Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy

Tarek Kanan and Edward A. Fox
Department of Computer Science, College of Engineering, Virginia Polytechnic Institute and State University (Virginia Tech), McBryde Hall Room 114 (0106), Blacksburg, VA 24061. E-mail: {tarekk, fox}@vt.edu

Arabic news articles in electronic collections are difficult to work with. Browsing by category is rarely supported. While helpful machine learning methods have been applied successfully to similar situations for English news articles, limited research has been completed to yield suitable solutions for Arabic news. In connection with a QNRF funded project to build digital library community and infrastructure in Qatar, we developed software for browsing a collection of about 237K Arabic news articles, which should be applicable to other Arabic news collections as well. We designed a simple taxonomy for Arabic news stories that is suitable for the needs in Qatar and other nations, is compatible with the subject codes of the International Press Telecommunications Council, and was enhanced with the aid of a librarian expert as well as five Arabic-speaking volunteers. We developed tailored stemming (i.e., a new Arabic light stemmer) and automatic classification methods (the best being binary SVM classifiers) to work with the taxonomy. Using evaluation techniques commonly used in the information retrieval community, including 10-fold cross-validation and the Wilcoxon signed-rank test, we showed that our approach to stemming and classification is superior to state-of-the-art techniques.

Keywords: Classification, Information Retrieval, Taxonomy, Stemming, Digital Libraries, Natural Language Processing.

Introduction

Motivation

To provide digital library services through an Arabic language information retrieval system, software must be reorganized, redesigned, or improved to satisfy the needs of particular stakeholder groups. Thus, while computational methods for automatic news classification exist, these have drawbacks, especially regarding quality. Accordingly, we have developed improved methods for Arabic text classification.

Arabic language information retrieval (IR) and natural language processing (NLP) are seen as difficult areas because of :

- Particular characteristics of the Arabic language, e.g., ambiguity;
- The limited research work in these two domains, as compared to that with the English language;
- The fact that most of the software packages used to process English texts – like indexing, cleaning, and tokenizing – do not support Arabic, or need tailored handling and addition of specific features when applied to Arabic text.

Problem statement

Online Arabic news article categorization is not of high quality. When articles are accessed in a collection rather than in a newspaper (site), it is hard to browse by category. Further, the lack of a suitable taxonomy restricts more in-depth browsing. Moreover, there is no simple and accepted taxonomy for Arabic news. International taxonomies are too complex. Taxonomies used by a particular news service are not general enough to apply to other news service collections. Even when a taxonomy is selected, there is no proven best method for classifying Arabic news stories. Also, there is no simple and proven best technique for Arabic word stemming that will enhance Arabic news classification based on a given taxonomy.

The Arabic Language

Arabic is a widely used global language that has major differences from the most popular, e.g., English, Spanish, and Chinese. The Arabic language has many grammatical forms, varieties of word synonyms, and different word meanings that vary depending on factors like word order and inclusion of diacritics. In spite of such complexities, there has been limited natural language processing work with Arabic, especially in comparison to the English language.

Most of the software packages, tools, and APIs for information retrieval and natural language processing do not address Arabic language requirements. To allow these software packages and tools to handle Arabic language data, modification and extra work are required.

According to (Nationsonline, 2014; Wikipedia, 2014), Arabic is the fifth spoken language in the world, with around 4.5% of the world population using this as their primary language (see

FIG. 1). Arabic is written from right to left, and consists of 28 different characters with different formulation and shapes for the same letter, based on the location of the letter in the word. Further, there are diacritics, i.e., small characters that can be attached to a letter either as superscript or subscript to add different pronunciation, grammatical formulation, and sometimes meaning to that letter as well as the whole word. These diacritics are commonly used in formal Arabic.

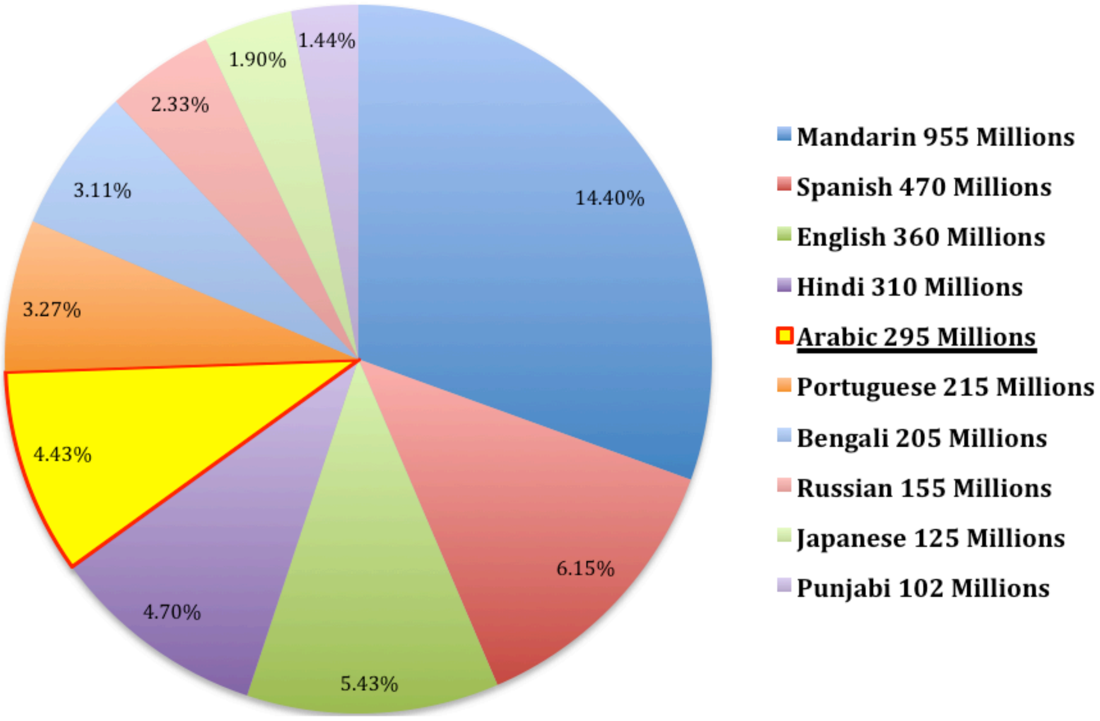


FIG. 1. The Top 10 Spoken Languages in the World with their Corresponding Percent

Related Work

Building Categorization Systems

Building Taxonomies for News. (Li, 2013) explains how news organizations, such as the NY Times and the Associated Press, build robust taxonomies that their computers use to automatically tag news content. (Woehler & Faerber, 2007) discuss how they generate a taxonomy from a collection of documents and how later they connect the taxonomy with document data. They mention that building the taxonomy deals with labeling and classifying documents to help users search and retrieve documents efficiently. (Uschold & King, 1995) outline key points toward building a methodology for ontologies; they define a technique that can be used to identify terms for the taxonomy. They describe the importance of knowing why the taxonomy is being built, what the use of it is, and who its users are. By identifying these key features we can create more focused taxonomies. (Fernández & Fernández, 2004) have developed tools to carry the Semantic Web into the journalism domain, including development of a taxonomy for news. What we are doing with our taxonomy is quite similar to the above-mentioned research. We built our taxonomy considering user needs, and defined the taxonomy terms based on the collection and data we studied, all applied to the news domain. Our taxonomy will cover Arabic news articles, so we built an Arabic taxonomy with terms chosen to help classify news-oriented Arabic textual data.

Building Taxonomies and Evaluation. Evaluating a taxonomy is still an emerging area of research. (Gómez, 2004) discusses ways to evaluate ontologies, considering consistency, completeness, conciseness, expandability, and sensitiveness. (Brank, Grobelnik & Mladenić, 2005) survey the state of the art in taxonomy evaluation. They highlight the need for evaluation to determine which ontology is best for a particular purpose. Our taxonomy is domain dependent since we built it for the Arabic news domain, but it should be general enough to cover any Arabic national news collection. To confirm this, a librarian taxonomy expert and Arabic native speaker volunteers helped to evaluate our taxonomy; they followed some of the evaluation techniques discussed above, considering completeness and consistency.

Arabic IR & NLP, Stemming, Text Classification, and Evaluation

Arabic Stemming. The Khoja stemmer (Khoja & Garside, 1999) is a well-known root-based Arabic stemmer used by many researchers due to its relative effectiveness as compared to other Arabic stemmers. Many researchers have tried to enhance its effectiveness, like (Al-Kabi, 2013). Most Arabic heavy (root-based) stemmers use patterns to extract the Arabic root from native Arabic words, but not all Arabic stemmers consider Arabic verb patterns. (Al-Sarhan, Al-Shalabi & Kannan, 2003) is one of the related stemming studies based on mathematical rules and relations between letters. Three phases have been used to develop a new Arabic root stemmer (Al-Kabi, Kazakzeh, Abu Ata, Al-Rababah & Alsmadi, 2014). They removed prefixes and suffixes in phase one, in phase two they compared the output of phase one to standard word sources, and they corrected the extracted root in the last phase. Their stemmer showed better results when compared with the Khoja stemmer (Khoja & Garside, 1999) and the Ghawanmeh stemmer (Ghawanmeh, Al-Shalabi, Kanaan, Khanfar, & Rabab'ah, 2009).

The effect of Arabic light stemming on the efficacy of information retrieval (IR) is presented by (Larkey, Ballesteros & Conell, 2007). Those researchers have built a number of light stemmers for Arabic, and evaluated their effectiveness for IR applications. They conclude that light stemming has a positive effect on Arabic IR. Also, (Kanaan, Al-Shalabi, Ababneh & Al-Nobani, 2008) have built another Arabic light stemmer and tested its effect on information retrieval of Arabic text. They compared the effect of their stemmer on Arabic IR relative to (Larkey, Ballesteros & Conell, 2007) and (Khoja & Garside, 1999). (Al-Omari & Abu Ata, 2014) developed an Arabic light stemmer that is not based on Arabic root patterns. Instead they used well-defined mathematical rules and some relations between letters to extract the stem. Their stemmer showed good results when tested on around 6225 Arabic words; it gives around 5733 correct results with around 92% accuracy.

