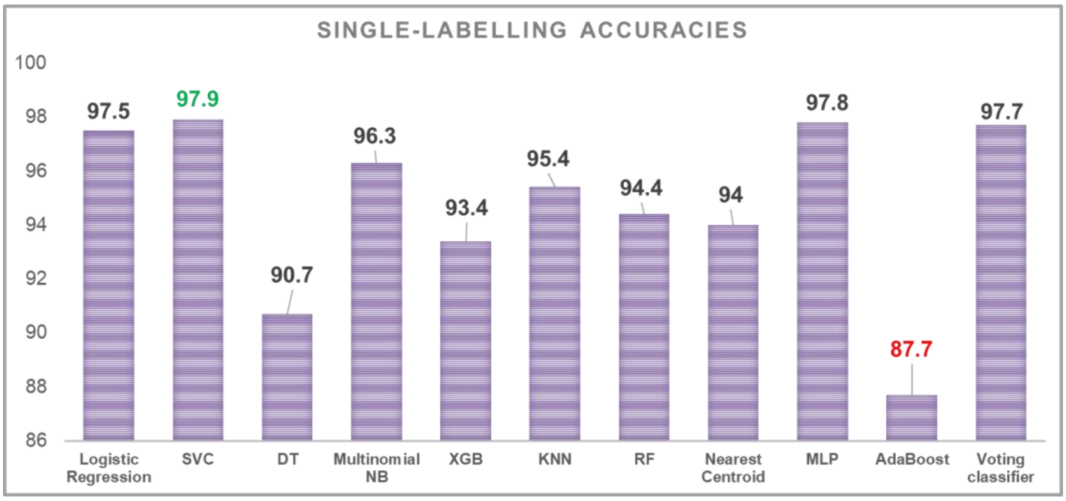


Figure 7. Accuracies for single-labeling classifiers using our dataset.

Producing the best result of 97.9% is the SVM classifier, while the worst percentage was produced by Ada-Boost. 4 classifiers achieved scores between 97.5% and 97.9% and the overall average of the percentages was 94.8%.

Two of the algorithms (KNeighbors and MultinomialNB) scored higher than the average, with percentages of 95.4% and 96.3%, respectively. The other models scored lower than the average with percentages ranging from 87.7% to 94.4%. The confusion matrix for SVM, the best classifier, is shown in Figure 8, while Figure 9 shows the matrix of the worst classifier Ada-Boost.

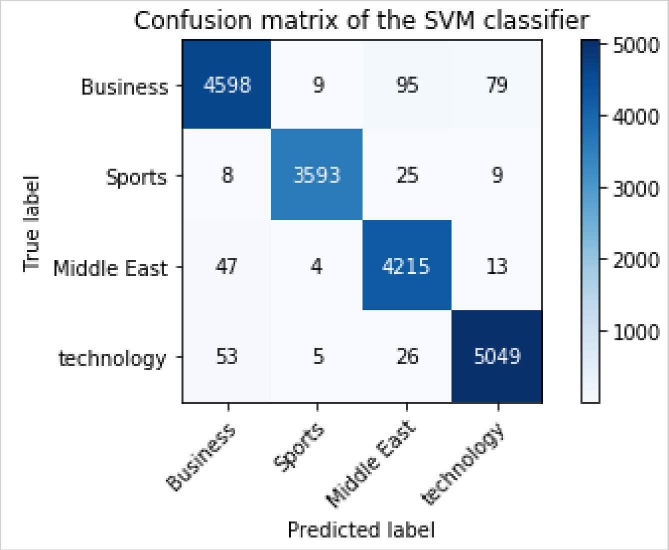


Figure 8. Confusion matrix for the best classifier.

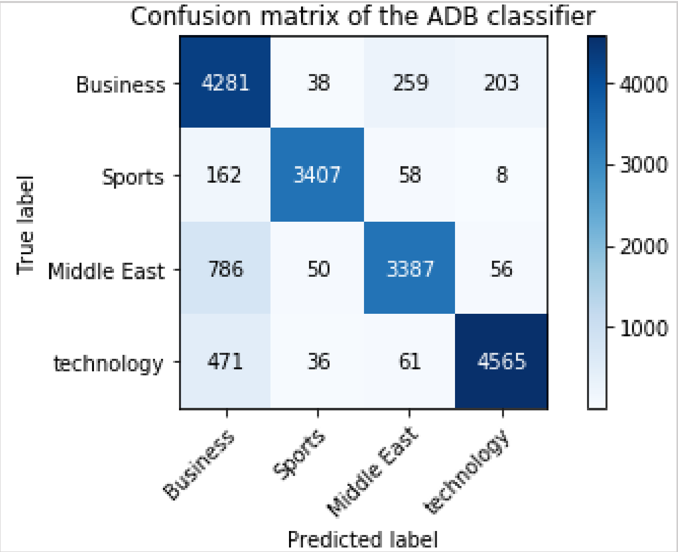


Figure 9. Confusion matrix for the worst classifier.

The matrices highlight that the 2 categories (Business and Middle East) had the greatest number of misclassifications and we believe that this is caused by feature similarity between the two. Table 2 shows the scores of the additional accuracy metrics used. The majority voting classifier and the SVM achieved the highest F1-score of 97.9%. The lowest score of 87.7% was produced by AdaBoost.

Table 2. Accuracy metrics for classifiers tested on our dataset.

Algorithms	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.98	0.98	0.98
SVC	0.98	0.98	0.98
DT Classifier	0.91	0.91	0.91
Multinomial NB	0.96	0.96	0.96
XGB Classifier	0.94	0.93	0.94
KNN Classifier	0.95	0.95	0.95
RF Classifier	0.95	0.94	0.94
Nearest Centroid	0.95	0.94	0.94
Ada-Boost Classifier	0.89	0.88	0.88
MLP Classifier	0.98	0.95	0.95
Voting Classifier	0.98	0.98	0.98

5.2.2 Multi-label Text Classification

Table 4 displays the evaluation metrics scores of both OVR-Logistic Regression and the OVR-XGBoost using “Dataste_2”. The average of the accuracies is 89.4%. XGBoost scored the highest of the two with a 91.3 % accuracy, while Logistic Regression scored an 87.6% accuracy. The hamming loss scores were low, where XGBoost scored the lowest with a percentage of 1.54% and Logistic Regression scored a percentage of 1.8%. The results prove that the XGBoost classifier was the better of the two.

Table 3. Evaluation metrics for multi-label classification using “Dataset_1”.

Evaluation metrics	OVR-Logistic Regression	OVR - XGBoost
Custom Accuracy	81.3%	84.7%
Hamming Loss	2.24%	2.22%

Table 4. Evaluation metrics for multi-label classification using “Dataset_2".

Evaluation metrics	OVR-Logistic Regression	OVR - XGBoost
Custom Accuracy	87.6%	91.3%
Hamming Loss	1.8%	1.54%

Lastly, we calculated the F1-scores, precision and recall for each label predicted by the XGBoost classifier. The results shown in Table 6 slightly vary due to the imbalance in support numbers.

