Arabic News Text Classification and Summarization: A Case of the Electronic Library Institute SeerQ (ELISQ)

Tarek Ghaze Kanan

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Edward A. Fox, Chair
Riyad F. Al-Shalabi
Weiguo (Patrick) Fan
Roger W. Ehrich
Clifford A. Shaffer

June 4, 2015
Blacksburg, Virginia

Keywords: Classification, Summarization, Arabic Language, Natural Language Processing, Digital Libraries

Arabic News Text Classification and Summarization: A Case of the Electronic Library Institute SeerQ (ELISQ)

By

Tarek Ghaze Kanan

ABSTRACT

Arabic news articles in heterogeneous electronic collections are difficult for users to work with. Two problems are: that they are not categorized in a way that would aid browsing, and that there are no summaries or detailed metadata records that could be easier to work with than full articles. To address the first problem, schema mapping techniques were adapted to construct a simple taxonomy for Arabic news stories that is compatible with the subject codes of the International Press Telecommunications Council. So that each article would be labeled with the proper taxonomy category, automatic classification methods were researched, to identify the most appropriate. Experiments showed that the best features to use in classification resulted from a new tailored stemming approach (i.e., a new Arabic light stemmer called P-Stemmer). When coupled with binary classification using SVM, the newly developed approach proved to be superior to state-of-the-art techniques. To address the second problem, i.e., summarization, preliminary work was done with English corpora. This was in the context of a new Problem Based Learning (PBL) course wherein students produced template summaries of big text collections. The techniques used in the course were extended to work with Arabic news. Due to the lack of high quality tools for Named Entity Recognition (NER) and topic identification for Arabic, two new tools were constructed: RenA for Arabic NER, and ALDA for Arabic topic extraction tool (using the Latent Dirichlet Algorithm). Controlled experiments with each of RenA and ALDA, involving Arabic speakers and a randomly selected corpus of 1000 Qatari news articles, showed the tools produced very good results (i.e., names, organizations, locations, and topics). Then the categorization, NER, topic identification, and additional information extraction techniques were combined to produce approximately 120,000 summaries for Qatari news articles, which are searchable, along with the articles, using LucidWorks Fusion, which builds upon Solr software. Evaluation of the summaries showed high ratings based on the 1000-article test corpus. Contributions of this research with Arabic news articles thus include a new: test corpus, taxonomy, light stemmer, classification approach, NER tool, topic identification tool, and template-based summarizer – all shown through experimentation to be highly effective.

DEDICATION

I dedicate this dissertation to my family, especially

to my mother and late father for their endless care and love;

to my wife for her love and patience;

to Bana and Noora my sweetheart daughters;

to Ghassan for his encouragement and assistance; and

to my brothers and sisters for their support.

ACKNOWLEDGMENT

We would like to acknowledge my committee's help and suggestions with this work, my advisor Dr. Edward Fox for all his support and advice over my years at Virginia Tech, and the group of students who volunteered to help build our baseline corpus and evaluate our topics and template summaries. Special thanks go to the 30 students enrolled in the computational linguistics class in the fall of 2014. We are grateful for Mary English's guidance regarding PBL. We thank Digital Library Research Laboratory researchers for their help. Special thanks goes to the co-authors of the publications we produced out of this work. We would also like to thank Mr. Philip Young, a scholarly communication librarian, for his expert assistance, and the Arabic native speakers, who served as experienced volunteers and helped with this work.

We acknowledge QNRF for their support. This Ph.D. dissertation 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 author. Thanks go to the US National Science Foundation for their support through grants DUE-1141209, IIS-0736055, IIS-0916733, and IIS-1319578.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

This chapter introduces the subsequent chapters and explains the motivation, problems, hypotheses, and research questions.

1.1 Background

The building of digital libraries has been widely adapted for many life situations. There are digital libraries for images, movies, health related issues, sports, disasters, and theses and dissertations. Digital library science allows multilingual content to serve users who prefer documents in different languages.

The idea behind the Electronic Library Institute - SeerQ (ELISQ) project (ELISQ, 2014) (see Figure 1-1) is to establish a digital library able to serve a whole country. That is, if anyone, for example, a scholar, visitor, or librarian, whether from inside or outside the country, is interested in any matter relevant to this country, such person could benefit from this service. In order to make the use of digital libraries broader and more responsive to the country's requirements, bilingual content was provided to serve any user who speaks either of the two popular languages, Arabic and English, and is interested in information relevant to Qatar. These languages have been selected based on the fact that Arabic is the formal language of the country, while English is the language most widely used in international communication, especially for Web content. The ELISQ project has two main objectives:

- The first is to build a digital library community, and to raise awareness of the importance of digital libraries. Accordingly, the project has held many workshops and seminars, and a consulting center has been established at Qatar University Library.
- The second objective is to build digital library infrastructure for the country by using advanced software packages to support and provide services to users. These services include searching and browsing of diverse collections – after crawling, indexing, and information extraction.

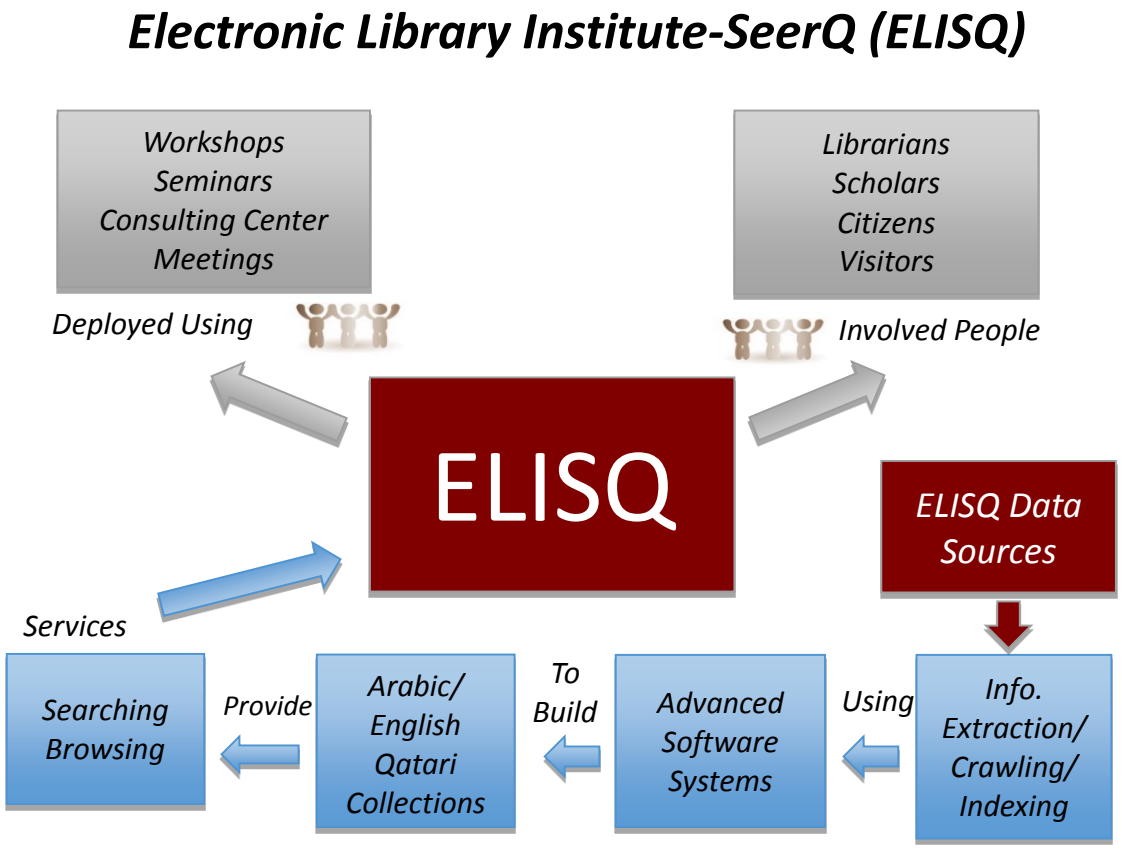

Figure 1-1: An Overview of ELISQ Project Activities

A very important part of this project is to adapt services and packages to Qatari needs, including through modifications, e.g., better managing Arabic texts.

1.2 Motivation

To provide services through ELISQ for the Arabic language, software must be reorganized, redesigned, or improved to satisfy the needs of different stakeholder groups. For example, computational methods for automatic news classification and summarization exist. However, for Arabic texts, these methodologies have drawbacks in the efficiency and quality of their outputs. To overcome these drawbacks, we developed methods to enhance Arabic text classification, and others to extend Arabic text summarization.

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

a. Arabic language complexities and ambiguities;
b. The limited research work on these two domains compared to work with the English language;
c. The limited support from available tools and software packages for Arabic; and
d. The fact that most of the software packages used to process English texts, like indexing, cleaning, and tokenizing, do not support Arabic, or need special handling and addition of specific features when applied to Arabic text.

The ELISQ project is building a digital library for a country that speaks Arabic. This digital library is expected to provide many information retrieval services, like searching and browsing. Hence, this library will require tailored natural language processing support to handle the Arabic language, which requires research in that domain.

1.3 Research Problems

For an overview of the research problems addressed, along with a breakdown of related solutions, see Figure 1-2.

- **Sub-Problem 1**: 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.
- **Sub-Problem 2**: There is no proven best method for classifying Arabic news stories according to a given taxonomy.
- **Sub-Problem 3**: There is no simple and proven best technique for Arabic word stemming that will enhance Arabic news classification based on a given taxonomy.
- **Sub-Problem 4**: There are no good natural language processing tools to extract key information from Arabic news articles, e.g., for Named Entity Recognition or Topic Identification.
- **Sub-Problem 5**: There is no good summarization framework for Arabic news articles that can satisfy a diverse user community through the use of fully automated methods. Moreover, there are no good automatically generated summaries for Arabic news articles. For users who lack such summaries,

determining whether an article is of interest without reading (or at least scanning) it is infeasible.

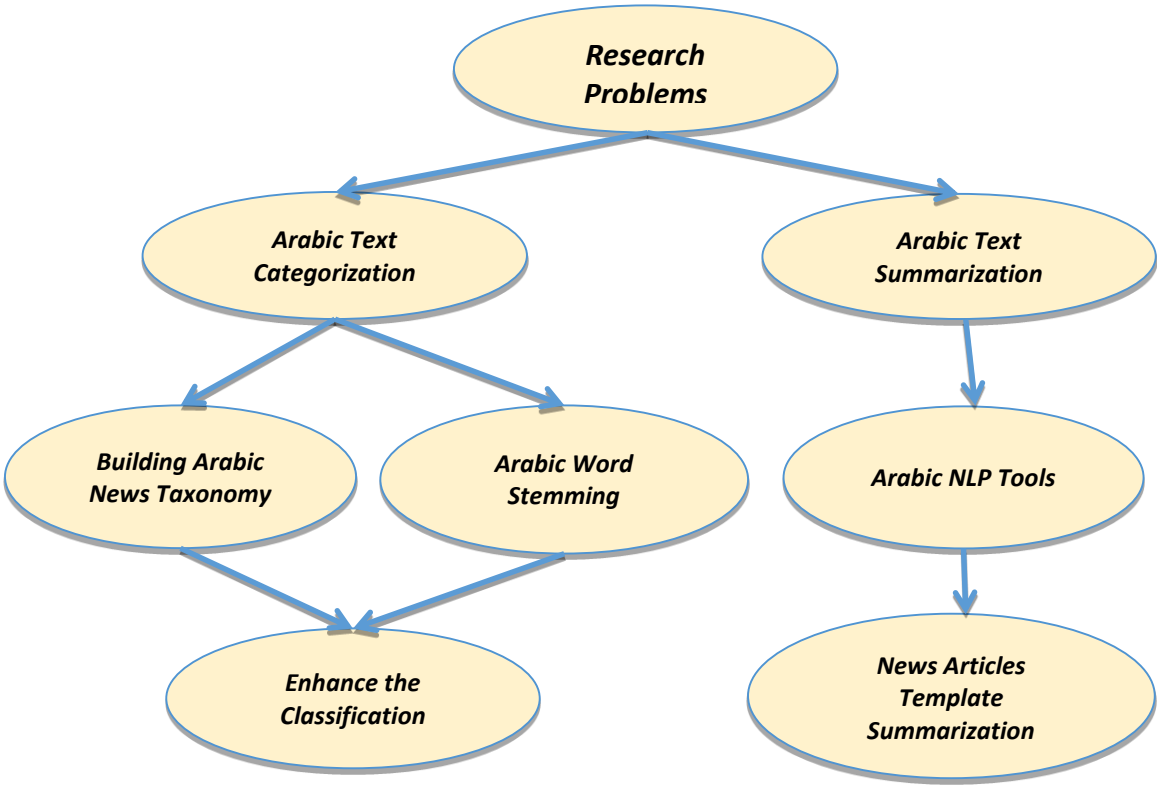

Figure 1-2: The Structure of our Research Problems and Techniques to Address those Problems

1.4 Hypotheses

Main Hypothesis 1:

The quality and accuracy of Arabic text classification using the proposed taxonomy and stemmer is better than with state-of-the-art approaches and systems.

- Hypothesis 1-1:

The proposed taxonomy is easy to use, works well with any Arabic newspaper, and is compatible with the International Press Telecommunications Council (IPTC) system.

- Hypothesis 1-2:

The proposed combined stemming and text classification method is more effective than state-of-the-art pairs of stemmers and text classification methods.

Main Hypothesis 2:

The proposed automatic Arabic text summarization approach, applied to news articles, will give accurate summaries that are relevant to the news articles.

- Hypothesis 2-1:

The proposed summarization approach will produce high quality Arabic news article summaries, by using text extraction methods to fill in a developed template, evaluated through human assessment.

1.5 Research Questions

The overall research question for this study is:

- How and to what extent can we classify and summarize Arabic language text resources into Arabic text article categories and Arabic readable summaries without direct human interaction while achieving high quality results?

From the above question, we can derive more specific research questions:

- How can we create a simple, but general, classification taxonomy for Arabic news articles?
- What is the most effective approach for classifying Arabic news articles that leads to high quality labeling?
- How can stemming enhance Arabic text classification?
- How can we apply natural language processing methods to create good summaries of Arabic news articles?
- Are the produced summaries as good as human-produced summaries?

1.6 Overview of Research

The research discussed in this dissertation has two main components, summarized in the next two subsections, and described in subsequent chapters.

1.6.1 Text Classification

- Text Classification for Arabic news articles using a generated taxonomy and applying a newly proposed stemmer called P-Stemmer:

Automatic English text classification has been studied for decades (Lewis, 1991). Some of that work addresses news articles and, recently, online news articles (Fernández & Fernández, 2004) and (Li, 2013). However, to date, there has been limited research with Arabic online newspapers. Such research must address many different factors, e.g., the different representations of features that are based on words in the data set.

Stemming is often used to prepare a feature set for a classifier, since it yields a common representation for multiple related word forms. Stemming thus may enhance classification results. Accordingly, our research considers approaches to Arabic stemming. That is an underpinning to research on taxonomies and classifiers.

1.6.2 Text Summarization

- Text Summarization for Arabic news articles after applying methods to extract the key information from the articles using different NLP tools, and filling in templates:

Automatic summarization aims to create brief overviews of longer texts, while keeping original ideas. There has been little research on automatic summarization of online Arabic news articles. Existing methods, though fast, tend to have deficiencies as to coverage, accuracy, and linguistic quality. Assessment of the various approaches is also a challenge, calling for judgments from multiple domain experts.

1.7 Dissertation Document Structure

The structure of this dissertation is as follows:
The front-matter includes the dissertation title, names of committee members, abstract, acknowledgments, table of contents, and lists of figures and tables.

Chapter 1 describes the research hypotheses, problems, and questions. The background, motivation, and overview of the research also are provided.

Chapter 2, related to Hypothesis 1, is a version of the journal paper that was submitted right after the preliminary exam. That submission has been accepted, conditional on minor changes, by the Journal of the Association for Information Science and Technology (JASIST).

Chapter 3, related to Hypothesis 2, paves the way for the research described in Chapter 4 and Chapter 5. It describes how undergraduate students taking a computational linguistics course in the fall of 2014 learned a variety of methods for text summarization of large English webpage collections; those methods were subsequently adapted to Arabic news articles. This chapter is a version of a paper presented at the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2015).

Chapter 4 covers two important parts of the research accomplished after the preliminary exam: Arabic Named Entity Recognition and Arabic Latent Dirichlet Allocation. This chapter is a version of an accepted paper in the proceedings of the ICSIC 2016 conference in Jordan.

Chapter 5, related to Hypothesis 2, describes the last part of this research. It addresses Arabic text summarization using filled in templates.

Chapter 6 summarizes the conclusions for all parts of this research, and describes future plans.

An important aspect of this research is its focus on experimentation. Accordingly, there is discussion of the development of baseline or test corpora, which may be of use for other researchers too. There also is description of evaluations of the tools and methods. Recall especially Chapter 2 and Chapter 3.

Appendices A&B include the Institutional Review Board (IRB) documents for the evaluation studies. Appendix C extends the discussion in Chapter 2 and includes full details of the significance test and of the modified standardized taxonomy.

Chapter 2: Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy

Abstract

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 to automatically classify the articles, limited research has been completed to yield suitable solutions for Arabic news. In connection with a Qatar National Research Fund (QNRF) funded project to build digital library community and infrastructure in Qatar, we developed a taxonomy (categorization structure) for browsing a collection of about 237K Arabic news articles, which should be applicable to other Arabic news collections as well. We designed this simple taxonomy for Arabic news stories capable of fulfilling the needs of Qatar and other nations, compatible with the subject codes of the International Press Telecommunications Council (IPTC), and enhanced with the expert aid of a librarian, as well as five Arabic-speaking volunteers. We developed tailored stemming (i.e., a new Arabic light stemmer called P-Stemmer) and automatic classification methods (the best being binary SVM classifiers) to work with the taxonomy. Using evaluation techniques commonly applied by the information retrieval community, including 10-fold cross-validation and the Wilcoxon signed-rank test, it was established 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, Arabic, IPTC, Machine Learning, SVM.

2.1 Introduction

2.1.1 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. The existing methods for automatic news classification have

drawbacks in quality, especially for Arabic. 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:

- Arabic has characteristics that make it difficult to work with in terms of Information Retrieval and Natural Language Processing. This includes a complex morphology, and a high level of ambiguity.
- Existing methods for processing English do not work well with Arabic; many modifications are required to handle Arabic.

2.1.2 Problem statement

Online Arabic news article categorization is not of high quality. Thus, when articles are accessed within a heterogeneous collection rather than within a specific newspaper site, it is hard to browse them by category. Further, there is no simple and accepted taxonomy for Arabic news. International taxonomies are too complex because they have many categories. Thus, the International Press Telecommunication Council (IPTC) has around 1,400 categories and subcategories in only one of their NewsCodes (i.e., SubjectCode). 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 automatic text classification method for classifying Arabic news stories. Further, though word stemming for English generally enhances text classification, there is no simple and proven technique for Arabic word stemming that has been shown to enhance Arabic news classification based on a given taxonomy.

2.1.3 The Arabic Language

Arabic is a widely used global language that has major differences from most popular languages, (e.g., English, Hindi, 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.

Most 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 most common spoken language in the world, with around 4.5% of the world population using it as their primary language, as shown in Table 2-1. 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 as subscript to add different grammatical formulation and sometimes meaning to that letter as well as the whole word. These diacritics are commonly used in the formal written version of Arabic known as Modern Standard Arabic.

Table 2-1. The Top 10 Spoken Languages in the World and Corresponding Percentage

Language	**Millions**	**Percentage**
Mandarin	955	14.40%
Spanish	470	6.15%
English	360	5.43%
Hindi	310	4.70%
Arabic	295	4.43%
Portuguese	215	3.27%
Bengali	205	3.11%
Russian	155	2.33%
Japanese	125	1.90%
Punjabi	102	1.44%

2.2 Related Work

2.2.1 Building Categorization Systems

2.2.1.1 Building Taxonomies for News

(Li, 2013) explains how news organizations, such as the New York Times and the Associated Press, build robust taxonomies that their computers use to automatically tag news content. (Woehler & Faerber, 2007) discusses 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, how it is used, 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; see also Appendix C. Our taxonomy will cover Arabic news articles, so we have built an Arabic taxonomy with terms chosen to help classify news-oriented Arabic textual data.

2.2.1.2 Taxonomies and Evaluation

Evaluating a taxonomy is still an emerging area of research. (Gómez-Pérez, 2004) discusses ways to evaluate ontologies, while 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 Arabic news in Qatar, 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.

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

2.2.2.1 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 & Kanaan, 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).

(Larkey, Ballesteros & Connell, 2007) present the effect of Arabic light stemming on the efficacy of information retrieval (IR). 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 compare the effect of their stemmer on Arabic IR relative to (Larkey, Ballesteros & Connell, 2007) and (Khoja & Garside, 1999). (Al-Omari & Abu Ata, 2014) develop an Arabic light stemmer that is not based on Arabic root patterns. Instead, they use well-defined mathematical rules and some relations between letters to extract the stem. Tested on around 6225 Arabic words, their stemmer shows good results, with around 5733 correct results and around 92% accuracy.

Arabic dialects have been used for many years. Dialects present more challenges than Modern Standard Arabic in the field of natural language processing since they add a new set of variational dimensions (Abu Ata & Al-Omari, 2014). This paper discusses a new rule-based stemming algorithm that can find stems for the Arabian Gulf Dialect. They prove that their algorithm performs poorly when applied to Modern Standard Arabic but performs 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 have studied Arabic stemmers with the goal of building a better understanding of stemming and with the goal of creating an improved Arabic stemmer. Our proposed stemmer is compared 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.

2.2.2.2 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, and Evens (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%.

2.2.2.3 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 compare automatic text classification using kNN, Rocchio, and Naïve Bayes on the Arabic language. The study concludes that Naïve Bayes is the best performer, followed by kNN and Rocchio. (El-Haless, 2006) used Maximum Entropy to aid classification of Arabic data sets. Results reveal that the average F1-measure increases from 68.13% to 80.41% using such pre-processing techniques. (Saad, 2011) compares the impact of text preprocessing on Arabic text classification using popular text classification algorithms. (Saad, 2011) applies different term weighting schemes and Arabic morphological analysis. The study attempts 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) proposes 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 suggested by some of the above studies, preprocessing steps are applied to our data set to enhance classification of Arabic text. Our study proves that using stemming as a preprocessing step enhances classification results, and that our stemmer outperforms some of the most popular Arabic stemmers. In this study, three kinds of classifiers are used along with both binary and multiclass classification. Comparison among classifiers themselves is followed by another comparison among different categories used.

2.2.2.4 IR Evaluation

(Kanaan, Al-Shalabi, Al-Zamil, and Saifan, 2004) use 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, the F1-measure calculated from recall and precision is used in our study.

2.3 Building a Standardized Categorization System for Arabic Newspapers

2.3.1 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-Rayah Newspaper, 2014) has both online and paper versions. From Figure 2-1, we observe a reasonable number of well-defined categories (Society, Locals, Politics, …) and subcategories (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.

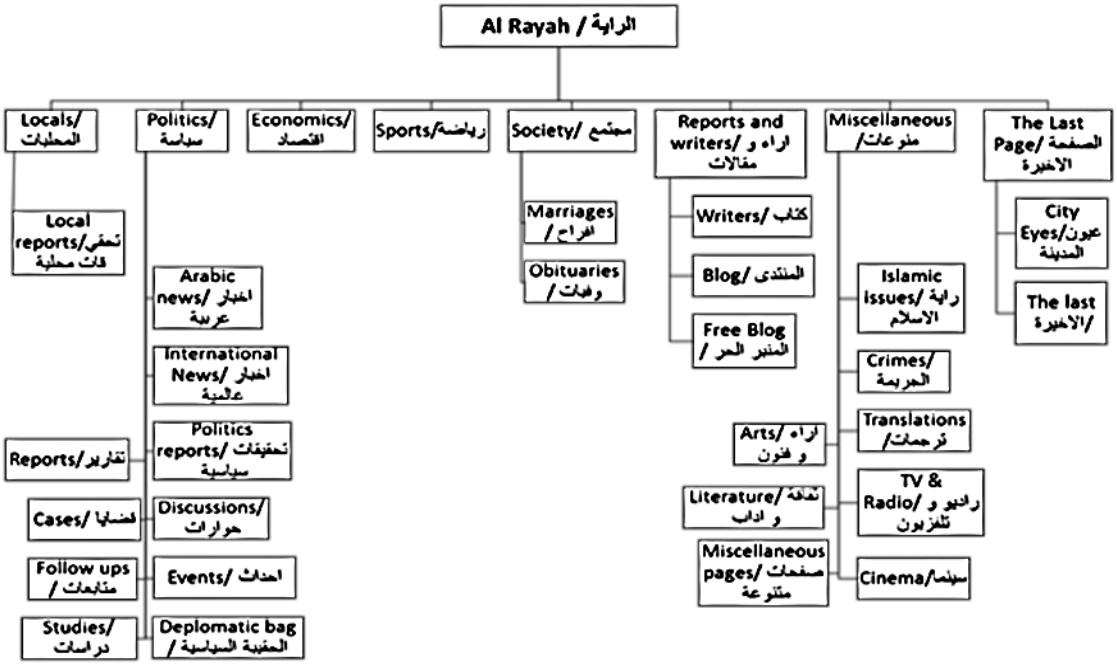

Figure 2-1. The Taxonomy for Al-Rayah Newspaper

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 Figure 2-2.

