On the Automatic Categorization of Arabic Articles Based on Their Political Orientation

Raddad Abooraig
Jordan University of Science and Technology
Irbid, Jordan

Ahmed Alwajeeh
Jordan University of Science and Technology
Irbid, Jordan

Mahmoud Al-Ayyoub
Jordan University of Science and Technology
Irbid, Jordan
maalshbool@just.edu.jo

Ismail Hmeidi
Jordan University of Science and Technology
Irbid, Jordan
hmeidi@just.edu.jo

ABSTRACT

The prevalence of the dynamic online web pages (such as the social networks, forums, personal Blogs, etc.) that are covering all fields (such as social events, economical events, political events, etc.) are allowing the Internet surfers to interact with their contents such as writing comments and articles. Regarding politics and political events, the Internet surfers post comments and articles based on their beliefs and ideologies. The ability to automatically determine the political orientation of an article can be of great benefit in many areas from Academia to security. This work addresses this important yet largely understudied problem for Arabic texts as a supervised learning problem. Aside from collecting and manually labeling a dataset of articles from different political orientations in the Arab world, the two most popular feature extraction approaches for such a problem (the TC approach and the stylometric features approach) are studied. Moreover, four classifiers are considered to study the effects of different kinds of feature reduction techniques, such as stemming and feature selection, on their effectiveness. Although the experimentation results show the superiority of the TC approach over the stylometric features approach, they also show that the latter approach can be significantly improved by adding new and more discriminating features.

Keywords

Arabic language processing; authorship authentication; stylometric features; bag of words; classification

1. INTRODUCTION

The prevalence of the dynamic online web pages (such as the social networks, forums, personal Blogs, etc.) that are covering all fields (such as social events, economical events, political events, etc.) are allowing the Internet surfers to interact with their contents such as writing comments and articles. Regarding politics and political events, the Internet surfers post comments and articles based on their beliefs and ideologies. They are especially encouraged by the anonymity provided by the inherent nature of the Internet. Consequently, these web pages receive millions of comments and articles daily and the process of analyzing them to extract useful information is very expensive task in terms of both time and effort.

Political articles (especially in the Arab world) are different from other articles due to their subjectivity. A political article is heavily influenced by the author's convictions and political affiliation. The ability to automatically determine the political orientation of an article can be of great benefit in many areas from Academia (e.g., to help with the studies on the development of the political life in certain societies) to security [27]. Moreover, this is an example of author profiling problems, which are useful for optimizing search engines, sentiment analysis and marketing intelligence [18]. This problem can be viewed as a special case of the text categorization (classification) problem with the categories being the major political ideologies in the Arab world such as: Liberal, Islamic Sunni (Brotherhood, Salafi, etc.), Islamic Shia (Hezbollah, Ansarollah, Jeaish Almahdi, etc.), Arab Nationalists (Baathi, Nasri, etc.), and Communist.

There have been several works on Arabic text categorization and authorship analysis. While typical text categorization usually focuses on identifying a text's domain (Sport, Politic, Economy, etc.) based on its contents (topic-based); *authorship analysis* focuses on authorship authentication and authorship characterization (style-based). *Authentication* (attribution) deals with verifying whether a text was written by a certain author or not based on stylometric and statistical similarities with other texts written by the same author. On the other hand, *authorship characterization* (profiling) tries to detect the characteristics of the author such as gender, age group, level of education, social class, cultural background, etc. [2].

The aim of this work is to analyze articles written in Modern Standard Arabic (MSA) to determine their political orientation. To the best of our knowledge, the only work that have addressed anything similar to our problem is that of Koppel et al. [27]. This particularly important yet largely understudied problem is worth further investigation. Moreover, the benefits of this work (and the intended follow-ups) will spill over the boundaries of analyzing the political orientation of an article into the general field of authorship characterization (or profiling) of Arabic text, which is another understudied field of great importance.

The problem at hand is addressed as a supervised learn-

ing problem and special attention is paid to the three main stages of such an approach: dataset collection, feature extraction and selection and classification. Accordingly, we start by collecting and manually labeling a dataset containing articles from different political orientations. Unfortunately, the lack of any standard dataset for this purpose forced us to spend considerable amount of time in collecting the dataset on our own. One of the benefits of this work is that we plan on making this dataset publicly available to interested parties to help in advancing the research efforts in this area. Another benefit of this work is the detailed study of the two most popular feature extraction approaches for such problems, the TC approach and the stylometric features approach. Finally, the effects of different kinds of feature reduction techniques, such as *stemming* (reducing words to their stem, base or root form) and feature selection, are investigated.