Arabic dialects have been used for many years. Dialects present more challenges than the Modern Standard Arabic in the field of natural language processing since they add a new set of variational dimensions (Abu Ata & Al-Omari, 2014). Abu Ata et al. discussed in their paper a new rule based stemming algorithm that can find stems for the Arabian Gulf Dialect. They showed in their results that their algorithm is performing poorly when applied to Modern Standard Arabic but performing well with the Arabian Gulf Dialect.

In our discussion below of using stemming (root/light) to enhance Arabic classification we provide an example to explain why using root stemming will not help classification as much as light stemming .

We studied Arabic stemmers toward building a better understanding of stemming and toward creating an improved Arabic stemmer. We compared our proposed stemmer with well-known stemmers by classifying our data set after applying the various kinds of stemmers. We hypothesized and experimented to show that using light stemming will enhance classification for Arabic text, especially in the news domain.

Arabic IR and NLP. The effects of Arabic stop word removal and term weighting on the accuracy of Arabic information retrieval systems is examined by (El-Khair, 2006). (Hmeidi, Kanaan & Martha, 1997) describe how to build an automatic indexing system for Arabic text with comparable accuracy to human indexing systems. In their study, (Abuleil & Evens, 1998) show how to automatically build a large Arabic integrated and comprehensive lexicon. They developed a part of speech (POS) tagger for Arabic text to extract features of the Arabic words encountered. POS tagging is the process of assigning low-level grammatical categories to words based on their context. (Kanaan, Al-Shalabi & Sawalha, 2003) designed a fully automatic tagging system for Arabic language text, and achieved an accuracy rate of about 93%.

Arabic Text Classification. Text classification is the task of deciding whether a piece of text belongs to any of a set of predefined classes (Lewis, 1991). The problem of classification has been widely studied in the database, data mining, and information retrieval communities (Aggarwal & Zhai, 2012). However, the nature of Arabic text is different from that of English text. (Kanaan, Al-Shalabi, Ghwanmeh & Al-Ma'adeed, 2009) implemented three automatic text classification techniques for the Arabic language. A corpus of 1445 Arabic text documents belonging to 9 categories underwent testing. They compared automatic text classification using kNN, Rocchio, and Naïve Bayes on the Arabic language. The study concludes that Naïve Bayes was the best performer, followed by kNN and Rocchio. (El-Haless, 2006) used Maximum Entropy to aid classification of Arabic data sets. Results revealed that the average F1-measure increased from 68.13% to 80.41% using such pre-processing techniques. (Saad, 2011) compared the impact of text preprocessing on Arabic text classification using popular text classification algorithms. (Saad, 2011) applied different term weighting schemes and Arabic morphological analysis. He attempted to estimate the performance of different classification approaches that yield simple "If-Then" knowledge in order to select the most applicable category during Arabic text classification. (Khreisat, 2006) results show that N-gram text classification using the Dice measure outperforms classification using the Manhattan measure for a corpus that is to be described with four categories (Elberrichi & Abidi, 2012). (Mesleh, 2006) proposed a conceptual representation for Arabic text representation, and discusses the SVM algorithm with the use of Chi-square as a feature selection method, to classify Arabic documents. The results show that SVM with Chi-square outperforms the Naïve Bayes and kNN classifiers in terms of F1-measure .

As some of the above studies suggested, we applied preprocessing steps to our data set to enhance classification of Arabic text. We proved that using stemming as a preprocessing step enhances classification results. Our stemmer outperformed some of the most popular Arabic stemmers. We used three kinds of classifiers, and both binary and multiclass classification. Then we compared the classifiers with each other, and compared the different categories with each other as well.

IR Evaluation. (Kanaan, Al-Shalabi, Mohammed & Ahmad, 2004) used average recall/precision to compare an ad-hoc retrieval system with a filtering retrieval system. The most commonly used measures of retrieval performance are precision and recall (Lassi, 2002). Recall, precision, and F1-measure have been widely used in the history of IR system evaluation (Zhou & Yao, 2010). In order to evaluate our classifiers, both binary and multiclass, we used the F1-measure, which is calculated from recall and precision.

Building a Standardized Categorization System for Arabic Newspapers

Building the Taxonomy: Five Arabic Newspapers

In order to build our general categorization system, we considered eight Arabic newspapers from five different countries. In particular, we studied and analyzed the category system for each of five Qatari Arabic newspapers:

Al-Rayah: This newspaper (Al-Rayah, Al-Raya Newspaper, 2014) has both online and paper versions. From FIG. 2, we observe a reasonable number of well-defined categories (Society, Locals, Politics, …) and sub-categories (Arabic news, Arts, Discussions, …) that can help identify news articles. We have crawled this collection and use it as our data set (see Section *Our Data Set: Al-Rayah Newspaper Collection*) for testing. This newspaper keeps their archive in PDF files, which was very helpful for processing.

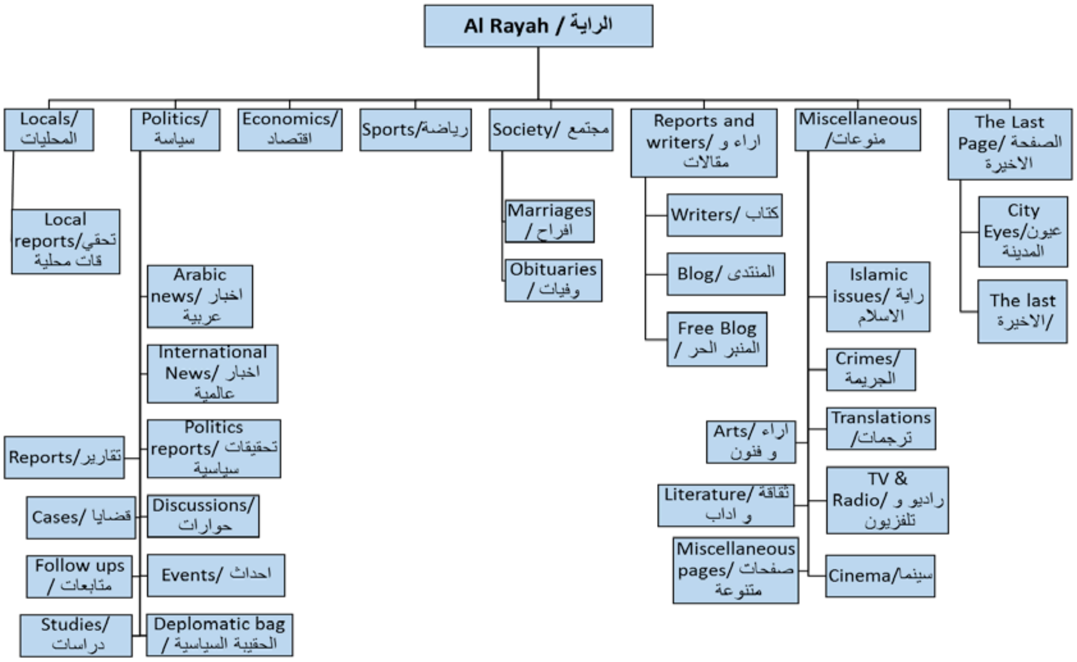


FIG. 2. The Taxonomy for Al-Rayah Newspaper

Qatar News Agency: This newspaper (QNA, Qatar News Agency, 2014) is published online only. The categorization system includes categories for local and foreign news and a good structure for sports. The whole categorization system is organized as illustrated in FIG. 3 below.

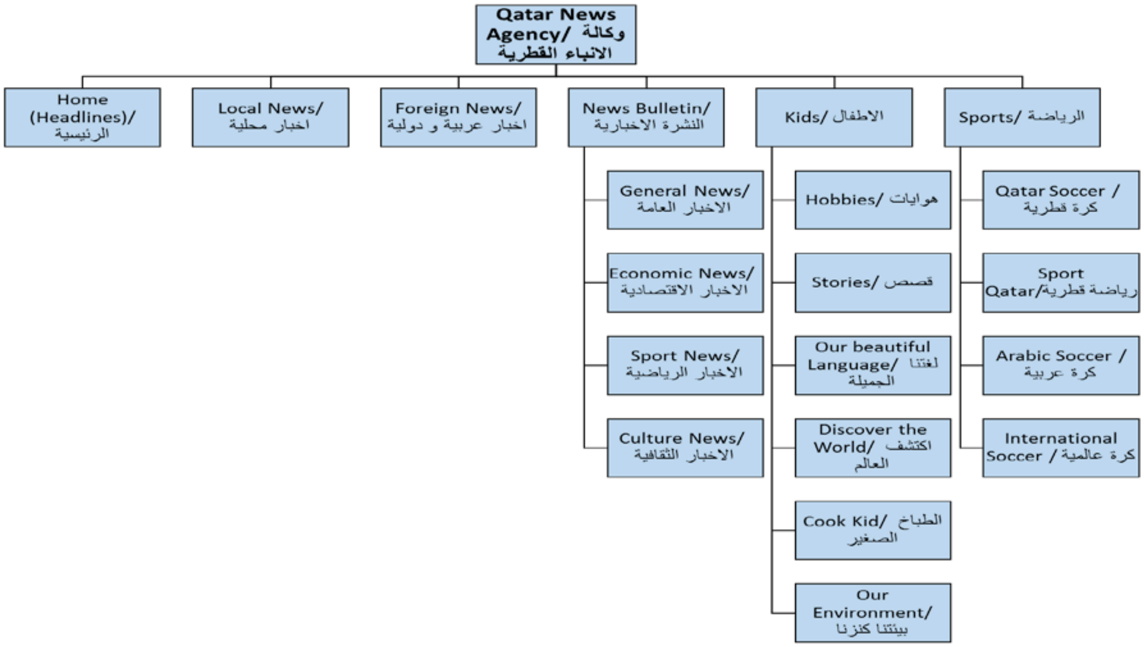

Qatar News Agency/ وكالة الانباء القطرية
Home (Headlines)/ الرئيسية
Local News/ اخبار محلية
Foreign News/ اخبار عربية و دولية
News Bulletin/ النشرة الاخبارية
General News/ الاخبار العامة
Economic News/ الاخبار الاقتصادية
Sport News/ الاخبار الرياضية
Culture News/ الاخبار الثقافية
Kids/ الاطفال
Hobbies/ هوايات
Stories/ قصص
Our beautiful Language/ لغتنا الجميلة
Discover the World/ اكتشف العالم
Cook Kid/ الطباخ الصغير
Our Environment/ بيئتنا كنزنا
Sports/ الرياضة
Qatar Soccer / كرة قطرية
Sport Qatar/رياضة قطرية
Arabic Soccer / كرة عربية
International Soccer / كرة عالمية

FIG. 3. The Taxonomy for Qatar News Agency

Al-Watan: This newspaper (Al-Watan, Al-Watan Newspaper, 2014) has both online and paper versions. As illustrated in FIG. 4, their taxonomy includes four categories (Economics, Sports, Citizens, and Al-Watan) in the first level and eight in the second level, all of them just under one main category called "Al-Watan".

