Arabic Text Mining Using Rule Based Classification

Fadi Thabitah
MIS Department, Philadelphia University
Amman, Jordan
Ffayez@philadelphia.edu.jo

Omar Gharabeh
Computing and IS Department,
Brunel University, UK
Ogharaibeh@brunel.edu.jo

Rashid Al-Zubaidy
CIS Department, Philadelphia University
Amman, Jordan
Rzubidy@philadelphia.edu.jo

Abstract. A well-known classification problem in the domain of text mining is text classification, which concerns about mapping textual documents into one or more predefined category based on its content. Text classification arena recently attracted many researchers because of the massive amounts of online documents and text archives which hold essential information for a decision-making process. In this field, most of such researches focus on classifying English documents while there are limited studies conducted on other languages like Arabic. In this respect, the paper proposes to investigate the problem of Arabic text classification comprehensively. More specifically the study measures the performance of different rule based classification approaches adopted from machine learning and data mining towards the problem of text Arabic classification. In particular, four different rule based classification approaches: Decision trees (C4.5), Rule Induction (RIPPER), Hybrid (PART) and Simple Rule (One Rule) are evaluated against the published Corpus of Contemporary Arabic Arabic text collection. This experimentation is carried out by employing a modified version of WEKA business intelligence tool. Through analysing the produced results from the experimentation, we determine the most suitable classification algorithms for classifying Arabic texts.

Keywords: Arabic text mining; classification; data mining; rule-based classifiers.

1. Introduction

In today’s information-oriented society, the number of online documents has been growing dramatically. These documents conceal useful information that can be utilised by decision makers in their key business activities. The process of finding and generating the hidden and useful information from the online document manually by domain experts is hard and time consuming. This is due to the fact that the available amounts of online textual data are massive and having large dimensionality. Consequently, employing intelligent tools to discover essential information automatically from textual documents grants companies to make right decisions that work for improving their competitive advantages.

Text mining (TM) is such a discipline that has been emerging to respond the need of discovering useful knowledge from unstructured text data format (Kroeze et al., 2003). Unlike, data mining which is mainly concerned about finding valuable patterns from highly structured data format (Weiss et al., 2005). Nevertheless, TM derives a great deal of its inspiration from seminal studies on data mining. As a result, TM and data mining applications demonstrate several high level architectural similarities (Feldman and Sanger, 2007). There are many domains surrounding TM including information retrieval, information extraction, summarisation, clustering and classification (Song, 2009).

Text classification (TC) is one of the significant domains of TM that is responsible about understanding, recognising, and organising various types of textual data collections (Guo et al., 2010). TC is concerned about predicting a specific category of incoming textual document. This type of prediction is called “supervised” learning in which the class category is predetermined within the input text collection (Kantardzie, 2003). TC is a multi phase process that includes, processing the textual
documents, learning the document by an algorithm and evaluation of output models (Shi and Kong, 2009). There are many classification approaches towards TC that has been adopted from data mining and machine learning, e.g. Decision trees, Naïve bayes, Support Vector Machine (SVM) and neural network (Feldman and Sanger, 2007). These approaches have been mainly investigated on classifying English documents (Tan, 1999). But, researchers paid little interest for applying these approaches to other languages such as Arabic.

A few researchers have applied a number of classification approaches which are solely applicable for the problem of Arabic TC, i.e. Naïve bayes (El-Kourdi et al., 2004), SVM (Harrag et al., 2009) and Decision tree (Harrag et al., 2009). However, researchers concluded that Arabic TC is a very challenging task due to language complexity. For instance in Arabic morphology, words have affluous meanings and contain a great deal of grammatical and lexical information. In syntactic structure, Arabic sentence formation differs from English. In this regard, the Arabic text documents are required extensive pre-processing routines to build accurate classification model.