The rest of this paper is organized as follows. The following section gives a general overview of the current literature on text categorization and authorship analysis with a focus on the Arabic language. Our work is discussed in Section 3 and the experimental results obtained are discussed in Section 4. Finally, concluding remarks along with a discussion of future work is discussed in Section 5.

2. RELATED WORKS

As discussed previously, the closest works to the problem at hand are those on text categorization and authorship analysis. We briefly discuss some of recent works on general Arabic text categorization. Unfortunately, to the best of our knowledge, the field of authorship analysis is still largely understudied for the Arabic language.

2.1 General Text Categorization

Several papers on Arabic text categorization have been published in the past two decades. We limit our discussion here to the most interesting/recent works. Another important thing to note is that the majority of the current works consider datasets collected from online sources such as news websites and start by preprocessing the text to remove punctuation marks, numbers, non-Arabic letters and stop words. We shall call this filtering and explicitly mention any technique that involves something different than what we have discussed.

Harrag et al. [24, 23] focused on the effect of different attribute selection methods on Arabic text classification. In [24], they showed that a hybrid approach of Document Frequency Thresholding using an embedded information gain criterion of the decision tree algorithm is preferable, whereas in [23], they studied the effect of stemming. They compared three stemming techniques: light, root-based and dictionary-lookup using a dataset of 453 Hadith documents. The authors used the artificial neural networks (ANN) and support vector machine (SVM) classifiers. They showed that light stemming produce better results than other stemming techniques and that ANN outperforms SVM.

A dataset of 1,445 articles collected from popular Arabic newspaper was used by Mesleh [29]. The author started by filtering the text (which included the removal of infrequent items), and used the χ^2 statistics for feature selection. The author showed that the SVM classifier outperforms the K nearest neighbor (K-NN) and Naive Bayes (NB) classifiers. In another work on the same dataset, Kanaan et al. [26] compared the K-NN, Rocchio and NB classifiers. Simple filtering and light stemming were applied to the dataset. To compute the features of each instance, several methods were considered such as Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), Weighted Inverse Document Frequency (WIDF) and Inverse Category Frequency (ICF). As for the similarity measure, the Jaccard measure was used. The results showed that the NB classifier generally outperformed the other classifiers.

The dataset of El-Halees [17] consisted mainly of news articles from Aljazeera website. The author started by filtering the text. He then applied stemming and part-of-speech (POS) tagging to remove anything but nouns and proper nouns. After that, the Maximum Entropy algorithm was used for classification.

Hadi et al. [21] used a dataset of newspaper articles to show that that the NB classifier outperforms the K-NN classifier with the cosine method as the similarity measure.

Alsaleem [9] used a dataset of 5,121 articles collected from Saudi newspapers. After filtering the text, the author shows that the SVM classifier outperformed the NB classifier.

Working on a dataset of 1,000 articles, Duwairi [16] started by filtering and stemming the text. The author then collects the keywords of every class into a single feature vector. To classify a new document, the author simply chooses the class with the most similar (based on the Dice measure [20]) feature vector to the document's feature vector.

Al-Harbi et al. [7] tested two classifiers (SVM and decision tree) on seven different datasets with a combined size of 17,658 articles. Similar to [29], the authors used the χ^2 statistics for feature selection after filtering the text. The results show that the decision tree classifier outperformed SVM for all datasets.

Recently, several interesting works addressing the text categorization problem in an innovative and contemporary way. One example is [19] in which the authors focus on the emergence of new writing styles due to the prevalence of online social networks and the effect of these styles on typical methods of text categorization. Other works used an innovative way to address the text classification problem by focusing on character-based features. One such example is the use of compression based techniques [28], which were only recently applied to Arabic text [36]. The results shows some advantages for such approaches over typical "word features" based approaches. These advantages are worth investigating for problems like authorship analysis.

All of the previously discussed works are topic-based since they focus on identifying an Arabic text's domain or topic. However, the same approaches can be used to identify an Arabic text's sentiment [3, 8, 5, 4, 6] and authors' characteristics such as identity [1, 2, 33, 11] and gender [18, 10].

2.2 Authorship Analysis

Compared to text categorization, authorship analysis is largely understudied in the Arabic language, to the best of our knowledge. Below, we discuss some of the works on this field starting with the more popular problem of authorship authentication (attribution). We then discuss the less popular but more related problem of authorship characterization (profiling). For both problems, computing stylometric features can have significant positive impact on the accuracy.