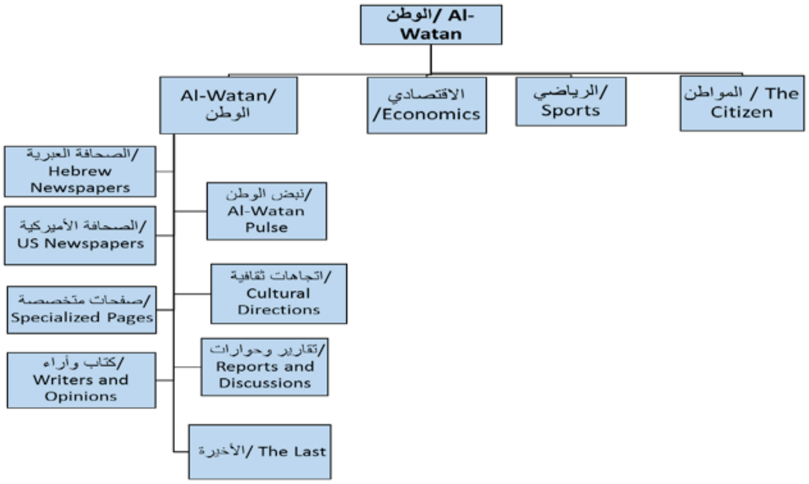

الوطن/ Al-Watan
Al-Watan/ الوطن
الاقتصادي /Economics
الرياضي/ Sports
المواطن / The Citizen
الصحافة العبرية/ Hebrew Newspapers
الصحافة الأميركية/ US Newspapers
صفحات متخصصة/ Specialized Pages
كتاب وأراء/ Writers and Opinions
نبض الوطن/ Al-Watan Pulse
اتجاهات ثقافية/ Cultural Directions
تقارير وحوارات/ Reports and Discussions
الأخيرة/ The Last

FIG. 4. The Taxonomy for Al-Watan Newspaper

Al-Arab: This newspaper (Al-Arab, Al-Arab Newspaper, 2014) has both online and paper versions, (see FIG. 5). We see in their taxonomy many categories (Sports, Arts, Economics, International, Qatar, …) and sub-categories (Local, Arabic, Accidents, …).

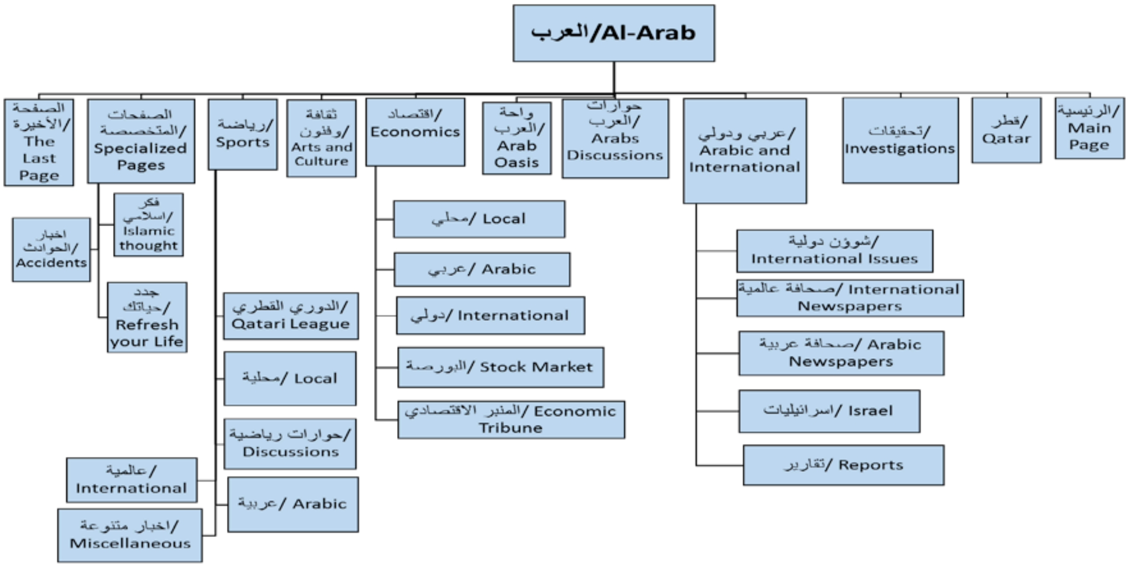

العرب/Al-Arab
الصفحة الأخيرة/ The Last Page
الصفحات المتخصصة/ Specialized Pages
رياضة/ Sports
ثقافة وفنون/ Arts and Culture
اقتصاد/ Economics
واحة العرب/ Arab Oasis
حوارات العرب/ Arabs Discussions
عربي ودولي/ Arabic and International
تحقيقات/ Investigations
قطر/ Qatar
الرئيسية/ Main Page
اخبار الحوادث/ Accidents
فكر اسلامي/ Islamic thought
جدد حياتك/ Refresh your Life
الدوري القطري/ Qatari League
محلية/ Local
حوارات رياضية/ Discussions
عالمية/ International
عربية/ Arabic
اخبار متنوعة/ Miscellaneous
محلي/ Local
عربي/ Arabic
دولي/ International
البورصة/ Stock Market
المنبر الاقتصادي/ Economic Tribune
شوؤن دولية/ International Issues
صحافة عالمية/ International Newspapers
صحافة عربية/ Arabic Newspapers
اسرائيليات/ Israel
تقارير/ Reports

FIG. 5. The Taxonomy for Al-Arab Newspaper

Al-Sharq: This newspaper (Al-Sharq, Al-Sharq Nespaper, 2014) has both online and paper versions,see FIG. 6). Based on their hierarchy, we see that their taxonomy has a reasonable number of categories (Sports, News, Economics, Accidents, …) and sub-categories Reports, Pictures, Investigations, …) that cover all of the important news articles.

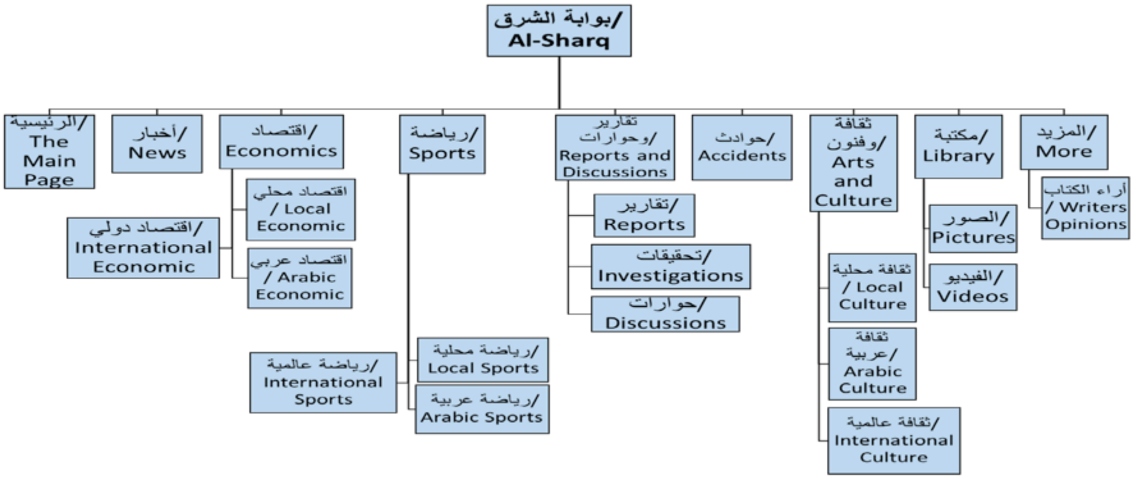

بوابة الشرق/ Al-Sharq
الرئيسية/ The Main Page
أخبار/ News
اقتصاد/ Economics
اقتصاد محلي / Local Economic
اقتصاد دولي/ International Economic
اقتصاد عربي / Arabic Economic
رياضة/ Sports
رياضة عالمية/ International Sports
رياضة محلية/ Local Sports
رياضة عربية/ Arabic Sports
تقارير وحوارات/ Reports and Discussions
تقارير/ Reports
تحقيقات/ Investigations
حوارات/ Discussions
حوادث/ Accidents
ثقافة وفنون/ Arts and Culture
ثقافة محلية / Local Culture
ثقافة عربية/ Arabic Culture
ثقافة عالمية/ International Culture
مكتبة/ Library
الصور/ Pictures
الفيديو/ Videos
المزيد/ More
أراء الكتاب / Writers Opinions

FIG. 6. The Taxonomy for Al-Sharq Newspaper

Building the General Categorization System

With the aim to enhance Arabic news article classification and improve online newspaper browsing, we created a categorization system called the General System (Taxonomy).

We studied the five newspapers discussed above, along with a number of other Arabic newspapers (i.e., Alghad: http://www.alghad.com/, AlAhram: http://www.ahram.org.eg/, AlKhabar: http://www.alkhabar.ma/, and AlQudsAlarabi: http://www.alquds.co.uk/) that we reviewed. We analyzed their online categorization systems, to gain a wider understanding of Arabic newspaper taxonomies. Based on the five hierarchies and the understanding we gained from the other reviewed newspapers, we created our unified categorization system. We identified common categories between the mentioned taxonomies, and considered topic coverage, as we developed our categorization system. The result should be applicable not only to the five newspapers, but also to any Qatari newspaper, and, in general, any Arabic newspaper. The generality of our taxonomy is inferred since we studied many Arabic newspapers, including five in depth, to find the common categories between them and find what are the general categories that should be included in a newspaper to cover any news article topic.

Graphical mapping methods were used to map the studied taxonomies toward getting a general taxonomy. We received help from a librarian expert and volunteers to identify the categories and the level of generality for each category and sub-category. We call this the "General Categorization System" or "General Taxonomy" (see FIG. 7). It contains twenty-seven different categories divided into three levels: seven categories in the main or first level, thirteen categories in the second level, and seven categories in the third and last level. By creating this taxonomy we aimed to enhance Arabic news article classification and improve online newspaper browsing.