Most of the previous works on Arabic TC attempt only to achieve the classification accuracy from the above mentioned learning approaches. Some of domain experts have also interests in understanding the produced models that are formed as “IF-Then” patterns (Kantardzic, 2003; Sebastiani, 2002). This is since such models are easy to interpret and can be modified by users manually (Song, 2009). Thus, this paper is devoted to investigate different rule based classification approaches on the problem of Arabic TC. Primarily, OneRule, PART, C4.5, and RIPPER learning methods are applied to Corpus of Contemporary Arabic (CCA) Arabic data collection to measure their performance and effectiveness with reference to different text evaluation metrics such as error rate, precision, recall, and the number of derived rules. This paper extends a recently accepted conference article (Thabtah et al., 2011).

The rest of the paper is organized as follows: Arabic TC related works is surveyed in Sec. 2. Section 3 is about the research methodology used, and Sec. 4 describes the Arabic data collection. Experimentations and results analysis are demonstrated in Sec. 5, and finally the conclusions and further research are given in Sec. 6.

2. Arabic Language and Related Works

2.1 Arabic language

Arabic is one of widely spoken languages of the world, which is a primary language in Arab states and a secondary language in many other countries. The alphabet of the language consists of 28 letters. Nevertheless, there is also character called Hamza (\(\dot{\alpha}\)), which is considered as a separate letter by some linguists. Unlike Latin and Cyrillic style languages, the writing direction in Arabic is from right to left. The style of writing letters in a word varies depending on the position of the letter within the word. So, if the letters comes at the beginning, middle or at the end of the word, the letter shapes changes. For example, the letter (\(\dot{\imath}\)) appears differently in the following words (\(\scriptstyle\dot{\imath}\)), (\(\dot{\imath}\)), (\(\imath\)) because of its locations. A further difference of Arabic from English is not-having a letter capitalisation in Arabic. Moreover, there are diacritics in Arabic that are symbols placed above or below the letters to double the letter in the pronunciation or to give short vowels (Duwairi, 2007).

2.2 Related works

Reviewing the existing related works proved that there are several methods which have been proposed by researchers towards Arabic TC. For classifying Arabic text sources the N-Gram Frequency technique is investigated by (Khreisat, 2006). This method is based on both Dice similarity and Manhattan distance measures in classifying an Arabic corpus. For this research the Arabic corpus was obtained from various online Arabic newspapers. The data is associated with four categories. After performing several pre-processing on the data, and experimentation, the results indicated that the “Sport” category outperformed the other categories with respect to recall evaluation measure. The least category was “Economy” with around 40% recall. In general the N-gram Dice similarity measure figures outperformed that of Manhattan distance.

A modified version of Artificial Neural Network (ANN) method is proposed for classifying Arabic texts by (Harrag and El-Qawasmeh, 2009). The authors have used a Singular Value Decomposition (SVD) for data representation, which is a new representation space of the observations. A collection of Prophet Mohammad’s “Peace Be Upon Him” Hadeeth was collected from the “Nine Hadeeth Book”. The data consists of 453 documents that are associated with 14 categories. A comparison between the proposed method (ANN with SVD) and the original ANN was carried out against the Arabic corpus with reference to F1 evaluation measure. The results revealed that (ANN with SVD) outperformed the classic ANN when dimensionality increased.

Decision trees classification approach and the effect of feature selection on the predictive accuracy were applied to the problem of Arabic TM in the work of (Harrag et al., 2009). Specifically, the authors have compared ID3
algorithm with the known statistical method of Naive Bayes for two Arabic corpuses collected from Arabian scientific encyclopaedia (Hal Taalam) and Prophet Mohammad’s “Peace Be Upon him” Hadeeths “Nine Hadeeth Book”. The initial results indicated an improvement on average 10% and 26% on the “Hadeeth” and “Scientific” datasets respectively when employing feature selection instead of the whole corpus. Moreover, the F1 has also improved by around 2.5% when using decision trees over that of Naive Bayes.