The authorship authentication (attribution) problem is a classification problem in which the goal is to determine the author of a certain text given a set of texts written by various authors. Obviously, the writing style is the most intuitive aspect on which to focus. Clark and Hannon [13] proposed a system for author attribution in the English language based on the idea that an author's choice of synonyms is idiosyncratic enough to uniquely identify the author. Pavelec et al. [31, 32] focused on the Portuguese language. In [31], they proposed to use stylometric features with SVM. They considered both the writer-dependent model (also known as the personal model) and the writer-independent model. The difference between the two is that in the former, a model is built for each author with only two classes: one representing authorship and the other representing forgery. The results show that the writer-dependent model is more accurate provided that the dataset is large enough. In [32], the same authors proposed a completely different approach of using a compression algorithm known as Prediction by Partial Matching (PPM) algorithm for feature extraction. They compared the performance of the proposed algorithm of [32] with the SVM classifier of [31] that is based on stylometric features. The results show that the accuracies of both systems were comparable while the proposed system of [32] avoided the computationally expensive process of feature definition, extraction, and selection incurred in [31]. Finally, for a more comprehensive coverage of the different approaches to address the authorship authentication problem, interested readers are referred to [25, 34].

Authorship authentication for the Arabic language has not been studied well. The only works we know of are [1, 2, 33, 11]. In [1, 2], Abbasi and Chen used different sets of features including lexical, syntactic, structural and content-specific features for the authorship identification problem. In [1], they applied their technique to web forum messages, whereas, in [2], they collected and analyzed messages posted on extremist groups' web forums both in Arabic and English. In [33], the authors focused on the problem of small and imbalanced datasets, which is a common problem with authorship identification datasets. They represented each document using the bag of character n-gram approach which is better than the bag of words approach as it can capture stylistic as well as thematic information more accurately.

In the authorship characterization (profiling) problem, the goal is to determine certain traits or characteristics of the author. Identifying the gender of the author is one example of such characteristics [15, 30, 12, 18]. In [18], the authors considered a deeper level of author profiling by considering two types of traits: demographic and psychometric. The demographic traits included the author's age, gender, and level of education whereas the psychometric traits are: extraversion, lie (or social desirability), neuroticism (or emotionality), and psychoticism (or tough mindedness). The Arabic dataset they consider consisted of more than 8,028 emails written by 1,030 authors. To collect information about their traits, the authors were asked to fill out a questionnaire.

The closest paper to our work is that of Koppel et al. [27], which consider the same problem as ours (classifying documents based on their political orientation), but with a slightly different set of classes. In fact, the authors consider two subtly different variations of the problem. In the first one, the authors classify the documents based on their "ideological affiliation," which refers to the doctrine underlying the documents. The classes considered for this problem are: Salafi-Jihadi, Mainstream Islam (apolitical), Muslim Brotherhood, and Wahhabi. In the second one, the authors classify the documents based on their "organizational affiliation," which refers to the religious group from which a particular document stems. The classes considered for this problem are: Hamas, Hezbollah, Al Qaeda and Muslim Brotherhood. The features they chose were surprisingly simple. With no stemming applied, they used the frequencies of the 1000 most common words (which included both function words and content words) in the entire corpus. Despite the simplicity of the selected features, the obtained accuracy was relatively high. For the ideological categorization problem, the accuracy was 73% whereas for the organizational categorization problem, the accuracy was 80%.

3. PROPOSED APPROACH

This section is dedicated to discussing the details of the proposed approach. We start with a discussion of the dataset collection and annotation process before going into the extracted feature sets.

3.1 Dataset Collection and Annotation

Like any other supervised learning text mining problem, the first step is to build a comprehensive and representative dataset. Since the number of political orientations (categories) to be considered can be high and collecting a large enough dataset for all of them is prohibitively time-consuming, we focus on the following five categories, which are among the most common political streams in the Arab world: Arab Nationalist (abbreviated as Nat), Islamic Shia (abbreviated as Shi), Socialist (abbreviated as Soc), Liberal (abbreviated as Lib) and Muslim Brotherhood (abbreviated as Bro). The articles/comments of these categories are collected mainly from Arabic speaking authors of different backgrounds and education levels. Such choices will provide more meaningful results and add to the level of challenge posed by this problem especially if we take into account the similarities between these categories in terms of the ways ideas are presented as well as the vocabulary and the general rhetoric used by the members of these categories.