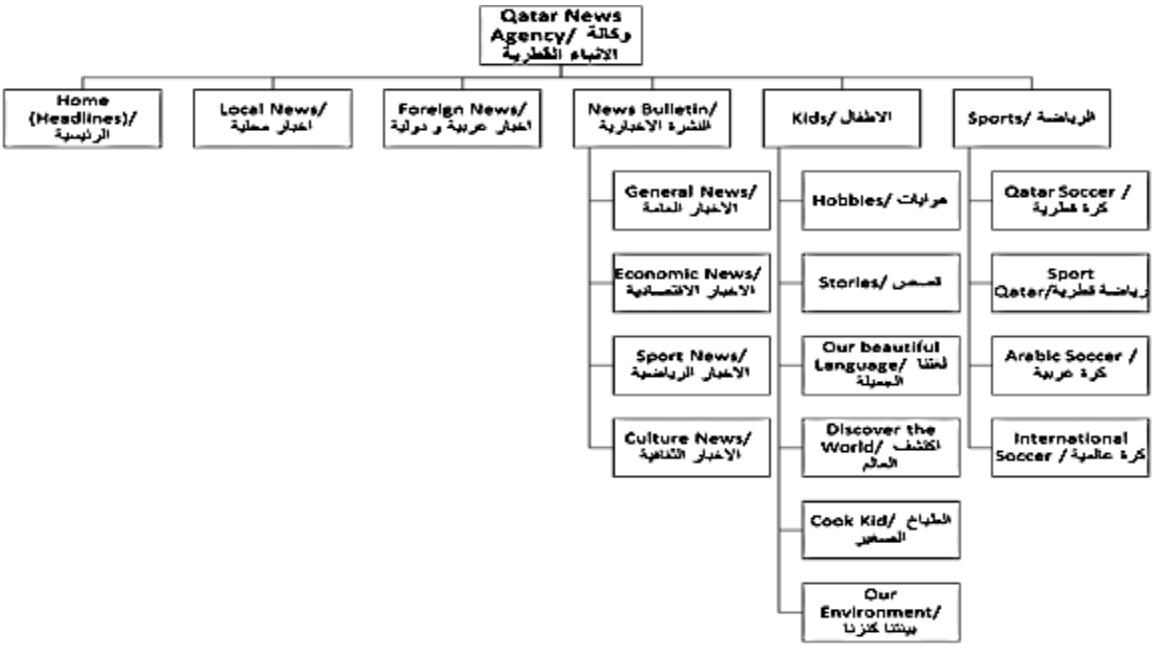

Figure 2-2. The Taxonomy for Qatar News Agency

Al-Watan: This newspaper (Al-Watan, Al-Watan Newspaper, 2014) has both online and paper versions. As illustrated in Figure 2-3, 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".

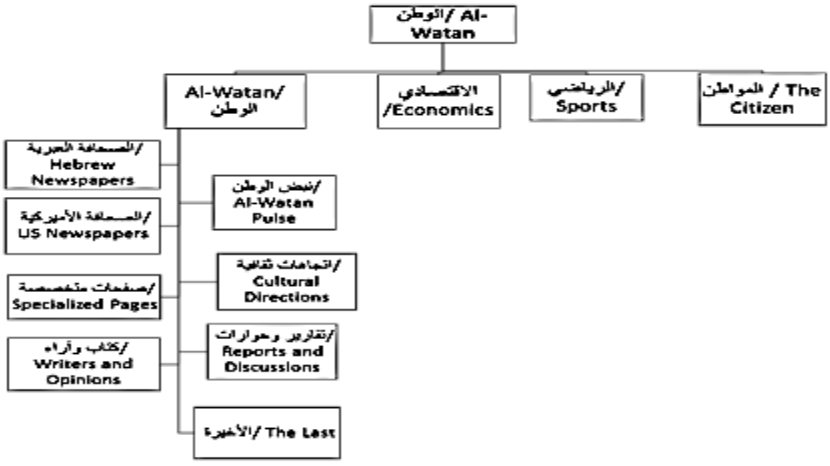

Figure 2-3. The Taxonomy for Al-Watan Newspaper

Al-Arab: This newspaper (Al-Arab, Al-Arab Newspaper, 2014) has both online and paper versions, as shown in Figure 2-4. We see in their taxonomy many categories (Sports, Arts, Economics, International, Qatar, …) and subcategories (Local, Arabic, Accidents, …).

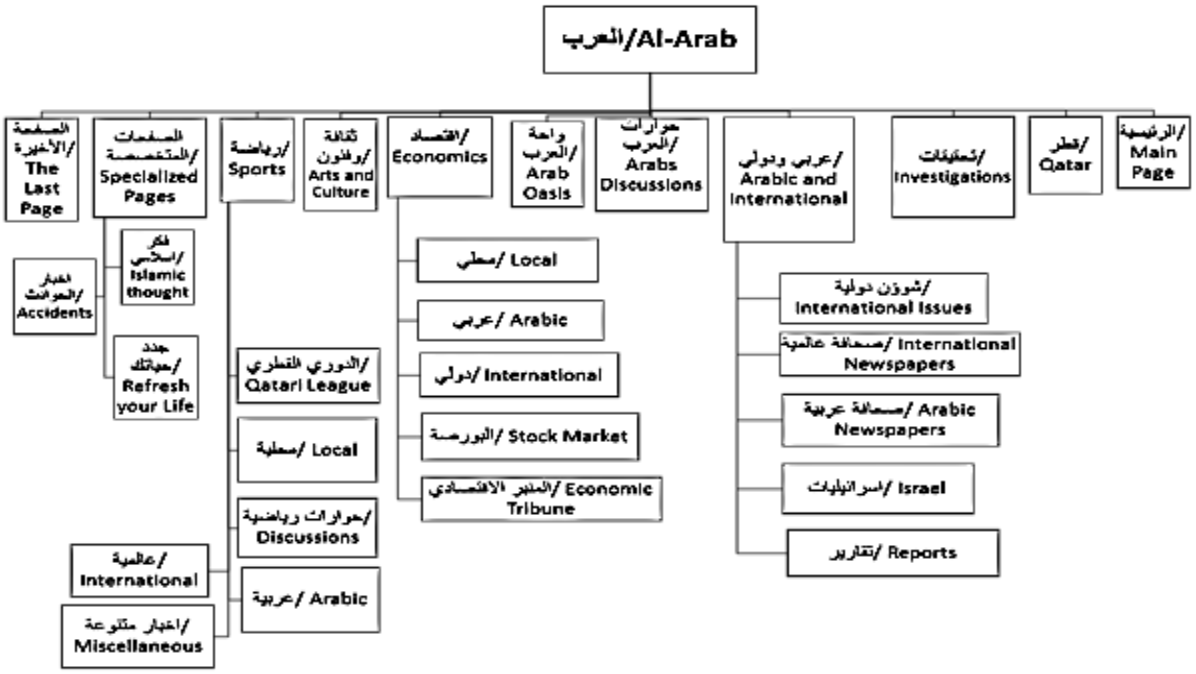

Figure 2-4. The Taxonomy for Al-Arab Newspaper

Al-Sharq: This newspaper (Al-Sharq, Al-Sharq Nespaper, 2014) has both online and paper versions, as shown in Figure 2-5. Based on their hierarchy, we see that their taxonomy has a reasonable number of categories (Sports, News, Economics, Accidents, …) and subcategories (Reports, Pictures, Investigations, …) that cover all of the important news articles.

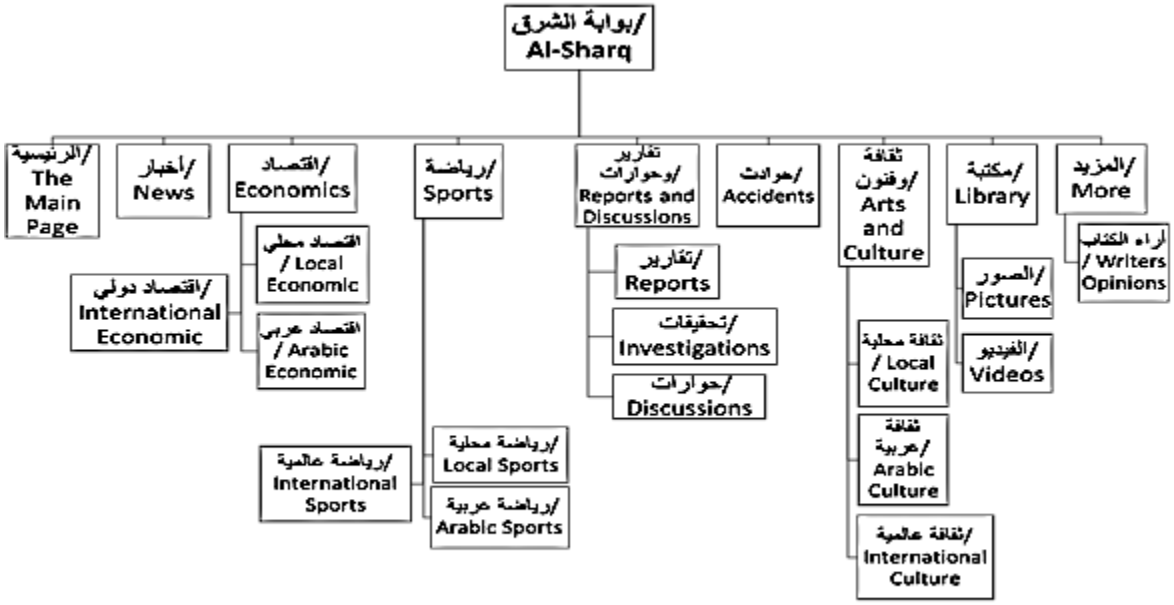

Figure 2-5. The Taxonomy for Al-Sharq Newspaper

2.3.2 Building the General Categorization System

With the aim of enhancing Arabic news article classification and improving online newspaper browsing, we created a categorization system called the General Categorization System or General Taxonomy. See also in Appendix C.

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/). 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. Thus, in order to ensure that our taxonomy is truly general, we have studied

many Arabic newspapers, including five in depth, to find the common categories between them and find the general categories that should be included in a newspaper taxonomy 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 subcategory. We call the result the "General Categorization System" or "General Taxonomy" (see Figure 2-6). 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. See also in Appendix C.

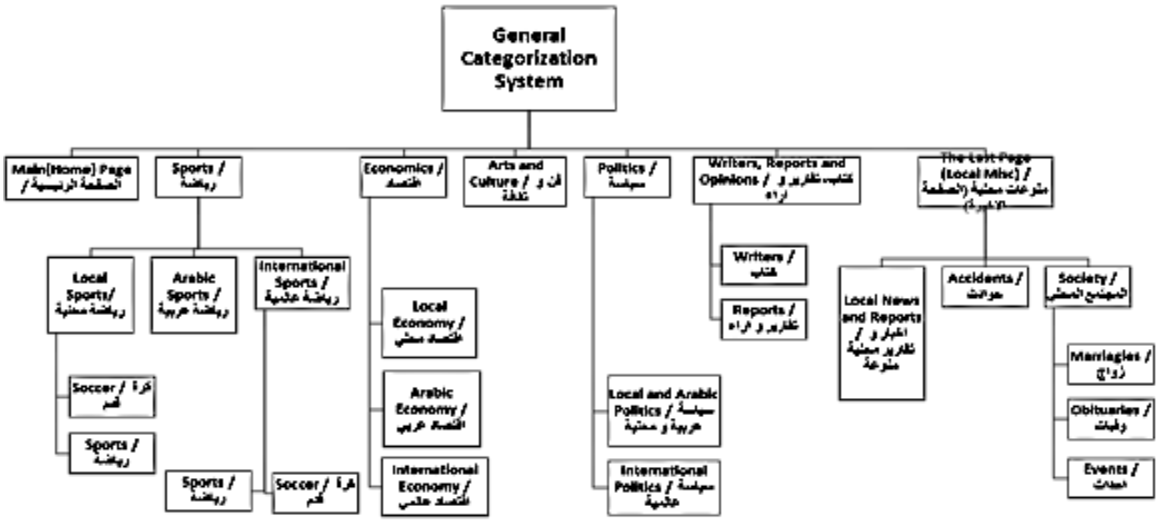

Figure 2-6. The General Categorization System (Taxonomy)

2.3.3 The IPTC System: International Press Telecommunications Council

Founded in 1965, IPTC (International Press Telecommunications Council) is a comprehensive worldwide standards body for news media, with its main office in London. Their mission is to make information dissemination easier and standards-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 studied the Subject Code part of IPTC Descriptive NewsCodes, because it is the main IPTC taxonomy, focused on text, and addresses subjects of newspaper articles. (IPTC, Interactive Diagram for the Subject NewsCodes, 2014) illustrates the process that we used in the creation and assessment of our new taxonomy. Please see Figure 2-7 for examples, noting that "Subjects NewsCodes", "Politics", and "Politics (general)" categories show which node in the tree is expanded further during interactive browsing.

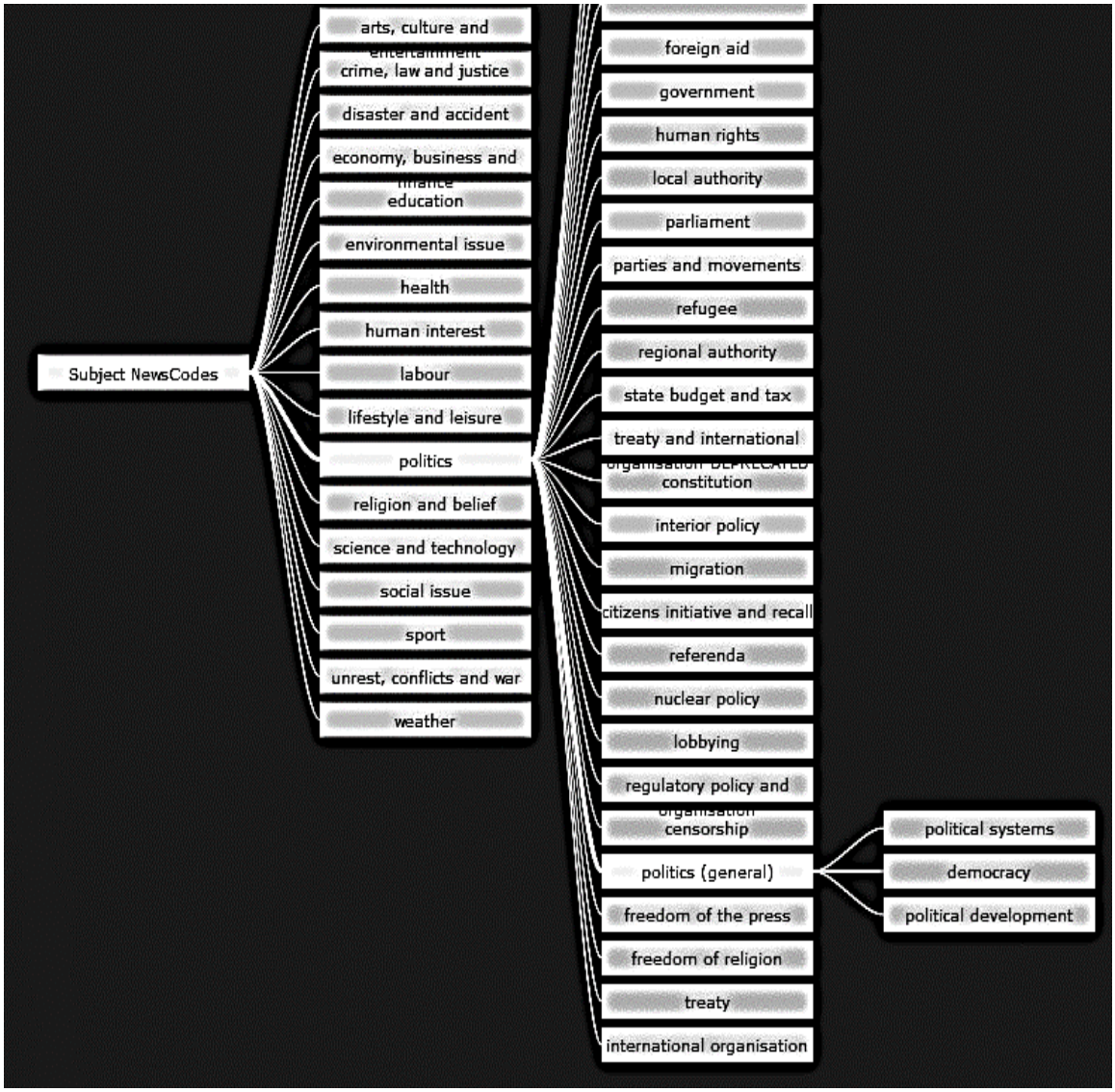

Figure 2-7. Example of Subject NewsCodes in the IPTC Taxonomy for "Politics" Category

2.3.4 The Standardized Categorization System

2.3.4.1 Revising According to IPTC

We decided to make yet another revision in our General Taxonomy to ensure consistency with the IPTC system. The main point of this taxonomy is to assist and enhance Arabic newspaper browsing. See also discussion in Appendix C.

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 subcategories in the second level of the general system (international sports, Arabic sports, and local sports) plus their four subcategories in the third level, into only two subcategories (sports general and soccer). Therefore, 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 subcategories are grounded in the IPTC system. Accordingly, we can say that our system is compatible with an accepted international news categorization system. These modifications have been approved by the librarian domain expert. Subsequent evaluation of the accuracy of our classifiers, discussed later, addressed whether the modifications were correct or not.

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 Categorization System for Arabic Newspapers" as shown in Figure 2-8. The main aim of this standardized system is to provide a taxonomy for browsing Arabic news articles. Table 2-2 shows results of an experiment to prove that our taxonomy can cover most of the news article topics.

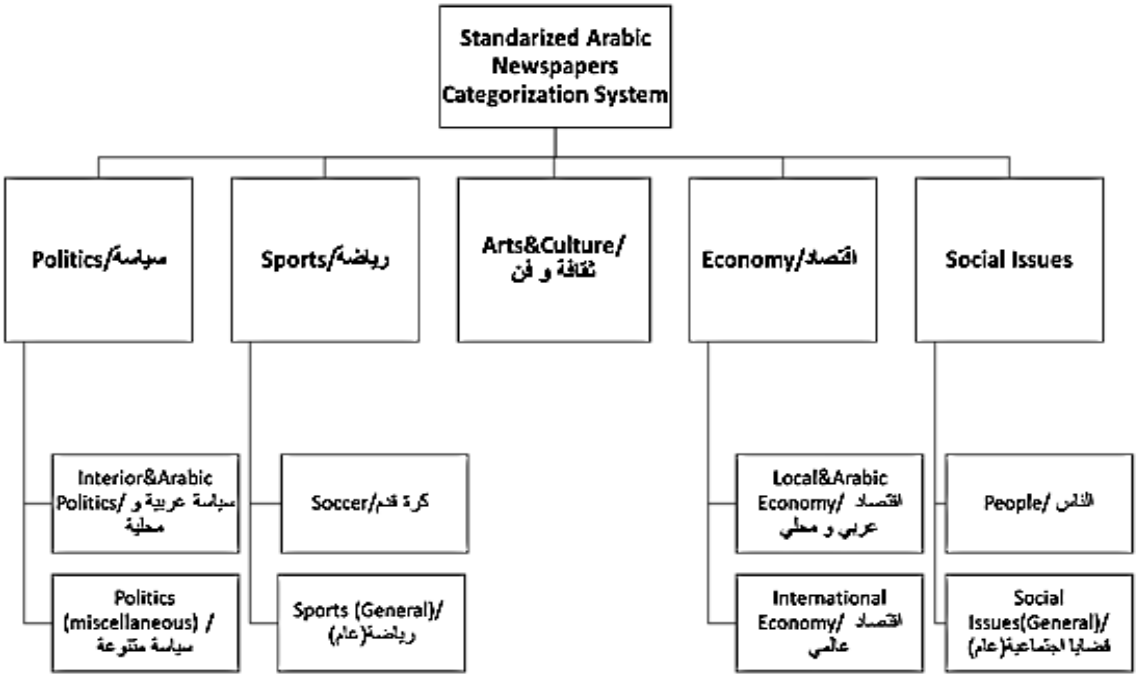

Figure 2-8. The Standardized Categorization System for Arabic Newspaper

2.3.4.2 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 by checking the appearance (mapping) of each category in our standardized taxonomy with the studied newspapers and cross-referenced it against the IPTC system by mapping it also with the IPTC taxonomy. 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).

The subsections below give a more detailed explanation of the taxonomy development research, covering in order: the data used, the problem formulation, applicable principles and constraints, the procedures and algorithms employed, and an overview discussion.

2.3.4.2.1 Data

We chose newspaper articles published in Qatar as a collection to construct, so we could demonstrate how to add value through advanced digital library services.

2.3.4.2.2 Problem

When collections of newspaper articles are made available online, as opposed to read on a daily basis as part of a newspaper, users face a number of challenges. One of these challenges also is faced with other collections, (e.g., being able to search based on a query). But many users also like to browse, or to combine browsing with searching, as with faceted search (Tunkelang, 2009). A familiar way to browse, employed by those who use libraries, is by way of a taxonomy / classification system, like the Dewey Decimal System (OCLC, 2015) or Library of Congress Subject Headings (Library of Congress Subject Headings, 2015), yet newspaper articles are rarely classified using either of those systems. Rather, the only categories used in the Arabic newspapers we studied are those that are given as names for the various sections of newspapers. Unfortunately, the names of those categories varied across the 13 newspapers we studied, which would complicate the work of those interested

in browsing through an integrated collection built as the union of the collections from the individual newspaper sources. Accordingly, we chose to address the problem of developing a standardized taxonomy made up of suitable categories to support faceted searching and browsing in the Qatari newspaper article collection.

2.3.4.2.3 Principles and Constraints

To guide the taxonomy developed, we identified a number of principles and constraints to follow:

1. Utilize concepts and best practices from computer science, library science, and information science (i.e., CS and LIS). Those include human-computer interaction and relevant areas of psychology, since taxonomies are usually implemented for online browsing using menus or information visualization schemes. Involving librarians was clearly appropriate, since working with taxonomies is a key part of many library efforts (Special Library Association, 2015). Thus, a domain expert from Virginia Tech library has helped with building the taxonomy.
2. Be consistent with relevant standards. Thus, we wanted our solution to be consistent with IPTC. Since that is a very large system with thousands of categories and subcategories, it means our solution should be a subset of the IPTC taxonomy. Since that system is hierarchical, our solution should be hierarchical as well.
3. According to Human Computer Interaction (HCI) and psychological studies related to the limitations of human short-term memory, have a branching ratio in the range seven plus or minus two, i.e., 5-9 (Miller, 1956). Thus, the taxonomy should have 5-9 children of the root, and each of those children should have no more than 9 children in turn.
4. According to LIS practices, utilize accepted schemes for classification where applicable. Thus, studying the category systems used in the 13 newspapers identified as relevant to Qatari needs was appropriate, and the problem could be viewed as extended taxonomy merging. So we first merged all of the studied taxonomies and then pruned the unnecessary categories.

5. According to CS and LIS practices, utilize best practices in closely related efforts. Since taxonomies are a constrained type of ontology (Tunkelang, 2009) and (Yang & Magdy, 2014), ontology merging practices are applicable, and we could draw on experience in doing that during the CTRnet project (Murthy, Fox, Ramakrishnan, Kavanaugh, Sheetz, Shoemaker, & Srinivasan, 2009). Likewise, since schemas and taxonomies are related, and schema mapping is an important process for which we have built and applied tools (Raghavan, 2005) and (Raghavan, Vemuri, Shen, Goncalves, Fan, & Fox, 2005), practices from schema mapping could be applied. So, after we merged the categories we tried to map between each category in each one of the studied taxonomies to find a common ground, following the schema and ontology mapping rules.
6. Apply principles from information retrieval and machine learning. Since our taxonomy would be connected with an information retrieval system (i.e., Solr) that supports facets (Tunkelang, 2009), and since it was not feasible in our project to have people assign categories to large numbers of newspaper articles, our taxonomy needed to be amenable to automatic classification techniques (Manning, Raghavan, & Schutze, 2008) and (Srinivasan & Angara, 2014). Thus, the taxonomic tree should be relatively balanced, with roughly equal numbers of articles assigned to each of the top level categories, and each lower level split also should be similarly balanced. Our results with text classification, discussed later in this chapter, should help with assessment of the quality of the taxonomy. Good classification results using our developed taxonomy should provide further confirmation regarding the quality of that taxonomy.

2.3.4.2.4 Algorithm/Procedure

To develop the needed taxonomy, we carried out the following steps, thereby applying the above-mentioned principles, best practices, and constraints.

1. Merged the taxonomies found in the 13 newspapers studied.
    a. When a category appeared in many of the taxonomies, and was largely disjoint from categories already identified, add it.

b. When similar words were used in different taxonomies for the same concept/content, pick the most frequently found, or else the most general (broadest, to be sure to cover the others).
c. Repeat the above, so all of the elements of all of the taxonomies were considered, and so the result had no more than 9 categories at a level (see #3 above in principles and constraints section).