(El-Halees, 2007) has applied a probabilistic supervised learning method called the Maximum Entropy on real Arabic corpus collected from Aljazeera news website. The author has tested the proposed method with and without pre-processing phase with regards to F1 evaluation measure. The results showed that F1 accuracy has increased from 68% to 84% proving that removing noise (stopwords, tokenisation, stemming, etc) improves the classification accuracy in the selected Arabic dataset.

The performance of three classification data mining algorithms (KNN, Naive Bayes, Distance-based learning) have been evaluated with respect to error rate, recall, and precision measures on Arabic data collection by (Duwairi, 2007). The benchmark used in the experiment consists of 1,000 documents and 10 categories. The author has processed the data before performing the training where punctuation marks and stopwords have been removed. The obtained results have shown a variety of the performance of the categories. For instance, “Internet” category achieved only 22% Recall, whereas “Economic” category achieved 98% Recall. Overall, the author has concluded that Naive Bayes algorithm outperformed the other considered classifiers with respect to the above mentioned measures.

(Al-Harbi et al., 2008) have evaluated the performance of two common data mining approaches mainly SVM and decision trees (C4.5) (Quinlan, 1993) on different Arabic benchmarks. The data which consists of 17,658 documents have been gathered from different sources, especially Saudi Press Agency, Saudi newspapers, Internet articles, and discussion forums. The experimental tests have revealed that C4.5 algorithm achieved on average 10% more in accuracy than SVM on the considered data.

2.3. Discussion

Most scholars (e.g. Khreisat, 2006; Harrag and El-Qawasmeh, 2009; Duwairi, 2007) consider classifying Arabic text documents as a hard task. This is due to complexity and richness of the Arabic morphological analysis (El-Halees, 2007). In general, most of the current research works conducted on Arabic TM is just simple comparison studies. These researches mainly focused on adopting some of the data mining and machine learning approaches such as probabilistic, decision trees and SVM which are solely designed for English text collections to that of Arabic corpuses. For instance, (Al-Harbi et al., 2008) have tested only two data mining methods and (El-Halees, 2007) has examined the performance of one probabilistic method with and without pre-processing.

Moreover, the lack of standardised published Arabic corpuses is also unavailable or rare. Such works can be used as key datasets for researchers in related fields to compare the results. In fact, most of the related research articles obtain data from online newspapers and websites. Such works usually do not publish their data for other researchers to utilise. Consequently, the confidence in the results derived from such experimental studies is not high enough. Furthermore, the performance of the adopted data mining approaches is biased to such datasets and sometimes ambiguous. For example, the research conducted by (Al-Harbi et al., 2008) showed that decision tree algorithms outperformed the SVM with respect to classification accuracy. However, the majority of international research on English TM proved that SVM is the best machine learning approach with reference to predictive accuracy (Joachims, 2001).

In general, comprehensive experimental and critical research studies that cover most of the common rule based business intelligence techniques are rare. This is one of key motivations for this research work. Thus, this paper aims to investigate four different rule based learning methods in data mining, which are used to derive simple “If-Then” rules. Further, the work proposes to establish the performance on published Arabic dataset collection (CCA) so future researchers might get access and utilise it in their experiments and compare their derived results with this work.

3. Research Methodology

Since the research is a particular form of goal — oriented acting, selecting a proper research strategy and technique is a crucial task to meet the main goal (Jonker and Pennink, 2010). The research under consideration is developed under the applied research category which manipulates data, techniques, processes, and concepts that usually produced from the collected information during the research step (Jonker and Pennink, 2010). Accordingly, applied research attempts to achieve more knowledge about a particular problem and to supply the enhancement of that problem. More specifically, the research design is
mainly experimental one since the performance of different data mining algorithms particularly rule based are applied to a collection of Arabic text documents in order to derive useful knowledge.