The sources of our dataset are mainly social networks and forums where heated discussions of current political events are taking place. Since the start of the Arab Spring in late 2010, these discussions gained more attention and diversity. Our dataset include articles as well as posts, comments and excerpts from such discussions. Due to its massive popularity and the pivotal role of its users in forming the events of the Arab Spring, Facebook represent probably one of the richest sources for a dataset like ours. In fact, Facebook

posts represent the majority of our dataset. Other sources for our dataset include specialized/dedicated websites and forums such as الفكر القومي العربي[1] for Nat articles/comments, قناة المنار[2] for Shi articles/comments and منتدى الاشتراكي الثوري[3] and ماركسي[4] for Soc articles/comments. The number of articles/comments per category is 300 and the average lengths of articles/comments (in terms of words and characters) as well as other statistics about the dataset are reported in Table 1.

Table 1 shows that Nat articles are the longest in terms of both the number of words/article and the number of characters/artcile, whereas Shi articles are the shortest. The same applies to the number of sentences/artcile except that Soc articles have slightly more sentences/article than Nat. Now, regarding the lengths of the sentences, Nat articles have the longest sentences in terms of both the number of words/sentence and the number of sentence, whereas Soc and Lib articles have the shortest sentences. The final length measure is the character lengths of the words in each category. It should be noted here that unlike English, Arabic words tend to be small and the concept that more educated people use bigger words does not apply [2, 11]. Hence, it is expected for all categories under consideration to have close characters/word averages.

Table1 reveals other interesting observations related to the performances of the two feature extraction approaches under consideration (the TC approach and the stylometric features approach) discussed in the following section. For example, the average number of unique words/article[5] for Nat articles is almost twice the averages for Shi articles. This is also true for the average number of words that occur once and the average number of words that occur twice. However, these observation are not as interesting and influential as the ones related to the percentages of these words. For example, on average, the percentages of unique words per article for each category range from 28% for Nat articles to 36.4% for Shi articles. As for the percentages of the words that occur once and the words that occur twice, they range, on average, from 20.5% for Nat articles to 27.8% for Shi and Bro articles and from 7.2% for Nat articles to 8.5% for Shi articles, respectively. Such observation give strong indications about the richness of the vocabulary of the authors of each category, which will have a positive effect on any classifier's ability to correctly identify articles of these classes. This might be the main justification for the surprisingly high accuracies obtained in Section 4.

3.2 Feature Extraction

The problem at hand is a special type of authorship characterization, where we intend to investigate whether the political orientation of the author can be automatically determined from an article in an efficient and accurate manner. As suggested by numerous papers in the Linguistics and NLP literature [25, 14, 1, 2, 37, 33, 34, 12], writing styles can be very useful for various authorship analysis problems including authorship characterization [12]. To capture the nuance differences in the writing styles, researchers have suggested a rich and diverse set of *stylometric* features. Another view of the problem at hand places it within the general context of TC where the classes to be predicted are the political orientations of the articles instead of the domain/topic of the article. We call this the TC approach. Since both approaches compute large numbers of features, a feature selection algorithm is used to reduce the number of features. The algorithm used for this purpose is the correlation-based subset evaluator algorithm proposed by Hall [22]. In the following subsections, we discuss each approach in details.

3.2.1 The TC approach

In the TC approach, the feature vector for each document depends on the occurances/frequencies of the tokens (words, phrases, etc.) within it. If one would take all possible tokens from each document, tens (or even hundreds) of thousands of features will be extracted since there are that many different tokens in a dataset of a size similar to ours. Obviously, such huge numbers would prohibit any classifier from working properly. So, we address this problem on three levels. The first level is to employ some techniques such as stemming to reduce the number of extracted features as shown in Table 2. To study the potentially significant effect of such techniques on the performance of the considered classifiers, we create four versions of our dataset as follows.

- **V1**: is the original dataset with no changes other than the manual correction of misspelled words.
- **V2**: is simply **V1** with Khoja stemmer applied.
- **V3**: is simply **V1** with light stemmer applied.
- **V4**: is simply **V1** with n-gram (where $n = 3$) approach applied.

The second level to address the high dimensionality of the addressed problem is to limit the number of features extracted from each document to the most frequent $W = 1,000$ tokens. This way, the total numbers of extracted features from each of the four versions of the dataset are significantly reduced to manageable numbers as shown in Table 2. One might be concerned whether we are sacrificing the accuracy with such feature reduction techniques; however, as shown in Section 4, this procedure has a positive effect on the accuracy.