Table 5. Accuracy metrics for OVR-XGBoost classifier using “Dataset_1”.

Labels	Precision	Recall	F1-score	Support
Business	0.99	0.98	0.98	5154
Oil	0.92	0.98	0.91	2026
Business-America	0.91	0.72	0.81	1549
Business-Egypt	0.93	0.92	0.92	1197
Business-Saudi	0.86	0.83	0.85	1154
Middle East	1.00	1.00	1.00	9164
Syria	0.95	0.91	0.93	3181
Egypt	0.96	0.90	0.93	2270
Yemen	0.95	0.87	0.91	1774
Saudi	0.88	0.82	0.85	1759
Iraq	0.94	0.86	0.90	1616
Sports	1.00	0.99	0.99	7428
Premier League	0.92	0.91	0.92	3193
Real Madrid	0.90	0.90	0.90	2120
Barca	0.91	0.90	0.90	2048
Football	0.86	0.43	0.57	1856
Technology	0.98	0.98	0.98	6529
Android	0.62	0.25	0.36	2187
Apple	0.55	0.14	0.22	1566
Google	0.61	0.33	0.43	2319
Social Media	0.74	0.19	0.30	1630

Table 6. Accuracy metrics for OVR-XGBoost classifier using “Dataset_2”.

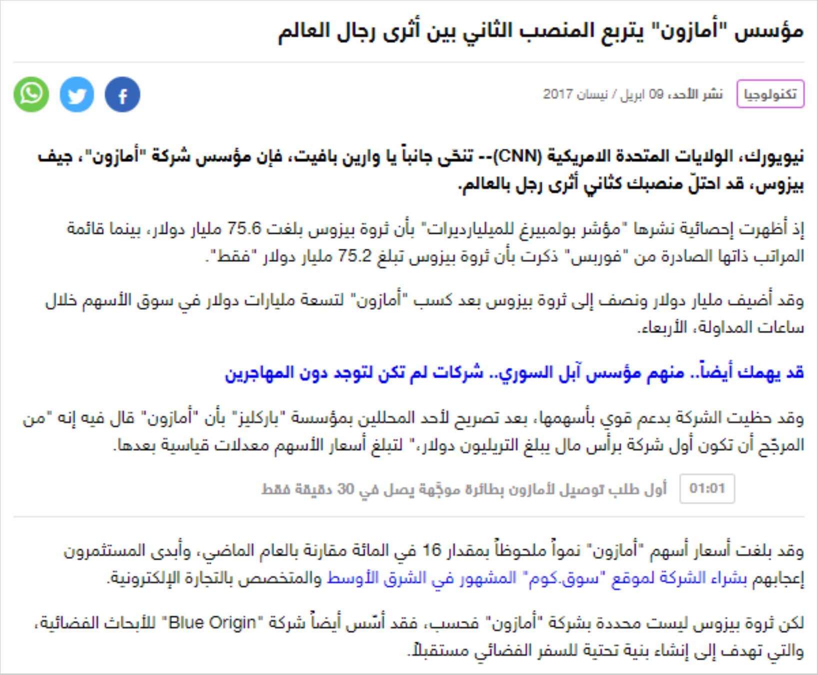
Labels	Precision	Recall	F1-score	Support
Business	0.97	0.94	0.95	5116
Oil	0.91	0.89	0.90	2041
Business-America	0.90	0.69	0.78	1557
Business-Egypt	0.91	0.84	0.87	1163

Business-Saudi	0.84	0.72	0.78	1162
Middle East	0.98	0.98	0.98	9197
Syria	0.94	0.90	0.92	3222
Egypt	0.93	0.86	0.90	2224
Yemen	0.94	0.87	0.90	1804
Saudi	0.83	0.76	0.79	1710
Iraq	0.92	0.86	0.89	1633
Sports	1.00	0.99	0.99	7309
Premier League	0.92	0.90	0.91	3136
Real Madrid	0.92	0.87	0.89	2169
Barca	0.90	0.88	0.89	2041
Football	0.81	0.43	0.56	1757
Technology	1.00	0.99	0.99	8054
Android	0.89	0.87	0.88	2493
Apple	0.93	0.86	0.89	2463
Google	0.87	0.87	0.87	1715
Social Media	0.94	0.90	0.92	2433

5.3 Sample Experiments

5.3.1 Single-label Text Classification

This phase is divided into 2 parts. The first part is to test the performance of the best classifier (SVM) by checking the predicted class of an article taken from the testing set.



مؤسس "أمازون" يتربع المنصب الثاني بين أثرى رجال العالم

تكنولوجيا نشر الأحد، 09 ابريل / نيسان 2017

نيويورك، الولايات المتحدة الامريكية (CNN)-- تنحّى جانباً يا وارين بافيت، فإن مؤسس شركة "أمازون"، جيف بيزوس، قد احتلّ منصبك كثاني أثرى رجل بالعالم.

إذ أظهرت إحصائية نشرها "مؤشر بولمبيرغ للميلياردريات" بأن ثروة بيزوس بلغت 75.6 مليار دولار، بينما قائمة المراتب ذاتها الصادرة من "فوربس" ذكرت بأن ثروة بيزوس تبلغ 75.2 مليار دولار "فقط".

وقد أضيف مليار دولار ونصف إلى ثروة بيزوس بعد كسب "أمازون" لتسعة مليارات دولار في سوق الأسهم خلال ساعات المداولة، الأربعاء.

قد يهمك أيضاً.. منهم مؤسس آبل السوري.. شركات لم تكن لتوجد دون المهاجرين

وقد حظيت الشركة بدعم قوي بأسهمها، بعد تصريح لأحد المحللين بمؤسسة "باركليز" بأن "أمازون" قال فيه إنه "من المرجّح أن تكون أول شركة برأس مال يبلغ التريليون دولار،" لتبلغ أسعار الأسهم معدلات قياسية بعدها.