2. Renamed or merged the result so it was a subset of IPTC, using the steps given above as needed.
3. Added a category called "Miscellaneous," as guided by the doctoral Advisory Committee, to cover articles not fitting into the existing categories, as is common in classification and clustering practice (Manning, Raghavan, & Schutze, 2008).
4. Improved and finalized the emerging solution with a librarian expert, to ensure best practices were followed, and with a representative group of Arabic speakers from the region, to ensure the result would be usable by the target community.

2.3.4.2.5 Discussion

One of the things we noticed when we built and refined our taxonomy is that building a taxonomy is a language and region dependent matter. For example, if we are trying to build a taxonomy for US newspapers, we can find that under the sports category the taxonomy will have 3-4 major subcategories (i.e., Basketball, Baseball, Football) while that was not the case with our taxonomy since sports only has two subcategories (Soccer and Not Soccer). Thus, building a taxonomy for newspapers depends in part on the culture of the area and the language. Another example is the Science and Technology category. If we study any newspapers from the west coast of the USA we will find that Science and Technology articles appear frequently; thus this category should be a part of the news taxonomy for that region. In our taxonomy we did not have this category since there were few science and technology articles in the newspapers studied. This provides further support for the observation that building taxonomies depends on the language, region, and culture. Thus, we found multiple categories that only had small numbers of articles, and decided to modify our standardized taxonomy to cover those scattered articles in a category called Miscellaneous; this category will include all of the articles that did not fit very well

in any of the 13 categories in our taxonomy. For more details about the modified taxonomy, please see Appendix C.

To confirm that we were on the right track with our taxonomy, we conducted a study with 10 Arabic native speakers, asking them to manually assign a category to each of articles we assigned to them. We randomly selected 1,000 articles from our dataset and gave each of our participants the same number of articles. Table 2-2 shows the frequencies of articles for each of the 6 categories in our taxonomy (including the miscellaneous category). From this table we see that our taxonomy fits the data from the studied region. This appears likely as well for the Arabic region, since the frequency of news articles that do not fit into any of the 5 main categories, and so go to the miscellaneous category, is very low (0.018). Further evidence regarding the quality of our taxonomy is provided by the classification results that appear later in this chapter. In other words, since we used our developed taxonomy in the automatic text classification, then good classification results should confirm that the taxonomy is useful and applicable.

Table 2-2. Frequencies per Category for 1,000 Randomly Selected Articles

Category	Frequency	Percent
Sports	232	23.2%
Art&Culture	165	16.5%
Politics	255	25.5%
Economics	173	17.3%
Social Issues	157	15.7%
Miscellaneous	18	1.8%

2.4 Arabic News Articles Text Classification with Stemming

2.4.1 Our Data Set: Al-Rayah Newspaper Collection

Al-Rayah is a Qatari newspaper published in Arabic. This newspaper was used in our experiments. We employed the open-source Heritrix Crawler (Internet Archive, Heritrix Web Crawler, 2014) installed on one of the ELISQ servers. The size of our crawled collection 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 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 2-3.

Table 2-3. News Data Set

File Format	PDFs	TXTs
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 File	125	125
Total Number of Articles	260K	237K
Total Number of Sentences	900K	900K
Total Number of Words	2.35M	2.35M

2.4.2 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 & Connell, 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. For example, from the root "قصد", we can produce the words "يقصد", "مقاصد", "اقتصادية", "الاقتصادي", etc., as shown in Table 2-4. The word in the right is the root of the word on the left.

Table 2-4. Examples of Root Stemming

Word	Root
الاقتصادي The economic	قصد Intended
مقاصد Purposes	قصد Intended

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 meanings 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" (i.e., "Intended"). 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 difference in 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"), as shown in Table 2-5.

Table 2-5. Examples of Light Stemming

Word	Light Stemmer Result
اقتصادي	اقتصاد
الاقتصاد	اقتصاد

We implemented a tool to help stem Arabic words, which 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 Figure 2-9).

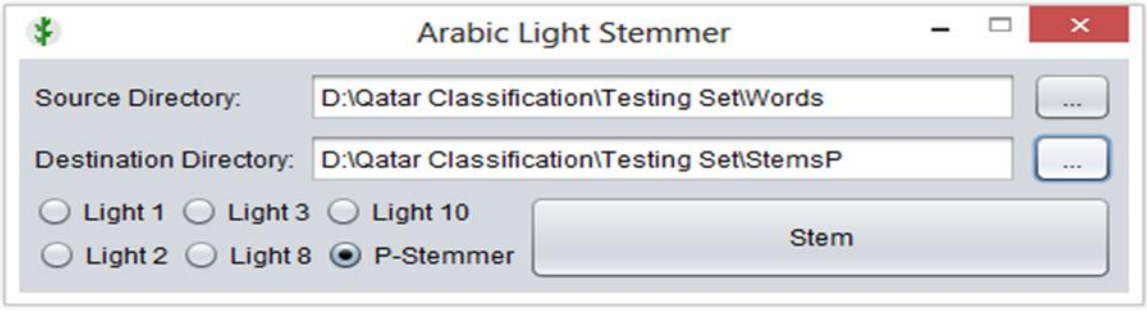

Figure 2-9. A Screenshot of the "Arabic Light Stemmer" Tool

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 2-6. 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 2-6. Five Larkey Versions of Arabic Light Stemmer, and the P-Stemmer.

Version	Prefixes to remove	Suffixes to remove
Light 1	"ال", "وال", "بال", "كال", "فال"	None
Light 2	"و", "ال", "وال", "بال", "كال", "فال"	None
Light 3	"و", "ال", "وال", "بال", "كال", "فال"	"ة", "ه"
Light 8	"و", "ال", "وال", "بال", "كال", "فال"	"ها", "ان", "ات", "ون", "ين", "ة", "ه", "ية", "يه", "ي"
Light 10	"و", "ال", "وال", "بال", "كال", "فال", "لل", "ل", "ول"	"ها", "ان", "ات", "ون", "ين", "ة", "ه", "ية", "يه", "ي"
P-Stemmer	"و", "ال", "وال", "بال", "كال", "فال", "لل", "ل", "ول"	None

2.4.3 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 as shown in Table 2-7.

Table 2-7. Light10 vs. P-Stemmer

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

To verify our claim, we have developed 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 2-5).

2.4.4 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 under one of the five classes at the first level of our generated taxonomy, see Figure 2-10. Figure 2-11 shows our Arabic Text Classification framework.

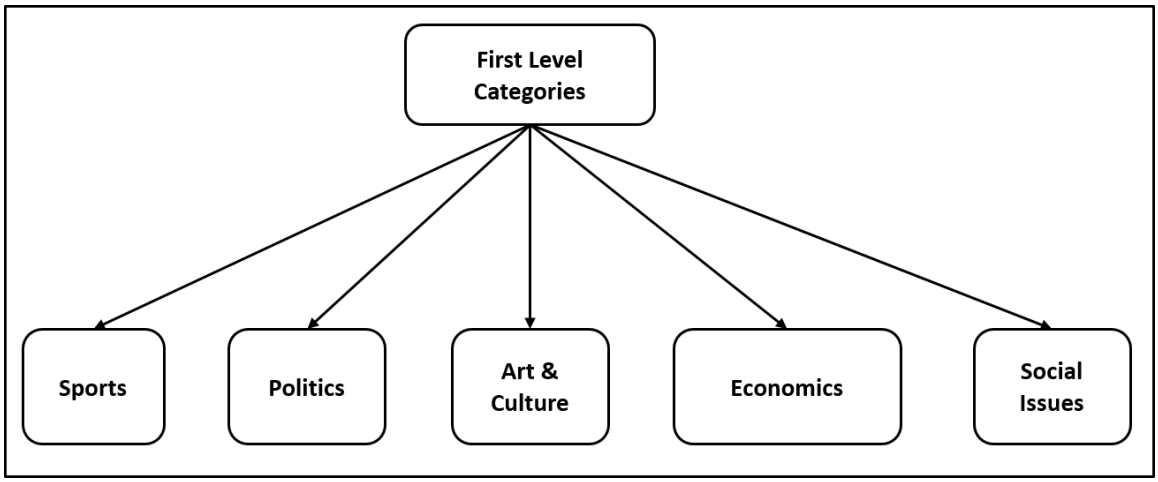

Figure 2-10. The Five Classes in the First Level of our Taxonomy

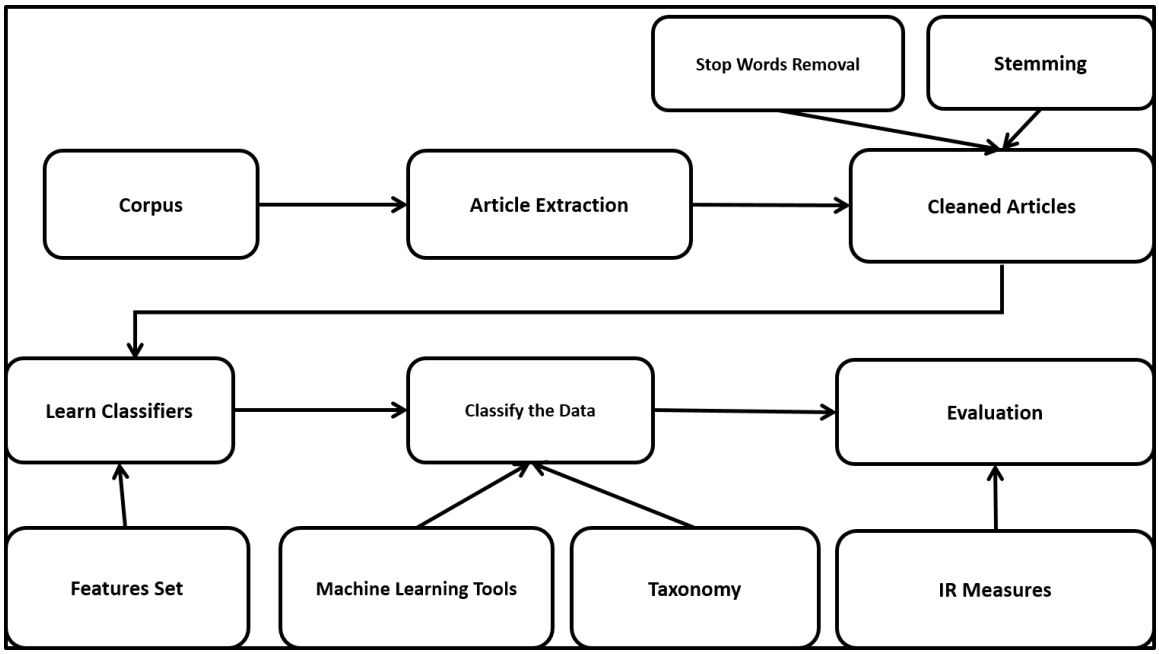

Figure 2-11. Arabic Text Classification Framework

In order to train and test different classifiers using different machine learning techniques, we use 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 of 750 tuples: (news-article-ID, class-label). Volunteers manually labeled the 750 documents in the training set 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 (Figure 2-9) is used to produce six different versions of the text files corresponding to the five versions of Larkey's light stemmers and 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 2-8. 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 2-8. 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	*P-Stemmer.

2.5 Results, Evaluation, and Discussion

2.5.1 Overview of Classification Experiments and Evaluation

In order to use Weka to classify our data, the training set must be converted to an Attribute-Relation File Format (ARFF) file. An ARFF 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 the existence of 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 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))*

2.5.2 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) – were 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 2-9 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 2-9. 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 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 2-10 shows the average recall, precision, and F1 measure values after running the three classifiers over our seven different data sets.

Table 2-10. 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	**Prec.**	**F1**	**Recall**	**Prec.**	**F1**	**Recall**	**Prec.**	**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

Figure 2-12 shows the F1-measure values for the three classification techniques with the seven word variations for the multiclass classification corresponding to Table 2-9.

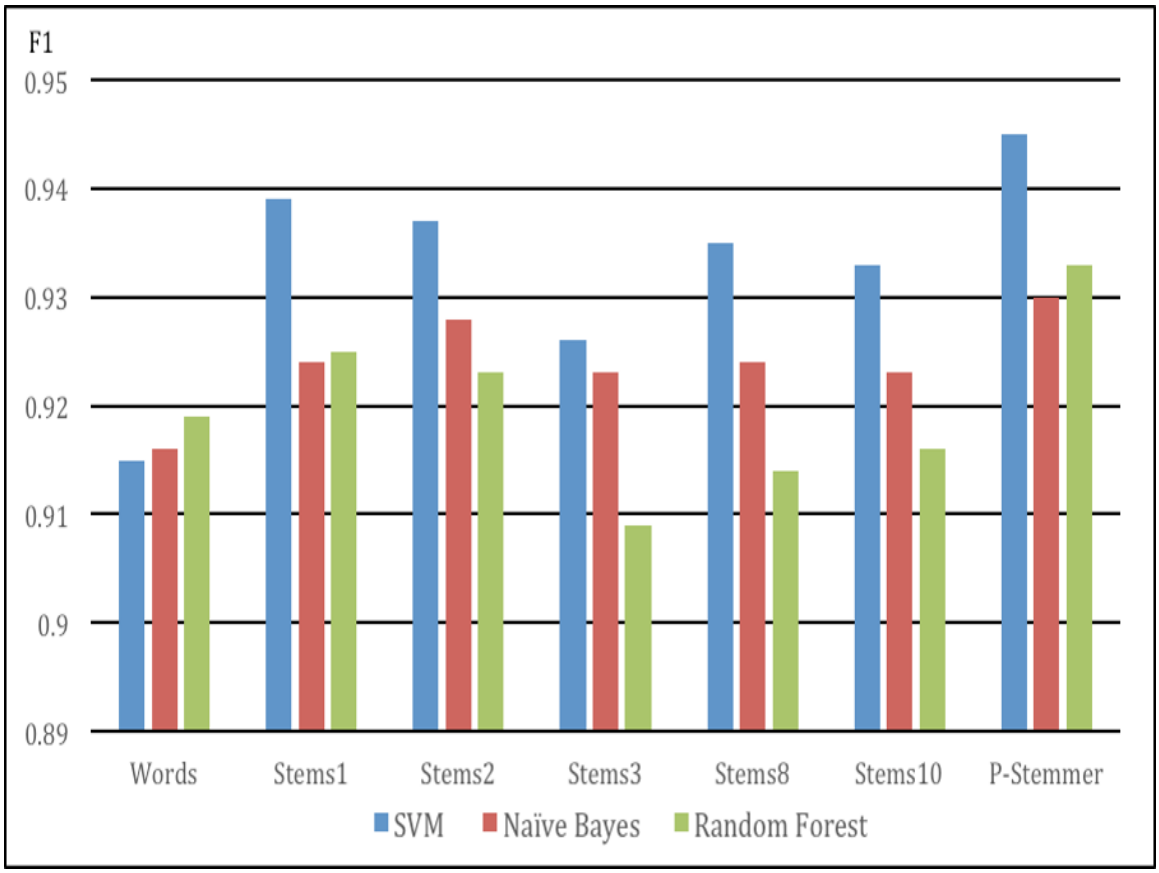

Figure 2-12. F1-Measure Values for the Three Classification Techniques with the Seven Word Variations for Multiclass Classification

2.5.3 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 2-11 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 2-11. 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 2-12. Figure 2-13 shows the F1-measure values for the three classification techniques with the five different categories, for Binary Classification.

Table 2-12. 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.

	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

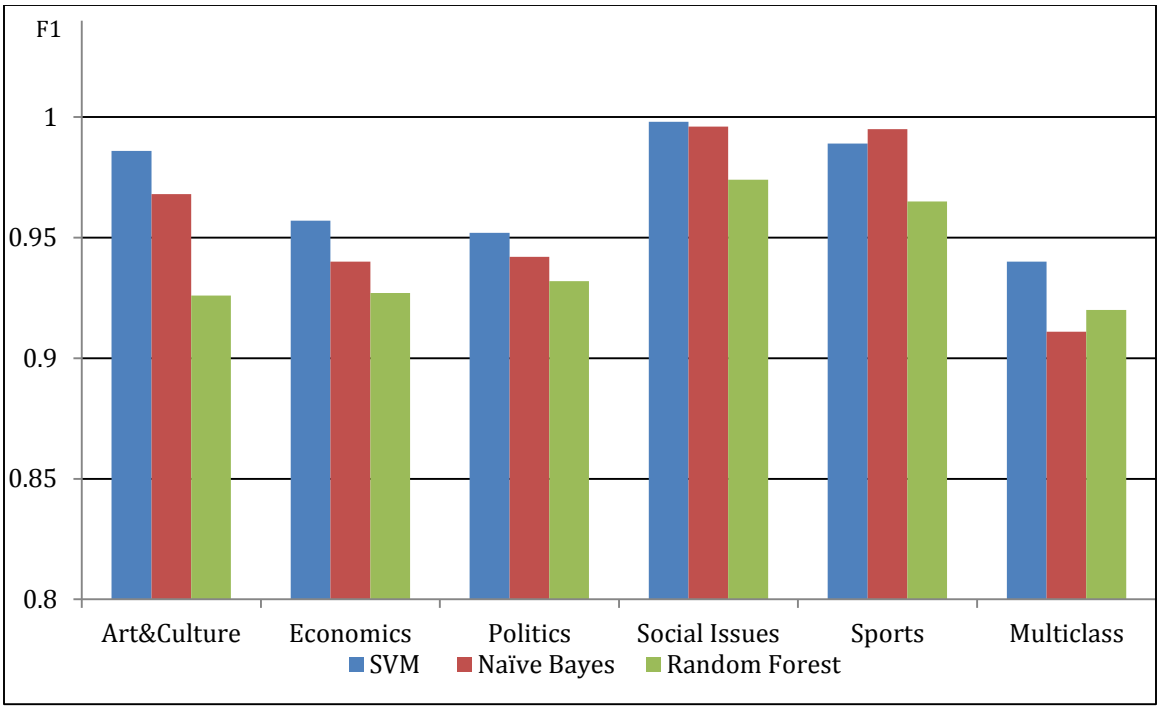

Figure 2-13. F2-Measure Values for the Three Classification Techniques with Respect to the Five Categories Training Sets, for Binary Classification vs. Multiclass

2.5.4 Discussion

In the multiclass classification experiments, we compared our P-Stemmer with six different word formulations as listed above in Table 2-8. We generated seven data sets and extracted seven feature sets with the intention of applying them to our classifiers. We used the SVM, NB, and RF classifiers to judge the performance of P-Stemmer for classification, and compared it with the other listed approaches. Out of the three classifiers, P-Stemmer gave better results than the full word and five Larkey 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 it can be concluded that using stemming enhances text classification. We calculated recall and precision for each of our datasets, as well as F1, see Table 2-10. Our P-Stemmer performed very well compared to the other stemmers and the full word option; see Figure 2-12.

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 2-11 and Figure 2-13 — 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.

2.5.5 Significance Test

We used the F1 measure results from the SVM classifier to do a statistical significance test between our P-Stemmer and each one of the five Larkey stemmers. Table 2-13 shows the F1 results using the SVM classifier for P-Stemmer and Stem1, Stem2, Stem3, Stem8, and Stem10.

Table 2-13. F1 Measure Results for the P-Stemmer and the Five Larkey 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 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 stemmers.

Figure 2-14 and Table 2-14 show a sample of the calculations and results, with the final Wilcoxon (W_{cal}) values; this sample is for the light Stem1 and the P-Stemmer test calculations and results. For more details on the significance test, please see Appendix C.

Let N be the number of pairs, sample size (5 in our case).
For i= 1, …, 5, let $X_{1,i}$ denote the 5 F1 measure values of one of the Larkey's stemmers (Stem1, Stem2, Stem3, Stem8, or Stem10) and $X_{2,i}$ denote the 5 F1 measure values of P-Stemmer.
Sgn is the sign of the value (+/-).
Abs is the absolute value | |.
R_i is the rank
Our Hypotheses are:
H_0: 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 zero
Vs.
H_1: The median difference of the F1 measure of P-Stemmer and each one of Larkey's stemmers (Stem1, Stem2, Stem3, Stem8, or Stem10) is greater than zero

Figure 2-14. Calculations and Formulas used with the Wilcoxon Signed-Ranked Test

Table 2-14. Values toward Calculating W_{cal} for the Wilcoxon Signed-Rank Test for the F1 Measure Ordered by Absolute Differences (abs) between P-Stemmer and Stem1

			X2,i - X1,i			
i	**X2,i**	**X1,i**	**sgn**	**abs**	**Ri**	**sgn * Ri**
4	0.991	0.990	1	0.001	1	1
3	0.915	0.913	1	0.002	2	2
1	0.918	0.915	1	0.003	3	3
5	0.964	0.960	1	0.004	4	4
2	0.935	0.919	1	0.016	5	5

2.6 Conclusion and Future Work

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

For 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 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 the five Larkey stemmers, as well as the original raw words. We found that using our proposed stemmer significantly enhanced classification results for Arabic textual data when three classifiers are used: 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 than 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: 1) Test our stemmer with another data set, to see if results match those with our data set; 2) Apply the stemmer and the classification methods on this new data set to confirm our findings; and, 3) Use different feature selection methods, like Chi-square, to see if they will enhance classification results.

Chapter 3: Big Data Text Summarization for Events: a Problem Based Learning Course

Abstract

Summaries can have particular value in digital libraries, especially to save time for those interested in searching and browsing, as well as to help with automatically filling in some of the fields in metadata records (which also might be called templates). Summarization and other NLP methods are difficult in Arabic, hence hard to learn. Problem/project Based Learning (PBL) is a highly effective student-centered teaching method, where student teams learn by solving problems. This chapter describes an instance of PBL applied to NLP education and some NLP techniques used for text summarization. Many of these techniques are discussed in subsequent chapters with regard to Arabic, so it is helpful to explain them in an English context first. Accordingly, we show that students learn how to do text summarization thorough studying and applying NLP methods. To provide context, we show the design, implementation, results, and partial evaluation of a Computational Linguistics course that provides students an opportunity to engage in active learning about adding value (through summaries) to digital libraries, and about NLP with large collections of text, i.e., one aspect of "big data". Students engage in PBL with the semester-long challenge of generating good English summaries of an event, given a large collection from our webpage archives. We later used most of the same NLP methods in our research with Arabic. Six teams, each working with a different type of event, and applying three different summarization methods, learned how to generate good summaries; these have reasonable precision relative to the Wikipedia page that describes their event. To get these results, this course focused on essential NLP methods (i.e., Named Entity Extraction and Topic generation using LDA) and important summarization techniques (i.e., template summaries). NER, LDA, and template summaries were essential in our later research, discussed in chapters 4 & 5.

Keywords: Problem based learning; Computational linguistics; summarization; Big data.

3.1 Introduction

As a student-centered teaching approach, Problem/project Based Learning (PBL) encourages students to learn by solving problems (Buck Institute for Education, 2015). To test our hypothesis that PBL is valuable as part of digital library education, we applied PBL to a new Computational Linguistics course in the fall semester of 2014, closely connected to a large digital library effort. We hope our findings will encourage other institutions to teach similar courses. The course goal is to produce a good summary of an event, given a text collection from the Integrated Digital Event Archiving and Library (IDEAL) project. Started in 2013, IDEAL aims to integrate archiving with digital library concepts (Fox, Akbar, Abdelhamid, Elsherbiny, Farag, Jin, Leidig & Neppali, 2014), along with appropriate technologies and applications (Fox & Leidig, 2014). A wide range of services is required to build a suitable information infrastructure, and to provide helpful methods of analysis, access, and visualization (Yang, Chung, Lin, Lee, Chen, Wood, Kavanaugh, Sheetz, Shoemaker, & Fox, 2013) for stakeholders (Fox & Leidig, 2014). Our contributions described in this chapter include: 1) a pedagogical approach that worked well in helping students learn about big data text summarization using computational linguistics methods, and 2) good results from applying these methods to corpora about events.

3.2 PBL Course Preparation

3.2.1 Computer Science Capstone Course

An undergraduate course in Computational Linguistics, focused on supporting digital libraries, was taught in the 2014 fall semester at the Department of Computer Science, Virginia Tech. It is a new course planned to be taught each fall. This course gives students the opportunity to engage in active learning about how to work with (i.e., process and analyze) large collections of text, one aspect of "big data".

Using methods employed in search engines, linguistic analysis, NLP, digital libraries, and statistical techniques, students are engaged in problem based learning (Buck Institute for Education, 2015), with the semester-long challenge of analyzing content collections automatically, extracting key information, and generating easily readable summaries of important events in English. Just-in-time learning will allow development of an

understanding of concepts, techniques, and toolkits, so students will master the key methods related to computational linguistics and digital libraries.

3.2.2 Course Learning Targets

The learning target is to summarize a big text collection regarding an event. Events are important to people/organizations, and are remembered based on time (when), location (where), person (who), and the subject or topic of interest (what). Given a set of text documents relevant to an event, students are asked to extract frequent/important words, topics, key sentences, and named entities, for summarization. To help students learn summarization techniques in a PBL approach, we designed a course structure with nine units. Ultimately, though, student teams are charged with discovering their own method for getting best results.

3.2.3 Dataset

We selected six types of disasters and one community event from the IDEAL archives. The community event differs from other types of events since it covers multiple topical areas. Though, for each category, both a small and large collection is analyzed, for our evaluation purposes, we choose to show the results and information for only six big collections, since the last collection (Community) has different characteristics than the other six. It does not deal with only one event like the six disaster event collections; also it is not about a disaster. Table 3-1 describes the events and the corresponding collection information.