The thousands of features kept from the first and second levels are still too large for many classifiers such as SVM and NB. So, we apply a third level of feature reduction by applying Hall's correlation-based feature selection algorithm [22]. Table 2 shows the numbers of extracted and selected features for each version of the dataset and for different values of W. The table confirms the expectation that the TC approach will compute thousands of features, which means that the Hall's feature selection algorithm will take a very long time

[1] http://www.alfikralarabi.org/
[2] http://www.manartv.com.lb/
[3] http://marxlistleninlist.topic-ideas.com/
[4] http://www.marxy.com/
[5] This is computed by taking the total number of unique words in each category (i.e., the ones that do not appear in other categories) and dividing it by the number of articles.

Table 1: Statistical features of the collected dataset.

	Nat	Shi	Soc	Lib	Bro	All
Avg. No. of articles	300	300	300	300	300	1500
Avg. No. of characters/article	2111.477	898.64	1522.423	1205.167	1030.387	1353.619
Avg. No. of characters/sentence	167.05	155.21	119.97	119.92	149.33	140.86
Avg. No. of characters/word	5.87	5.81	6.02	5.95	5.84	5.91
Most frequent letter ا	311.42	127.79	225.5	174.05	147.28	197.21
Least frequent letter ظ	4.03	1.79	4.14	2.76	2.4	3.03
Avg. No. of words/article	359.45	154.59	253.057	202.39	176.487	229.195
Avg. No. of unique words/article	100.66	56.21	75.47	67.36	63.56	-
Avg. No. of words/sentence	28.44	26.7	19.94	20.14	25.58	23.85
Avg. No. of words that occur once	73.69	42.96	54.26	51.72	48.98	-
Avg. No. of words that occur twice	25.9	13.19	18.88	15.19	13.83	-
Avg. No. of sentences/article	12.64	5.79	12.69	10.05	6.9	9.61

Table 2: Selected features of each dataset version for the TC approach.

Dataset Version	W	Extracted Features	Selected Features
V1	1,000	2,678	60
	3,000	9,793	61
	5,000	16,982	Out of memory
V2	1,000	1,494	17
	3,000	2,672	17
	5,000	2,672	17
V3	1,000	2,530	46
	3,000	9,182	49
	5,000	28,163	Out of memory
V4	1,000	2,724	62
	3,000	10,784	64
	5,000	28,931	Out of memory

to finish or even fail to finish at all as shown by the table. The table also show that the **V2** dataset produces a significantly smaller sets of extracted and selected features compared to the other versions which produce sets of comparable sizes. This is expected since the Khoja stemmer is a root-based one known for reducing words representing different inflections or derivations of the same root into their root. Finally, the table shows that, in terms of the number of selected features, it makes little difference whether the value of W is large or not.

3.2.2 The Stylometric Features Approach

As common in the literature [14, 37], the stylometric features are divided into four categories: lexical, syntactic, structural and content-specific. We follow the same organization and propose to use 334 features divided as follows.

- The set of lexical features (96 features) including both character-based and word-based features. The character-based features are 67 features including language independent features such as the number of white-space characters as well as features specific for the Arabic language such as the number of elongation characters, while the word-based features are 29 features including Yule's K measure, the Entropy measure, etc.
- The set of syntactic feature (13 features) mainly covering aspects like the usage pattern of punctuation marks (such as the number of commas, question marks, double quotation marks, etc.).
- The set of structural features (22 features) covering structural aspects of the articles such as the average number of characters/words in each sentence, paragraph, article, etc. Table 1 lists few of the structural features we use.
- The set of content-specific features (203 features) including the numbers of the different stop words in the Arabic language, the number of negative emotions, etc.

Similar to what is done for the TC approach, four versions of our dataset are created as follows.

- **F1**: is the dataset with the lexical features.
- **F2**: is the dataset with the lexical and syntactic features.
- **F3**: is the dataset with the lexical, syntactic and structural features.
- **F4**: is the dataset with the lexical, syntactic, structural and content-specific features.

As for the feature selection part, the same algorithm is applied to the stylometric features approach to determine the set of most discriminating features. Table 3 shows the number of selected features of each type. The table shows that only a small portion of the stylometric features possess the most discriminating power. Due to space limitations, the actual selected features are not presented. Instead, we discuss their types. The table shows that 11 lexical features are selected in **F1**. Moving from **F1** to **F2** (by adding the syntactic features) reveals that lexical features are more powerful than syntactic ones as 12 out of the 15 selected features in **F2** are lexical. In **F3**, the 15 selected features are equally divided between lexical and structural features (with only

Table 3: Selected features of each dataset version for the stylometric features approach.