01:01 أول طلب توصيل لأمازون بطائرة موجّهة يصل في 30 دقيقة فقط

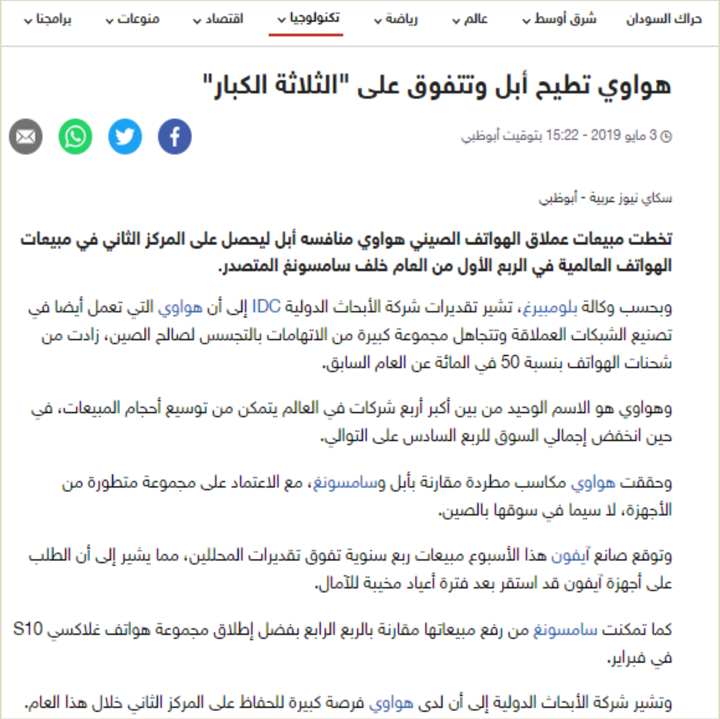
وقد بلغت أسعار أسهم "أمازون" نمواً ملحوظاً بمقدار 16 في المائة مقارنة بالعام الماضي، وأبدى المستثمرون إعجابهم بشراء الشركة لموقع "سوق.كوم" المشهور في الشرق الأوسط والمتخصص بالتجارة الإلكترونية.

لكن ثروة بيزوس ليست محددة بشركة "أمازون" فحسب، فقد أسّس أيضاً شركة "Blue Origin" للأبحاث الفضائية، والتي تهدف إلى إنشاء بنية تحتية للسفر الفضائي مستقبلاً.

Figure 10. Example of a correctly classified news article.

Figure 10 shows an example of an article grabbed from the testing set and originally tagged as Technology. The SVM model assigned the same tag for the article with a confidence of 95.7%. For the record, our model has clearly shown how robust and coherent it is by showing a confidence of 99.6%.

Moreover, we further studied the predictions of the model on the test set by checking a portion of the misclassified articles. In our investigation, the model proved that some of the articles were a good fit under the predicted categories more than the originally assigned categories by the news website. In Figure 11, we show an article that is tagged under the “Technology” category. However, after checking the article, we are convinced that the “Business” category is more suited for this article which is the same category that was predicted by the SVM classifier. This proves again that the model is precise and trustworthy.



هواوي تطيح أبل وتتفوق على "الثلاثة الكبار"

3 مايو 2019 - 15:22 بتوقيت أبوظبي

سكاي نيوز عربية - أبوظبي

تخطت مبيعات عملاق الهواتف الصيني هواوي منافسه أبل ليحصل على المركز الثاني في مبيعات الهواتف العالمية في الربع الأول من العام خلف سامسونغ المتصدر.

وبحسب وكالة بلومبيرغ، تشير تقديرات شركة الأبحاث الدولية IDC إلى أن هواوي التي تعمل أيضا في تصنيع الشبكات العملاقة وتتجاهل مجموعة كبيرة من الاتهامات بالتجسس لصالح الصين، زادت من شحنات الهواتف بنسبة 50 في المائة عن العام السابق.

وهواوي هو الاسم الوحيد من بين أكبر أربع شركات في العالم يتمكن من توسيع أحجام المبيعات، في حين انخفض إجمالي السوق للربع السادس على التوالي.

وحققت هواوي مكاسب مطردة مقارنة بأبل وسامسونغ، مع الاعتماد على مجموعة متطورة من الأجهزة، لا سيما في سوقها بالصين.

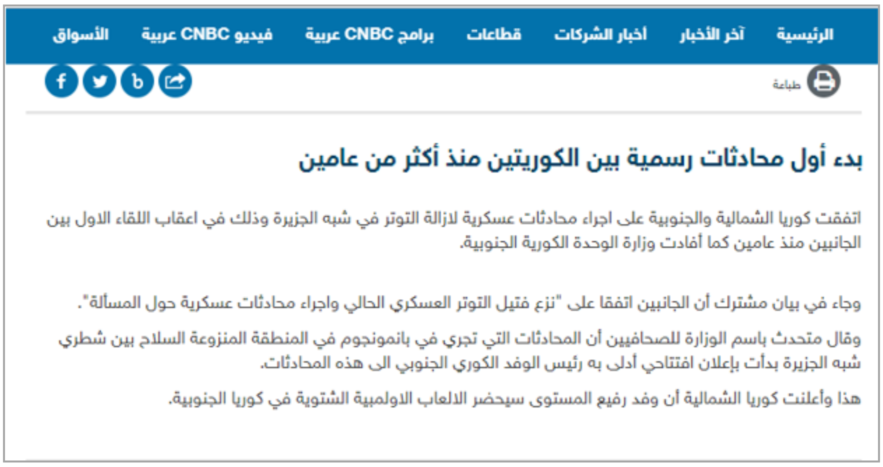
وتوقع صانع آيفون هذا الأسبوع مبيعات ربع سنوية تفوق تقديرات المحللين، مما يشير إلى أن الطلب على أجهزة آيفون قد استقر بعد فترة أعياد مخيبة للآمال.

كما تمكنت سامسونغ من رفع مبيعاتها مقارنة بالربع الرابع بفضل إطلاق مجموعة هواتف غلاكسي S10 في فبراير.

وتشير شركة الأبحاث الدولية إلى أن لدى هواوي فرصة كبيرة للحفاظ على المركز الثاني خلال هذا العام.

Figure 11. Example of an incorrectly classified news article as “Business”.

Finally, the article in Figure 12, taken from (cnbcarabia.com), was originally tagged as “Business”. The SVM model was more biased for the “Middle East” tag with a probability percentage of 40.3%. The model was also giving a percentage of 37.7% for the “Business” tag. The small difference in the confidence was toward the more suited tag.



بدء أول محادثات رسمية بين الكوريتين منذ أكثر من عامين

اتفقت كوريا الشمالية والجنوبية على اجراء محادثات عسكرية لازالة التوتر في شبه الجزيرة وذلك في اعقاب اللقاء الاول بين الجانبين منذ عامين كما أفادت وزارة الوحدة الكورية الجنوبية.

وجاء في بيان مشترك أن الجانبين اتفقا على "نزع فتيل التوتر العسكري الحالي واجراء محادثات عسكرية حول المسألة".

وقال متحدث باسم الوزارة للصحافيين أن المحادثات التي تجري في بانمونجوم في المنطقة المنزوعة السلاح بين شطري شبه الجزيرة بدأت بإعلان افتتاحي أدلى به رئيس الوفد الكوري الجنوبي الى هذه المحادثات.

هذا وأعلنت كوريا الشمالية أن وفد رفيع المستوى سيحضر الالعاب الاولمبية الشتوية في كوريا الجنوبية.

Figure 12. Another example of an article classified by SVM.