3.1. Mixed methodological approach

The research is systematic process in which defining the objective, controlling the data, and communicating the findings take place within recognised frameworks and associated with existing guidelines (Williams, 2007). The research success depends on selecting an appropriate research methodology. Employing both qualitative and quantitative methods meet the purposes of this research. Quantitative approach is concerned about data collecting in quantitative form which can be subject to extensive quantitative analysis in a formal method (Kothari, 2009). This research strategy is applied for analysing the experimentation results namely; precision, recall, error rate, and the number of generated rules derived from different classification approaches. Qualitative research involves subjective evaluation of approaches, opinions, and behaviour in which results are gained either in non-quantitative type or in the type which are not subjected to detailed quantitative analysis (Kothari, 2009).

Since one of this paper’s objectives is to conduct a literature review on Arabic TC, the qualitative approach is considered as a suitable research approach to carry out this task effectively. Integrating both of these approaches into a single study is known as a mixed research methodology (Williams, 2007). The importance of employing this type of methodological approach is to achieve strong points and to reduce weak points of quantitative and qualitative research approaches.

4. The CCA Data Collection

The data used in the experimentation was obtained from the online sources of Leeds University which is offered as free of charge for researchers and students. This data collection is called CCA which is considered as a real Arabic text collection since it was collected from various Arabic resources by its author. However, this dataset is considered as a small set in terms of size comparing to English collections such as UCI. Unlike such English datasets, the availability of standard Arabic data is limited. In this respect, this study uses CCA which is a small scale of Arabic text collection.

The CCA dataset consists of 14 different Arabic files associated with certain class labels i.e. Politics, Religion, Sports, Science, etc. The author gathered this text collection from different online Arabic sources over a two years period. The total number of documents in all datasets is 415. These documents vary with reference to the number of terms. Moreover, some of the document categories (classes) are associated with a limited number of documents whereas; others are associated with large numbers. For instance, the category “Sport” includes only four Arabic documents, while “Autobiography” category takes in 73 documents. Detailed sample of the dataset is shown in Fig. 1.

![number of instances in each category](image)
4.1. Data processing and generation
The results in experimentation are generated through employing intelligence tools for the CCA dataset. Mainly two tools: WEKA (Weka, 2000) and Khoja Stemmer are chosen to accomplish this task. The research employs the WEKA business intelligence tool that is available online and has an open-access for everyone. Prior to conducting the experimentation, this tool has been modified to interpret Arabic language. In particular, the pre-processing filters within WEKA tool are not applicable to handle Arabic text. In this regards, a common Arabic stemmer (called Khoja) is utilised, which is a stemming application designed to process Arabic dataset. This stemmer is an open source for academic researcher, and accessible online.

For the CCA Arabic text collection, the data is organised with reference to document categories in which each document is stored as a separate text file in its related category folder. Each document is represented in a numerical vector where terms in the document correspond to numerical values according to their frequency in that document by utilising StringToVector method in the WEKA tool. To conduct pre-processing operations such as stemming and stopwords elimination, approaches from (El-Kourdi et al., 2004) and (Syiam et al., 2006) have been adopted. This includes the following stages:

(1) Each document in the CCA dataset is processed to discard the numerical data as well as punctuation marks.
(2) All the non-Arabic texts and function words are deleted.
(3) The documents in the CCA dataset are stemmed in which all Arabic word derivatives are transformed into their single common root.
(4) The non useful Arabic frequent list of words (stopwords) was eliminated from the CCA data.

4.2. Data balancing
In the previous section, the number of documents associated with each category was plotted in a Fig. 1. In that figure, seven categories are associated with a number of documents in the range of 25–32, and three categories are associated with a number of documents in the range of 9–10. The rest of the categories are associated with different number of documents in which some have limited numbers and others high numbers. Therefore, it can be stated that the CCA dataset in this research paper is quit balanced as two groups of categories (10 categories) have close range with reference to their number of documents. However, there are four categories within the dataset which are either have limited number of documents such as “Sport” or have large number of documents such as “Autobiography”, “Science” and “Tourist and Travel”. This is since some real world datasets are unevenly distributed where there are majority and minority classes. Some researchers, e.g. (Chawla et al., 2004) tend to apply data balancing methods such as over sampling, under sampling, random sampling, etc. in order to balance such unevenly distributed categories which consequently improves the performance of outputted classification models.