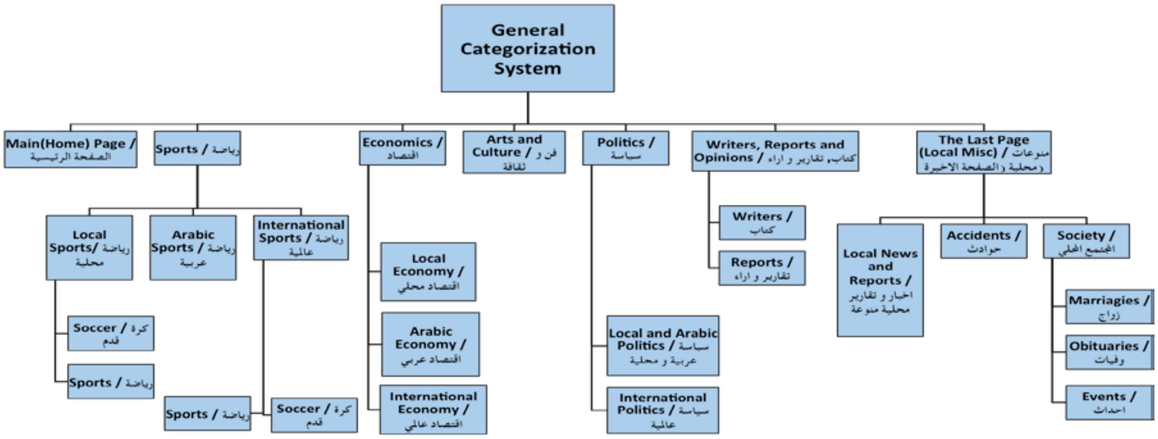

General Categorization System
Main(Home) Page / الصفحة الرئيسية
Sports / رياضة
Local Sports/ رياضة محلية
Soccer / كرة قدم
Sports / رياضة
Arabic Sports / رياضة عربية
International Sports / رياضة عالمية
Sports / رياضة
Soccer / كرة قدم
Economics / اقتصاد
Local Economy / اقتصاد محلي
Arabic Economy / اقتصاد عربي
International Economy / اقتصاد عالمي
Arts and Culture / فن و ثقافة
Politics / سياسة
Local and Arabic Politics / سياسة عربية و محلية
International Politics / سياسة عالمية
Writers, Reports and Opinions / كتاب, تقارير و اراء
Writers / كتاب
Reports / تقارير و اراء
The Last Page (Local Misc) / منوعات محلية (الصفحة الاخيرة)
Local News and Reports / اخبار و تقارير محلية منوعة
Accidents / حوادث
Society / المجتمع المحلي
Marriagies / زواج
Obituaries / وفيات
Events / احداث

FIG. 7. The General Categorization System (Taxonomy)

The IPTC System: International Press Telecommunications Council

Founded in 1965, IPTC is a comprehensive worldwide standards body for news media, with main office in London. Their mission is to make information dissemination easier and standard-based. They create technical standards to advance information management and exchange among news media providers and customers. IPTC further provides open standards and makes them accessible and available to users worldwide, free of charge. IPTC generates and preserves sets of concepts in the form of a controlled vocabulary or taxonomy, through IPTC NewsCodes.

For our taxonomy, we have focused on the Subject Code part of IPTC Descriptive NewsCodes, because it is the main IPTC taxonomy, focused on text, and addresses subjects of items. (IPTC, Interactive Diagram for the Subject NewsCodes, 2014) illustrates the taxonomy that we use in the creation and assessment of our proposed taxonomy. Please see FIG. 8 below for examples, noting that yellow shows which node in the tree is expanded further during interactive browsing.

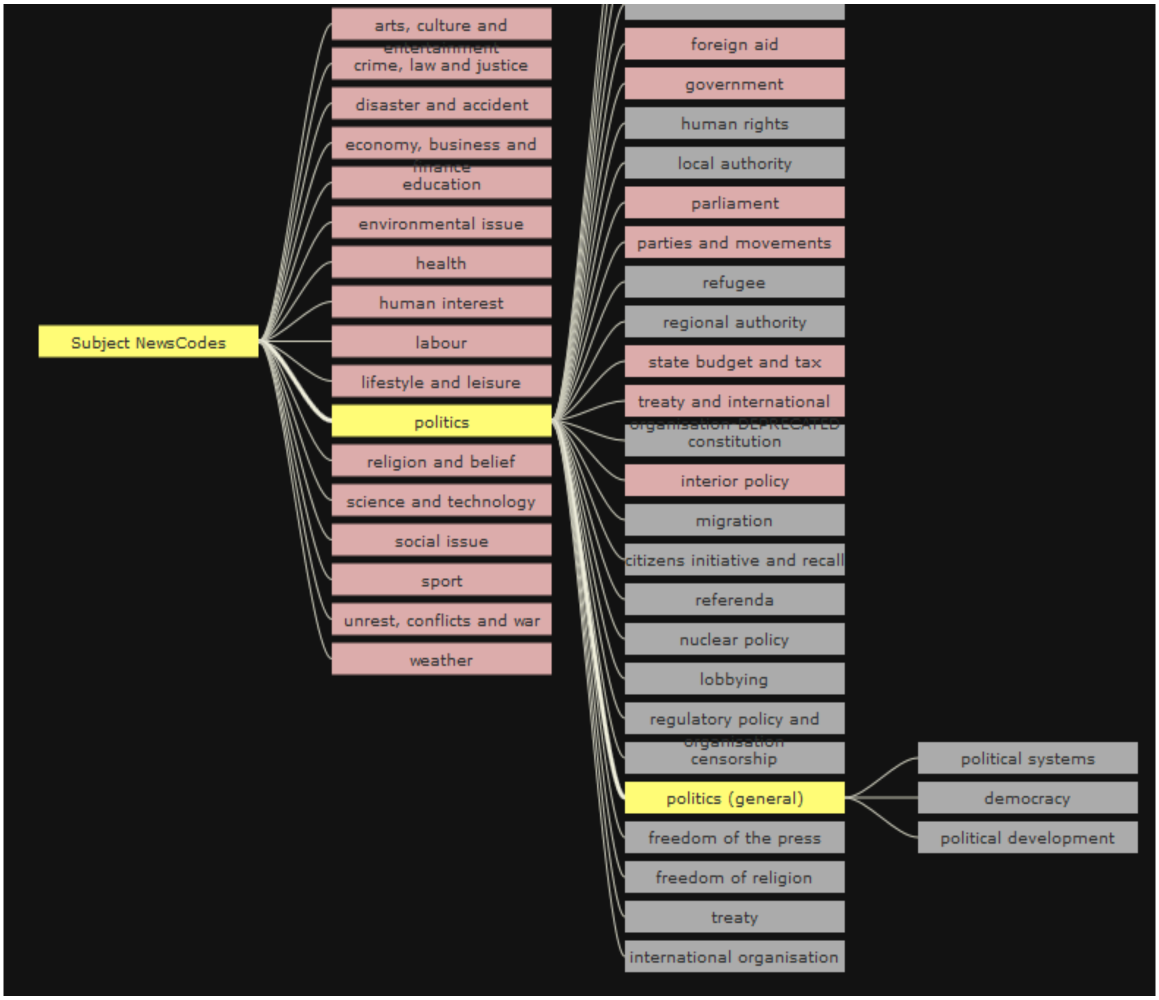


FIG. 8. Example of Subject NewsCodes in the IPTC Taxonomy for "Politics" Category

The Standardized Categorization System

Revising According to IPTC. To make our general categorization system more accepted, standardized, and widely usable with Arabic newspapers, we modified our initial general categorization system and made it compatible with the IPTC system (see the section above on The IPTC System: International Press Telecommunications Council). The main point of this taxonomy is to assist and enhance Arabic newspaper browsing.

Making our system compatible with a worldwide system like the IPTC should ensure wider acceptance. To adjust for the IPTC system, we had to trim some of the categories in our general system and in the IPTC system. Reducing the number of categories by combining or generalizing them to make them cover more topics adds generality to the system. We also reduced the levels in our general system from three to two, to make it even more broad, workable, and compatible with Arabic newspapers. For example, we combined the three sub-categories in the second level of the general system (international sports, Arabic sports, and local sports) plus their four sub-categories in the third level, into only two sub-categories (sports general and soccer). So, we reduced the number of categories in the sports class from seven to two; this will give more generality to those two categories. After giving the system broader generality of topical coverage, we made sure that all of the categories and their sub-categories are grounded in the IPTC system. Accordingly, we can say that our system is compatible with an accepted international news categorization system.

Thus, based on the IPTC system and the general system we initially created, we devised a new modified categorization system with thirteen categories and two levels. There are five categories in level one and eight categories in level two, with at most two in the second level under any category in level one. We called this the "Standardized Arabic Newspapers Categorization System" (see FIG. 9). The main aim of this standardized system is to provide a better taxonomy for browsing Arabic news articles.