For the second part of the testing, we experimented with a recently reported dataset (Akhbarona). It’s an unbalanced dataset that consists of seven categories [Medicine, Politics, Sports, Religion, Culture, Technology and Business] holding 46,900 articles. The dataset is cleaned by removing elongation, punctuation, Arabic digits, isolated chars, qur’anic symbols, Latin letters and other marks. We split the dataset into 80% training and 20% testing.

It can be stated that lower accuracies are to be expected by the models for 2 reasons; the increase in the number of categories will lead to a higher possibility of misclassifying an article and the unbalanced dataset may steer the classifier to be biased to a certain class.

Table 7 demonstrates the accuracy results that have been obtained on Akhbarona dataset. The SVM classifier proved to be the best classifier by producing an accuracy of 94.4% on the test set. In contrast, the Ada-Boost classifier produced no more than 87.7%. The average of the accuracies is 90%

only. Furthermore, four classifiers were producing a close result to the best classifier with a range from 94.4% to 93.9%. The KNeighbors classifier performed above the average with an accuracy of 90.8%. The other six classifiers performed below the average with accuracy scores ranging from 77.9% to 88.4%. Figures 13 and 14 show the confusion matrix of the best classifier (SMV) and the worst classifier (Ada-Boost).

Table 7. Classifiers' accuracies on "Akbarona".

Algorithms	**Accuracy %**
Logistic Regression	93.9
SVC	**94.4**
DT Classifier	83.0
Multinomial NB	88.0
XGB Classifier	88.4
KNN Classifier	90.8
RF Classifier	87.8
Nearest Centroid	86.2
Ada-Boost Classifier	77.9
MLP Classifier	94.1
Voting Classifier	94.3

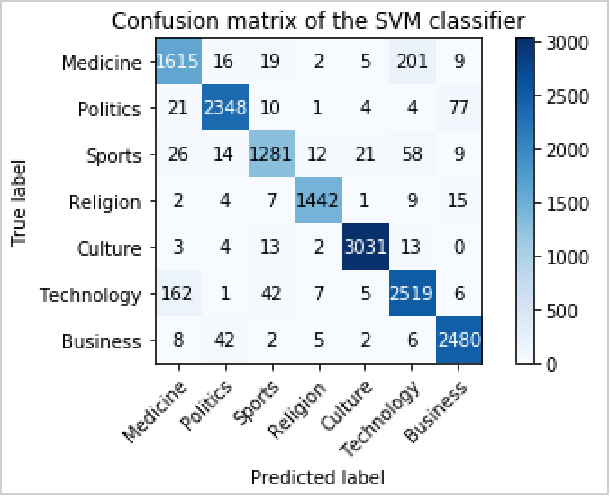


Figure 13. Confusion matrix for the best classifier on "Akhbarona".

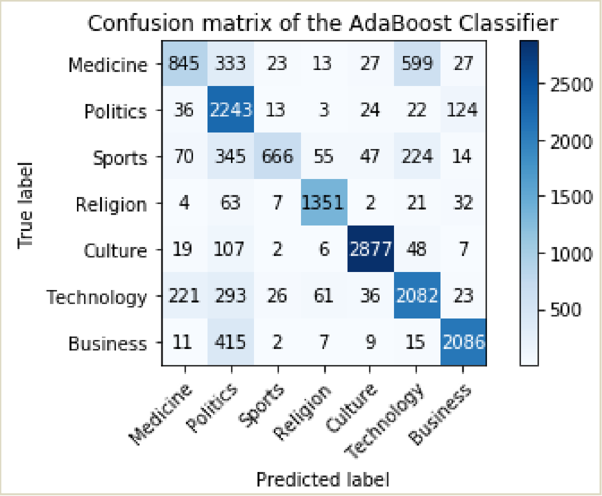


Figure 14. Confusion matrix for the worst classifier on "Akhbarona".

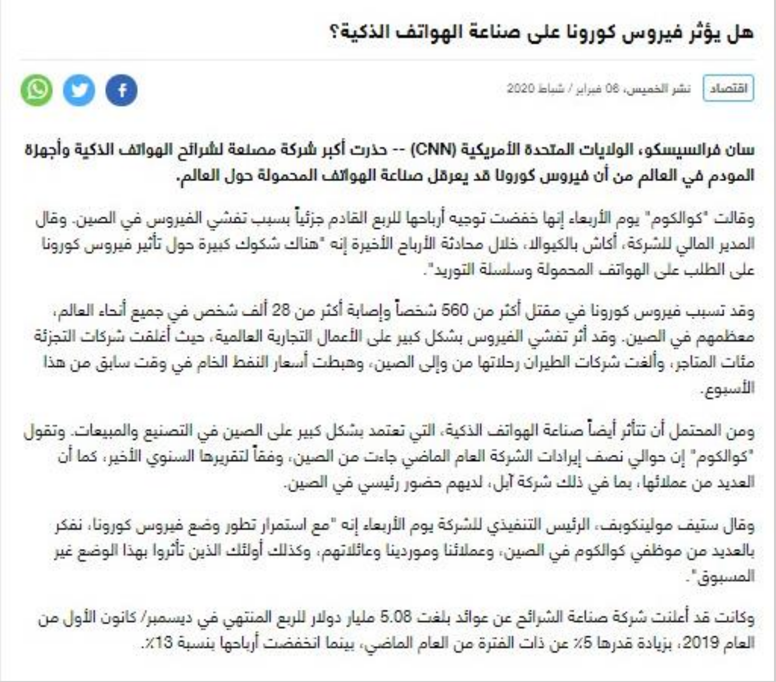
In Table 8, four different classifiers: (SVM, Logistic Regression, MLP and the voting classifier) scored the highest F1-score of 94%, while Ada-Boost scored the lowest score of 78%.

Table 8. Accuracy metrics for classifiers testing on “Akbarona”.

Algorithms	Precision	Recall	F1-score
Logistic Regression	0.94	0.94	0.94
SVC	0.94	0.94	0.94
DT Classifier	0.83	0.83	0.83
Multinomial NB	0.91	0.88	0.88
XGB Classifier	0.89	0.88	0.88
KNN Classifier	0.91	0.91	0.91
RF Classifier	0.88	0.88	0.88
Nearest Centroid	0.89	0.86	0.87
Ada-Boost Classifier	0.80	0.78	0.78
MLP Classifier	0.94	0.94	0.94
Voting Classifier	0.94	0.94	0.94

5.3.2 Multi-label Text Classification

During this phase, we test the performance of the OneVsRestXGBoost Classifier on recently published articles. Figure 15 shows an article taken from “Arabic.cnn.com”. The article discusses how the coronavirus could possibly impact the smartphones industry, since most of these factories are located in China. It’s originally tagged under “Business” only. Looking at the content of the article, it seems as it heavily talks about Technology as well. The XGBoost classifier picked up on both topics and predicted “Technology” and “Business” labels.