Table 3-1. Characteristics of the Seven Corpora

Event collection type	Event location	Event date	Collection size	Event collection name
Disease Outbreak	World Wide	2014	15,000	Ebola
Earthquake	Virginia USA	2011	8,765	Virginia earthquake
Fire	Brazil	2013	690,281	Brazil club fire
Flood	Pakistan	2011	20,416	Pakistan flood
Hurricane	East Coast USA	2012	75,929	Hurricane Sandy
Shooting	Tucson, AZ USA	2011	37,829	Tucson shooting
Community	Blacksburg, VA USA	2011-2012	16,024	Blacksburg events

3.2.4 Computational Resources

A tailored Cloudera virtual machine (VM), and an 11-node Hadoop cluster (Apache Hadoop, 2015), along with other supporting computing resources (shown in Table 3-2), aid the handling of over 11 terabytes of webpages. After the course, based on student needs, the cluster was expanded to 20 nodes, thanks to student lab fees.

Table 3-2. Hadoop Cluster Specification

Nodes	11 cluster nodes + 1 manager node
CPU	Intel Xeon, Intel i5
RAM	208 GB = 2 * 32 + 9 * 16
HDD	51.3 TB =1* 12TB + 1* 6TB + 7* 3TB + 2* 2TB + 8.3TB NAS backup

3.3 Summarization Methods, Results, and Evaluation

3.3.1 Methods

The class has engaged 30 students, assembled into seven teams, in active learning, through its adoption of problem based learning (Buck Institute for Education, 2015). Students have one goal during the course, i.e., to create good summaries of a large corpus of webpages. They are obliged to do that in a way that would be applicable to other corpora about events of the same kind. To provide scaffolding that will aid their solving of this challenging problem, the course materials provide content related to a set of units. Table 3-3 shows the 9 units that served as scaffolding for the course.

Table 3-3. Summarization Goals

Unit	Desired Results
1	A set of frequent words
2	A set of WordNet synsets that cover all entries in the set of frequent words
3	A set of words constrained to POS tagging
4	A set of features, and classifiers to classify the documents in the collection
5	A set of N most frequent & important named entities
6	A set of the most important topics**,** by using LDA (Apache Mahout, 2015)
7	A set of indicative sentences, identified by clustering
8	A set of values for each of a template's slots
9	A generated English readable summary based on the filled-in template from Unit 8

The course started by introducing; a) PBL (Buck Institute for Education, 2015), b) course goals, c) units overview, d) supporting facilities, and e) applicable resources such as the Python NLTK (Bird, Klein & Loper, 2009). Below we introduce the nine units in more detail.

In Unit 1, students find the frequency of words (single words and collocations) and pick the most frequent words in their collection. They should discuss the advantages/disadvantages of the approach they use; this is required in all the units.

In Unit 2 students are asked to propose a list of indicative words for their event. Students improve their summary by calculating the word length distribution and automated Readability Index (ARI). After that, students learn how to extend their indicative words with WordNet synsets. Parallel processing methods (MapReduce and Hadoop) are introduced in this unit, and used in subsequent units as well.

Part-Of-Speech (POS) analysis helps with the selection of appropriate words. Unit 3 introduces the POS tagger of NLTK. Students are asked to extract nouns (and verbs if they feel them useful) that can best reflect the characteristics of their events.

Classification is used for eliminating noise and categorizing documents. In Unit 4, students label (i.e., as relevant or not) a small set of text files out of each collection, select appropriate lexical features, build training sets, and experiment with multiple classifiers (Naive Bayes, Decision Tree, Maximum Entropy, and Support Vector Machine) to categorize text files.

Named entities often carry important information about documents. In Unit 5, students learn to extract named entities by NER, e.g., the Stanford NER (Stanford Natural Language Processing Group, 2015) and the NLTK chunking utility. Then, they use high frequency named entities to improve their summaries.

Topic modeling can extract the main topics from a text through the Latent Dirichlet Allocation algorithm, which helps users identify the key themes appearing in corpora. In

Unit 6, students learn how to extract topics from their collections with two LDA tools: the Gensim Python, and Mahout LDA for Hadoop (Apache Mahout, 2015).

Unit 7 focuses on clustering, which helps group similar content. The instances to be clustered may include sentences, paragraphs, or documents. Students learn how to cluster content units and select the best representative instances with Mahout K-Means (Apache Mahout, 2015), such as, the best sentence in each cluster.

Unit 8 requires students to work with templates carefully crafted for each type of event. The students need to design grammars or other methods for finding values appropriate for each template slot. Then, they use tools such as regular expressions to extract candidate values for the slots; the candidate values are sorted by frequency. Finally, to fill a slot, the best of the candidates is selected.

Unit 9 requires students to explore automatic methods for English text generation based on their filled-in templates. They are to devise automatic methods for ensuring broad coverage, cohesion, and coherence in the generated text.

3.3.2 Sample Results

3.3.2.1 The Community Collection

Since the Blacksburg community collection covers multiple small events and news, rather than focusing on a single event as in the other collections, summarization was difficult. Clustering the content was found to yield the best results. First, contents were filtered based on the frequent words and named entities. Next, the collection was clustered to produce four main clusters; for each of those, the sentence closest to the centroid was selected as the cluster summary. Thus, Table 3-4 shows three cluster names along with their summary sentence.

Table 3-4. The Representative Sentence for Three Clusters

Cluster (Event)	**A Sample Sentence**
Police-Crime	Michael Edwards and Johnny Worrell were arrested on scene and charged with manufacturing less than 227 grams of mixture containing a detectable amount of meth conspiracy to possess meth and possession of precursors to manufacture meth.
Weather	The National Weather Service in Blacksburg placed most of Southwest Virginia under a winter storm warning today and tonight with snow accumulations of 4 to 8 inches.
Virginia Tech	At the Aspirations in Computing ceremony the honorees heard from two speakers: Letitia Long director of the National Geospatial-Intelligence Agency and a 1982 graduate of Virginia Techs electrical engineering and Diane Reineke vice president of business development.
Local Festival	The Christiansburg High School music department will be hosting Night on Broadway March 30 and 31 in the Christiansburg High School auditorium.

3.3.2.2 Disaster Collections

Due to the limitation of space, we show results for the other six collections, for only three methods, i.e., those used in Units 5, 6, and 9. Table 3-5 gives sample student results. As results show, for their collection, the students were able to extract important and representative named entities and topics, and to generate a meaningful English text summary. However, the student team working with data about Hurricane Sandy had trouble with the NER processing, hence the "N/A" shown in the table.

Table 3-5. Ex ample Results for each Collection and for 3 Main Methods Extracted from Student Submissions

Collection / Method	Unit 5, Named Entities	Unit 6, LDA	Unit 9, Summary
Ebola	*Organizations*: Doctors Without Borders, National Centre for Disease, UN Security Council, World Health Organization, Infectious Disease Research. *Locations*: Africa, India, Kenya, Gorakhpur, Mumbai, USA *Person*: Modi, Desai, Patrick Sawyer, Anna Hazare, Sri Sene.	ebola, news, health, India, encephali-tis, new, us, virus, world, people, news, health, people, disease	There has been an outbreak of Ebola reported in the following locations: Liberia, West Africa, Nigeria, Guinea, and Sierra Leone. In January 2014, there were between 425 and 3052 cases of Ebola in Liberia, with between 2296 and 2917 deaths. Additionally, In January 2014, there were between 425 and 4500 cases of Ebola in West Africa. In January 2014, there were between 425 and 3052 cases of Ebola in Guinea, with between 2296 and 2917 deaths.
Virginia earthquake	*Organization*: Sports, Health, News, Business *Location*: Virginia, East, Washington, U.S, NY, *Person*: Alexander, Kearney, Vervaeck, Armand, Paulm	Voting, Virginia, print election, god, rabbi, global, Washing-ton, power, market	On 23 August, 2011 at 1:51, a 5.8 magnitude earthquake struck Virginia, The epicenter of the quake was located at Louisa. There were aftershocks that followed the earthquake and no tsunami was caused by the earthquake. There are no reports of landslides due of this earthquake. A total of 140 deaths occurred.
Brazil club fire	*Organization*: Post, North, Americas, News, People, World, News, Europe *Location*: East, India, Pakistan, Central, Brazil *Person*: Mark, Clinton, Hollande, Obama	Fire, brazil, people, Santa, club, news, Maria, sign, nightclub, news, youtube, ago, Maria	In January 2013 there was a fire started by indoor fireworks in Santa Maria. This fire, fueled by ignited foam, grew to the size of the building, engulfed the club and ended up killing 309. One exit was made unavailable for a period of time. Compared to previous fires in the city was a fast-moving fire.
Pakistan flood	*Organization*: Time, Gwal, Sahiwal, Chagai, Jalapur, Shaikh, Ahmadpur, Bazar, Daulat, Nawan, Mithrau, Shah, *Location*: Islamabad, Khairpur, Pakistan	Flood, destroyed, damaged, Pakistan, shahdad-kot, water, Indus, jaffarabad,	In August 2010 a flood spanning 600 miles caused by heavy monsoon affected the Indus river in Pakistan, The total rainfall was 200 millimeters and the total cost of damage was 250 million dollars. The flood killed 3000 people, left 809 injured, and approximately 15 million people were affected. The cities

	Person: Khan, Chauki, Tando, Tangi,Toba, Shahpur, Ziarat, Fort, Kalat, Dera, Garhi, Haji	sindh, Hyderabad, internatio-nal, disaster, relief	of Nasirabad Badheen and Irvine were affected most by flooding, in the province of Sindh Mandalay and Punjab, finally nearly all of the flood damage occurred in the state of Pakistan.
Hurricane Sandy	N/A	Nicolas, sea, supplies, central, evacuation, grave, forecast, power, nation, food, typhoon	The storm, Hurricane Sandy, hits in New York on October 2012. The hurricane was a Category 1. Furthermore, the hurricane had a wind speed of 75 mph. hurricane Sandy formed in the Atlantic. Also, Hurricane Sandy had a size of 1000 miles wide. Hurricane Sandy caused 10 inches of rain. For more information. Search for hurricane sandy.
Tucson shooting	Organization: news, Inc., Abc, cnn, supermarket Location: Tuscan, Arizona, Casas, Person: Giffords, Chief, John, Tylor, Loughner	Tuscan, shooting, news, gabrielle, killed, police, loighner, public, January, business	On the night of Sunday, January 9, Jared lee opened fire in Tucson. The suspect fired 5 rounds out of his rifle. 6 people lost their lives. 32 of the people were hurt, and are being treated for their injuries. The victims were between the ages of 40 and 50

More details about the students' summarization results can be found at https://vtechworks.lib.vt.edu/handle/10919/50956

3.3.3 Evaluation

An evaluation involved two surveys of student ratings of aspects of the course, along with free-form comments. Students reported learning much, and liking both the course and the PBL approach. Clearly, they were engaged when other teams presented approaches and outputs; a good deal of sharing and adoption of software and methods resulted. Evaluating results of the student work is much harder. There are no "gold standards" for our data, since the class engaged in authentic activities to aid IDEAL. Our first analysis, for each of the six groups, compared Named Entities, LDA Topics, and Summaries, as reported, with related Wikipedia pages for six events. Table 3-6 shows Precision (Words in both summary

and Wikipedia entry / Words in summary), for the three methods, for each collection. Though a lengthy explanation of this table could be provided if space allowed, and other comparisons would be helpful, nevertheless it is clear students were on the right track.

Table 3-6. Precision Results for Six Collections Using Three Methods

Collection/Method Names	Unit 5, Named Entities	Unit 6, LDA	Unit 9, Summary
Ebola	0.209	0.571	0.159
Virginia Earthquake	0.173	0.117	0.541
Brazil club fire	0.0465	0.545	0.297
Pakistan flood	0.022	0.289	0.283
Hurricane Sandy	N/A	0.190	0.393
Tucson shooting	0.421	0.388	0.26

3.4 Conclusion

Problem based learning was applied in a computational linguistics class to help students learn how to build automatic text summaries for big collections. Different methods were applied to produce multiple types of summaries.

Results demonstrate that the 30 students, in seven teams, have been able to learn and apply big data and computational linguistics methods to produce reasonable corpus summaries. Through active learning and PBL, students generally unfamiliar with computational linguistics or using a Hadoop cluster to handle large digital library collections, mastered a broad range of valuable skills. Feedback from students, and review of student deliverables by the teaching assistants, has all been very positive. Accordingly, we offer our approach, corpora, and course details to others interested in working with big data summarization.

More evaluation is needed to better understand the role and impact of PBL in such courses. PBL also was used in the spring semester of 2015 in a graduate Information Retrieval class. Further, as is explained in subsequent chapters, the author of this dissertation, who served as one of the graduate teaching assistants for the PBL course, has extended the approach and methods to aid with Arabic summarization research on news corpora.

Chapter 4: Extracting Named Entities Using Named Entity Recognizer and Generating Topics Using Latent Dirichlet Allocation Algorithm for Arabic News Articles

Abstract

This chapter explains, for the Arabic language, how to extract named entities and topics from news articles. Due to the lack of high quality tools for Named Entity Recognition (NER) and topic identification for Arabic, we have built an Arabic NER (RenA), and an Arabic topic extraction tool using the popular Latent Dirichlet Algorithm (LDA) algorithm (ALDA). NER involves extracting information and identifying types, such as name, organization, and location. LDA works by applying statistical methods to vector representations of collections of documents. Though there are effective tools for NER and LDA for English, these are not directly applicable to Arabic. Accordingly, we developed new methods and tools (i.e., RenA and ALDA). To allow assessment of these, and comparison with other methods and tools, we built a baseline corpus to be used in NER evaluation, with help from volunteer graduate students who understand Arabic. RenA produces good results, with accurate name, organization, and location extraction from news articles collected from online resources. We compared the RenA results with a popular Arabic NER, and demonstrated an enhancement. We also carried out an experiment to evaluate ALDA, again involving volunteer graduate students who understand Arabic. ALDA showed very good results in terms of topic extraction from Arabic news articles, achieving high accuracy, based on an experimental evaluation with participants using a Likert scale.

Keywords: Arabic Language; Named Entity Recognizer; Topic Extraction; Latent Dirichlet Allocation, Natural Language Processing

4.1 Introduction

4.1.1 Arabic: Language, Encoding and Morphology

4.1.1.1 Arabic Language

Arabic is a widely used global language that has major differences from most popular languages, e.g., English and Chinese. The Arabic language has many grammatical forms, varieties of word synonyms, and different word meanings that vary depending on factors, among which is word order. In spite of such complexities, limited work has been devoted to natural language processing involving Arabic, especially in comparison to the English language, which has been addressed by numerous studies. 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, substantial modification and extra work would be required for tailoring to Arabic.

Unlike most languages, Arabic is written from right to left, with no capitalization, and with 28 alphabetical characters as well as diacritics. According to (Habash, 2010), there are multiple forms of the Arabic language such as:

- Classical Arabic – This form is used in reading / reciting the holy books.
- Modern Standard Arabic (MSA) – Standard Arabic, which is commonly used in writing, speech, interviewing, broadcasting, etc. It should be noted that throughout this report, implementation is based on MSA.
- Spoken – oral dialects that vary significantly from region to region.

Arabic also employs Vowel Marks (Tashkeel or Harakat (Habash, 2010) (known as diacritics)), such as those shown in Table 4-1 for one of the letters.

Table 4-1: Diacritics for the Letter "Alef"

ا	آ	أ	إ

Diacritics for letters such as “Alef” are used to signify or distinguish sounds that are not fully specified by the Arabic letters. These characters can be used interchangeably, and change the meaning of the word. Since they are mostly used in the context of verbal exchanges or recitation, they hold very little value in the analysis carried out on texts in connection with computational linguistics.

4.1.1.2 Arabic Encoding

One of the first challenges faced while working with texts is the ability to recognize the characters programmatically (with a computer program). Encoding tends to be problematic; the most common and effective way to solve this difficulty is to use Unicode (UTF8 for example). Alternatives include Windows CP-1256 or X-MacArabic

4.1.1.3 Arabic Morphology

The Arabic language has a complex morphology due to its derivational and inflectional nature (Benajiba, 2009). Arabic verbs and nouns are derived from a root word, and usually consist of the root word followed by a pattern to form a lemma (Benajiba, 2009). Consider the words in Table 4-2.

Table 4-2: Derivation Forms of the Word “Read” in Arabic

Read	قرأ
Reading	قراءة
Reader	قارئ
I read	قرأت
Peruse	قرأ بتمعن
Legible	مقروء

Both the word “read,” and the other forms shown in Table 4-2, have the same base root. This is very common in the language. In Figure 4-1, observe how both words share the

same root word, but semantically, have two different meanings. The first word is the root word, with the second word derived from it, but the meaning is changed by inflection, due to the suffix indicating singularity or plurality. In other cases, it may indicate gender or both, i.e., singularity/plurality and gender.

Figure 4-1: Two Words Share the Same Root in Arabic

4.1.2 Named Entity Recognizer – NER

Consider the English quote,

“Go back, Sam. I’m going to Mordor alone.”

— Frodo, The Lord of the Rings: The Fellowship of the Ring

In theory, an NER should be able to extract “Sam” as a name and “Mordor” as a location (and arguably an organization). This is very useful as it extracts useful keywords in context.

Named Entity Recognition of Arab(ic) names of persons, organizations, and locations requires modification of available tools, e.g., the Stanford Named Entity Recognizer (SNER), or creation of new tools to extract names of entities from text, e.g., from news articles. Extracting the named entities for any text may help point out key elements. We believe these three main entities (persons, organizations, and locations) reflect the most important entities in the text and serve as the main features for future work in Arabic news article text summarization. Toward extracting the appropriate Arabic named entities, we have modified one of the available Arabic NER tools, i.e., the one created by Yasine Benajiba that is called ANER (Arabic Name Entity Recognition) (Benajiba, 2009).

4.1.3 Latent Dirichlet Allocation – LDA

As more information becomes available, it becomes more important to access what we are interested in. It requires advanced implementations to help us organize, search, and understand these large amounts of information. Topic Modeling, for example LDA, can

assist with automatic organization, understanding, searching, and summarization of massive amounts of electronic data. A collection of documents may cover a variety of topics and sometimes each document addresses a mixture of those topics. In order to determine or "select" topics through a computational algorithm, topic modeling will be required, for example using the Latent Dirichlet allocation (LDA) algorithm.

Suppose we have a document that consists of various subjects. This document contains words that may refer or correspond to a specific subject. Basically the LDA algorithm attempts to map each word to its corresponding topic(s). LDA is a form of probabilistic model that will collect a set of words that establish a statistical relation based on the document collection. This generative model uses Bayesian inference, and involves collapsed Gibbs sampling to collect topics from a collection of documents (Blei, 2012; Griffiths & Steyvers, 2004).

Figure 4-2 below appears in Blei's work (Blei, 2012) and provides further explanation of the LDA algorithm.

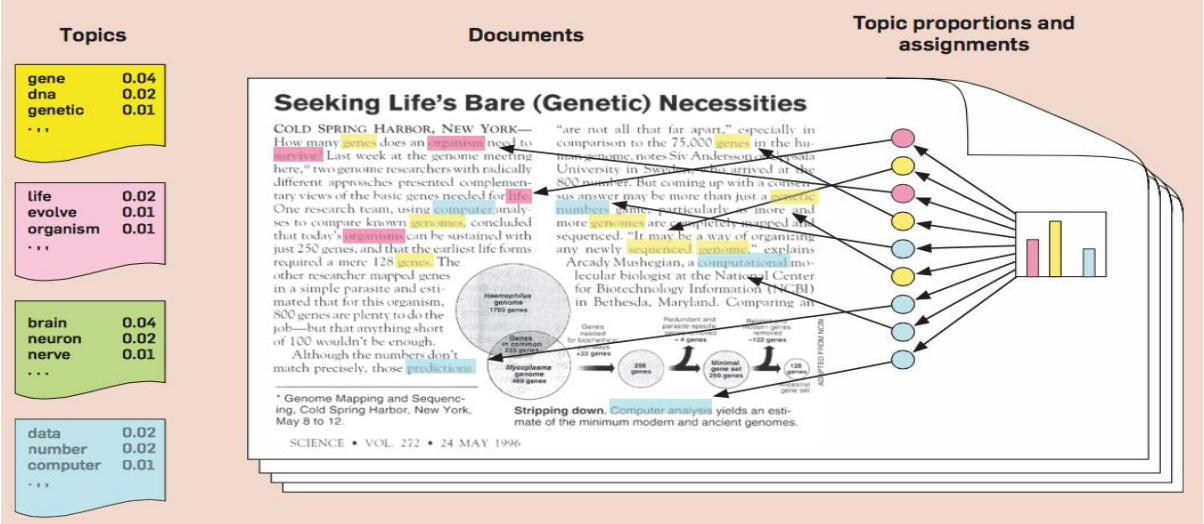

Figure 4-2: LDA Algorithm Demonstration (adapted from (Blei, 2012))

There is no implementation of the LDA algorithm that is publicly available and supports the Arabic language for topic modeling. So, to support the Arabic language, we have extended an open source LDA implementation (written in C Sharp) for the Chinese language that is publicly available on (GitHub, 2014).

4.2 Literature Review

4.2.1 Named Entity Recognizer – NER

The Stanford Named Entity Recognizer (SNER) is a Java implementation of a Named Entity Recognizer as defined by Manning et al. (Manning, Raghavan & Schütze, 2008). A Named Entity Recognizer (NER) labels sequences of words in a text, namely proper nouns, such as person and company names, or gene and protein names. It comes with feature extractors for Named Entity Recognition, and with many options for defining additional feature extractors. Included with the download are good named entity recognizers for English, particularly for the three classes (PERSON, ORGANIZATION, and LOCATION) (Manning, Raghavan & Schütze, 2008). In his dissertation "Arabic Name Entity Recognition", Benajiba describes a system he has developed to extract Arabic name entities within an open domain Arabic text. In order to create his ANER system, he examines the different aspects of the Arabic language related to NER tasks and the state-of-the-art of NERs (Benajiba, 2009). (Abuleil & Evens, 2004) paper describes a new technique to extract names from Arabic text by Abuleil et al. They build graphs to describe relationships between words. The proposed technique extracts some names, but misses others; they believe if they re-run the technique on more articles, the system will extract the missing names. Kanaan et al. use an existing tagger to identify proper names and other crucial lexical items and build lexical entries (Kanaan, Al-Shalabi & Sawalha, 2003). Shaalan develops a Named Entity Recognition system for Arabic (NERA) using a rule-based approach (Shaalan & Raza, 2009). He uses a whitelist to represent a dictionary of names, and he includes a grammar, in the form of regular expressions. NERA has been evaluated using special tagged corpora, yielding satisfactory results in terms of precision, recall, and F1-measure. In his work, Kareem Darwish tries to enhance Arabic named entity extraction by using cross-lingual resources (Arabic/English) for Wikipedia links (Darwish, 2013). He shows a positive effect on recall using his method compared with (Benajiba, 2009).

4.2.2 Latent Dirichlet Allocation – LDA

Allan et al. (Allan, Gupta & Khandelwal, 2001) define temporal summaries of news stories as extracting as few sentences as possible from each event within a news topic, where the stories are presented one at a time. They define an evaluation strategy and describe simple language models for capturing novelty and usefulness in summarization and they show that

their simple approaches work well (Allan, Gupta & Khandelwal, 2001). Topic discovery based on text mining techniques is discussed by Pons-Porrata et al. (Pons-Porrata, Berlanga-Llavori & Ruiz-Shulcloper, 2007). The authors present a topic discovery system that aims to reveal the implicit knowledge present in news streams. This knowledge is expressed as a hierarchy of topic/subtopics, where each topic contains the set of documents related to it. A summary is then extracted from these documents. The summaries they build are useful to browse, with topics of interest selected from the generated hierarchies (Pons-Porrata, Berlanga-Llavori & Ruiz-Shulcloper, 2007). An approach to building topic models based on a formal generative model of documents, Latent Dirichlet Allocation (LDA), is frequently discussed in the machine learning literature, but its practicality and effectiveness in information retrieval are mostly unidentified (Wei & Croft, 2006). Liu et al. (Liu, Zhou, Pan, Qian, Cai & Lian, 2009) consider two aspects in their paper. These aspects are the design and development of a time-based visual text summary that effectively conveys complex text summarization results produced by the Latent Dirichlet Allocation model. They have applied their work to a number of text corpora and their evaluation shows promise, especially in support of complex text analysis. Brahmi et al. observe that the topic model judges each document (considered as a bag of words) as a combination of topics defined by a probability distribution over words (Brahmi, Ech-Cherif & Benyettou, 2012).