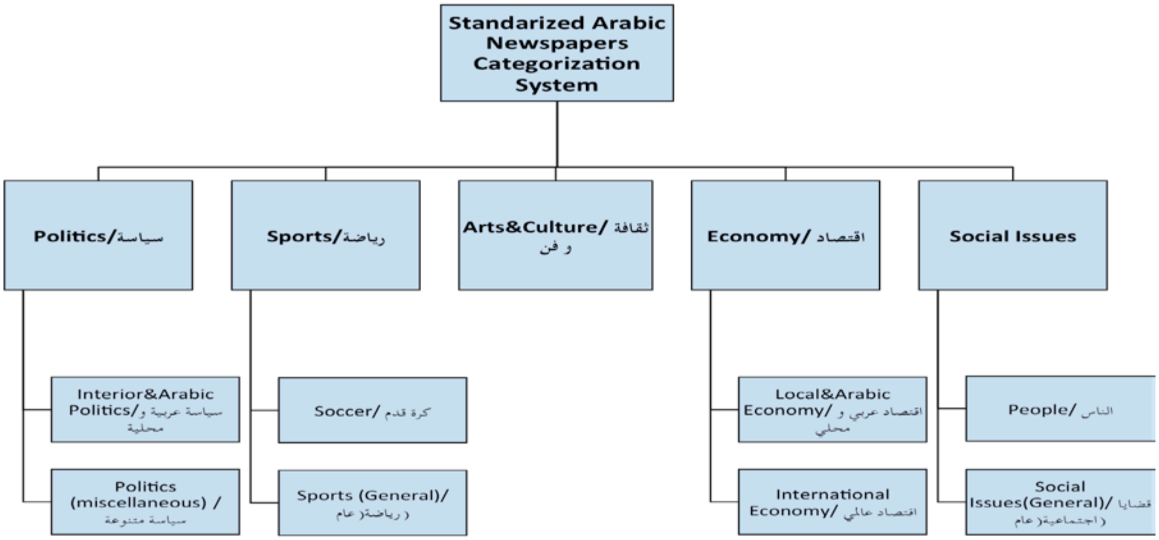

Standarized Arabic Newspapers Categorization System
Politics/سياسة
Sports/رياضة
Arts&Culture/ ثقافة و فن
Economy/ اقتصاد
Social Issues
Interior&Arabic Politics/سياسة عربية و محلية
Politics (miscellaneous) / سياسة متنوعة
Soccer/ كرة قدم
Sports (General)/ (رياضة(عام
Local&Arabic Economy/ اقتصاد عربي و محلي
International Economy/ اقتصاد عالمي
People/ الناس
Social Issues(General)/ قضايا (اجتماعية(عام

FIG. 9. The Standardized Arabic Newspapers Categorization System (Taxonomy)

Evaluation and Refinement. An ontology librarian expert and five native Arabic speaking volunteers helped evaluate versions of our categorization systems. Volunteers aided by creating categories for the general and the standardized taxonomy, based on studying the original taxonomies for the online Qatari newspapers. They confirmed that each category was indeed representative of topics in news articles. The librarian expert went through the new system, validated and approved its coverage, and cross-referenced it against the IPTC system. The domain expert librarian discussed the need and the importance of each category in our taxonomy. He suggested modifying, removing, or adding categories to make the system more general and compatible with the IPTC. For example, he suggested modifying the name and the coverage of the Last Page (Local Miscellaneous) category in the general system to call it Social Issues in the standardized system. This modification helps make this part of the system match better with IPTC, since IPTC has some categories talking about Social Issues but does not have anything called Last Page. Another example of an enhancement is removing the Main (Home) Page category from the general system since this category is vague, and it is not clear what topics can be included under it; also, it is not well matched with IPTC categories. This Main (Home) Page category usually contains different kinds of topics (e.g., like the first page of any newspaper) that can easily fit in one of the other categories, so there is no need to keep it. By deleting this category we again made our system more generalized and a better match with IPTC. Finally, the theory of "Magical Number Seven, Plus or Minus Two" for numbering (Miller, 1956) has guided the selection of the number of categories in the two levels in our taxonomy (5 categories in level 1, 8 categories in level 2).

Arabic News Articles Text Classification with Stemming

Our Data Set: Al-Rayah Newspaper Collection

Al-Rayah is a Qatari newspaper published in Arabic. This newspaper is used in our experiments. We employed the open-source Heritrix Crawler (Internet Archive, Heritrix Web Crawler, 2014) installed on one of the ELISQ servers. Our crawled collection's size is around 8.3 GB. The number of PDF files is around 2,200 (full newspapers) with more than 125 articles per newspaper, eventually totaling to around

237K articles. Newspapers in our initial collection are from March 2004 (the earliest provided) through July 2013 (the date of crawling). For more details, see TABLE 1.

TABLE 1. Our Data Set

File Format	PDF	Text
Original Number of Files	2,200	2,000
Number of Cleaned Files	2,100	1,900
Size on Disk	8.4GB	750MB
Avg. Number of Articles per Newspaper (File)	125	125
Total Number of Articles	260K	237K
Avg. Number of Words per Article	180	180
Avg. Number of Words per File	23,000	23,000

Stemming to Enhance Arabic Text Classification for News Articles

The main goal of a stemmer is to map different forms of the same word to a common representation called the "stem". Stemming can significantly improve the performance of text classification systems by reducing the dimensionality of word vectors. Generally, there are two main categories of Arabic stemmers: root extraction stemmers and light stemmers (Kanaan, Al-Shalabi, Ababneh & Al-Nobani, 2008). The most widely used stemmers for Arabic, one from each of these categories, respectively, are (Khoja & Garside, 1999) and (Larkey, Ballesteros & Conell, 2007).

In Arabic, each word has a root that acts as its basic form. We can obtain several words, including nouns, verbs, and adjectives, by adding certain letters at the beginning, end, or within the root letters, see **FIG. 10**. For example, from the root "قصد", we can produce the words "يقصد", "مقاصد", "اقتصادية", "الاقتصادي", etc., see TABLE 2. The word in the right is the root of the word on the left. The blue letters represent the word prefixes, the red letters represent the word infixes, and the green letters represent the postfixes.

TABLE 2. Examples of Root Stemming

Word	Root
الاقتصادي The economic	قصد Meant
مقاصد Purposes	قصد Meant

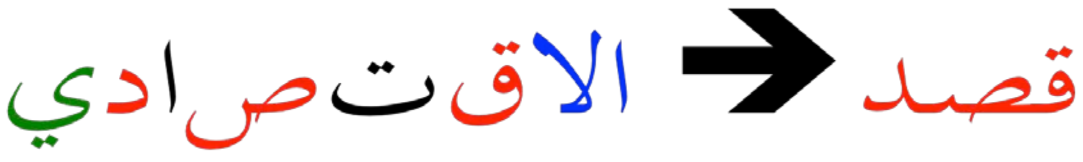


FIG. 10. An Example of Arabic Root Stemming that Shows the Different Types of Affixes in Arabic (blue for prefixes, red for infixes, and green for postfixes)

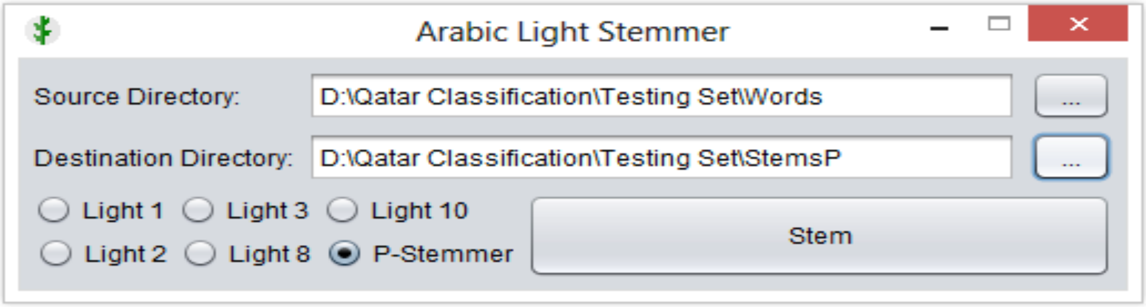
The goal of a root-based stemmer is to extract the basic form for any given word. The problem with extracting the root is that it is far more abstract than the word. Different words with completely different meaning can originate from the same root. For example, the word “مقاصد” (i.e., “purposes”) and the word “الاقتصادي” (i.e., “The economic”) both originate from the root “قصد”. Consequently, using root stemmers can result in poor classification effectiveness and problems with cross-lingual retrieval (Larkey, Ballesteros & Conell, 2007), since it will give the deep abstract concept of the word and that will lead sometimes to a different meaning between the word and the root, as in the example above.

The goal of a light stemmer is to find a canonical form of an Arabic word by removing prefixes and suffixes, while maintaining infixes. Usually, the meaning of the word remains intact, which results in improved classification effectiveness. For example, the stem for the words “اقتصادي” (i.e., “economic”) and “والاقتصاد” (i.e., “and the economy”) is “اقتصاد” (i.e., “economy”), rather than the root “قصد” (i.e., “intended”), see **TABLE 3**.

TABLE 3. Examples of Light Stemming

Word	Light Stemmer Result
اقتصادي	اقتصاد
الاقتصاد	اقتصاد

We implemented a tool to help stem Arabic words, that takes two arguments: the path of a directory containing the raw text files as input source and the path of a destination directory for output (see FIG. 11).



Arabic Light Stemmer

Source Directory: D:\Qatar Classification\Testing Set\Words

Destination Directory: D:\Qatar Classification\Testing Set\StemsP

○ Light 1 ○ Light 3 ○ Light 10

○ Light 2 ○ Light 8 ● P-Stemmer

FIG. 11. A Screenshot of the “Arabic Light Stemmer” Tool

The “Arabic Light Stemmer” tool includes five versions of a light-stemming algorithm (Larkey, Ballesteros & Conell, 2007). Each version of the algorithm strips off certain prefixes and suffixes as shown in TABLE 4. For example, the Light10 stemmer (Larkey & Ballesteros & M. Conell 2007) reduces the word “المدرسون” (i.e., “the teachers”) to the stem “مدرس” (i.e., “teacher”) by removing the “ال” (i.e., “the”) prefix and the “ون” suffix (which indicates a male plural). Although Light10 is the most used and best performing version (Otair, 2013) of Larkey’s light stemmer, we have included all other versions for evaluation and comparison purposes.

TABLE 4. The Five Larkey’s Versions of Arabic Light Stemmer and the P-Stemmer.

Version	Prefixes to remove	Suffixes to remove
Light 1	“ال”, “وال”, “بال”, “كال”, “فال”	nothing
Light 2	“و”, “ال”, “وال”, “بال”, “كال”, “فال”	nothing
Light 3	“و”, “ال”, “وال”, “بال”, “كال”, “فال”	“ه”, “ة”
Light 8	“و”, “ال”, “وال”, “بال”, “كال”, “فال”	“ها”, “ان”, “ات”, “ون”, “ين”, “يه”, “ية”, “ه”, “ة”, “ي”
Light 10	“و”, “ال”, “وال”, “بال”, “كال”, “فال”, “لل”, “ول”, “ل”	“ها”, “ان”, “ات”, “ون”, “ين”, “يه”, “ية”, “ه”, “ة”, “ي”
P-Stemmer	“و”, “ال”, “وال”, “بال”, “كال”, “فال”, “لل”, “ول”, “ل”	nothing

P-Stemmer (Prefix Stemmer)

We argue that just removing word prefixes can give better results than removing both prefixes and suffixes (as with the Larkey stemmers), and hence can improve the effectiveness of text classifiers. For example, the Light10 stemmer reduces the word “المباحثات” (i.e., “the talks”) to the stem “مباحث” (i.e., “investigation”) by removing the “ال” (i.e., “the”) prefix and the “ات” suffix (which indicates a female plural). It is clear that the two words have completely different meanings. Thus, light stemmers that remove word suffixes can suffer from the same abstraction problem found in root stemmers, which is especially troublesome in text classification (see **TABLE 5**).