هل يؤثر فيروس كورونا على صناعة الهواتف الذكية؟

اقتصاد نشر الخميس، 06 فبراير / شباط 2020

سان فرانسيسكو، الولايات المتحدة الأمريكية (CNN) -- حذرت أكبر شركة مصنعة لشرائح الهواتف الذكية وأجهزة المودم في العالم من أن فيروس كورونا قد يعرقل صناعة الهواتف المحمولة حول العالم.

وقالت "كوالكوم" يوم الأربعاء إنها خفضت توجيه أرباحها للربع القادم جزئياً بسبب تفشي الفيروس في الصين. وقال المدير المالي للشركة، أكاش بالكيوالا، خلال محادثة الأرباح الأخيرة إنه "هناك شكوك كبيرة حول تأثير فيروس كورونا على الطلب على الهواتف المحمولة وسلسلة التوريد".

وقد تسبب فيروس كورونا في مقتل أكثر من 560 شخصاً وإصابة أكثر من 28 ألف شخص في جميع أنحاء العالم، معظمهم في الصين. وقد أثر تفشي الفيروس بشكل كبير على الأعمال التجارية العالمية، حيث أغلقت شركات التجزئة مئات المتاجر، وألغت شركات الطيران رحلاتها من وإلى الصين، وهبطت أسعار النفط الخام في وقت سابق من هذا الأسبوع.

ومن المحتمل أن تتأثر أيضاً صناعة الهواتف الذكية، التي تعتمد بشكل كبير على الصين في التصنيع والمبيعات. وتقول "كوالكوم" إن حوالي نصف إيرادات الشركة العام الماضي جاءت من الصين، وفقاً لتقريرها السنوي الأخير، كما أن العديد من عملائها، بما في ذلك شركة آبل، لديهم حضور رئيسي في الصين.

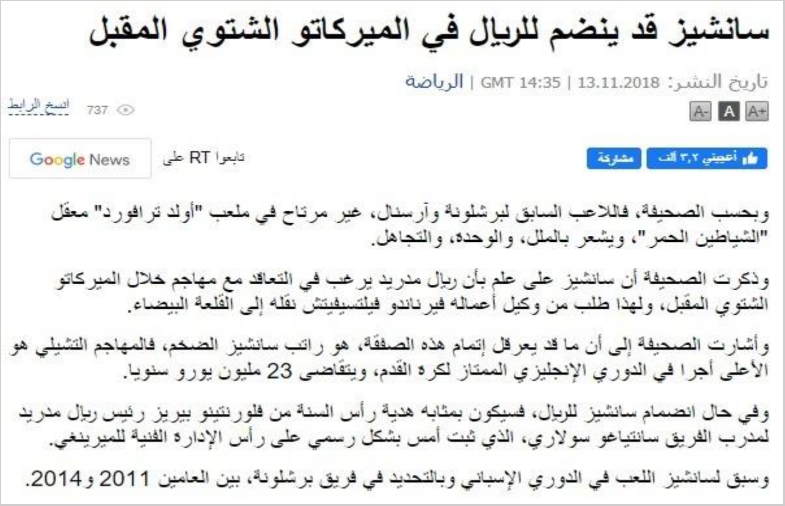
وقال ستيف مولينكوبف، الرئيس التنفيذي للشركة يوم الأربعاء إنه "مع استمرار تطور وضع فيروس كورونا، نفكر بالعديد من موظفي كوالكوم في الصين، وعملائنا وموردينا وعائلاتهم، وكذلك أولئك الذين تأثروا بهذا الوضع غير المسبوق".

وكانت قد أعلنت شركة صناعة الشرائح عن عوائد بلغت 5.08 مليار دولار للربع المنتهي في ديسمبر/ كانون الأول من العام 2019، بزيادة قدرها 5٪ عن ذات الفترة من العام الماضي، بينما انخفضت أرباحها بنسبة 13٪.

Figure 15. Example of a news article classified by OVR-XGBoost.

Another interesting article that has been correctly predicted by the model is shown in Figure 16. This time, it is taken from the “arabic.rt.com” website, where it was originally tagged with “Sports”, “Premier League”, “Real Madrid” and “Barcelona”. All of them were accurately predicted, without missing a single one. In Figure 17, the article originally is tagged under the “Business” category. Our trained classifier added to this tag two other labels: “Oil” and “Saudi Business”. This shows how specific the predictions by the model are.

Lastly, in Figure 18, we show an example of a misclassified article grabbed form the testing set. The article is originally tagged as “Middle East”, “Egypt” and “Yemen”. The predicted results show that the model has agreed with the author to a certain level, by predicting “Middle East”, “Yemen”, in addition to two other additional tags: “Iraq” and “Syria”. Along with the absence of the tag “Egypt”,



سانشيز قد ينضم للريال في الميركاتو الشتوي المقبل

تاريخ النشر: 13.11.2018 | GMT 14:35 | الرياضة

وبحسب الصحيفة، فاللاعب السابق لبرشلونة وآرسنال، غير مرتاح في ملعب "أولد ترافورد" معقل "الشياطين الحمر"، ويشعر بالملل، والوحده، والتجاهل.

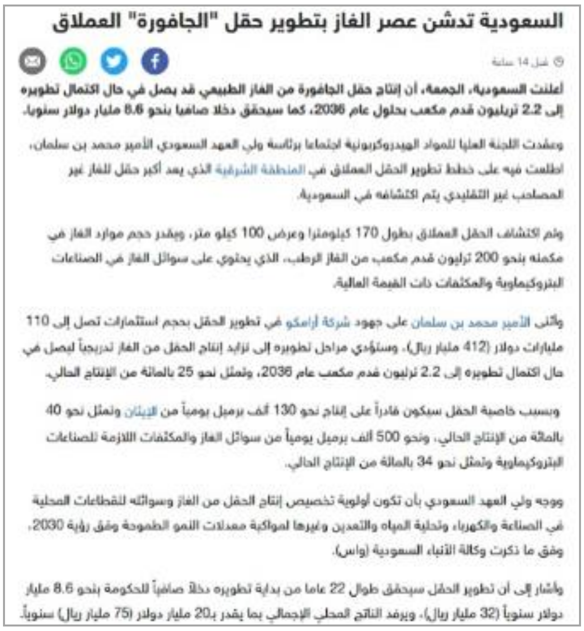
وذكرت الصحيفة أن سانشيز على علم بأن ريال مدريد يرغب في التعاقد مع مهاجم خلال الميركاتو الشتوي المقبل، ولهذا طلب من وكيل أعماله فيرناندو فيلتسيفيتش نقله إلى القلعة البيضاء.

وأشارت الصحيفة إلى أن ما قد يعرقل إتمام هذه الصفقة، هو راتب سانشيز الضخم، فالمهاجم التشيلي هو الأعلى أجرا في الدوري الإنجليزي الممتاز لكره القدم، ويتقاضى 23 مليون يورو سنويا.

وفي حال انضمام سانشيز للريال، فسيكون بمثابه هدية رأس السنة من فلورنتينو بيريز رئيس ريال مدريد لمدرب الفريق سانتياغو سولاري، الذي ثبت أمس بشكل رسمي على رأس الإدارة الفنية للميرينغي.