The LDA model has been introduced within a general Bayesian framework where the authors have developed a vibrational method and expectation–maximization (EM) algorithm for learning the model from the aggregation of discrete data (Blei, Ng & Jordan, 2003). Since the original Prolog version of the LDA model, several contributions have been proposed. However, few studies on finding latent topics in Arabic text have been identified. For integration with works related to Arabic topic detecting and tracking (Oard & Gey, 2002; Larkey, Feng, Connell & Lavrenko, 2004), a segmentation method that utilizes Probabilistic Latent Semantic Analysis (Hofmann, 1999) has been applied to an AFP_ARB corpus for monolingual Arabic document topic analysis (Brants, Chen, & Farahat, 2002). In (Larkey, Feng, Connell & Lavrenko, 2004), the researchers compare different topic tracking methods. They claim that the utilization of a separate language for building concrete topic models is preferred. Good topic models are obtained when native Arabic stories are available. However, Arabic topic tracking has not been satisfactory in texts

translated from English stories. In fact, studies on Arabic IR are insufficient and the few works carried out for topic modeling lack strong evaluation. Considering the high inflectional morphology in Arabic, it seems more opportune to learn an LDA model in a mono-language context, taking more care with linguistic aspects.

4.3 Methodology

4.3.1 Building a Baseline Dataset

Since we could not find a judged Arabic news article corpus to be used as a baseline corpus to test and evaluate the NER results and to compare the implementation of our NER with other existing NER systems, we decided to build a new baseline judged Arabic news article corpus.

We began with roughly 5,200 PDF archived documents from Al-Raya, a Qatari news site covering various topics such as sports, politics, etc. Since each document contains multiple news articles, we analyzed the files to separate the articles. Files that would require OCR or where encoding was problematic were discarded. Remaining documents were processed to extract individual news articles. However, since some encoding issues were not caught by the first filter, another layer of filtering was added to discard illegible articles. The result was roughly 120,000 articles. We randomly selected one thousand articles as a test sample toward building our judged baseline corpus, as shown in Figure 4-3.

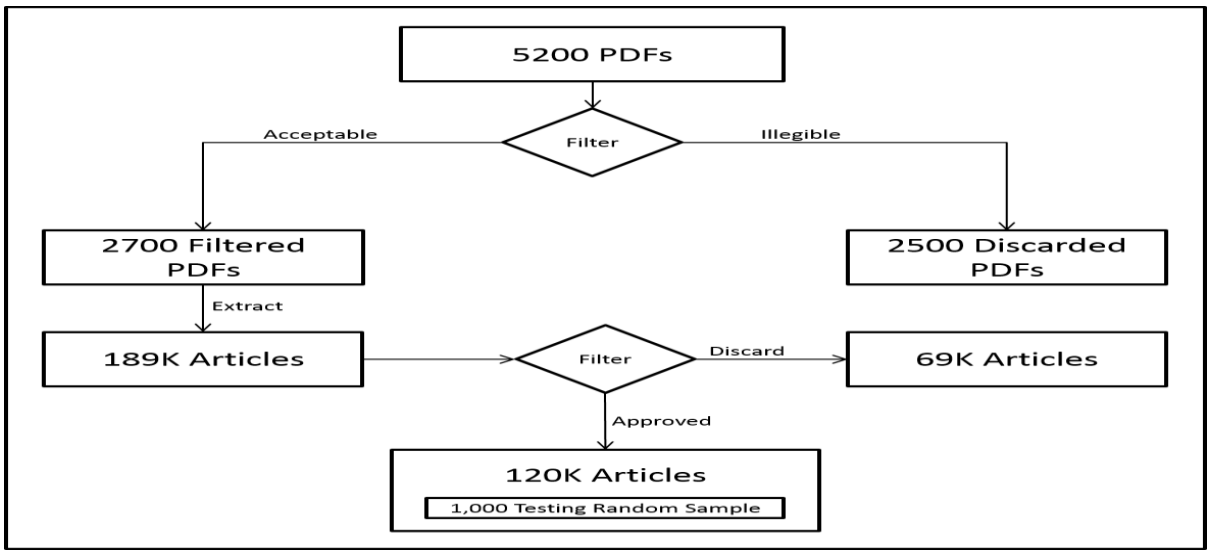

Figure 4-3: Flow of Dataset Extraction

We recruited graduate students who understood Arabic, because user experiences, behaviors, and task performances could differ depending on the participants' academic level and familiarity with Arabic reading and writing. For this study, we recruited ten participants; each participant was assigned one hundred articles from the one thousand random sample we collected during our dataset extraction. For each article, they were tasked to read the article and then label the named entities. Later, they were to evaluate the articles' topics.

4.3.1.1 Dataset Examples

Figure 4-4 below introduces an example news article from our dataset. This article has been chosen from one of the many documents we collected.

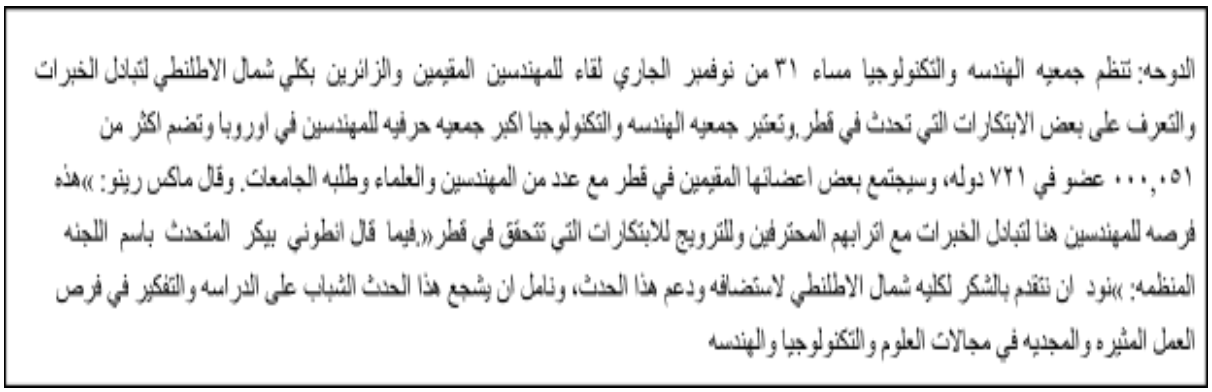
الدوحه: تنظم جمعيه الهندسه والتكنولوجيا مساء ٣١ من نوفمبر الجاري لقاء للمهندسين المقيمين والزائرين بكلي شمال الاطلنطي لتبادل الخبرات والتعرف على بعض الابتكارات التي تحدث في قطر .وتعتبر جمعيه الهندسه والتكنولوجيا اكبر جمعيه حرفيه للمهندسين في اوروبا وتضم اكثر من ٠٠٠,٥١ عضو في ٧٢١ دوله، وسيجتمع بعض اعضائها المقيمين في قطر مع عدد من المهندسين والعلماء وطلبه الجامعات. وقال ماكس رينو: »هذه فرصه للمهندسين هنا لتبادل الخبرات مع اترابهم المحترفين وللترويج للابتكارات التي تتحقق في قطر«.فيما قال انطوني بيكر المتحدث باسم اللجنه المنظمه: »نود ان نتقدم بالشكر لكليه شمال الاطلنطي لاستضافه ودعم هذا الحدث، ونامل ان يشجع هذا الحدث الشباب على الدراسه والتفكير في فرص العمل المثيره والمجديه في مجالات العلوم والتكنولوجيا والهندسه

Figure 4-4: An Arabic News Article from our dataset

For better understanding, we include an English news article example to explain how our process works. See Figure 4-5.

AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, "has all the qualities" required to take the helm of the club, Ancelotti told a news conference. "I enjoy Zidane's work, he's doing very well," Ancelotti said.
After a difficult start of the season, Castilla are top of Spain's third tier league. "He's doing very well in his first year in charge. He's taken Castilla to first place and he needs to keep up the good work.
"It's pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid," said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao's reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Figure 4-5: Example of English News Article

4.3.1.2 Stopword Removal

Similar to the English language, Arabic includes words that can be treated as stopwords. It is necessary to filter out those stopwords. However, due to language requirements, sometimes a stopword can have multiple meanings, such as the Arabic word "ال". In English, this translates to "the". Obviously, this keyword is considered as a stopword in English, which gives us a reasonable reason to discard it. However, this is not the case for Arabic, as the word can be used to represent a family name. It is important to consider these special stopwords when processing and extracting key entities.

There are multiple, freely available Arabic stopword lists. For this research we have merged two lists of Arabic stopwords to produce a richer collection of stopwords to meet the requirements of our work; it is important to reduce noise by avoiding stopwords when identifying named entities or topics. One of the lists we use is adopted from Université de Neuchatel (UniNE, 2015) and is also used by Lucene Apache (Apache Software Foundation, 2013). The other list is from a Google project called "Stop-Words", where stopwords are provided for 28 different languages, including Arabic (Google Code, 2014).

4.3.1.3 Stemming

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 topic extraction systems by reducing the dimensionality of word vectors. The goal of an Arabic light stemmer is to find the representative form of an Arabic word by removing

prefixes and suffixes, while maintaining infixes. Thus, the meaning of the word remains intact, which results in improved topic identification effectiveness.

For our experiments we used the stemmer that come with Al-Khalil Morphological System; an open source Arabic analyzer, to stem our corpus (SourceForge, 2011). In future work we plan also to test with P-stemmer.

4.3.1.4 Normalizing the Text

As discussed in Section 4.1.1, we need to consider every possible form of each word, since we rely on a knowledge base to help with NER. For example, consider the following in Table 4-3.

Table 4-3: Examples of Harakat (Diacritics)

Word with Vowel	**Root Word**	**English**
كَتَبَ	كتب	Write
كُتُب	كتب	Books
قَرَأَ	قرأ	Read
قارِئ	قرأ	Reader
رَكَضَ	ركض	Run
رَكّاض	ركض	Runner

In order to produce more precise results, the collection of content that needs to be processed should be normalized. However, in the example provided in Table 4-3 above, when such words as “write” and “books” are normalized to the same root form, the meaning changes.

4.3.2 Arabic Named Entity Recognizer - RenA

4.3.2.1 Building the NER

There are 3 ways to build an NER, as shown below:

- Knowledge Base – A collection of words used to identify entities based on a predefined dataset; such a collection contains a set of words mapped to a specific entity.
- Machine Learning – Using statistical models to classify and identify grammars to deterministically identify entities, with Conditional Random Fields (CRF) (Benajiba, 2009) being the plausible choice.
- Training – Manually or automatically generate a classifier that will identify the entities.

Initially, this research relies on a knowledge base, but it is later improved by training.

4.3.2.2 Knowledge Base

We are using an open source knowledge base (ANERCorp, 2010), which is Benajiba's (Benajiba, 2009) freely distributed corpus for Arabic, Table 4-4, which consists of roughly 150,000 tokens that are tagged.

Table 4-4: Ratio of sources used to build the ANERCorp (ANERCorp, 2010; Benajiba, 2009)

Source	Ratio
http://www.aljazeera.net	34.8%
Other newspapers and magazines	17.8%
http://www.raya.com	15.5%
http://ar.wikipedia.org	6.6%
http://www.alalam.ma	5.4%
http://www.ahram.eg.org	5.4%
http://www.alittihad.ae	3.5%
http://www.bbc.co.uk/arabic/	3.5%
http://arabic.cnn.com	2.8%
http://www.addustour.com	2.8%
http://kassioun.org	1.9%

4.3.2.3 Approach

4.3.2.3.1 Stage 1 – Building RenA

For the initial implementation of the NER, the knowledge base is used to classify the entities of the provided texts. There is a need to build a dictionary to map the words in the knowledge base to their entity type (PERS, ORG, and LOC). Once the collection of words has been mapped to the appropriate entities, we can use the populated dictionary and a chunker (tokenizer) to classify a collection of text and determine the words' entities. The chunker will tokenize based on whitespace. For each word, the chunker will identify possible results of the tags; for example, the word, "Washington" can be added to the dictionary with the tag (PERS) to indicate a person; we can also add "Washington" as a (LOC) to indicate a location. Once applied, every occurrence of the word "Washington" will return two different tags, [PERS, LOC].

The results produced are reasonable. Some of the keywords are tagged, while others are not. However, some of the words are tagged incorrectly. The most obvious problems in this stage are related to the stopwords and Harakat (diacritics). These problems are addressed in stage 2.

4.3.2.3.2 Stage 2 – Improving RenA

In this stage, the main focus is to properly filter out stopwords and normalize words. Removing stopwords is considered a simple approach as a list of stopwords can be used to filter out words of little significance. Normalization helps in reducing the words' dimensionality and preventing duplicate results when chunking.

At this point, results are improved by filtering out stopwords and by normalization to reduce the varieties of the words. However, this still raises the issues of some words being improperly tagged and some words missing tags. Indeed, outcomes for organization named entities lead to inaccurate results due to the complexity of organization naming. Often, it introduces ambiguities with person and location entities. Some Arabic NERs report low accuracy result for organizations (Benajiba, 2009; Darwish, 2013). Surprisingly, this issue

appears to be a continuing problem for Arabic NER. However, persons and locations produce better results, with defects that can be addressed.

In order to improve the precision of each entity based on the knowledge base, we must train the corpus. For this collection, the use of inclusion and exclusion lists yields an exceptional increase in precision. In our experiment, we have selected a random sampling of news articles that came from 10 different newspapers published in 10 different countries, those news articles contain a substantial number of keywords to help improve our training by enriching it with names of persons, organizations, and locations. Table 4-5 shows the list of resources and their regions, which reflect sources used to enrich our knowledge base.

Table 4-5: News articles used to train the NER collection via inclusion/exclusion

Sites	**Region**
http://www.aljazeera.net	Saudi Arabia
http://alwatan.com	Qatar
http://www.ahram.org.eg	Egypt
http://www.alarabalyawm.net	Jordan
http://www.al-akhbar.com	Lebanon
http://www.alanwar.com/	Lebanon
http://www.albayan.ae	UAE
http://www.alquds.co.uk	London (Universal)

With the use of inclusion and exclusion lists, our knowledge base starts to improve by the addition/removal of words for their appropriate entity type classes. For each training (enhancing) phase, we are relying on the current state of our NER and enriching the knowledge base by adding new tagged or removing wrong tagged entities for each phase.

An interesting note is, once the NER is trained on an article, it usually does better on the next set of articles, i.e., the training is generalizable.

Figure 4-6 below summarizes the steps used to build our NER (RenA). The inner part of RenA deals with building the knowledge base (dictionary) by importing the ANERCorp knowledge base, after normalizing it, and then adding our inclusion/exclusion list. All together this will build our NER dictionary. We used the normalizer and stopword removal to preprocess our article; then the chunker will act as a tokenizer and mapper. The chunker tokenizes the preprocessed article based on the article's whitespaces, and for each token it will attempt to check if an entity exists in the dictionary based on the token. Finally, the entities extraction function will produce the results from the mapped tokens.

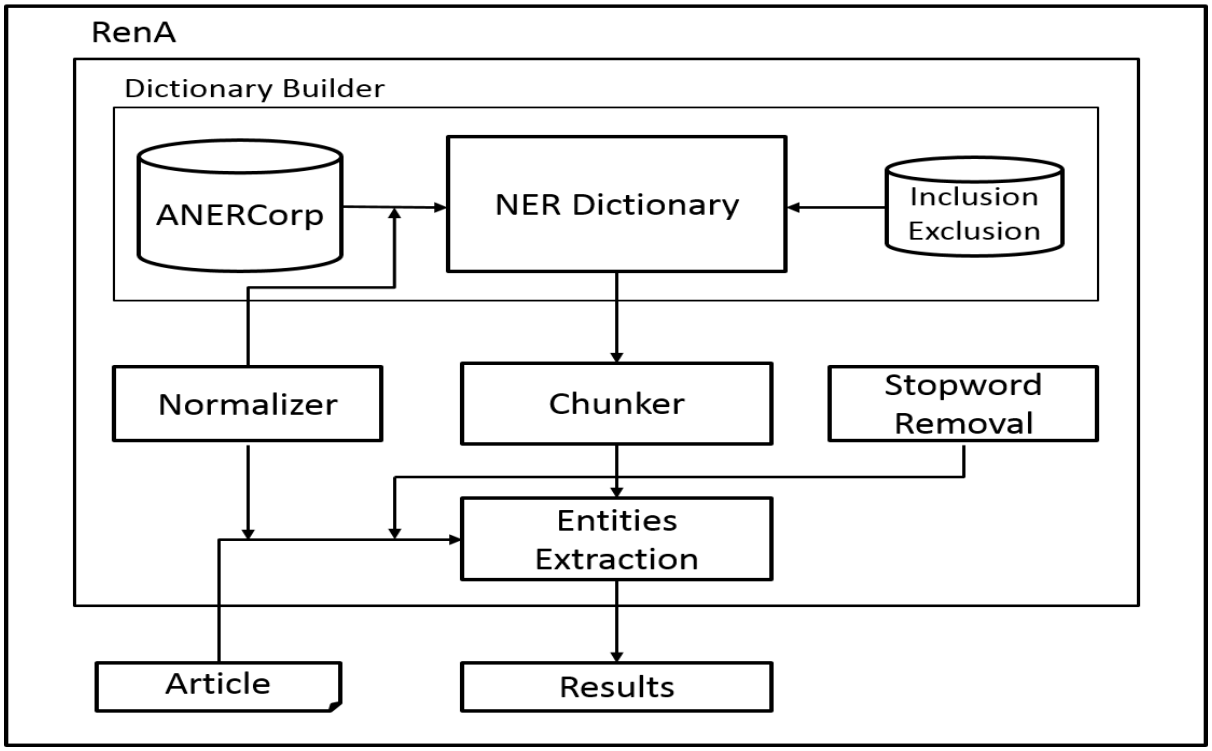

Figure 4-6: RenA Architecture

4.3.2.4 Results

Using the same articles from dataset examples in Section 4.3.1.1, we show results after extracting the named entities in the article. The results of the named entity extraction are shown in bold in the figures below. See Figure 4-7 for the Arabic example.

الدوحه: تنظم جمعيه الهندسه والتكنولوجيا مساء ٣١ من نوفمبر الجاري لقاء للمهندسين المقيمين والزائرين بكلي شمال الاطلنطي لتبادل الخبرات والتعرف على بعض الابتكارات التي تحدث في قطر .وتعتبر جمعيه الهندسه والتكنولوجيا اكبر جمعيه حرفيه للمهندسين في اوروبا وتضم اكثر من ٥١٠,٠٠٠ عضو في ٧٢١ دوله، وسيجتمع بعض اعضائها المقيمين في قطر مع عدد من المهندسين والعلماء وطلبه الجامعات. وقال ماكس رينو: »هذه فرصه للمهندسين هنا لتبادل الخبرات مع اترابهم المحترفين وللترويج للابتكارات التي تتحقق في قطر«.فيما قال انطوني بيكر المتحدث باسم اللجنه المنظمه: »نود ان نتقدم بالشكر لكليه شمال الاطلنطي لاستضافه ودعم هذا الحدث، ونامل ان يشجع هذا الحدث الشباب على الدراسه والتفكير في فرص العمل المثيره والمجديه في مجالات العلوم والتكنولوجيا والهندسه

Person: ماكس، رينو، انطوني، بيكر

Organization: جمعيه الهندسه والتكنولوجيا، كليه شمال الاطلنطي

Location: الدوحه، قطر،اوروبا

Figure 4-7: Arabic News Article and Extracted Named Entities

For more illustration, an English example is included to explain how our process works. Figure 4-8 shows an example of an English news article and its named entities.

AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, "has all the qualities" required to take the helm of the club, Ancelotti told a news conference. "I enjoy Zidane's work, he's doing very well," Ancelotti said.
After a difficult start of the season, Castilla are top of Spain's third tier league. "He's doing very well in his first year in charge. He's taken Castilla to first place and he needs to keep up the good work.
"It's pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid," said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao's reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Person: Zinedine, Zidane, Carlo, Ancelotti, Martin, Odegaard
Organization: Real, Madrid, Castilla, Athletic, Bilbao
Location: Spain, Madrid, Bilbao, Norway, Europe

Figure 4-8: English News Article and Extracted Named Entities

Figures 4-7 and 4-8 show articles where an NER was used to extract named entities (shown in bold). These extracted entities can be used to identify the underlying context of the article and show some of its key elements. For example, in Figure 4-8, a football fan will be able to recognize that this article is about the Real Madrid soccer team (Organization named entity) in regards to a player named Zidane (Person named entity).

4.3.2.5 Evaluation

Table 4-6 shows us the results between RenA and the basic NER from LingPipe (LingPipe, 2008). These results give recall, precision, and F1 measure.

Table 4-6: Recall, Precision, and F1 Values for RenA and LingPipe NERs

	RenA NER			**LingPipe Toolkit NER**		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
PERSON	0.826	0.497	0.539	0.582	0.371	0.374
ORGANIZATION	0.813	0.421	0.446	0.39	0.377	0.329
LOCATION	0.77	0.558	0.564	0.55	0.338	0.356
Average	0.803	0.492	0.516	0.507	0.362	0.353

In Figure 4-9 below, the precision values of both NERs are displayed for each entity. It shows that RenA produces higher precision results for each entity.

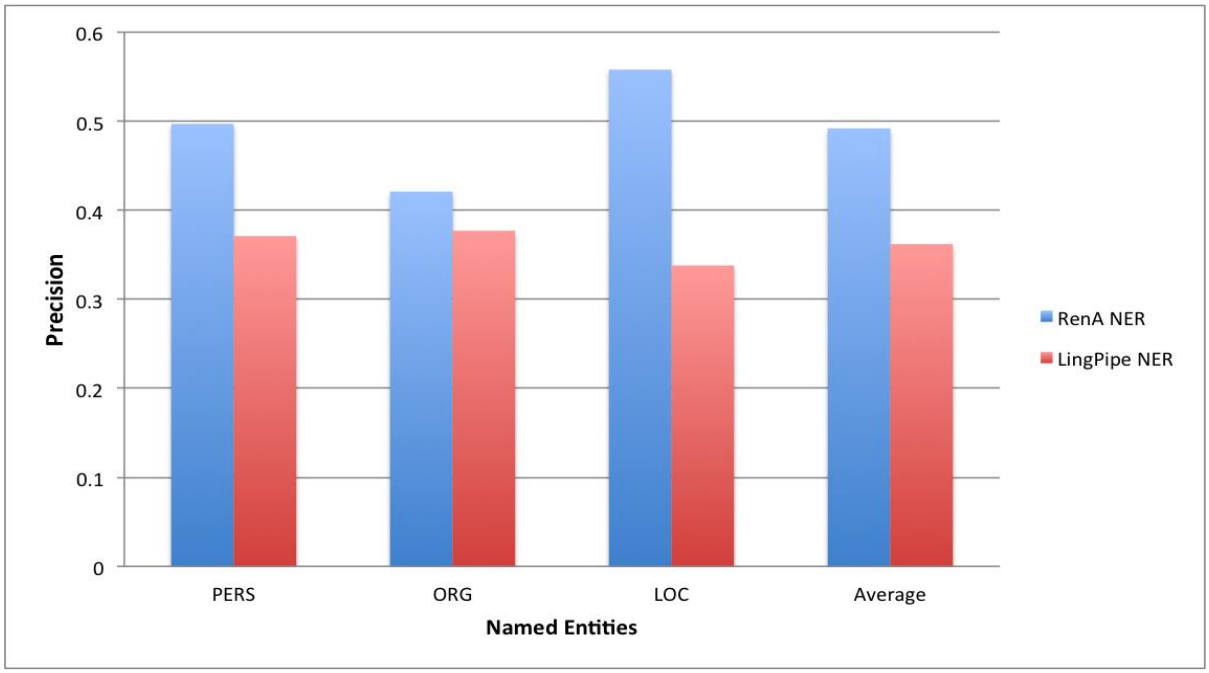

Figure 4-9: Precision Values for RenA and LingPipe NER

In Figure 4-10 below, the recall of both NERs is displayed for each entity. RenA showed better results than its counterpart NER in terms of recall.

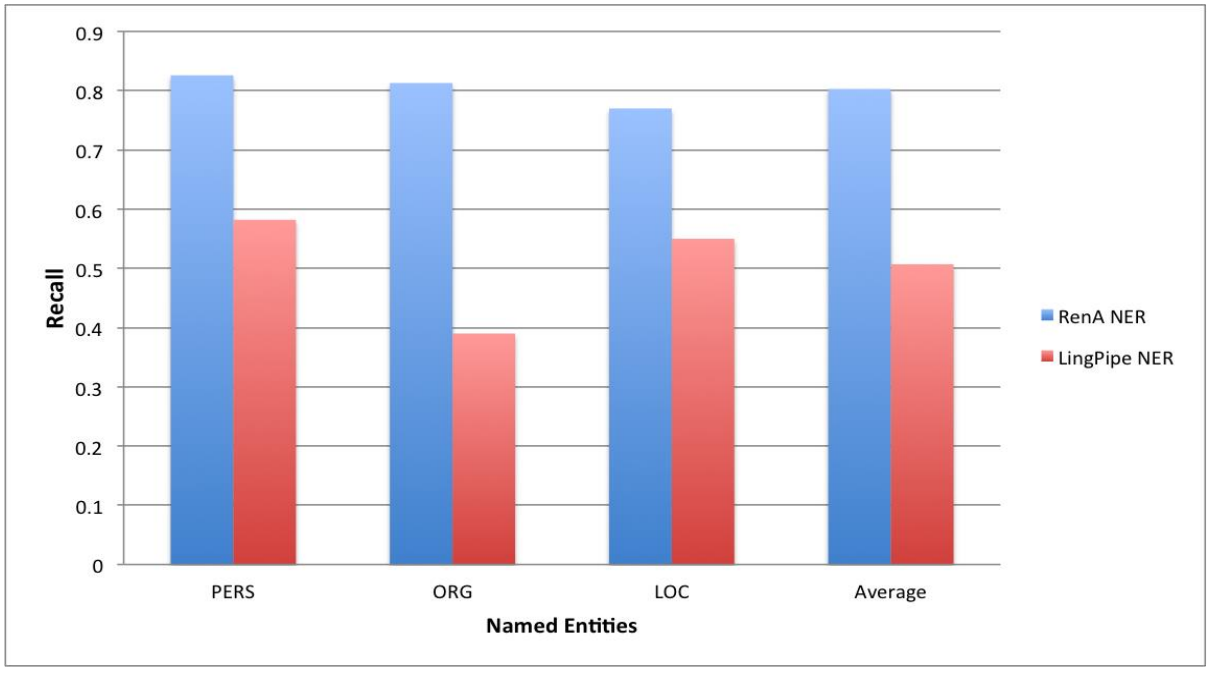

Figure 4-10: Recall Values for RenA and LingPipe NER

In Figure 4-11 below, the F1 measure of both NERs is displayed for each entity (the greater the value, the better it is). F1 measure shows that RenA is retrieving better results. Table 4-6 shows that RenA is on average more effective in retrieving results; calculations show approximately 15% improvement.