TABLE 5. Light10 vs. P-Stemmer

Word	Light10	P-Stemmer
كالصادرات As the imports	صادر Took	صادرات Imports
والوحدات And the units	وحد Aggregate	وحدات Units
المكتبات The libraries	مكتب Office	مكتبات The library
المباحثات The negotiations	مباحث Investigation	مباحثات Negotiations

To prove our argument, we have developed the P-Stemmer, a customized stemmer that removes word prefixes only and keeps all of the suffixes and infixes. This is used in our experiments along with five versions of Larkey’s light stemmer (recall TABLE 4).

Machine Learning Tools and Methods to Classify Arabic Text

The goal of a text classifier is to map documents into a fixed number of predefined classes. A text classifier can be either a binary classifier or a multiclass classifier.

In binary classification, a document can be in exactly one of two classes. In multiclass classification, a document can be in one and only one of multiple classes. When we apply our trained classifier to our test data, every document (instance) will be classified into only one class. Given a new instance, the classifier calculates a probability for each class, chooses the class with the highest probability, and classifies the new instance into this class.

Using supervised machine learning, classifiers can learn from examples and perform class assignments automatically. Several text classification algorithms have been proposed. We have chosen to use three of the most widely used state-of-the-art text classification approaches: Support Vector Machines (SVM), Naïve Bayes (NB), and Random Forest (RF).

Our goal is to develop a text classifier that can categorize a given document into one of the five classes at the first level of our generated taxonomy, see **FIG. 12**. FIG. 13 shows our Arabic Text Classification framework.

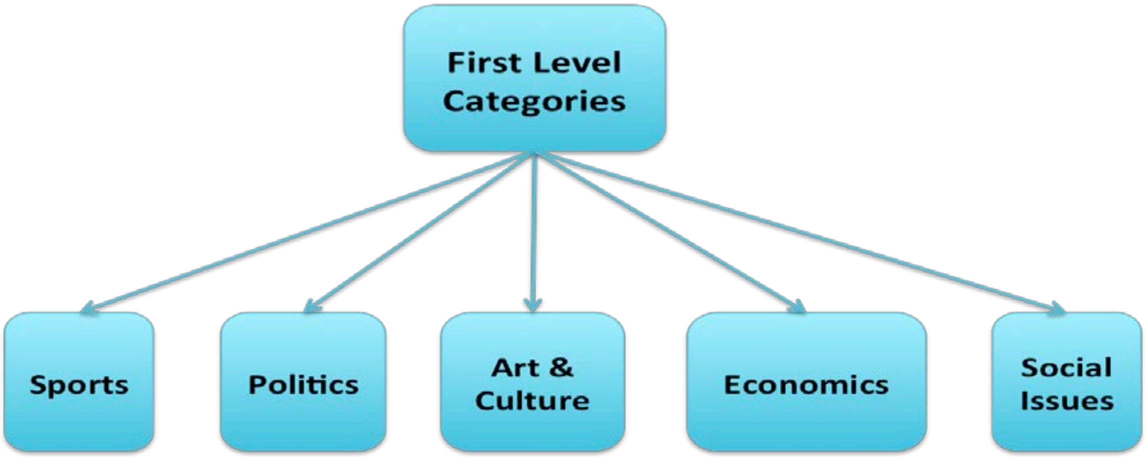


FIG. 12. The Five Classes in the first level of our Taxonomy

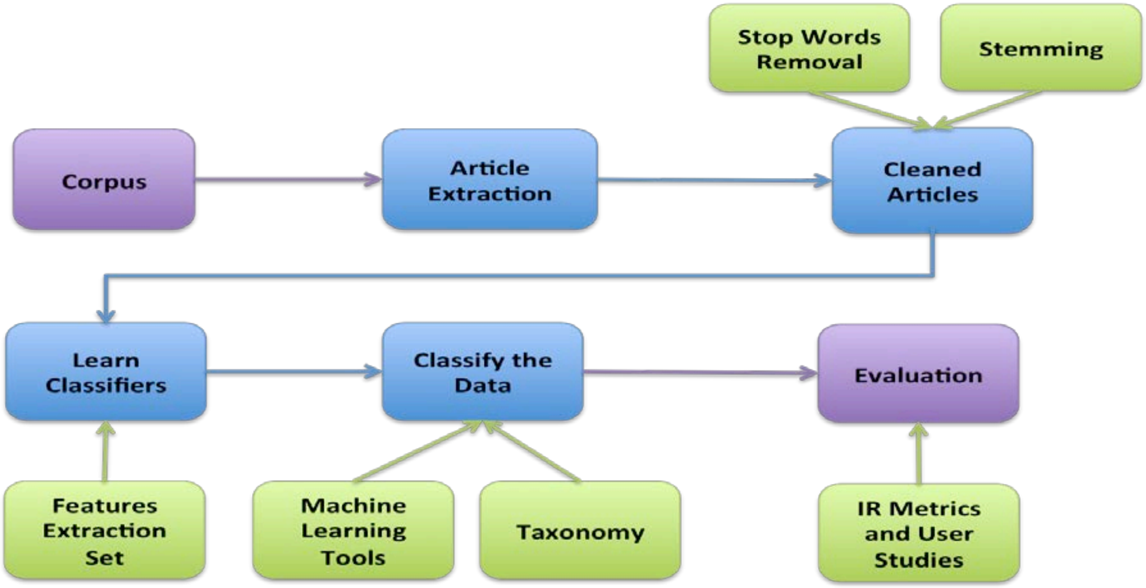


FIG. 13. Arabic Text Classification Framework

In order to train and test different classifiers using different machine learning techniques, we used Weka, a machine learning toolkit (Hall, Frank, Holmes, Pfahringer, Reutemann & Witten, 2009), which includes a suite of software written in Java, developed at the University of Waikato, New Zealand.

We have one training set, which is a set of 750 tuples: (news-article-ID, class-label). Volunteers manually assembled that so there are 150 instances for each of the 5 class labels, in the form of five text files for the 750 instances. For each news article, we have 7 different vectors, each for a different vector space; one for the full word, one for our proposed stemmer, and five for Larkey's stemmers. Each vector space is defined by its features, which are determined according to one of the seven methods of handling words (full word, P-stemmer, and Larkey's five stemmers). Our Arabic Light Stemmer tool (**FIG. 10**) is used to produce six different versions of the text files corresponding to the five versions of Larkey's light stemmers and the proposed P-Stemmer. Furthermore, the original word formulation is included to be compared with our results. Thus, we obtained seven versions of the same training set as shown in TABLE 6. Then we ran a classification experiment for each vector space, using the training set and each of three classifier techniques using the WEKA machine-learning tool.

TABLE 6. The Seven Vectors for the Same Training Set

Training set	Obtained from the word set after applying*
Words	Raw text words after cleaning text and removing stop-words.
Stem1	*Version 1 of the light stemmer.
Stem2	*Version 2 of the light stemmer.
Stem3	*Version 3 of the light stemmer.
Stem8	*Version 8 of the light stemmer.
Stem10	*Version 10 of the light stemmer.
P-Stemmer	*The proposed P-Stemmer.

Results, Evaluation, and Discussion

Overview of Classification Experiments and Evaluation

In order to use Weka to classify our data, the training set must be converted to an ARFF file. An Attribute-Relation File Format text file describes a list of instances sharing a set of attributes. Weka provides a Java tool, named TextDirectoryLoader, which can convert a set of text files into an ARFF file. TextDirectoryLoader takes two parameters: a directory and the output file name. It assumes that there are subdirectories within the provided directory, each corresponding to a given class and containing the text files representing the instances of that class. TextDirectoryLoader produces a single ARFF file that contains all instances, with two attributes per instance: text and class. For a given instance, the value of the text attribute is the content of the text file corresponding to that instance, while the value of the class attribute is the name of the subdirectory that contains this instance.

Recall, precision, and F1 (i.e., the harmonic mean of precision and recall) values are IR evaluation measures. The following formulas show how to calculate recall, precision, and F1-measure.

- Recall= (No. of Relevant and Retrieved Documents) / (No. of Relevant Documents)
- Precision= (No. of Relevant and Retrieved Documents) / (No. of Retrieved Documents)
- F1= 2 * ((Recall * Precision) / (Recall + Precision))

- The Relevant Documents are the articles assigned correctly to a specific class.
- The Retrieved Documents are the articles processed by the classifier.
- The Relevant and Retrieved documents are the articles that are correctly categorized by the classifier.

Multiclass Classification: Results and Evaluations

As mentioned previously, three of the most widely used classification approaches – Support Vector Machines (SVM), Naïve Bayes (NB), and Random Forest (RF) – are tested. Weka provides a classifiers list tree that includes these. We used an equal number of training instances with multiclass classification, 150 for each category (750 for the training set in total).

TABLE 7 shows the number of features for each of the seven training sets. In the table, “Distinct words” refers to the number of features after applying the “StringToWordVector” Weka filter, while “Selected features” refers to the number of features after applying the “StringToWordVector” Weka filter followed by the “AttributeSelection” Weka filter.

TABLE 7. Number of Features for each of the Training Set Versions to be used with Multiclass Classification

	Words	**Stems1**	**Stems2**	**Stems3**	**Stems8**	**Stems10**	**P-Stemmer**
Distinct Words	28,704	23,703	21,283	19,282	15,899	15,124	20,457
Selected Features	2,145	1,771	1,590	1,441	1,188	1,130	1,529

After selecting the feature sets and creating the word vectors for each of our data sets, we have built three text classifiers corresponding to the three classification techniques (SVN, NB, and RF) for each of the seven training set versions. 10-fold cross-validation was used to evaluate each of the 21 classifiers (three classifiers with seven different data sets). **TABLE 8** shows the average recall, precision, and F1 measure values after running the three classifiers over our seven different data sets.