وسبق لسانشيز اللعب في الدوري الإسباني وبالتحديد في فريق برشلونة، بين العامين 2011 و2014.

Figure 16. Example of a correctly classified news article by OVR-XGBoost.



السعودية تدشن عصر الغاز بتطوير حقل "الجافورة" العملاق

قبل 14 ساعة

أعلنت السعودية، الجمعة، أن إنتاج حقل الجافورة من الغاز الطبيعي قد يصل في حال اكتمال تطويره إلى 2.2 تريليون قدم مكعب بحلول عام 2036، كما سيحقق دخلا صافيا بنحو 8.6 مليار دولار سنويا.

وعقدت اللجنة العليا للمواد الهيدروكربونية اجتماعا برئاسة ولي العهد السعودي الأمير محمد بن سلمان، اطلعت فيه على خطط تطوير الحقل العملاق في المنطقة الشرقية الذي يعد أكبر حقل للغاز غير المصاحب غير التقليدي يتم اكتشافه في السعودية.

وتم اكتشاف الحقل العملاق بطول 170 كيلومترا وعرض 100 كيلو متر، ويقدر حجم موارد الغاز في مكمنه بنحو 200 ترليون قدم مكعب من الغاز الرطب، الذي يحتوي على سوائل الغاز في الصناعات البتروكيماوية والمكثفات ذات القيمة العالية.

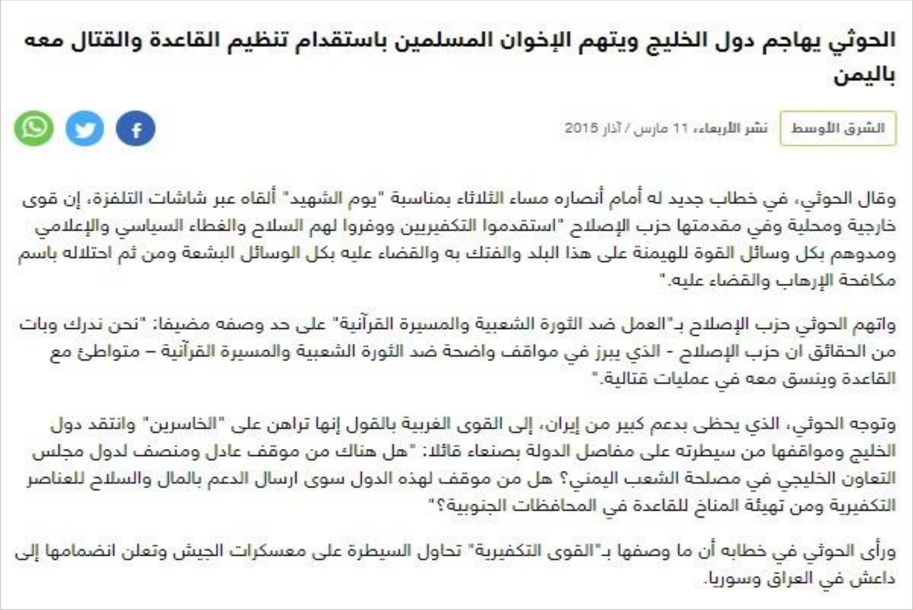
وأثنى الأمير محمد بن سلمان على جهود شركة أرامكو في تطوير الحقل بحجم استثمارات تصل إلى 110 مليارات دولار (412 مليار ريال)، وستؤدي مراحل تطويره إلى تزايد إنتاج الحقل من الغاز تدريجياً ليصل في حال اكتمال تطويره إلى 2.2 ترليون قدم مكعب عام 2036، وتمثل نحو 25 بالمائة من الإنتاج الحالي.

وبسبب خاصية الحقل سيكون قادراً على إنتاج نحو 130 ألف برميل يومياً من الإيثان وتمثل نحو 40 بالمائة من الإنتاج الحالي، ونحو 500 ألف برميل يومياً من سوائل الغاز والمكثفات اللازمة للصناعات البتروكيماوية وتمثل نحو 34 بالمائة من الإنتاج الحالي.

ووجه ولي العهد السعودي بأن تكون أولوية تخصيص إنتاج الحقل من الغاز وسوائله للقطاعات المحلية في الصناعة والكهرباء وتحلية المياه والتعدين وغيرها لمواكبة معدلات النمو الطموحة وفق رؤية 2030، وفق ما ذكرت وكالة الأنباء السعودية (واس).

وأشار إلى أن تطوير الحقل سيحقق طوال 22 عاما من بداية تطويره دخلاً صافياً للحكومة بنحو 8.6 مليار دولار سنوياً (32 مليار ريال)، ويرفد الناتج المحلي الإجمالي بما يقدر بـ20 مليار دولار (75 مليار ريال) سنوياً.

Figure 17. Example of a news article classified by OVR-XGBoost.



الحوثي يهاجم دول الخليج ويتهم الإخوان المسلمين باستقدام تنظيم القاعدة والقتال معه باليمن

الشرق الأوسط

نشر الأربعاء، 11 مارس / آذار 2015

وقال الحوثي، في خطاب جديد له أمام أنصاره مساء الثلاثاء بمناسبة "يوم الشهيد" ألقاه عبر شاشات التلفزة، إن قوى خارجية ومحلية وفي مقدمتها حزب الإصلاح "استقدموا التكفيريين ووفروا لهم السلاح والغطاء السياسي والإعلامي ومدوهم بكل وسائل القوة للهيمنة على هذا البلد والفتك به والقضاء عليه بكل الوسائل البشعة ومن ثم احتلاله باسم مكافحة الإرهاب والقضاء عليه."

واتهم الحوثي حزب الإصلاح بـ"العمل ضد الثورة الشعبية والمسيرة القرآنية" على حد وصفه مضيفا: "نحن ندرك وبات من الحقائق ان حزب الإصلاح - الذي يبرز في مواقف واضحة ضد الثورة الشعبية والمسيرة القرآنية – متواطئ مع القاعدة وينسق معه في عمليات قتالية."

وتوجه الحوثي، الذي يحظى بدعم كبير من إيران، إلى القوى الغربية بالقول إنها تراهن على "الخاسرين" وانتقد دول الخليج ومواقفها من سيطرته على مفاصل الدولة بصنعاء قائلا: "هل هناك من موقف عادل ومنصف لدول مجلس التعاون الخليجي في مصلحة الشعب اليمني؟ هل من موقف لهذه الدول سوى ارسال الدعم بالمال والسلاح للعناصر التكفيرية ومن تهيئة المناخ للقاعدة في المحافظات الجنوبية؟"

ورأى الحوثي في خطابه أن ما وصفها بـ"القوى التكفيرية" تحاول السيطرة على معسكرات الجيش وتعلن انضمامها إلى داعش في العراق وسوريا.