According to (Chawla et al., 2004) it is hard to determine whether the under-sampling or over-sampling will affect the general performance of the classification models. Thus, there are no balancing methods utilised against the original CCA dataset. In more details, the next section highlights the impact of the category document frequency on the performance of the generated classifiers with respect to the error rate.

5. Experimental Results
5.1. Introduction
This section presents a comparison of the rule based classification algorithms used to produce the results in this paper. In particular, four different classification approaches — decision tree (C4.5), rule induction, simple Rule, and hybrid, are evaluated for CCA text collections according to different evaluation measures.

5.2. Experiments details
Table 1 demonstrates classification methods employed in the experimentation such as the algorithm name, its learning approach, and the implementation name in WEKA. The choice of the classification algorithms in Table 1 is based on different learning approaches which are applied to produce classification models. In particular, C4.5 classification algorithm (Quinlan, 1993) uses divide and conquer approach, i.e. decision tree, to construct the

<table>
<thead>
<tr>
<th>Classification approach</th>
<th>Classification learning algorithm</th>
<th>Algorithm implementation name in WEKA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision trees</td>
<td>C4.5</td>
<td>Trees.J48</td>
</tr>
<tr>
<td>Simple rule</td>
<td>One Rule</td>
<td>Rules.OneR</td>
</tr>
<tr>
<td>Rule induction</td>
<td>RIPPER</td>
<td>Rules.JRip</td>
</tr>
<tr>
<td>Hybrid</td>
<td>PART</td>
<td>Rules.Part</td>
</tr>
</tbody>
</table>
The performance of comparison of the above selected algorithms is carried out through employing a number of known TM evaluation measures including recall, precision and error rate. The investigation also covers the numbers of rules derived by the algorithms in the experimentation phase. Finally, a depth analysis of generated results are also conducted in order to discover the most applicable rule based classification approach towards the problem of Arabic TC.

In the experimentation, 10-fold cross validation testing method is employed to produce the classification models and their performance measure results. Ten-fold cross validation is a popular testing method in data mining that works as follows according to (Witten and Frank, 2000):

“The data is arbitrary divided into ten blocks where the class label is represented in the same proportions as in each data block. Each block is carried out in turn and the learning approach trained on the remaining nine-tenths blocks. Afterwards the error rate is computed on the holdout block. As a result, the learning approach is performed a total of ten times on different training datasets. Lastly, the 10 error rates are averaged in order to generate the final error rate.

One more point is that the experimentation is carried out on a Windows Vista Pentium IV 2.1 Ghz, 3 GB RAM machine.

5.3. Results and analysis

Table 2 represents the average precision and recall results for the selected classification algorithms. The figures in that table give a clear indication that C4.5 algorithm outperformed the remaining classification algorithm on the CCA dataset. In particular, C4.5 achieved more precision on average 77.6%, 1%, and 1.2% respectively than OneRule, RIPPER, and PART algorithms. Additionally, this algorithm gained respectively more recall on average 59.7%, 1.6%, 1.1% than OneRule, RIPPER, and PART algorithms.