Feature Type	Extracted Features	Selected Features
F1	96	11
F2	109	15
F3	131	15
F4	334	14

one syntactic feature selected). Finally, the 14 selected features in **F4** are mostly content-specific (nine) features with few lexical (two) and structural (three) features. This shows that the features with the most discriminating power are the content-specific features and the features with the least discriminating power are the syntactic features. It also shows that lexical and structural features posses almost the same discriminating power.

4. RESULTS AND EVALUATION

In this section, we discuss the several experiments we conduct on the collected dataset with the objective of studying and comparing the two feature extraction techniques for the problem at hand. As for the classification models under consideration, we choose to focus on the following four widely used classifiers for text mining: Naive Bayes (NB), Discriminative Multinomial Naive Bayes (DMNB) [35], Support Vector Machine (SVM) and Random Forest (RF). For testing purposes, the 10-fold cross-validation technique is employed. Due to space constraints, only the accuracy of each classifier (the percentage of correctly classified instances) is reported.

The main objective is to compare the effect of the two feature extraction approaches on the accuracies of different classifiers. Moreover, the effects of feature selection on each classifier and each feature extraction approach are evaluated. Hence, this section is divided into two parts each dedicated to a feature extraction approach.

4.0.3 Results of the TC Approach

The first set of experiments is dedicated to the TC approach. Table 4 shows the results of applying the classifiers under consideration on the four versions of the dataset generated for the TC approach. Moreover, since the TC approach is affected by the parameter W, we report the results for two different values of W, 1,000 and 3,000. As can be seen from the table, the number of features extracted for $W = 3,000$ (even after applying stemming) is large enough to cause a crash in NB and SVM. As for its effect on accuracy, increasing W has almost no positive effect. The only classifier benefitting from it (sometimes) is DMNB which supports the claim that DMNB is one of the text classifiers favoring relatively larger feature sets (to a certain extent). Another classifier with interesting behavior with regards to changing W is RF. The table shows that increasing W without applying feature selection caused a significant degeneration in RF's accuracy. The degradation gets larger with dataset versions whose feature sets are larger reaching up to 7.2% for **V4**. However, after applying feature selection which reduces the number of features from order of thousands to order of tens, RF's accuracy actually increases with the increase of W. Even though the increase is small, the shift in behavior might suggest that increasing W gives the feature selection technique a higher chance of selecting features that are more suitable for RF.

Table 4: Results for the TC approach.

DS	Feat.	W	NB	DMNB	SVM	RF
V1	Ext.	1,000	66.20%	82.53%	76.07%	59.60%
		3,000	Out of memory	83.93%	Out of memory	53.33%
	Sel.	1,000	68.00%	71.07%	70.33%	64.33%
		3,000	68.00%	71.13%	70.40%	64.20%
V2	Ext.	1,000	52.27%	72.00%	66.60%	46.00%
		3,000	54.20%	71.87%	67.20%	44.27%
	Sel.	1,000	52.67%	54.33%	53.33%	51.67%
		3,000	52.67%	54.33%	53.20%	52.13%
V3	Ext.	1,000	63.47%	80.73%	72.53%	57.87%
		3,000	Out of memory	81.07%	Out of memory	54.13%
	Sel.	1,000	65.33%	67.53%	65.33%	63.27%
		3,000	65.33%	67.87%	65.87%	64.33%
V4	Ext.	1,000	67.07%	81.87%	75.80%	64.20%
		3,000	Out of memory	Out of memory	Out of memory	57.00%
	Sel.	1,000	68.27%	71.07%	70.93%	65.00%
		3,000	68.27%	71.07%	71.13%	66.60%

Comparing the results across the table, several interesting trends are observed related to effect of applying stemming and the n-gram technique as follows. The table shows that both stemming techniques under consideration have a negative effect on accuracy with the Khoja stemmer having stronger negative effect than light stemming. This is due to the fact that words representing different inflections or derivations of the same stem/root might have different relative importance in different categories, and thus considering them has better discriminative power that only considering their stem. Now, since the Khoja stemmer is better than the light stemmer at reducing such words into their stems, using it caused a loss in important information leading to a degradation in the accuracy. Comparing the results for dataset versions **V1** and **V4**, one can see that applying the n-gram technique on the dataset has small but positive effect on the accuracies of most classifiers. This is probably due to its ability to capture nuance differences in the literature of different political schools of thought. For example, while all categories might include the words عالم, عربي and اسلامي separately, the usage of the phrase العالم العربي والاسلامي is more likely to be used by authors with Islamic mentality such as those of the Shi and Bro categories. Finally, the table shows the superiority of DMNB and inferiority of RF compared with the other classifiers.