Figure 18. Example of a misclassified news article by OVR-XGBoost from the test set.

we found that the model's overall prediction was much more suitable for the article and its content, proving that it can rectify human errors.

6. CONCLUSIONS

In summary, this paper has presented both a multi-class text classifier system for Arabic news articles and a multi-label classifying system. We propose a single-labeled dataset that holds over 89k Arabic news articles divided into 4 categories, from seven websites. Another dataset is also proposed that contains 293k multi-labeled Arabic articles along with their tags, scraped from 10 different websites. The first dataset was examined by 11 different classifiers, all trained and tested using the single-labeled dataset. From 87% to 97% is the range of final accuracies, where the SVM classifier scored the top accuracy and F1-score. The voting classifier was used in hopes of improving the accuracy, but the resulting percentage is comparable to the SVM classifier's score. To further explore how robust our proposed system is, we conducted additional experiments on a recently reported dataset "Akhbarona". The dataset has 7 classes and the results of using it for training and testing the same classifiers were as good as on our dataset. The highest accuracy was scored by the SVM classifier as well. The second dataset was examined by implementing and comparing the results of two different classifiers. A custom accuracy metric was implemented to evaluate the performance along with hamming loss metric. The OVR-XGBoost classifier performed better than OVR-Logistic Regression classifier, scoring 91.3% accuracy, while Logistic achieved 87.6%.

In future, we intend to increase the number of classes in the single-labeled dataset and the number of labels in the multi-labeled dataset, which will require more scraping scripts. We would also like to compare and study our results with some deep learning methods.

ACKNOWLEDGEMENT

We would like to thank the University of Sharjah for funding this work under targeted research project no.: 1702141151-P.

REFERENCES

[1] A. Elnagar and O. Einea, "BRAD 1.0: Book Reviews in Arabic dataset," Proc. of the IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pp. 1-8, DOI: 10.1109/AICCSA.2016.7945800, Agadir, Morocco, 2016.

[2] A. Elnagar, Y. Khalifa and A. Einea, "Hotel Arabic-reviews Dataset Construction for Sentiment Analysis Applications," Book Chapter in Intelligent Natural Language Processing: Trends and Applications, pp. 35-52, DOI: 10.1007/978-3-319-67056-0_3, 2017.

[3] A. Elnagar, L. Lulu and O. Einea, "An Annotated Huge Dataset for Standard and Colloquial Arabic Reviews for Subjective Sentiment Analysis," Procedia Computer Science, vol. 142, pp. 182-189, 2018.

[4] N. Boudad, R. Faizi, R. O. Thami and R. Chiheb, "Sentiment Analysis in Arabic: A Review of the Literature," Ain Shams Engineering Journal, vol. 9, pp. 2479-2490, 2017.

[5] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud and P. Duan, "Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification," Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING), pp. 2418–2427, Osaka, Japan, 2016.

[6] H. Almuaidi, S. Alqrainy and A. Ayesh, "Automated Tagging System and Tagset Design for Arabic Text," International Journal of Computational Linguistics Research, vol. 1, pp. 55-62, 2010.

[7] A. Al-Alwani and M. Beseiso, "Arabic Spam Filtering Using Bayesian Model," International Journal of Computer Applications, vol. 79, pp. 11-14, 2013.

[8] Y. Li, X. Nie and R. Huang, "Web Spam Classification Method Based on Deep Belief Networks," Expert Syst. Appl., vol. 96, pp. 261-270, 2018.

[9] S. Malmasi and M. Dras, "Language Identification Using Classifier Ensembles," Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, Association for Computational Linguistics, pp. 35–43, Hissar, Bulgaria, 2015.

[10] M. El-Haj, P. Rayson and M. Aboelezz, "Arabic Dialect Identification in the Context of Bivalency and Code-Switching," Proceedings of the 11th International Conference on Language Resources and

Evaluation (LREC 2018), European Language Resources Association (ELRA), pp. 3622-3627, Miyazaki, Japan, 2018.

[11] N. Y. Habash, Introduction to Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies, Edited by Graeme Hirst, [Online], Available: https://doi.org/10.2200/S00277ED1V01Y201008HLT010, 2010.

[12] S. C. Dharmadhikari, M. Ingle and P. Kulkarni, "Empirical Studies on Machine Learning Based Text Classification Algorithms," Advanced Computing: An International Journal, vol. 2, pp. 161-169, 2011.

[13] C. C. Aggarwal and C. Zhai, "A Survey of Text Classification Algorithms," Mining Text Data, pp. 163-222, 2012.

[14] V. Korde and C. N. Mahender, "Text Classification and Classifiers: A Survey," International Journal of Artificial Intelligence & Applications, vol. 3, pp. 85-99, 2012.

[15] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig and N. A. Mahyoub, "Automatic Arabic Text Categorization: A Comprehensive Comparative Study," Journal of Information Science, vol. 41, no. 1, pp. 114-124, 2015.

[16] A. M. Sbou, "A Survey of Arabic Text Classification Models," International Journal of Informatics and Communication Technology, vol. 8, pp. 25-28, 2019.

[17] M. Saad and W. Ashour, "Arabic Text Classification Using Decision Tree," Proc. of the 12th International Workshop on Computer Science and Information Technologies (CSIT'2010), vol. 2, 2010.

[18] F. Harrag, E. El-Qawasmeh and P. Pichappan, "Improving Arabic Text Categorization Using Decision Trees," Proc. of the 1st International Conference on Networked Digital Technologies, pp. 110-115, Ostrava, Czech Republic, 2009.

[19] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. S. Khorsheed and A. Alrajeh, "Automatic Arabic Text Classification," JADT 2008: 9es Journées Internationales d'Analyse Statistique des Données Textuelles, pp. 77-83, 2008.

[20] M. E. Kourdi, A. Bensaid and T. Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," Workshop on Computational Approaches to Arabic Script-based Languages, DOI: 10.3115/1621804.1621819, 2004.

[21] H. M. Noaman, S. Elmougy, A. Ghoneim and T. T. Hamza, "Naive Bayes Classifier-based Arabic Document Categorization," Proc. of the 7th International Conference on Informatics and Systems (INFOS), pp. 1-5, Cairo, Egypt, 2010.

[22] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," International Arab Journal of e-Technology, vol. 2, no. 2, pp. 124-128, 2011.

[23] M. J. Bawaneh, M. Alkoffash and A. I. Rabea, "Arabic Text Classification Using K-NN and Naive Bayes," Journal of Computer Science, vol. 4, no. 7, pp. 600-605, 2008.

[24] T. F. Gharib, M. B. Habib and Z. T. Fayed, "Arabic Text Classification Using Support Vector Machines," International Journal of Computers and Their Applications, vol. 16, no. 4, pp. 192-199, 2009.

[25] F. Harrag and E. Al-Qawasmah, "Improving Arabic Text Categorization Using Neural Network with SVD," Journal of Digital Information Management, vol. 8, no. 4, pp. 233-239, 2010.