Table 2 also demonstrates that an OneRule algorithm is the least applicable classification approach towards the CCA Arabic dataset due to the low results of precision and recall. This is since OneRule algorithm seeks to derive the rules that covers the majority class label in the dataset. In the CCA Arabic corpus, most of the categories (class labels) are associated with a limited number of documents, apart from “Autobiography” and “Tourist and Travel” classes. In this regards, only two rules have been derived using such an algorithm that correctly covers only 129 documents out of 427 documents. In other words, OneRule learning approach showed high sensitivity towards the dataset where it particularly was unable to derive rules for the balanced class labels and only produced rules for the majority class label which in fact the cause of the unbalancing in the CCA dataset. On the other hand, C4.5 algorithm produced 24 rules that represent most of the classes in the training dataset. This method correctly covered 384 out of 427 documents. This means C4.5 learning approach is somehow not impacted with the unbalanced categories. In general, all classification algorithms except OneRule showed very competitive performance with regards to precision and recall on the CCA corpus, as their generated results are very close to each other.

The confusion matrices for the selected algorithms in the experimentation have been derived as shown in Fig. 2. The confusion matrix in the figure below represents the distribution of documents for each class in the training dataset. For instance, for the C4.5 (J48) algorithm, class “Autobiography” have correctly covered 66 documents and incorrectly covered seven documents (three to “Health and Medicines”, two to “Science” and two to “Sociology”). Further, the C4.5 confusion matrix demonstrates 100% classification accuracy for four class labels that correctly covered all of their documents (“Children stories”, “Education”, “Politics” and “ScienceB”). On the other hand, for PART algorithm, class “Tourist and Travel” represents 60 documents in the training dataset in which 52 of them are classified correctly and eight of them are classified incorrectly (four by class “Children Stories”
and four by “Sociology”). In general, the confusion matrix is a helpful evaluation measure that reveals the performance of class labels with regards to the available documents in the training dataset. In other words, it displays the number of documents covered correctly by the class and the number of documents incorrectly classified to other classes.

Figure 3 depicts the error rate produced by the selected classification algorithms for the CCA dataset. This figure reveals that C4.5, RIPPER and PART have almost a similar error rate. However, C4.5 algorithm achieved a slightly less error rate by (1%) than RIPPER and PART. Further, the error rate of OneRule algorithm on the CCA dataset is very high (69.78%) which definitely makes this algorithm the least applicable to such dataset. The reason behind the high error rate is that OneRule classification model represents a very small portion of the dataset that belongs to just two categories (“Autobiography” and “Tourist and Travel”) from a total of 15 categories. Thus, this classification model often predicts the documents to either “Autobiography” or “Tourist and Travel”, which consequently lead to misclassification.

Figure 4 displays the number of rules derived by the selected algorithms on the CCA dataset. The figure clearly indicates that decision tree and rule induction (C4.5, PART, J48) tend to produce a smaller number of rules compared to the rule mining algorithms (RIPPER and OneR) which have a tendency to overfit. This is a common issue in rule-based classifiers, where they tend to learn the noise in the data rather than the underlying patterns. The decision tree algorithms, on the other hand, tend to have a more generalized view of the data, which results in fewer rules but with better generalization to unseen data.
RIPPER) approaches produced the largest number of rules, whereas OneRule algorithm produced the least. In fact, the size of the classifiers generated by C4.5, RIPPER, and PART are of a moderate size and with a good quality with respect to classification accuracy in which domain experts can easily understand. On the other hand, the classification model generated by OneRule is small in size. Consequently, it lacks the most fundamental objective of classification models, which is to predict the new documents into its corresponding category.

An intensive analysis has been conducted on the performance of each selected classification approach for the document categories. Table 3 represents precision and recall evaluation results for C4.5, RIPPER, PART, and RuleOne algorithms respectively. As it is shown in Table 3, there are three categories “Politics”, “Science”, and “Education” that achieved nearly 100% prediction accuracy. The majority categories’ performance for C4.5 algorithm with reference to precision and recall measures is very good. The only exception is category “Sport” which achieved a 50% recall. Further, almost similar performance to RIPPER algorithm has been achieved by the PART algorithm document categories. Lastly, the “Sports” category produced less precision and recall.

Table 3. Precision and recall results per class for the classification algorithms.