TABLE 8. The Recall, Precision, and F1 Measure Values for the three Classification Techniques with Respect to the Seven Versions of the Training Set, for Multiclass Classification

Data Set Version	**SVM**			**Naïve Bayes**			**Random Forest**		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Words	0.917	0.913	0.915	0.92	0.913	0.916	0.918	0.92	0.919
Stems1	0.936	0.943	0.939	0.928	0.92	0.924	0.924	0.926	0.925
Stems2	0.935	0.94	0.937	0.93	0.926	0.928	0.924	0.923	0.923
Stems3	0.93	0.923	0.926	0.922	0.924	0.923	0.91	0.909	0.909
Stems8	0.933	0.938	0.935	0.928	0.921	0.924	0.915	0.913	0.914
Stems10	0.934	0.933	0.933	0.928	0.919	0.923	0.917	0.915	0.916
P-Stemmer	**0.942**	**0.949**	**0.945**	**0.932**	**0.928**	**0.93**	**0.936**	**0.931**	**0.933**

FIG. 14 shows the F1-measure values for the three classification techniques with the seven word variations for the multiclass classification corresponding to **TABLE 8**. Bold indicates best value in a column.

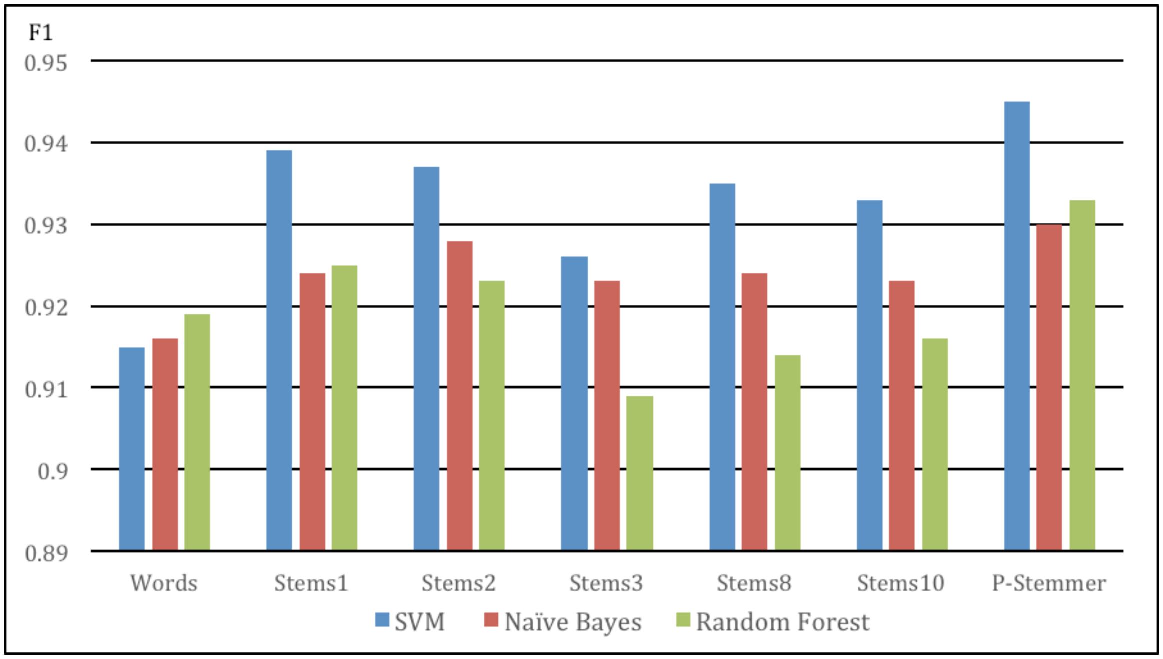


FIG. 14. F1-Measure Values for the Three Classification Techniques with the Seven Word Variations for Multiclass Classification

Binary Classification: Results and Evaluations

Processing the word instances with the P-Stemmer, we created five training sets corresponding to the five top-level classes (i.e., Art, Economy, Politics, Social Issues, and Sports). Each training set has 150 positive instances and 150 negative instances (around 35 randomly selected from each of the other four categories). We used Weka's "TextDirectoryLoader" tool to create the ARFF files for the five training sets.

TABLE 9 shows the number of features for each of the five training sets. In the table, "Distinct words" refers to the number of features after applying the "StringToWordVector" Weka filter, and "Selected features" refers to the number of features after applying that filter followed by the "AttributeSelection" Weka filter.

TABLE 9. Number of Features for each Training Set Version to be used with Binary Classification

	Art & Culture	**Economics**	**Politics**	**Social Issues**	**Sports**
Distinct Words	20,457	20,457	20,457	20,457	20,457
Selected Features	1,529	1,529	1,529	1,529	1,529

Afterwards, Weka is employed to build the three text classifiers for each of the five training sets. The results of 10-fold cross-validation of the three classifiers, for the training sets Art&Culture, Economics, Politics, Social Issues, and Sports, using the recall, precision and F1 measures, are shown in **TABLE 10**. **FIG. 15** shows the F1-measure values for the three classification techniques with the five different categories, for Binary Classification.

TABLE 10. The Recall, Precision and F1 Measure Values for the Three Classification Techniques with Respect to the Five Categories Training Sets using the P-Stemmer, for Binary Classification. Bold is best F1 in a row.

	SVM			Naïve Bayes			Random Forest		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**	**Recall**	**Precision**	**F1**
Art&Culture	0.989	0.983	0.986	0.971	0.965	0.968	0.93	0.923	0.926
Economics	0.956	0.958	0.957	0.936	0.945	0.94	0.925	0.929	0.927
Politics	0.948	0.956	0.952	0.937	0.947	0.942	0.928	0.937	0.932
Social Issues	0.999	0.997	0.998	0.995	0.997	0.996	0.98	0.969	0.974
Sports	0.992	0.986	0.989	0.996	0.994	0.995	0.959	0.972	0.965

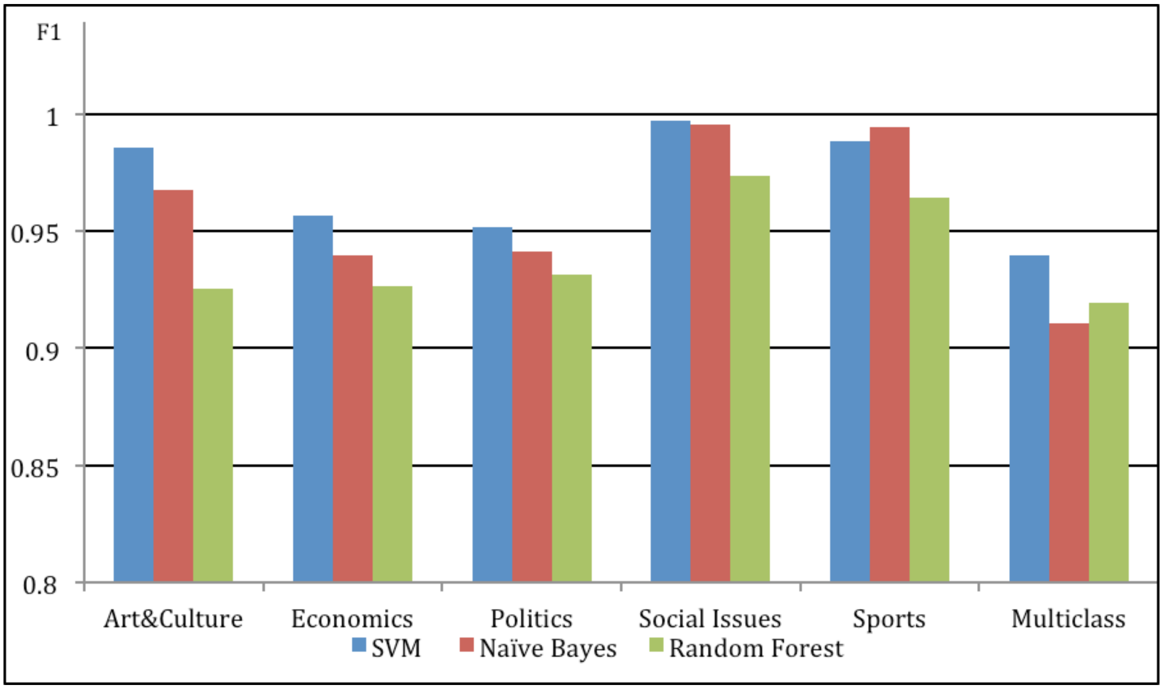


FIG. 15. F1-Measure Values for the Three Classification Techniques with Respect to the Five Categories Training Sets, for Binary Classification vs. Multiclass

Discussion

In the multiclass classification experiments, we compared our proposed P-Stemmer with six different word formulations as listed above in TABLE 6. We generated seven data sets and extracted seven feature sets toward applying them to our classifiers. We used the SVM, NB, and RF classifiers to judge the performance of the P-Stemmer for classification, and compared it with the other listed approaches. For each of the three classifiers, P-Stemmer gave better results than the full word and five Larkey's stemmers. We also found that SVM gave the best results, relative to the other two classifiers, when stemming is employed. Using the full words resulted in the lowest performance, so we can say that using stemming enhances text classification. We calculated recall and precision for each of our datasets, as well as F1, see TABLE 8. Our P-Stemmer performed very well compared to the other stemmers and the full word option, see FIG. 14.

In the binary classification experiments, we applied the P-Stemmer to our original data set and then extracted the set of features. Those were used when building our three classifiers (SVM, NB, and RF). We compared the results for the five different categories and the three different classifiers. We first calculated the recall and precision and then calculated the F1-measure. From these results, see **TABLE 10** and **FIG. 15**, we noticed that the Social Issues class had the best classification accuracy over the five categories while the Politics class had the least F1 value over the five categories. The Social Issues class with the SVM classifier gave the best result over the categories and classifiers. We learned also that SVM generally performed the best, compared to the other two classifiers. When we compared the binary classification results with the multiclass classification, all using the P-Stemmer, we found that the binary classification method gave better results.

Significance Test

We used the F1 measure results from the SVM classifier to do a statistical significance test between our proposed P-Stemmer and each one of the five Larkey's stemmers. **TABLE 11Error! Reference source not found.** shows the F1 results using the SVM classifier for P-Stemmer and Stem1, Stem2, Stem3, Stem8, and Stem10.