[26] I. Hmeidi, B. Hawashin and E. El-Qawasmeh, "Performance of KNN and SVM Classifiers on Full Word Arabic Articles," Advance Engineering Informatics, vol. 22, no. 1, pp. 106-111, 2008.

[27] S. Boukil, M. Biniz, F. E. Adnani, L. Cherrat and A. E. Moutaouakkil, "Arabic Text Classification Using Deep Learning Techniques," International Journal of Grid and Distributed Computing, vol. 11, pp. 103-114, 2018.

[28] F. A. Zaghoul and S. Al-Dhaheri, "Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks," Proc. of the 15th International Conference on Computer Modelling and Simulation (UKSim), pp. 485-490, Cambridge, UK, 2013.

[29] M. M. Al-Tahrawi and S. N. Al-Khatib, "Arabic Text Classification Using Polynomial Networks," Journal of King Saud University- Computer and Inform. Sciences Archive, vol. 27, pp. 437-449, 2015.

[30] L. Lulu and A. Elnagar, "Automatic Arabic Dialect Classification Using Deep Learning Models," Procedia Computer Science, vol. 142, pp. 262-269, 2018.

[31] A. A. Altowayan and A. Elnagar, "Improving Arabic Sentiment Analysis with Sentiment-specific Embeddings," Proc. of IEEE International Conference on Big Data (Big Data), pp. 4314-4320, Boston, MA, USA, 2017.

[32] A. Elnagar, R. Ismail, B. Alattas and A. Alfalasi, "Automatic Classification of Reciters of Quranic Audio Clips," Proc. of the 15th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA), pp. 1-6, Aqaba, Jordan, 2018.

[33] A. Elnagar and M. Lataifeh, "Predicting Quranic Audio Clips Reciters Using Classical Machine Learning Algorithms: A Comparative Study. In book: Recent Advances in NLP: The Case of Arabic Language, vol. 874, pp. 187-209, DOI: 10.1007/978-3-030-34614-0_10, 2020.

[34] A. Elnagar, O. Einea and R. A. Debsi, "Automatic Text Tagging of Arabic News Articles Using Ensemble Deep Learning Models," Proceedings of the 3rd International Conference on Natural Language and Speech Processing (ICNLSP), pp. 59-66, Trento, Italy, 2019.

[35] A. Elnagar, R. Al-Debsi and O. Einea, "Arabic Text Classification Using Deep Learning Models," Information Processing and Management, vol. 57, no. 1, pp. 102-121, 2020.

[36] A. El-Halees, "A Comparative Study on Arabic Text Classification," Egyptian Computer Science Journal, vol. 30, 2008.

[37] R. Al-Shalabi and R. Obeidat, "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing," Proc. of the 6th International Conference on Informatics and Systems, pp. 108-112, Cairo, Egypt, 2008.

[38] G. I. Raho, R. Al-Shalabi, G. Kanaan and A. Nassar, "Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study," International Journal of Advanced Computer Science and Applications, vol. 6, no. 2, pp. 192-195, 2015.

[39] A. Mesleh, "Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System," Journal of Computer Science, vol. 3, no. 6, pp. 430-435, 2007.

[40] B. Hawashin, A. Mansour and S. A. Aljawarneh, "An Efficient Feature Selection Method for Arabic Text Classification," International Journal of Computer Applications, vol. 83, no. 17, pp. 1-6, 0975-8887, 2013.

[41] N. Alalyani and S. L. Marie-Sainte, "NADA: New Arabic Dataset for Text Classification," International Journal of Advanced Computer Science and Applications, vol. 9, DOI: 10.14569/IJACSA.2018.090928, 2018.

[42] I. A. El-Khair, 1.5 Billion Words Arabic Corpus, Computer Science, Computation and Language, ARXIV, ABS/1611.04033, 2016.

[43] T. Gonçalves and P. Quaresma, "The Impact of NLP Techniques in the Multi-label Text Classification Problem," Intelligent Information Systems, vol. 25, pp. 424-428, 2004.

[44] L. A. Qadi, H. E. Rifai, S. Obaid and A. Elnagar, "Arabic Text Classification of News Articles Using Classical Supervised Classifiers," Proc. of the 2nd International Conference on New Trends in Computing Sciences (ICTCS), pp. 1-6, DOI: 10.1109/ICTCS.2019.8923073, Amman, Jordan, 2019.

[45] A. M. Hassanein and M. Nour, "A Proposed Model of Selecting Features for Classifying Arabic Text," Jordanian Journal of Computers and Information Technology, vol. 5, no. 3, 2019.

ملخص البحث:

يتمثل تصنيف النصوص في الوسم الآلي لوثيقة نصية بأنسب وصفٍ أو أكثر، له (لها) علاقة بتلك الوثيقة. يهدف هذا العمل الى وسم الوثائق بما يلائمها من أوصافً بناءً على خصائص المفردات المستخدمة فيها. ولتحقيق هذا الهدف، تم تكوين مجموعتي بيانات كبيرتين من المنافذ الخاصة بالأخبار باللغة العربية. تتكون مجموعة البيانات الأولى من مقالاتٍ موسومةٍ بوصفٍ واحد من بين 4 حقول هي: (الأعمال، والشرق الأوسط، والتكنولوجيا، والرياضة). أما مجموعة البيانات الثانية، فهي أكبر من الأولى، وتحتوي على مقالاتٍ موسومة بعدة أوصافٍ. وستُتاح مجموعتا البيانات لمجتمع البحث فيما يتعلق باللغويات الحاسوبية باللغة العربية.

ولفحص مدى الفائدة من مجموعتي البيانات، تم تطبيق مصفوفةٍ من عشرة مصنِّفات ضحلة التعلُّم. بالإضافة الى ذلك، تم تطبيق نموذج تجميعٍ من أجل جمع أفضل المصنِّفات معاً في مصنِّف واحدٍ اعتماداً على تصويتٍ الأغلبية. وتراوح أداء المصنِّفات على مجموعة البيانات الأولى بين 87.7% لمصنِّف (Ada-Boost) و97.9% لمصنِّف (SVM).

وقد أكدّ تحليل بعض المقالات التي فشل النموذج في تصنيفها بشكل صحيح الحاجة الى تصنيفٍ متعدد الأوصاف على العكس من التصنيف القائم على وصفٍ واحد؛ للحصول على نتائج تصنيف أفضل. وقد جرى استخدام مصنِّفاتٍ متوافقة مع مهمات التصنيف متعدد الأوصاف؛ مثل (LR) و (XGBoost)، وتم اختبار هذه المصنِّفات على مجموعة البيانات الثانية. وتم تطوير مقياس للدقة لتقييم الأداء، الى جانب مقياس لفقْد الطرق. وأثبت المصنِّف (XGBoost) أنه الأفضل؛ إذ سجّل دقة بلغت 91.3%، في حين كانت دقة المصنِّف (LR) 87.6%.