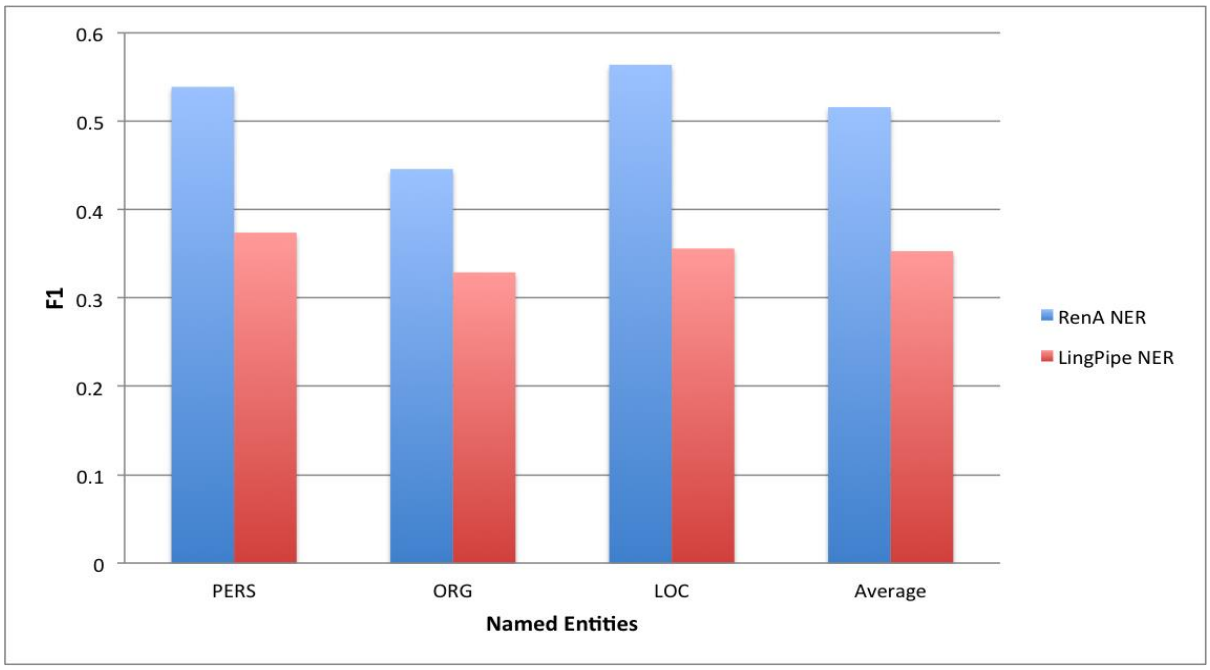

Figure 4-11: F1 Values for RenA and LingPipe NER

4.3.3 Arabic Latent Dirichlet Allocation – ALDA

4.3.3.1 Algorithm

In order for the LDA to identify the different topics – since a corpus may consist of multiple topics and various words, which involves various distributions – Bayesian inference must be applied. The inference techniques that have been used in the LDA algorithm involve collapsed Gibbs sampling in order to marginalize the distribution and probabilistically determine the topics of a corpus. By using a Dirichlet Multinomial distribution (also referred to as categorical distribution) in the sampling process, the algorithm can marginalize and determine the topical model. It should be noted that since LDA is a generative model using Gibbs sampling, iteration is required to generalize the steady-state (or Markov chain) model; the sampling algorithm is random in the initial state, whereas each iteration will further improve the topic choices. One of the characteristics of the LDA

implementation is modularity; this allows preprocessing of the data, which can further enhance the topic selection, for example, stemming and stopword removal.

The open source code we have used consists of two main parts: 1) Core, which involves the LDA generative model and statistics algorithms that have been modified to support the Arabic language, stemming, stopword removal, and normalization; and 2) Viewer, for which we have modified and implemented an interface in order to include extra interactive variables, such as number of topic, words per topic, and number of inference iterations using the LDA model in the Core library. We also modified the Viewer to make it handle all the input variables through the interface. The result of each model will be displayed as a table and also saved in a CSV file for further evaluation. The following sections show screenshots and examples produced by our Arabic Latent Dirichet Allocation (ALDA). The main changes we made were to create the interface and make the use of the tool easier through the Viewer plus adding the text preprocessing steps on the Core to help enhance the results. Our contribution to Arabic was to make this tool read Unicode text, apply preprocessing steps to filter the text, and finally provide the results in an easier readable format.

4.3.3.2 Result

Figure 4-12 shows a news article example and the results after applying the ALDA algorithm. It shows the main topic covered by this article as a list of words with their corresponding probability.

طرابلس – رويترز: قال مسؤول ان رجال قبائل ليبيين انهوا حصارهم لحقل الشراره النفطي لكن لا يتسنى استئناف الانتاج لحين انتهاء احتجاج منفصل عند خط انابيب مرتبط بالحقل. وكان قبليون وحراس امن اغلقوا الحقل ال ذي تبلغ طاقته ٠٤٣ الف برميل يوميا بجنوب البلاد في فبراير شباط للضغط من اجل مطالب ماليه وسياسيه وهو ما زاد من حده الحصار المفروض على موانئ نفط في الشرق. وقال حسن الصديق مدير حقل الشراره لرويترز ان المحتجين الذين اغلقوا الحقل تركوا المكان لكن لا يمكن استئناف العمل به نظرا لان الصمامات ما زالت مغلقه. واضاف ان هناك مفاوضات تهدف الى انهاء اغلاق صمامات خط الانابيب في الجبال الغربيه ويامل المهندسون باستئناف الضخ في غضون اسبوع. واغلقت مجموعه اخرى من المحتجين في منطقه ال زن تان في الغرب خطوط الانابيب من اجل مطالب ماليه وسياسيه. واغلق محتجون حقل الشراره اكثر من مره.وكان انتاج ليبيا يبلغ نحو ٤،١ مليون برميل يوميا حتى منتصف عام ٣١٠٢ حين بدات الاحتجاجات التي قلصته الى اكثر قليلا من ٠٠٢ الف برميل يوميا.حقل الشراره الليبي ما زال مغلقا رغم انتهاء الاحتجاج بكتيريا ماصه للغازات الطبيعيه لمواجهه التسرب النفطي

Probability – Topic
0.0361768646717284, الشراره
0.0272443054935239, النفطي
0.0272443054935239, احتجاج
0.0272443054935239, انابيب
0.0272443054935239, برميل
0.0272443054935239, المحتجين
0.0183117463153193, حصارهم
0.0183117463153193, استئناف
0.0183117463153193, الانتاج
0.0183117463153193, انتهاء

Figure 4-12: News Article with its Corresponding Topic Using ALDA

Figures 4-13 and 4-14 show two screen shots for ALDA and the related results. Different parameters are used for each of the two shots.

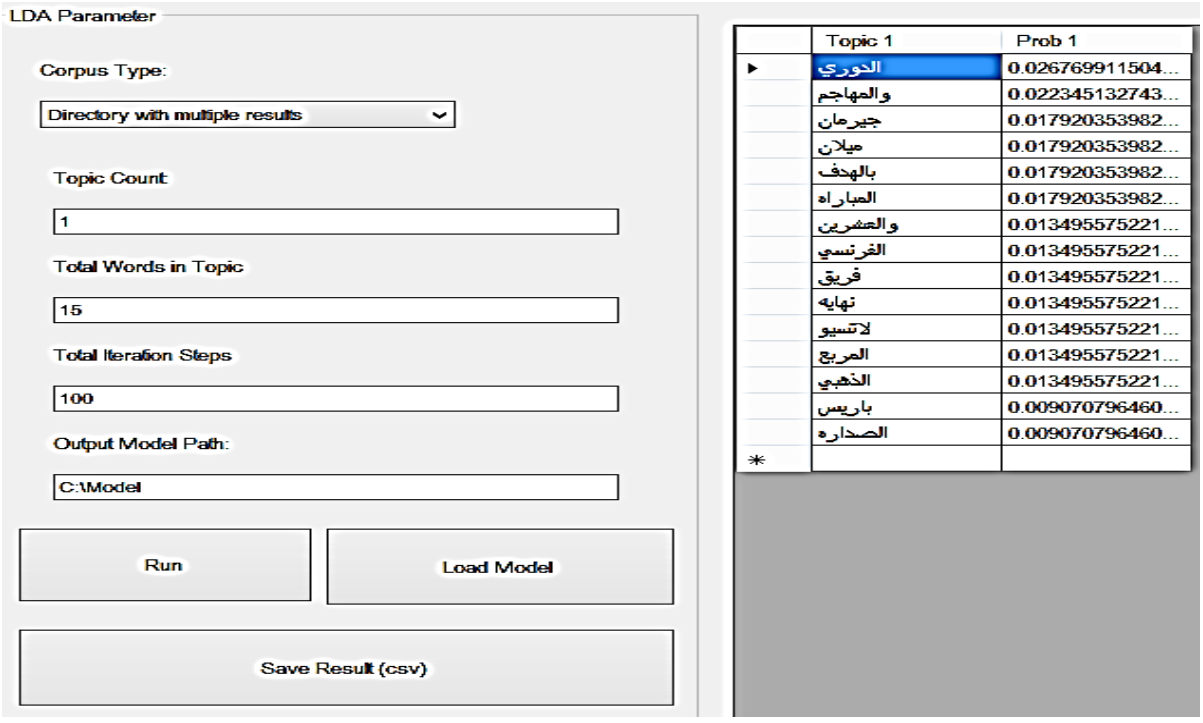

Figure 4-13: ALDA Screen Shot Showing One Topic

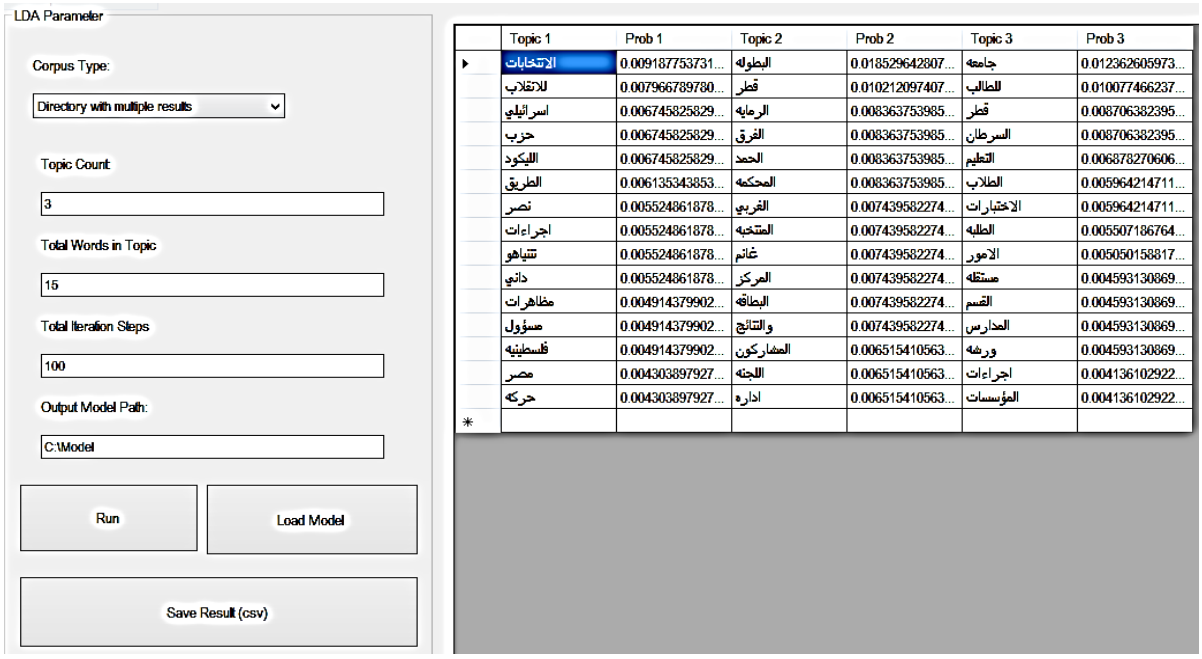

Figure 4-14: ALDA Screen Shot Showing Multiple Topics

4.3.3.3 Evaluation

For evaluating ALDA, we recruited graduate students who understand Arabic. This is mainly because user experiences, behaviors, and task performances differ according to the academic level of participants and their familiarity with Arabic reading and writing. Each participant is assigned a set of articles from the randomly sampled one thousand articles we have collected in our baseline dataset. For each article, students are tasked to read the article and its corresponding topic generated by ALDA. After that, students are asked to evaluate the topic of each article using a Likert scale, indicating their view of the relevance of the topic to the article. Each topic/article pair must be assigned a number between 0-10 for relevance, where 0 means the topic is not relevant to the article and 10 means the topic is highly relevant to the article. This study is different than the study we did to create the NER baseline corpus. Two different groups of people did the two studies. Please see Appendices A and B for more detail on the two studies.

Figure 4-15 and Table 4-7 show the evaluation results for one thousand articles. Articles are divided into eleven categories (0-10), based on their rates (scores). Ten participants

each evaluated two hundred random articles; each of the one thousand articles was evaluated by two different participants. We averaged the evaluation score of each article and counted the frequency of each resulting score value. As we explained in the previous paragraph, a high topic score for an article indicates that the topic is more relevant to the article. Table 4-7 demonstrates that the majority of the articles are scored between 7 and 10. From Figure 4-15, it can be concluded that applying the LDA algorithm over Arabic news articles leads to achieving very good results in terms of generating accurate and relevant topics.

Table 4-7: Number of Articles for each Score

Rate Value	0	1	2	3	4	5	6	7	8	9	10
Number of Articles	0	0	0	15	49	71	74	125	269	241	156

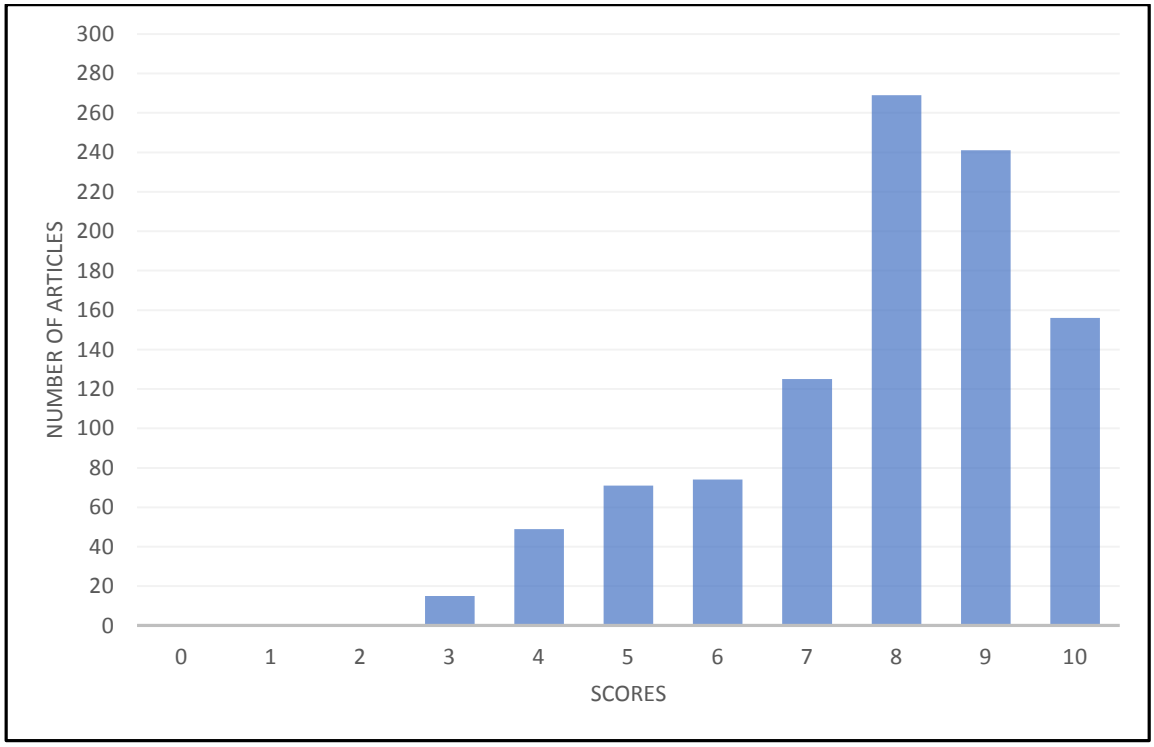

Figure 4-15: 11 Categories Used to Show ALDA Evaluation Results

4.4 Conclusion

Natural Language Processing (NLP) research involving the Arabic language is relatively hard, compared to other popular languages, such as English. Adding to that, available free NLP tools and resources are rare. These issues form the source of inspiration for this research. There is no substantial research addressing the extraction of named entities from Arabic news articles. The same applies to generation of topics from articles.

In this study, we aim at developing a Named Entity Recognizer that can extract, with good accuracy, the named entities from Arabic news articles, called RenA, by enriching the knowledge base though including and excluding steps we created for this purpose. We also modify the popular topic extraction model, LDA, to enable it to handle and generate topics from Arabic news articles, called ALDA, by applying some preprocessing steps to clean up the Arabic text and by creating a new interface for the tool. Due to the lack of free resources for a judged news articles corpus, we have resorted to building a corpus, with the help of graduate students fluent with Arabic, to be used with RenA and ALDA evaluations, and later by other researchers.

We use Information Retrieval evaluation measures to evaluate and compare our RenA NER with another NER that is available through the LingPipe toolkit. We considered three types of named entities: Person, Organization, and Location. Our results show that using the proposed RenA enhances the named entity extraction results for the three mentioned types of entities, compared to the LingPipe toolkit NER.

A second experiment, with graduate students who understand Arabic, helped to evaluate our ALDA tool. Using a Likert scale for assessment with our Arabic news article corpus, evaluation results confirmed that our developed tool generates highly relevant topics.

4.5 Future Work

Our future plan is to expand this research by using the RenA and ALDA results to fill in templates. We also are planning to extract more attributes to fill in templates, towards generating improved Arabic news article summaries. In addition, we aim to use another Arabic stemmer, e.g., P-Stemmer, and to compare the ALDA results with the two stemmers.

Chapter 5: Arabic News Articles Template Summarization

Abstract

This chapter explains for the Arabic language how to extract key information like person and organization named entities, titles, writers, publishing dates, categories, and topics, from news articles. Due to the lack of high quality tools for Named Entity Recognition (NER) and topic identification for Arabic, we have built an Arabic NER (RenA) and an Arabic topic extraction tool using the popular LDA algorithm (ALDA) to extract writers, person and organization named entities, and topics. Dates of publication are extracted using regular expressions, while categories are identified using machine learning methods. Titles, on the other hand, are extracted using simple text extraction methods. All of the previous seven attribute values are extracted for each news article, with the seven values together serving as the template summary for the article. With help from volunteer graduate students who understand Arabic, we carried out an experiment to evaluate our summaries. Template summaries showed very good results in terms of text extraction from Arabic news articles, achieving high accuracy, based on an experimental evaluation with participants using a Likert scale.

Keywords: Arabic Language; Text Extraction; Template Summarization; Natural Language Processing

5.1 Introduction

5.1.1 The Arabic Language

Arabic is a broadly used language with key differences from most other widespread languages like English; recall Section 2.1.3. The Arabic language has many structural, grammatical, and linguistic forms, diversities of word synonyms, and different word meanings. Because of these difficulties, there has been limited natural language processing work with the Arabic language, especially in comparison with other languages. Most of the computational linguistics tools do not address Arabic language necessities. To make these tools work with Arabic language data, changes and extra efforts are required for adapting to Arabic.

Arabic is the fifth most spoken language in the world (Nationsonline, 2015; Wikipedia, 2015), with around 300 million speakers (see Figure 5-1). Arabic consists of 28 different characters with different shapes for the same character, depending on the place of the character in the word.

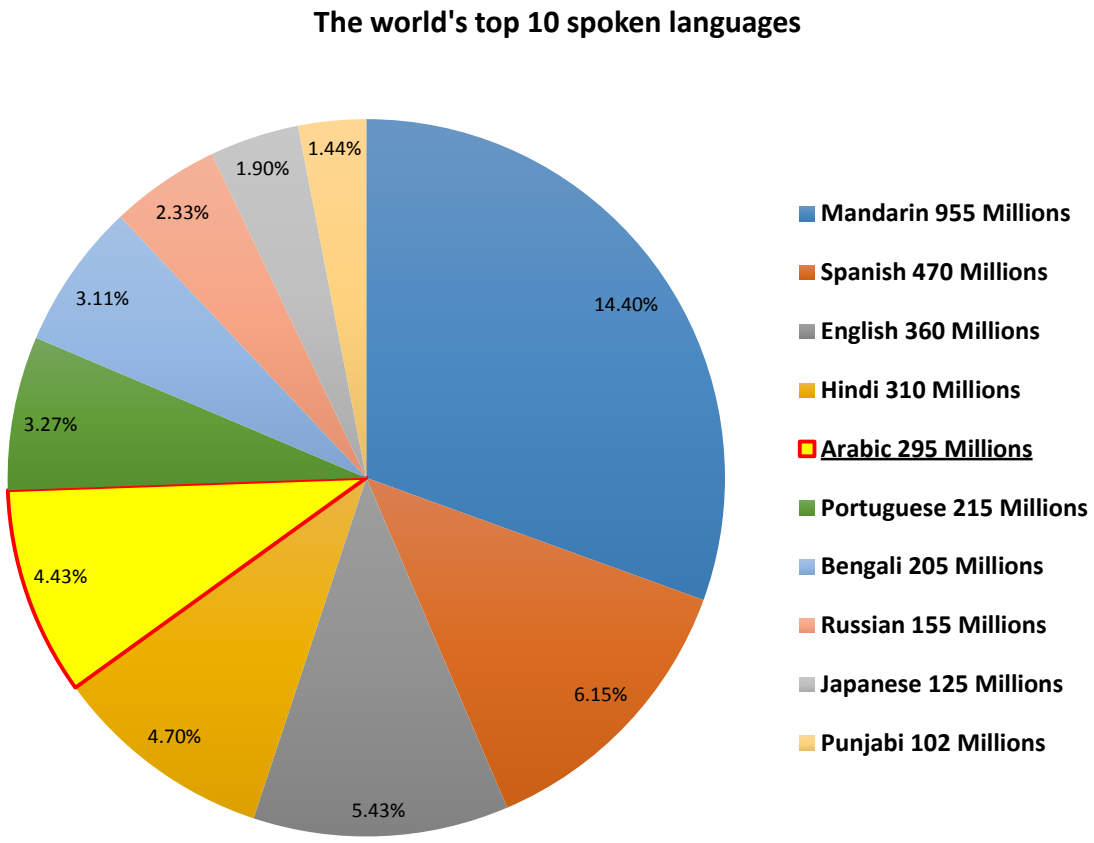

Figure 5-1: Top 10 Spoken Languages in the World with their Corresponding Percentages

5.1.2 Stopword Removal

Similar to the English language, Arabic includes stopwords that can be removed; recall Section 4.3.1.2. While it is important to clean the text by removing stopwords, some language necessities require keeping those that may hold valuable meanings, like the Arabic word "ابن", which means "son" in English. This word is considered in general a

stopword, which we usually remove. But, this is not the case for Arabic, since the word is considered a person named entity. Therefore, it is essential to study and keep these stopwords when extracting key elements from articles. There are various freely available Arabic stopword lists. For this study we decided to merge two lists of Arabic stopwords to make sure that we have a richer list that contains more words to be removed. One of the lists we use is adopted from UniNE (Savoy, 2015) and is also used by Lucene Apache (Apache Software Foundation, 2015). The other one is taken from a Google project called "Stop-Words", where stopwords lists are provided for 28 different languages, including Arabic (Google Code, 2015). The stopword list from the "Stop-Words Project" contains 163 words while the UniNE list contains 161 stopwords. The final merge after uniting both sets (lists) resulted in a total of 185 unique stopwords.