<table>
<thead>
<tr>
<th>Category/Class</th>
<th>OneRule</th>
<th>C4.5</th>
<th>RIPPER</th>
<th>PART</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Autobiography</td>
<td>0.216</td>
<td>1</td>
<td>0.88</td>
<td>0.904</td>
</tr>
<tr>
<td>Children’s stories</td>
<td>0</td>
<td>0</td>
<td>0.964</td>
<td>1</td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>0.76</td>
<td>0.655</td>
</tr>
<tr>
<td>Education</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Health and medicine</td>
<td>0</td>
<td>0</td>
<td>0.906</td>
<td>0.906</td>
</tr>
<tr>
<td>Interviews</td>
<td>0</td>
<td>0</td>
<td>0.88</td>
<td>0.957</td>
</tr>
<tr>
<td>Politics</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Recipes</td>
<td>0</td>
<td>0</td>
<td>0.944</td>
<td>0.895</td>
</tr>
<tr>
<td>Religion</td>
<td>0</td>
<td>0</td>
<td>0.35</td>
<td>0.844</td>
</tr>
<tr>
<td>Science</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Science</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sports</td>
<td>0</td>
<td>0</td>
<td>0.794</td>
<td>0.9</td>
</tr>
<tr>
<td>Sociology</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Sports</td>
<td>0</td>
<td>0</td>
<td>0.629</td>
<td>0.933</td>
</tr>
</tbody>
</table>

Fig. 5. Success rate of the class labels for each classification algorithm.
results than the rest of the categories in most of the algorithms. This is due to the fact that this category is the least representative one in the training dataset in term of document numbers. In particular, there are only four documents associated with “Sport” category in the whole dataset.

The performance of each algorithm is investigated considering the class accuracy as well. Figure 5 shows the success rate for each class in predicting documents and for each classification algorithm. The success rate (accuracy) is computed as the average per class performance of the classification models (Sokolova and Lapalme, 2009), and it is a complement of the error rate. The figure clearly shows that the performance of the success rate for C4.5, RIPPER, and PART algorithms are slightly different. All of these learning methods have outperformed OneRule algorithm with respect to the success rate. Further, the figure shows that most of the document categories are wrongly classified by OneRule method. In fact only two categories have achieved a good success rate. The rest of the categories for OneRule have generated a zero success rate. In this regards this algorithm derives a very small number of rules, which are represented two categories (“Autobiography” and “Children Stories”).

6. Conclusions

TM is becoming increasingly important because of the huge number of documents available offline and online. TC is one of the significant tasks in the field of TM. This task involves classifying text documents into a number of predefined classes based on their content. In this paper, the problem of Arabic TC is investigated using different classification learning algorithms (C4.5, PART, RIPPER, OneRule) from data mining. These classification algorithms are known as rule based since the output classification model is in the form of rules. WEKA, the open business intelligence tool, is employed in order to test the performance of these algorithms for the published CCA Arabic corpus. The bases of the comparison in the experimentation are different text evaluation metrics, including error-rate, precision, and recall. The results indicated that the least applicable learning algorithm towards the chosen Arabic dataset is OneRule. This is due to the fact that the CCA dataset contains categories with a limited number of documents. In this regards OneRule produced only two rules. Moreover, the most applicable algorithm to the Arabic dataset is C4.5 in which it derived higher results in all evaluation criteria than RIPPER, and PART, respectively. Further, the confusion matrix which represents the distribution of documents per category is derived for all the learning algorithms. The confusion matrices for the learning algorithms indicate that the “Sport” category achieved the least results with respects to precision and recall. Since Arabic language is extremely rich and require careful treatment for its text, we intend in near future to design and implement an Arabic language specific rule based classification approach that can overcome some of the main challenges faced by the current used rule based classification data mining algorithms including morphological analysis and order verbs.

References


Harrag, F and E El-Qawasneh (2009). Neural Network for Arabic text classification¹, pp. 778–783, IEEE.