The table discussed in this section reports only the accuracies, which can be misleading. To avoid such issues, researchers often report the precision and recall values. Due

Table 5: Results for the stylometric features approach.

DS	Feat.	NB	DMNB	SVM	RF
F1	Ext.	24.53%	37.07%	39.07%	43.20%
	Sel.	24.20%	35.47%	30.60%	42.53%
F2	Ext.	26.00%	40.00%	41.87%	44.67%
	Sel.	25.87%	37.00%	34.47%	46.13%
F3	Ext.	25.33%	43.00%	46.07%	47.93%
	Sel.	33.20%	38.20%	36.67%	50.53%
F4	Ext.	27.80%	56.13%	56.47%	52.33%
	Sel.	36.80%	43.87%	50.13%	60.20%

to space constraints, such values are not explicitly reported for each classifiers. Instead, we only report and discuss classifiers with interesting precision/recall values. Specifically, we are looking for classifiers for which the difference between precision and recall is significant. For DMNB, SVM and RF, this difference rarely goes above 1%, which means that Type-I errors are as frequent as Type-II errors for these classifiers. The only interesting case is when we apply NB on the **V2** dataset where the difference reached 5.5%. However, since the accuracy is very low for this case (less than 55%), such an observation is not alarming.

4.0.4 Results of the Stylometric Features Approach

The second set of experiments is dedicated to the stylometric features approach. As mentioned in Section 3.2.2, 334 features are computed to measure different aspects related to the styles of different authors. These features are divided into four categories and we test the effect of gradually adding more more stylometric features on the performance of the considered classifiers.

Table 5 shows the results of applying the classifiers under consideration on the four versions of the dataset generated for the stylometric features approach. The accuracies reported in this table are very low compared to those shown in the previous section. In fact, NB's accuracies are not better than random guessing. RF produces the best accuracies; however, they barely cross the 60% accuracy level. The only bright side of these results is the noticeable increase in accuracy with the increase in the number of stylometric features. This suggests that researchers interested in improving this approach need to worry about computing new and more discriminating features, especially, content-specific ones.

5. CONCLUSION AND FUTURE WORK

In this paper, we addressed the problem of automatically determining the political orientation of an Arabic article/comment. This important yet largely understudied problem has great benefits in many areas from Academia to security. We presented our efforts to address this problem by collecting and manually labeling a dataset containing articles from different political backgrounds. We built more than one version of our dataset to study and compare the two most popular feature extraction approaches for such a problem (the TC approach and the stylometric features approach). We considered four classifiers and studied the effect of feature selection on their effectiveness. The experimentation results showed the superiority of the TC approach (with 83.93% accuracy using DMNB) over the stylometric features approach (with 60.20% accuracy using RF). Another interesting observation is related to RF, which was the best classifier for the stylometric features approach and the worst classifier for the TC approach. Finally, the results showed that different kinds of feature reduction techniques, such as stemming and feature selection, affected the accuracies negatively.

As part of our future work, we plan on extending our dataset to include more categories. Moreover, we plan on focusing more on stylometric features and their effect on the accuracy of the considered classifiers. Examples of such features include font types, font sizes, colors, richness of vocabulary, etc. Finally, experimenting with different classifiers such as decision trees, neural networks, etc. can be of great benefit to gain more insight into this problem.

6. REFERENCES

[1] A. Abbasi and H. Chen. Applying authorship analysis to arabic web content. In *Intelligence and Security Informatics*, pages 183–197. Springer, 2005.

[2] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, 20(5):67–75, 2005.

[3] N. Abdulla, N. Mahyoub, M. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Corpus-based and lexicon-based. In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2013.

[4] N. Abdulla, R. Majdalawi, S. Mohammed, M. Al-Ayyoub, and M. N. Al-Kabi. Automatic lexicon construction for arabic sentiment analysis. In *The 2nd International Conference on Future Internet of Things and Cloud (FiCloud)*, 2014.

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

[6] M. Al-Ayyoub, S. Bani Essa, and I. Alsmadi. Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining (IJSNM)*. To appear.