TABLE 11. F1 Measure Results for the P-Stemmer and the Five Larkey's Stemmers

	SVM					
	P-Stemmer	**Stem1**	**Stem2**	**Stem3**	**Stem8**	**Stem10**
Art&Culture	0.918	0.915	0.912	0.912	0.921	0.920
Economics	0.935	0.919	0.918	0.904	0.910	0.900
Politics	0.915	0.913	0.908	0.864	0.889	0.896
Society	0.991	0.990	0.993	0.993	0.993	0.992
Sports	0.964	0.960	0.955	0.962	0.962	0.959

We used the Wilcoxon signed-ranked test (Wilcoxon, 1945) to compare our proposed stemmer and each one of Larkey's stemmers, with the P-value less than or equal to 0.05. This test is very popular for information retrieval evaluation (Smucker, Allan & Carterette, 2007). We did the test five times and successfully rejected our null hypothesis, for each one of the five tests: that "The median difference of the F1 measure of P-Stemmer and each one of Larkey's stemmers (Stem1, Stem2, Stem3, Stem8, or Stem10) is less than or equal to zero". We concluded that, using the F1 measure for evaluation; our P-Stemmer is statistically significantly better than each one of the five Larkey's stemmers.

Conclusion and Future Work

Online Arabic news articles are not consistently categorized, so when accessed in an aggregate collection rather than in a newspaper (site), they are hard to browse by category. Taxonomies used by a particular news service are not general enough to apply to other news service collections. Further, it was not known what would be the best method for classifying Arabic news stories according to a given taxonomy. Nevertheless, preprocessing steps are supposed to enhance the classification process, and stemming is supposed to be part of these preprocessing steps.

For the purpose of enhancing Arabic information retrieval and natural language processing, we have developed a standardized Arabic categorization system (taxonomy) to support browsing services for online Arabic newspapers. The same hierarchy aids us to classify our data. This taxonomy was evaluated by an expert in this domain with help from volunteers, and was further validated by mapping from a worldwide news taxonomy, i.e., the IPTC system. In order to classify our data using the taxonomy, we built three classifiers, and used a newly developed stemmer, i.e., P-Stemmer, a modified version of one of the Larkey's light stemmers that we hypothesized would enhance Arabic text classification. Then we ran classification experiments using binary and multiclass classification methods .

We used information retrieval evaluation measures to compare our classification results using P-Stemmer with those from each of six variations of Larkey's stemmer, as well as the original raw words. We found that using our proposed stemmer significantly enhanced classification results for Arabic textual data, when using three classifiers: Naïve Bayes, SVM, and Random Forest. We noticed that SVM performed better than the other two classifiers. We also found that using binary classification gave better results compared to multiclass classification. We did a Wilcoxon signed-rank test to test if the observed improvements with P-stemmer were statistically significant, and concluded they are.

In the future, we plan to:

- Test our stemmer with another data set, to see if results are compatible with our news data set.
- Apply the stemmer and the classification methods on this new data set to confirm our findings.
- Use different feature selection methods, like Chi-square, for the classifiers, to see if they will enhance classification results.

Acknowledgments

We acknowledge QNRF for their support. This research was made possible by NPRP grant # 4-029-1-007 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. We thank Digital Library Research Laboratory researchers for their help.

References

Abu Ata, B. M. and Al-Omar, A. A Rule-Based Stemmer for Arabic Gulf Dialect. Accepted for Publication in "Journal of King Saud University- Computer Information Sciences".

Abuleil, S., & M., Evens, (1998). Discovering lexical information by tagging Arabic newspaper text, in Proceedings of the Workshop on Computational Approaches to Semitic Languages. Association for Computational Linguistics: Montreal, Quebec, Canada. p. 1-7.

Aggarwal, C., & C., Zhai, (2012). Mining text data. Springer, 533 pages.

Al-Kabi, M., (2013). Towards improving Khoja rule-based Arabic stemmer. In Applied Electrical Engineering and Computing Technologies (AEECT), IEEE Jordan Conference on. IEEE.

Al-Kabi M. N., Kazakzeh S. A., Abu Ata B. M., Al-Rababah S. A., Alsmadi I. M. A Novel Root Based Arabic Stemmer. Accepted for Publication in "Journal of King Saud University- Computer Information Sciences".

Al-Omari, A. and Abuata, B. M. (2014). Arabic Light Stemmer (ARS). Journal of Engineering Science and Technology. Vol. 9, No. 6, 702-717.

Al-Rayah. Al-Rayah Newspaper. [Cited 10/15/2014]; available from: http://www.raya.com/portal.

Al-Serhan, H., & Al-Shalabi, R., & Kannan, G., (2003). New approach for extracting Arabic roots. In Arab Conference on Information Technology.

Al-Sharq. Al-Sharq Newspaper. [Cited 10/15/2014]; available from: http://www.al-sharq.com/.

Al-Watan. Al-Watan Newspaper, [Cited 10/15/2014]; available from: http://www.al-watan.com/.

Brank, J., & M. Grobelnik, & D. Mladenić, (2005). A survey of ontology evaluation techniques . In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD).

Elberrichi, Z., & K., Abidi, (2012). Arabic text categorization: a comparative study of different representation modes. Int. Arab J. Inf. Technol., 9(5): 465-470.

El-Halees, A., (2006). Mining Arabic association rules for text classification. In Proceedings of the First International Conference on Mathematical Sciences.

El-Khair, I., (2006). Effects of stop words elimination for Arabic information retrieval: a comparative study. International Journal of Computing & Information Sciences, 4(3): 119-133.

Fernández-Garcıa, N., & L. Sánchez-Fernández, (2004). Building an Ontology for news Applications. Poster Session of the 3rd International Semantic Web Conference, ISWC.

Ghawanmeh, S., Al-Shalabi, R., Kanaan, G., Khanfar, K., and Rabab'ah, S. Enhanced Algorithm for Extracting the Root of Arabic Words. In Proceeding of the Sixth International Conference on Computer Graphics, Imaging and Visualization. China (2009). 388-391.

Gómez-Pérez, A., (2004). Ontology Evaluation, in Handbook on Ontologies, S. Staab and R. Studer, Editors. Springer Berlin Heidelberg. p. 251-273.

Hall Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten (2009); the WEKA Data Mining Software: SIGKDD Explorations, Volume 11, Issue 1. http://www.cs.waikato.ac.nz/ml/weka.

Hmeidi, I., & G., Kanaan, & M., Evens, (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents. JASIS. 48(10): 867-881.

Internet Archive, Heritrix, Internet Archive Web Crawler. 09 June 2011 10/15/2014]; available from: http://crawler.archive.org/index.html.

IPTC. Interactive Diagram for the Subject NewsCodes in the IPTC system. [Cited 10/15/2014]; available from: http://show.newscodes.org/index.html?newscodes=subj&lang=en-GB&startTo=Show.

Kanaan, G., Al-Shalabi, R., Ababneh, M., Al-Nobani, A., (2008). Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. In Proceeding of Innovations in Information Technology, IIT, International Conference on, IEEE.

Kanaan, G., R., Al-Shalabi, M., Sawalha, (2003). Full automatic Arabic text tagging system. In Proceedings of the International Conference on Information Technology and Natural Sciences, Amman/Jordan.

Kanaan, G., & R., Al-Shalabi, & MOHAMMED, A., & Ahmad, A., (2004). Comparison between Ad-hoc Retrieval and Filtering Retrieval Using Arabic Documents. International Journal of Computer Processing of Oriental Languages. 17(03): 181-199.

Kanaan, G., & R., Al-Shalabi, & s., Ghwanmeh & H., Al-Ma'adeed, (2009). A comparison of text‐classification techniques applied to Arabic text. Journal of the American society for information science and technology. 60(9): 1836-1844.

Khoja, S., & R., Garside, (1999). Stemming Arabic text. Lancaster, UK, Computing Department, Lancaster University, Lancaster, U.K. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps

Khreisat, L., (2006). Arabic text classification using N-gram frequency statistics a comparative study. In Proceedings of the International Conference on Data Mining.

Larkey, L., & L., Ballesteros, & M., Connell, (2007). Light stemming for Arabic information retrieval, in Arabic Computational Morphology, Springer. p. 221-243.

Lassi, M., (2002). Automatic thesaurus construction. University College of Boras, Sweden., http://www.academia.edu/506142/Automatic_thesaurus_construction.

Lewis, D., (1991). Evaluating Text Categorization. In Proceedings of the Workshop on Speech and Natural Language (HLT '91). Association for Computational Linguistics, Stroudsburg, PA, USA.

Li, A., 2014. How taxonomies help news organizations understand and categorize their content. Sep. 2, 2013, [Cited 10/15/2014]; available from: http://www.poynter.org/how-tos/digital strategies/222187/how-taxonomies-help-news-organizations-understand-and-categorize-their-content/.

Mesleh, A., (2006). Chi square feature extraction based SVMs Arabic language text categorization system. Journal of Computer Science, 3(6): p. 430.

Miller, George A. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, no. 2 (1956): pp. 81.

Nationsonline. Most widely spoken Languages in the World. [Cited 11/24/14]; available from: http://www.nationsonline.org/oneworld/most_spoken_languages.htm

Otair, M., (2013). Comparative analysis of Arabic stemming algorithms. International Journal of Managing Information Technology. 5(2).

QNA. Qatar News Agency, [Cited 10/15/2014]; available from: http://www.qna.org.qa/

Saad, M., (2011). Arabic Text Classification. Lap Lambert Academic Publishing. 172 pages.

Smucker, Mark D., James Allan, and Ben Carterette, (2007). A comparison of statistical significance tests for information retrieval evaluation. In Proceedings of the sixteenth ACM conference on information and knowledge management, pp. 623-632. ACM.

Uschold, M., & M., King, (1995). Towards a methodology for building ontologies. In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95.

Wikipedia. List of languages by number of native speakers. 12 October 2014 [Cited 10/15/2014]; available from: http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.

Wilcoxon, F. Individual comparisons by ranking methods. Biometrics Bulletin (1945): Vol. 1, No. 6, pp. 80-83.

Woehler, J. & F., Faerber, (2007). Taxonomy generation for electronic documents. Patent number US7243092 B2.

Zhou, B., & Y., Yao, (2010). Evaluating information retrieval system performance based on user preference. Journal of Intelligent Information Systems. 34(3): 227-248.