5.1.3 Research Problem and Proposed Solution

The problem that has prompted this research is the absence of good summaries for Arabic news articles. For users, determining whether an article is of interest to them without reading (or at least scanning) it is unfeasible. Also there is no good summarization framework for Arabic news articles that can satisfy a diverse user community through the use of fully automated methods. We have proposed an automatic Arabic text summarization approach, applied it to news articles, and tried to prove that it is better than existing state-of-the-art approaches, with quality not significantly less than those generated by humans. Relative to existing Arabic text summarizers, the proposed approach is more efficient and produces similar quality Arabic news article summaries, by using text extraction methods to fill out a developed template and generate the summary from the template.

5.1.4 Text Summarization

Automatic summarization aims to create brief overviews of longer texts, while keeping original ideas and flow. There has been little research on automatic summarization of online Arabic news articles. Existing methods, though fast, tend to have deficiencies as to coverage, accuracy, and linguistic quality (El-Haj & Hammo, 2008) and (Kanaan,

Hammouri, Al-Shalabi & Swalha, 2009). Assessment of the various approaches is also a challenge, calling for judgments from multiple domain experts.

Producing readable Arabic news article summaries may lead to the following dilemma:

Users may not like to read the whole article unless they are sure it is interesting. Instead, they prefer to skim through something that will give them an idea about the article so they can decide if the article is interesting enough to read. The problem is to find a suitable summary for a news article to enable the reader to decide if they want to read the whole article or not. However, such summaries are rare, and too costly to produce manually.

5.2 Literature Review

5.2.1 Text Summarization

Automatic text summarization is an essential tool to overcome the so-called information overload phenomenon, and is part of the area of computational linguistics. (Bird, Klein, & Loper, 2009) published a useful textbook on natural language processing, the field that supports a variety of language technologies, from predictive text and email filtering to automatic summarization and translation. How to write Python programs that can work with large collections of unstructured text is the main idea of this book, which also discusses how to extract information from text. A summarization procedure based on the application of trainable machine learning algorithms is addressed in (Neto, Freitas, & Kaestner, 2002). Neto et al. compare results using well-known text databases and some baseline summarization procedures. Schlesinger et al. presents an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic (query)-driven summaries for single documents or clusters of documents (Schlesinger, O'Leary & Conroy, 2008). Bender's book is a compact reference work aimed at researchers and students in the area of NLP (Bender, 2013). It is concerned with morphology and sentence structure. Hovey et al. (Hovey & Lin, 1998) discuss the task of a text summarizer. The level of sophistication of the summary can vary from a simple list of keywords that indicates the main content of the document(s), through a list of independent sentences that together reflect the main content, to a coherent, fully generated

compact text that covers the document(s). The more complex the summary, the more effort it generally takes to produce.

Kim et al. describe summarization as a process of reducing information to a smaller size, and to its most important points (Kim, Medelyan, Kan & Baldwin, 2010). They discuss that various kinds of summaries (e.g., headlines, abstracts, key-phrases, outlines, previews, reviews, biographies, and bulletins) can be read with limited effort in a shorter time. Therefore, people tend to read summaries before they choose to read the whole text. The textual content of online statement objects is a significant source of information about social relationships. Ryan Richardson in his dissertation hypothesized that concept maps can work as a summary of large documents such as Thesis and Dissertations (Richardson, 2007). (Goldstein, Kantrowitz, Mittal & Carbonell, 1999) presents the authors' analysis of news-article summaries generated by sentence selection. Sentences are ranked for potential presence in the summary using a weighted combination of statistical and linguistic features. The statistical features are adapted from standard IR methods. The potential linguistic ones result from an analysis of news-wire summaries. Alguliev et al. present a document summarization model, which extracts key sentences from given documents while reducing redundant information in the summaries. A useful first step in the automatic or semi-automatic generation of summaries from source texts is the selection of a small number of 'meaningful' sentences from the source text. To reach this, each sentence in the text is scored according to some degree of importance, and the best-rated sentences are selected. This results in collections of the most 'meaningful' sentences, in the order in which they appear in the source text (Alguliev, Aliguliyev & Mehdiyev, 2011).

A step toward generating summaries is to utilize templates. Ma et al. suggest first deciding: What is a template? One view defines a template as a consecutive group of text tokens that appear in every page applicable to that template, share the same geometrical location and size within the webpages, and serve primarily as navigation, trademark, or advertising without providing other information (Ma, Goharian, Chowdhury & Chung, 2003). Chambers et al. (Chambers & Jurafsky, 2011) claim that a template defines a specific type of event (e.g., a bombing) with a set of semantic roles for the usual entities involved in such an event. Standard algorithms for template-based information extraction (IE) require

predefined template schemas, and often-labeled data, to learn to extract slot fillers. They describe an approach to template-based IE that removes this requirement and performs extraction without knowing the template structure in advance. The algorithm instead learns the template structure automatically from raw text, inducing template schemas as sets of linked events (Chambers & Jurafsky, 2011).

5.2.2 Named Entity Recognizer-NER

A Named Entity Recognizer (NER) labels sequences of words in a text, namely proper nouns, such as person and company names, or gene and protein names; recall Sections 4.1.2, 4.2.1, and 4.3.2. It comes with feature extractors for Named Entity Recognition, and with many options for defining additional feature extractors. Manning et al. provided a downloadable named entity recognizer for English, particularly for three classes (PERSON, ORGANIZATION, and LOCATION) (Manning, Raghavan & Schutze, 2008). In his dissertation "Arabic Name Entity Recognition", Benajiba describes a system he has developed to extract Arabic name entities within an open domain Arabic text. In order to create his ANER system, he examines the different aspects of the Arabic language related to NER tasks and the state-of-the-art of NERs (Benajiba, 2009). Abuleil et al. describe a new technique to extract names from Arabic text. They build graphs to describe relationships between words. The proposed technique extracts some names, but misses others; they believe if they re-run the technique on more articles, the system would extract the missing names (Abuleil & Evens, 2004).

5.2.3 Topic Generation-LDA

Recall Sections 4.1.3, 4.2.2, and 4.3.3. Allan et al. define temporal summaries of news stories as extracting as few sentences as possible from each event within a news topic, where the stories are presented one at a time. They define an evaluation strategy and describe simple language models for capturing novelty and usefulness in summarization. They show that their simple approaches work well (Allan, Gupta & Khandelwal, 2001). The LDA model has been introduced within a general Bayesian framework where the authors have developed an expectation–maximization (EM) algorithm for learning the model from the aggregation of discrete data (Beli, Ng & Jordan, 2003). Since the original

Prolog version of the LDA model, several contributions have been proposed. However, few studies on finding latent topics in Arabic text have been identified. For integration with work related to Arabic topic detecting and tracking (Ord & Gey, 2002; Larkey, Feng, Connell & Lavrenko, 2004), a segmentation method that utilizes Probabilistic Latent Semantic Analysis (Hofmann, 1999) has been applied to an AFP_ARB corpus for monolingual Arabic document topic analysis (Brants, Chen & Farahat, 2002). In (Larkey, Feng, Connell & Lavrenko, 2004), researchers compare different topic tracking methods. They claim that the utilization of a separate language for building concrete topic models is preferred. Good topic models are obtained when native Arabic stories are available. However, Arabic topic tracking has not been satisfactory in texts translated from English stories. In fact, studies of Arabic IR are insufficient and the few works carried out for topic modeling lack strong evaluation. Considering the highly inflected morphology in Arabic, it seems more opportune to learn an LDA model in a mono-language context, taking more care with linguistic aspects.

5.2.4 Arabic Text Summarization

The amount of natural language text available in electronic form is overwhelming and is increasing every day. Yet, the complexity of natural language can make it very difficult to access the information in that text. The state-of-the-art research in NLP is still far from being able to build general-purpose representations of meaning from text (Bird, Klein & Loper, 2009). Douzidia et al. describe, develop, and evaluate an Arabic summarization system on the very short summary of noisy text of DUC2004 (Douzidia &Lapalme, 2004). Compaction techniques are used to produce ten word summaries of news articles. El-Haj et al. attempted to produce a query-oriented summary for a single Arabic document (El-Haj & Hammo, 2008). They implemented an Arabic Query-Based Text Summarization System. Results were short summaries, indicating a promising simple approach for text summarization. Kanaan et al. depicted the architecture of a question answering system and methodically evaluated contributions of different system components to accuracy (Kanaan, Hammouri, Al-Shalabi & Swalha, 2009).

5.3 Methodology

5.3.1 Dataset

We collected our dataset from online Arabic newspapers. We first crawled approximately 5000 full newspapers in PDF format, then, after multiple filtrations and parsing processes, we ended up extracting around 120K Arabic news articles in text format. For more details, see Section 4.3.1 and Figure 4-3.

We extracted all of the attributes' values for each news article, combined the attributes' values, and then made them serve as a template summary for the news article. Finally, we attached each template summary to its corresponding article and saved both of them in the same file. We randomly selected one thousand articles as a sample set toward conducting the evaluation experiment. Table 5-1 shows the main characteristics and statistics of our dataset.

Table 5-1: Dataset Characteristics

Article Languages	Arabic
Encoding	UTF-8
File format	Text File
Size on disk	860MB
Number of Articles	117,753
Number of Sentences	875,951
Number of Words	6,182,621
Number of Characters	25,976,445
Avg. Sentences per Article	7.4
Avg. Words per Article	52.5
Avg. Characters per Article	220.6

5.3.2 Template Summarization Approach

We are extracting the values of seven attributes from each news article. Together, the seven values will serve as a template summary for the news article. Different approaches have

been used to extract these attributes' values. Figures 5-2 and 5-3 show the template attributes with values to be integrated in the summaries. We included the English figure for better illustration.

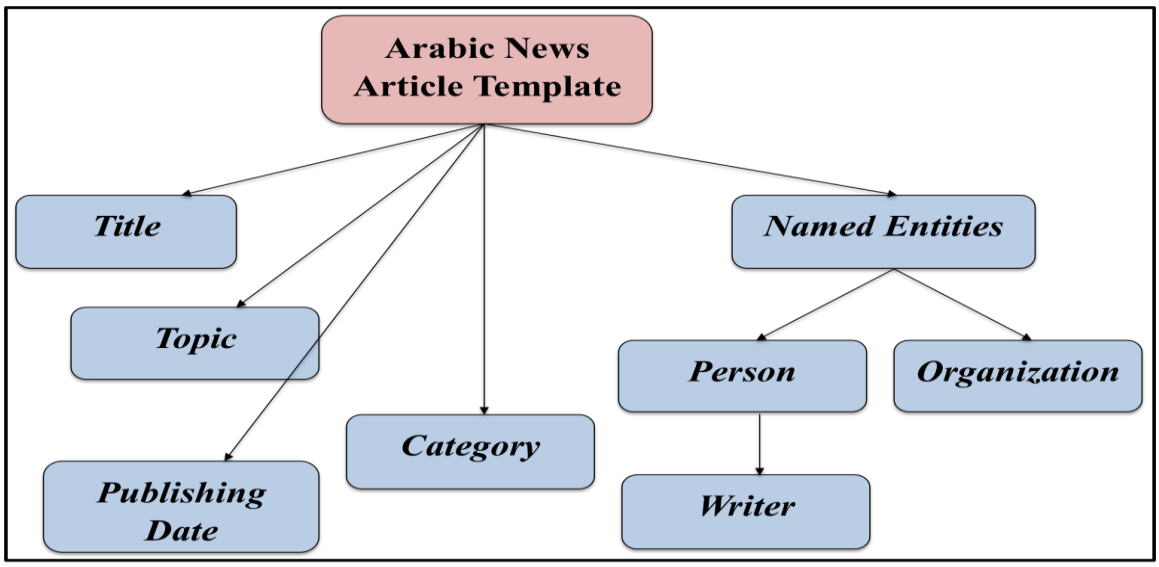

Figure 5-2: Template Attributes in English

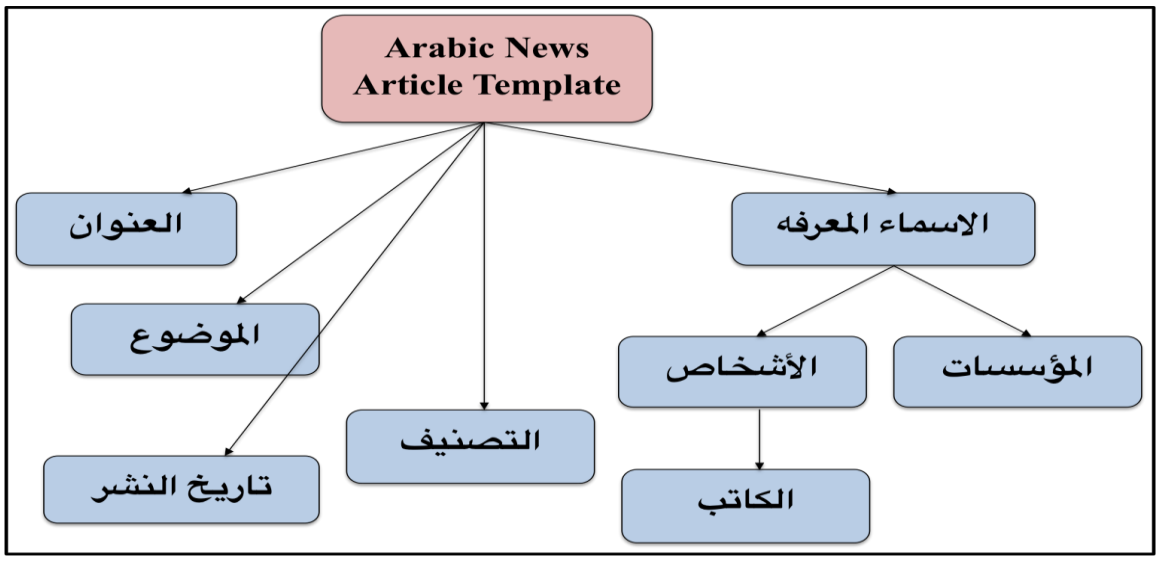

Figure 5-3: Template Attributes in Arabic

The approaches and tools we used to extract the attributes' values are:

1. Named Entity Recognizer for Arabic Persons' and Organizations' names. This is used to extract values for the Writer, Person, and Organization attributes.
2. Topic Identification tool for the Arabic language: a version of the Latent Dirichlet Allocation algorithm that can handle Arabic. This is used to extract the values for the Topic attribute.
3. Simple text extraction method using punctuation. This is used to extract the Title attribute.
4. Regular expressions are used to extract the Date attribute, see Table 5-2.
5. Machine learning methods and tools are used to classify news articles, then to extract the Category attribute.

Table 5-2: Part of the Regular Expressions Used to Extract Date Attribute

Regex Description	**Regex**
Arabic Months	سبتمبر|اكتوبر||يونيو|يونيه|يوليو|يوليو|اغسطس|يناير|فبراير|مارس|ابريل|مايو نوفمبر|ديسمبر
Indian-Arabic Numerals	[٠|١|٢|٣|٤|٥|٦|٧|٨|٩]
Indian-Arabic Years	(?<Year>[IA_NUMERAL]{2}(?:٠٢|٩١))
Arabic (YYYY_MM_DD)	(?<Day>(\\d|[IA_NUMERAL]){1,2}).(?<Month>[AR_MONTH]).(?<Year>(\\d{2}(?:91|02)|[IA_YEAR2]))
Arabic (MM_DD_YYYY)	(?<Month>[IA_NUMERAL]]{1,2}).(?<Day>[IA_NUMERAL]{1,2}).[IA_YEAR]
Punctuation	\\.|,|\\(|\\)|:|;|،|؟|؛|\\?|»|«|-|\\\\\|/|

Table 5-3 describes information related to the template attributes:

- Column one shows the attribute name.
- Column two shows the attribute description and the method used to extract/generate the values for the attribute.

Table 5-3: Attributes' Name and Description

Template Attribute	Description
Writer	The first Person named entity extracted using NER
Date	The publishing Date extracted using regular expressions
Title	The article title, probably the first line in the article
Person(s)	The Person(s) named entity(ies) extracted using NER
Organization(s)	The Organization(s) named entity(ies) extracted using NER
Topic	The main Topic in the article generated using Arabic LDA tool
Category	The Category of the article generated using a classification algorithm

After applying all summarization steps and extracting all template components, we generate Arabic news article summaries, aiming for high quality, see Figure 5-9.

5.3.3 Sample Summarization Results and Statistics

This section shows and discusses some samples and statistics from our attributes' fields and summaries. Table 5-4 shows the seven attributes and the number of missing values for each attribute. Table 5-5 shows the opposite values of Table 5-4, being the number of filled-in attributes. Note that the total number of summaries obtained is almost 120,000, i.e., 117,751. If some attributes introduce 117,751 values under category, topic, and date attributes, it means we are able to extract all of the values relevant to these fields.

Table 5-4: Frequency of Missing Values for the Summary Attribute

Attributes	Count
Title	1042
Date	0
Writer	3420
Person	5462
Organization	13553
Category	0
Topic	0

Table 5-5: Frequency of Filled-in Attribute Values for Summaries

Attribute	Count
Title	116709
Date	117751
Writer	114331
Person	112289
Organization	104198
Category	117751
Topic	117751

Table 5-6 shows the number of articles in each of the five top-level categories, and their percentages. Figure 5-4 portrays each category frequency and percentage. From both the table and the figure, it is evident that the Politics category has the highest percentage with 26.67%, while the Social Issues category has the lowest.

Table 5-6: Category Attribute Frequency and Percentage of Overall Summaries

Category	Count	Percent
Art & Culture	18438	15.66
Economy	21400	18.17
Politics	31399	26.67
Social Issues	17935	15.23
Sport	28579	24.27

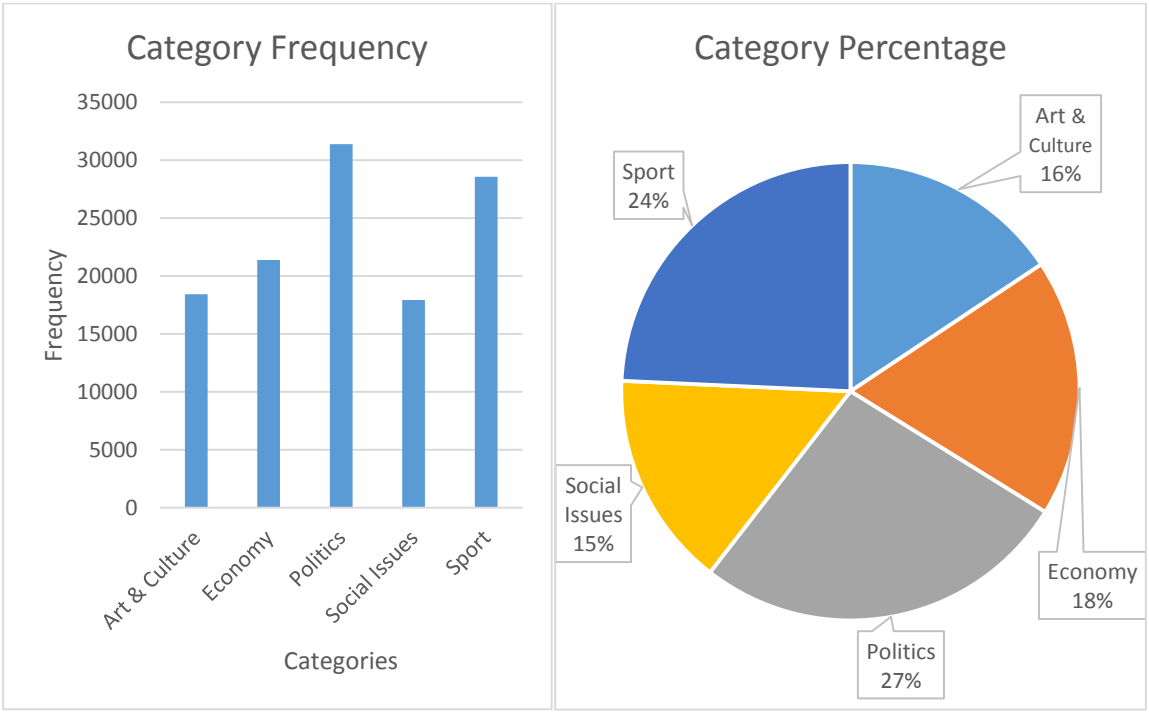

Figure 5-4: Category Attribute Frequency and Percentage

Table 5-7 shows the Publication Date attribute value frequencies per year plus the percentage of the appearance years. The table further shows that 2009 has the least number of articles. This may be attributed to the fact that most of the 2009 newspapers collected contain images, which are discarded in the filtering processes.

Table 5-7: Publication Date Attribute by Year – Overall Frequency and Percentage

Years	Frequency	Percent
2009	123	0.1
2010	10035	8.52
2011	15696	13.33
2012	37305	31.68
2013	35297	29.98
2014	19295	16.39

Table 5-8 shows the most frequent 20 publication dates appearing in article summaries. It also indicates that the first date in the table appears in around 300 different summaries.

Table 5-8: Top 20 Publication Dates and per Article Frequency

Publication Dates	Frequency
٤/مايو/٢٠١٠	292
٦/سبتمبر/٢٠١٣	254
٣/ابريل/٢٠١٠	251
٥/يونيو/٢٠١٣	229
٢٦/يوليو/٢٠١٢	228
٢/يونيو/٢٠١٠	221
٦/مايو/٢٠١٠	209
١٨/يونيو/٢٠١٤	206
١٧/مايو/٢٠١٠	206
٢٧/ابريل/٢٠١٠	205
٢٦/ابريل/٢٠١٢	204
١٦/اغسطس/٢٠١٢	202
٢٠/ابريل/٢٠١٠	200
١٢/ابريل/٢٠١٠	199
١٨/ديسمبر/٢٠١٣	197

١٣/مايو/٢٠١٤	197
١٥/مايو/٢٠١٣	197
٢٦/ابريل/٢٠١٠	195
٣١/يوليو/٢٠١٢	195
٢٤/يناير/٢٠١٢	194

Table 5-9 shows person and organization named entities with their frequency appearance in summaries.

Table 5-9: Distinct Frequency of Person and Organization Named Entities

Entities	**Frequency**
Person	133790
Organization	35557

Table 5-10 and Figure 5-5 list the top 20 person named entities frequently appearing in summaries, while Table 5-11 and Figure 5-6 show the frequencies of organization named entity appearances.

Table 5-10: Top 20 Frequent Person Names in Summaries

Person	**Frequency**
الشيخ حمد بن خليفه ال ثاني	3041
الشيخ تميم بن حمد ال ثاني	2849
بشار الاسد	2760
عبد الله بن محمد	1658
صالح بن احمد	997
نداء صالح	989
صلاح عبد الغني	929
رمضان مسعد	903
دلال قناوي	898
محمود عباس	819

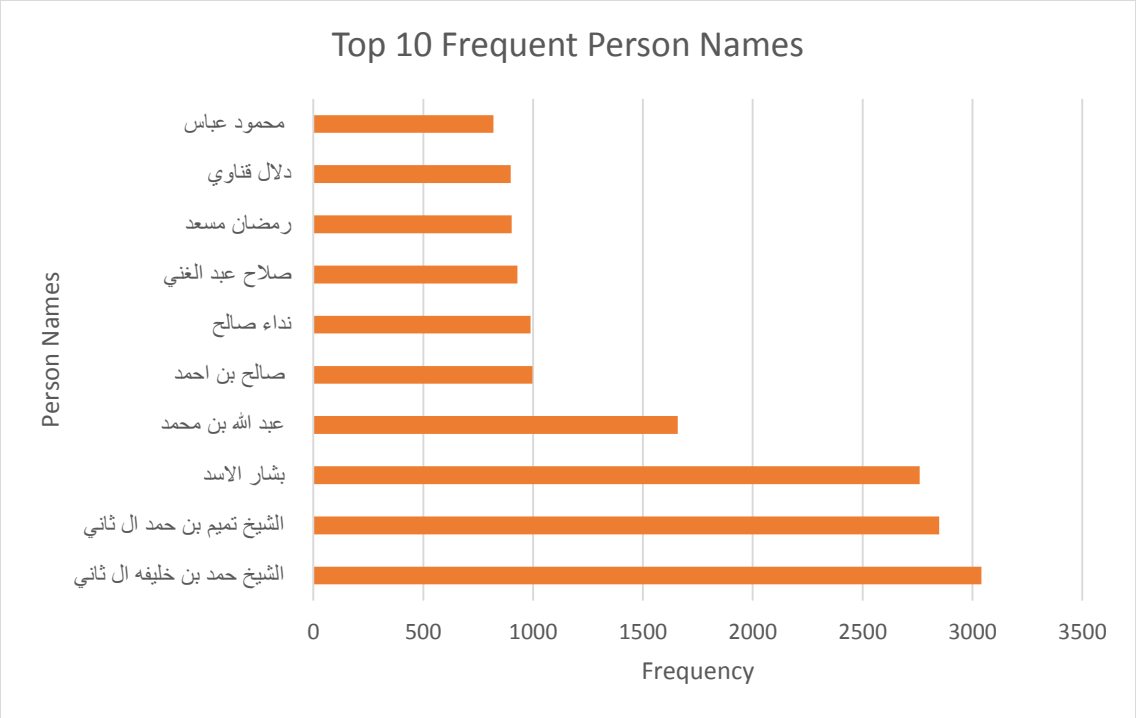

Figure 5-5: Top 10 Frequent Person Names

Table 5-11: Top 10 Frequent Organization Names in Summaries

Organization	Frequency
الشرق الاوسط	6833
الولايات المتحدة	6470
الامم المتحدة	6194
مجلس التعاون الخليجي	2829
مجلس الوزراء	2411
الاتحاد الاوروبي	1904
الاتحاد الدولي	1818
الاتحاد القطري	1716
قطر الوطنية	1709
مجلس الامن	1610

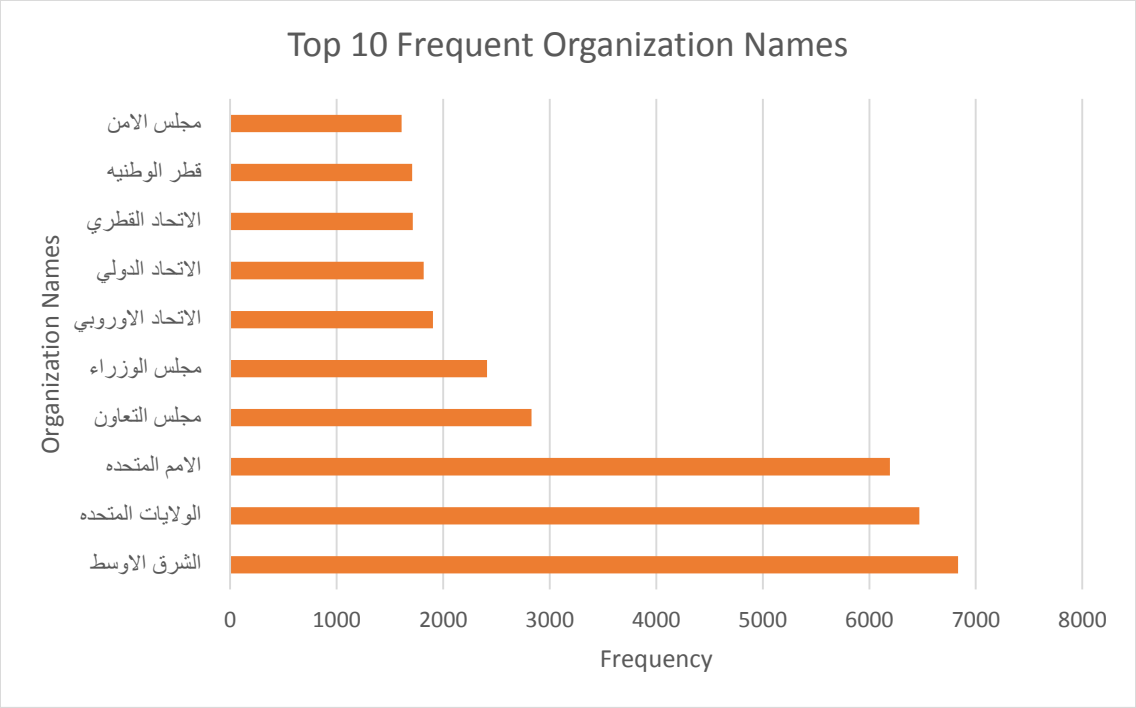

Figure 5-6: Top 10 Frequent Organization Names

Table 5-12 shows a random sample of 20 topics used in our template summaries. As previously discussed, each topic consists of 10 different words.