[7] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh. Automatic arabic text classification. In *The 9es Journées internationales d'Analyse statistique des Données Textuelles*, pages 77–83, 2008.

[8] M. N. Al-Kabi, N. A. Abdulla, and M. Al-Ayyoub. An analytical study of arabic sentiments: Maktoob case study. In *The 8th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 89–94. IEEE, 2013.

[9] S. Alsaleem. Automated arabic text categorization using svm and nb. *Int. Arab J. e-Technol.*, 2(2):124–128, 2011.

[10] K. Alsmearat, M. Al-Ayyoub, and R. Al-Shalabi. An extensive study of the bag-of-words approach to gender identification of arabic articles. In *The ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 2014.

[11] A. Alwajeeh, M. Al-Ayyoub, and I. Hmeidi. On authorship authentication of arabic articles. In *The fifth International Conference on Information and Communication Systems (ICICS 2014)*, 2014.

[12] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[13] J. H. Clark and C. J. Hannon. An algorithm for identifying authors using synonyms. In *Current Trends in Computer Science, 2007. ENC 2007. Eighth Mexican International Conference on*, pages 99–104. IEEE, 2007.

[14] O. De Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[15] W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu. Author gender prediction in an email stream using neural networks. *Journal of Intelligent Learning Systems and Applications*, 4:169–175, 2012.

[16] R. M. Duwairi. Machine learning for arabic text categorization. *Journal of the American Society for Information Science and Technology*, 57(8):1005–1010, 2006.

[17] A. El-Halees. Arabic text classification using maximum entropy. *The Islamic University Journal (Series of Natural Studies and Engineering)*, 15:157–167, 2007.

[18] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson. Tat: an author profiling tool with application to arabic emails. In *Proceedings of the Australasian Language Technology Workshop*, pages 21–30, 2007.

[19] M. Faqeeh, N. Abdulla, M. Al-Ayyoub, Y. Jararweh, and M. Quwaider. Cross-lingual short-text document classification for facebook comments. In *The 2nd International Conference on Future Internet of Things and Cloud (FiCloud)*, 2014.

[20] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)*, 21(1):64–93, 2003.

[21] W. Hadi, F. Thabtah, S. ALHawari, and J. Ababneh. Naive bayesian and k-nearest neighbour to categorize arabic text data. In *European Simulation and Modeling Conference*, pages 196–200, 2008.

[22] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[23] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman. Stemming as a feature reduction technique for arabic text categorization. In *Programming and Systems (ISPS), 2011 10th International Symposium on*, pages 128–133. IEEE, 2011.

[24] F. Harrag, E. El-Qawasmeh, and P. Pichappan. Improving arabic text categorization using decision trees. In *Networked Digital Technologies, 2009. NDT'09. First International Conference on*, pages 110–115. IEEE, 2009.

[25] P. Juola. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334, 2006.

[26] G. Kanaan, R. Al-Shalabi, S. Ghwanmeh, and H. Al-Ma'adeed. A comparison of text-classification techniques applied to arabic text. *Journal of the American society for information science and technology*, 60(9):1836–1844, 2009.

[27] M. Koppel, N. Akiva, E. Alshech, and K. Bar. Automatically classifying documents by ideological and organizational affiliation. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, pages 176–178. IEEE, 2009.

[28] Y. Marton, N. Wu, and L. Hellerstein. On compression-based text classification. In *Advances in Information Retrieval*, pages 300–314. Springer, 2005.

[29] A. M. Mesleh. Chi square feature extraction based svms arabic language text categorization system. *Journal of Computer Science*, 3(6):430, 2007.

[30] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 163–167, 2006.

[31] D. Pavelec, E. Justino, L. V. Batista, and L. S. Oliveira. Author identification using writer-dependent and writer-independent strategies. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 414–418. ACM, 2008.

[32] D. Pavelec, L. S. Oliveira, E. Justino, F. D. N. Neto, and L. V. Batista. Author identification using compression models. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 936–940. IEEE, 2009.

[33] E. Stamatatos. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799, 2008.

[34] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[35] J. Su, H. Zhang, C. X. Ling, and S. Matwin. Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th international conference on Machine learning*, pages 1016–1023. ACM, 2008.

[36] H. Ta'amneh, E. Abu Keshek, M. Bani Issa, M. Al-Ayyoub, and Y. Jararweh. Compression-based arabic text classification. In *The ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 2014.

[37] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.