Table 5-12: 10 Randomly Selected Topics

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
الصناعية	مطار	تراخيص	المصرية	انتاج	مليار	بالحقل	الغازات	تصنيف	سعادة
الدولي	الدولي	والغاز	الوقود	للشهر	شركه	الشرارة	للاحتباس	جلسات	وزير
جويك	القطرية	نظريان	مصفاه	النفط	بقيمه	النفطي	الحراري	المالية	المواصلات
الخليج	الخطوط	جوله	دولار	تراجع	سنوات	احتجاج	البكتيريا	السوق	السفير
الشريعة	الجوية	الحكومة	الشركة	الروسي	لاستيراد	انابيب	المسببة	قطر	الدوحة
مجلس	قطر	مهله	مصر	مليون	الغاز	برميل	تقرير	التداول	استقبل
داماك	حمد	العروض	الديزل	برميل	الطبيعي	المحتجين	تسرب	الدولي	جاسم
الإسلامية	الجديد	بحريه	مليار	الرابع	المسال	حصارهم	الميكروب	الاسلامي	سيف
معرض	مياه	للنفط	مشروع	التوالي	دولار	استئناف	نوعا	الائتماني	السليطي
سوق	والتكييف	اغسطس	نفط	والغاز	عقد	الانتاج	يمتص	البنوك	كريستوف

In Table 5-13, we can see the most and the least frequent topics in the summaries with the frequencies for each one of the two topics.

Table 5-13: Most and Least Frequently Appearing Topics

Most Frequent Topic	**Least Frequent Topic**
للمال	لاعبي
قطر	الدوحة
مركز	دوري
نجيم	قطر
تنظيم	المشاركة
ورئيس	الإصابة
محكمه	الجهاز
شركه	وبالتالي
نيروبي	الاتحاد
دولار	الجاري
Frequency: 89	**Frequency: 20**

Table 5-14 illustrates the frequency distribution for the top 20 words appearing in the set of topics extracted from the collection. According to the table, "قطر" -"Qatar" in English- scores highest in the set of topics used, which is sensible since our dataset is collected from Qatar.

Table 5-14: Frequency Distribution for Top 20 Topic Words

Topic	**Frequency**
قطر	314
الدوحة	149
الامير	136
دول	104
مجلس	101
فريق	99
دور	65

لاعب	65
مركز	57
المباراة	55
ريال	53
دولار	52
الدولي	51
اتحاد	51
مباراة	51
شرك	50
شركه	50
العربية	49
اداره	48
العالم	48

5.3.4 Template Summarization Examples

Figure 5-7 shows an example of one news article from our dataset.

الدوحه 25/8/2012- انور الخطيب:

تحقيق الامن والاستقرار واعاده بناء مصر من جديد على اساس المساواه والعداله الاجتماعيه. مهام ملحه ارتاى عدد من الكتاب والمحللين السياسيين في قطر ان على الدكتور محمد مرسي رئيس مصر المنتخب ان يركز عليها في بدايات عهده الجديد.. مؤكدين ان نجاح الدكتور مرسي كاول رئيس مصري منتخب من الشعب يؤسس لبناء مصر الجديده وسينعكس ايجاب ا على باقي الدول العربيه. ونوهوا بما وصفوه ب »الحمل الثقيل« الذي ورثه مرسي، الذي لن يتمكن دون تكاتف كل القوى السياسيه في مصر، من بناء مصر الجديده على قواعد جديده تضمن لجميع المواطنين المساواه في الفرص وتحقيق العداله الاجتماعيه واعاده عجله التنميه الى الدوران في البلاد.

فمن جانبه راى الدكتور عبد الحميد الانصاري عميد كليه الشريعه والقانون السابق في جامعه قطر ان المهمه الاساس للرئيس الجديد تحقيق الاصلاح وتنفيذ الوعود التي قطعها خلال حملته الانتخابيه. مضيفا ان تحقيق الاستقرار في مصر سينعكس ايجاب على الوضع المصري الداخلي وبالضروره ان ذلك سينعكس على الدول العربيه. واكد الدكتور الانصاري ان انتخاب الرئيس الجديد تم بصوره ديمقراطيه وان صندوق الاقتراع هو من حسم هويه الفائز وان على جميع المصريين ان يتقبلوا هذه النتيجه ويتقبلوا نتيجه صندوق الاقتراع. فالفائز اصبح رئيسا لجميع المصريين مهما كانت انتماءاتهم السياسيه، وفي المقابل فان على الرئيس الفائز والذي كان مرشحا لجماعه الاخوان المسلمين ولحزب الحريه والعداله ان يمارس مهامه كرئيس لمصر وليس رئيسا لحزب والا يفرق بين ابناء الوطن حسب انتماءاتهم السياسيه او الدينيه. وكرر التاكيد على ضروره ان يبادر الرئيس الجديد بتحقيق الاصلاح الداخلي والتركيز على قضيه الاقتصاد الذي وصل في مصر الى مرحله الحضيض وان يعمل على دفع المصريين لمزيد من الانتاج والعمل لانقاذ اقتصاد البلاد المتهاوي. من جهتها عبرت الكاتبه الدكتوره موزه المالكي عن سعادتها بنجاح الثوره المصريه ونجاح اول تجربه ديمقراطيه حقيقيه في العالم العربي من حيث الاحتكام الى صندوق الاقتراع والتاكيد على الديمقراطيه وسياده القانون. وتمنت التزام الرئيس الجديد بفترته المحدده باربع سنوات وعدم السعي الى تغيير القانون وان يعود للشعب المصري مره اخرى ان اراد تجديد ولايته لرئاسه ثانيه. وقالت د. المالكي ان المهام التي تنتظر الرئيس الجديد خاصه على الصعيد الداخلي شاقه وصعبه فهو ورث تركه ثقيله وهما ثقيلا والمهمه الملحه امامه هي اعاده ترتيب البيت الداخلي.

واعتبر عضو المجلس البلدي السابق ابراهيم ال ابراهيم ان فوز الدكتور محمد مرسي بالرئاسه اثبت نجاح مصر في اجتياز المرحله الصعبه التي مرت بها ولا تزال تمر بها وجنبها الوقوع في مزالق جديده. وراى ان المرحله المقبله بعد فوز المرشح الاسلامي الدكتور محمد مرسي لن تكون سهله، لكنه رجح قدره الرئيس الجديد على تجاوزها وبناء الانموذج الذي نامل ان يمتد تاثيره الى جميع الدول العربيه. واضاف ان نجاح الثوره في تحقيق اهدافها في مصر سينعكس ايجاب ا على جميع الدول العربيه وعلى البلدان العربيه التي شهدت ثورات الربيع العربي معربا عن الامل في ان تتغير الامور في العالم العربي نحو الافضل.

وراى الكاتب الصحفي عيسى ال اسحاق ان فوز الدكتور محمد مرسي بالرئاسه في مصر جنب الشعب المصري وقوع حرب اهليه في البلاد. وقال ال اسحاق نحن نامل ان ينجح الرئيس الجديد في انتشال مصر من الازمات التي تمر بها، وان يفي بوعوده في تشكيل حكومه وحده وطنيه تضم كافه اطياف المجتمع المصري. معتبرا ان ذلك هو التحدي الاساس امامه الان.واضاف ان جماعه الاخوان المسلمين طرحت خلال العقود الماضيه شعار »الاسلام هو الحل« وهم وصلوا الى سده الحكم في مصر الان وهي مطالبه بتحويل هذا الشعار الى برامج اجتماعيه واقتصاديه وسياسيه ليرى الشعب المصري نتائجه على ارض الواقع.ودعا ال اسحاق الرئيس المصري الجديد الى التركيز على القضايا الداخليه المصريه. مؤكدا ان قوه مصر وقوه الشعب المصري ستنعكس ايجابا بالضروره على الوضع العربي باكمله.

واعتبر الدكتور ربيعه الكواري ان تجربه الانتخابات الرئاسيه في مصر كانت ناجحه بدليل فوز مرشح الاخوان المسلمين. وقال الدكتور ال كواري ان التغيير والتجديد مطلوب وان على الرئيس الجديد ان يسعى في هذه المرحله الى تحقيق مبدا العداله الاجتماعيه بين المواطنين والقضاء على مشاكل الفقر والبطاله واعاده الحياه للاقتصاد المصري وان ينفذ البرنامج الذي انتخبه الشعب المصري على اساسه.

من جهته دعا الدكتور عيسى مطر الاستشاري في مؤسسه حمد الطبيه الرئيس الجديد الى ان يكون ملتصقا بهموم ومشاكل الناس وان ينزل الى الشارع ليسمع مطالبهم وقضاياهم. معبرا عن الامل ان يكون في نجاح الرئيس الجديد خير لمصر وللامه العربيه والاسلاميه. فوز مرسي يؤسس لبناء مصر الجديده دعوا الرئيس المصري لتحقيق الامن والاستقرار.

Figure 5-7: Arabic News Article Example

Figure 5-8 shows the template to be used for the generation of Arabic news articles summaries and their attributes. This figure shows an Arabic/English template for better illustration.

{Title} : {العنوان}
{Publication Date} : {تاريخ النشر}
{Writer} : {الكاتب}
{People Mentioned} : {الأشخاص المشار اليهم}
{Organizations Mentioned} : {المؤسسات المشار إليها}
{General Topical Category} : {التصنيف العام}
{Words in Main Topic} : {الكلمات في الموضوع الرئيسي}

Figure 5-8: Example of an Arabic/English Empty News Article Template

Figure 5-9 shows an Arabic news article summary using a filled-in template. We collected and extracted the attributes' information from the original article shown in Figure 5-7, and then filled the empty template shown in Figure 5-8 (the Arabic part) with that information.

{العنوان}: تحقيق الامن والاستقرار واعاده بناء مصر من جديد على اساس المساواه والعداله الاجتماعيه
{تاريخ النشر}: ٢٥/آغسطس/٢٠١٢
{الكاتب}: أنور الخطيب
{ الأشخاص المشار اليهم}: محمد مرسي, عبد الحميد الانصاري, موزه المالكي, ابراهيم ال ابراهيم, عيسى ال اسحاق, ربيعه الكواري
{ المؤسسات المشار إليها}: رئيس مصر, جامعه قطر, حزب الحريه والعداله , جماعه الاخوان ا لمسلمين, مؤسسه حمد الطبيه
{التصنيف العام}: السياسية
{الكلمات في الموضوع الرئيسي}: حكم, محمد, مرسي, الاخوان, المسلمين, مصر, الرئيس، والاستقرار ، السياسية، رئيس

Figure 5-9: Example of a Filled in News Article Template in Arabic

For more clarification, Figure 5-10 shows an English translation for the filled in template results shown in Figure 5-9.

{Title}: Achieve security and stability and the rebuilding of new Egypt on the basis of equality and social justice

{Publication Date}: 25 August 2012

{Writer}: Anwar Al-Khateeb

{People Mentioned}: Mohammed Mursi, Abdul Hamid Ansari, Moza al-Maliki, Ibrahim Al-Ibrahim, Isa Al Isaac, Rabia Al-Kuwari

{Organizations Mentioned}: Qatar University, Freedom and Justice Party, The Muslim Brotherhood, Hamad Medical Corporation

{General Topical Category}: Politics

{Words in Main Topic}: Leader, Muhammad, Morsi, Brotherhood, Muslims, Egypt, President, stability, political, Chairman

Figure 5-10: Example of a Filled in News Article Template in English

5.3.5 Summarization Results and Evaluation

For our summary evaluation, we engaged a group of graduate students as volunteer participants, selecting those who understand Modern Standard Arabic. This group helped evaluate the summaries. Toward that, we provided each participant with a number of articles, along with the corresponding summaries (that we generated automatically). We asked them to read the article first, then read the summary, and evaluate the quality of the summary using a Likert (rating) scale. We asked each participant to assign a rating for each summary based on its quality, between 0-10, with 0 for extremely bad quality or even irrelevant summary, and 10 for excellent quality or best summary. Two different participants evaluated each summary/article pair, and then we averaged the results. After collecting the data from participants, we assessed the quality of our summaries and thus of our template method.

Figure 5-11 and Table 5-15 show the evaluation results of one thousand articles. Articles are divided into eleven categories (0-10), based on their ratings. We had ten participants;

each evaluated two hundred random articles. Each of the one thousand articles was evaluated by two different participants. We averaged the evaluation score of each article and counted the frequency of each score. As explained in the previous paragraph, receiving a high summary relevance score indicates that the summary is more relevant to the article. Table 5-15 demonstrates that the majority of the articles score around 8. From the figure below, it can be concluded that applying the template summarization methods over Arabic news articles leads to achieving very good results in terms of generating relevant summaries.

Table 5-15: Number of Articles for each Score

Rate Value	0	1	2	3	4	5	6	7	8	9	10
Number of Articles	0	0	0	4	24	99	269	300	237	65	2

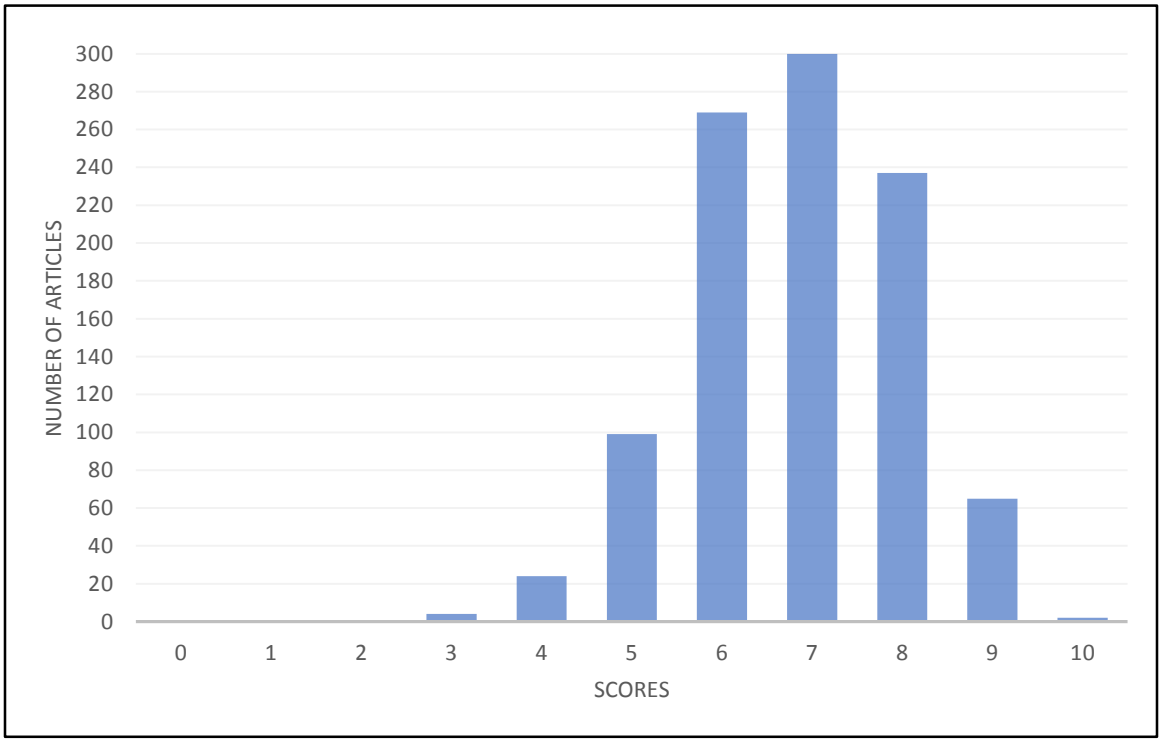

Figure 5-11: 11 Categories Used to Show the Summary Evaluation Results

5.4 Conclusion

The number of open source natural language processing tools, software packages, and resources for Arabic is insufficient. Furthermore, natural language processing research involving the Arabic language is relatively difficult due to the language complexity when compared to other popular languages. These matters inspired this research. There is no substantial research addressing the extraction of key features from Arabic news articles toward building template summaries.

In this study, we use a developed Arabic Named Entity Recognizer that can extract, with good accuracy, the named entities from news articles. We also use an Arabic topic generation tool to generate topics from news articles. Regular expressions, simple text extraction, and machine learning methods are used to extract dates, titles, and categories as key elements from the Arabic news articles. All these attributes are put together to construct a template summary for the news article.

An experiment, with graduate students who understand Arabic, helped to evaluate our extraction methods and summaries. Using a Likert scale for assessment with our Arabic news article data, evaluation results confirmed that our methods, used to extract key elements from articles, generate highly relevant template summaries.

5.5 Future Work

Our future plan is to enhance the quality of our summaries. We will run more evaluation and a failure analysis on the results, trying to determine which attributes give the least accurate values (and so reduce the quality of the summaries). After figuring out the weakest attributes, we can study how to enhance the extraction of these attributes. Finally, we will generate new summaries after enhancing some of the attributes, and again will evaluate the quality.

Chapter 6: Conclusions and Future Work

Natural Language Processing (NLP) research involving the Arabic language is relatively hard due to many reasons, like the language complexity. Free Arabic NLP tools and resources for research are relatively rare, compared to other popular languages, such as English. Moreover, online Arabic news articles are not consistently categorized; therefore, they are hard to browse by category when accessed in an aggregate collection rather than a site. Taxonomies used by a particular news service are not general enough to apply to other news service collections. Indeed, what would be the best method for classifying Arabic news stories according to a given taxonomy is still unknown. Nevertheless, preprocessing steps are supposed to enhance the classification process, while stemming is supposed to be part of these preprocessing steps. With the aim to enhance Arabic information retrieval and natural language processing, a standardized Arabic categorization system (taxonomy) is developed in this study to support browsing services for online Arabic newspapers, using the same hierarchy for the classification of 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 types of classifiers and used a newly developed stemmer (i.e., P-Stemmer), a modified version of one of the Larkey light stemmers that we hypothesized would enhance Arabic text classification. After that, classification experiments were run 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, i.e., the five Larkey stemmers, as well as the original raw words. We found that using our proposed stemmer significantly enhanced classification results for Arabic textual data, when using any of three types of classifiers: Naïve Bayes, SVM, and Random Forest. We found that SVM performed better than the other two types of classifiers. We also found that using binary classification gave better results compared to multiclass classification. We did a Wilcoxon signed-rank test to check if the observed improvements with P-stemmer were statistically significant, and concluded that they were.

For the purposes of this study, project based learning (PBL) was applied in a computational linguistics class to help students learn how to build automatic text summaries for big collections, using different methods, to produce multiple types of summaries. Results demonstrated that 30 students, distributed among seven teams, were able to learn and apply big data and computational linguistics methods to produce reasonable corpus summaries. Through active learning and PBL, students generally unfamiliar with computational linguistics, or with using a Hadoop cluster to handle large digital library collections, mastered a broad range of valuable skills. This, in fact, inspired the idea of template summarization for Arabic news articles, given the positive feedback received from students, and the results of the teaching assistants' review of student deliverables. With this in mind, we offer our approach, corpora, and course details to those interested in working with big data summarization.

There is no substantial research addressing the extraction of named entities and generation of topics from Arabic news articles. In this study, we aimed at developing a Named Entity Recognizer able to extract, with good accuracy, the named entities from Arabic news articles, RenA. A popular topic extraction model, LDA, was modified to enable it to handle and generate topics from Arabic news articles, ALDA. Given the lack of free resources for a judged news article corpus, a focus of this study was the building of a corpus to be used with RenA and ALDA evaluations, and later by other researchers. To assess performance, the help of graduate students fluent in Arabic was sought, and information retrieval evaluation measures were used to evaluate and compare our RenA NER with another NER available through the LingPipe toolkit. Three types of named entities are extracted: Person, Organization, and Location. Results indicate that the use of RenA enhances the named entity extraction results for the three mentioned types of entities, eventually producing better results than the LingPipe alternative. To evaluate our ALDA tool, a second experiment was conducted with graduate students who understand Arabic. Using a Likert scale for assessment with our Arabic news article corpus, evaluation results confirmed that our developed tool generates highly relevant topics.

Research addressing the extraction of key features from Arabic news articles for building template summaries is quite limited. As mentioned earlier, a developed Arabic Named

Entity Recognizer that can extract, with good accuracy, the named entities from news articles is used in this study, along with an Arabic topic generation tool to generate topics from news articles. Regular expressions, simple text extraction, and machine learning methods are used to extract dates, titles, and categories from Arabic news articles. All these attributes are combined to generate a template summary for each news article. The study also involves another experiment completed with help from graduate students who understand Arabic, to evaluate our extraction methods and summaries. Using a Likert scale for assessment with our Arabic news article data, evaluation results confirmed that our methods can extract key elements from articles and generate highly relevant template summaries.

Our overall contribution to the research were:

1. A developed standardized taxonomy that helps with browsing any Arabic newspaper
2. A newly developed Arabic stemmer that helps enhance Arabic text classification
3. A named entity recognizer for Arabic language that can extract three types of named entities (Location, Person, and Organization)
4. A topic identification tool using LDA algorithm that helps generate topics from Arabic news articles
5. Finally, an automated way to extract key information from Arabic news articles and fill in a developed template toward generating automatic template summaries for the articles.

In the future, we plan to test our stemmer with another data set to determine whether results are similar to those generated when using our data set. We also plan to apply the stemmer and classification methods on this new data set to confirm our findings. Using different feature selection methods, e.g., Chi-square based, to see if they enhance classification results, will also be part of our future research. We have future plans for the ALDA and RenA tools too, for example, to expand this research by using the RenA and ALDA results to fill in template summaries for other types of Arabic documents. We also plan to extract more attributes from the articles to fill in templates towards generating improved Arabic

news article summaries. In addition, we aim to use another Arabic stemmer and compare the ALDA results with our proposed P-Stemmer and the other stemmer. Another future plan in relation to template summaries is to enhance the outcomes of this research by enhancing the quality of our summaries. We plan to run more evaluation and analysis on the results, trying to determine which attributes give the least accurate values (and so reduce the quality of the summaries). After finding the weakest attributes, we plan to study how to enhance the value extraction of these attributes. Finally, we will produce new summaries after enhancing some of the attributes, and then will evaluate the quality of the new results.

REFERENCES

Abu Ata, B. M. & Al-Omari, A. (2014). “A Rule-Based Stemmer for Arabic Gulf Dialect”. Journal of King Saud University - Science 09/2014; 50(2). DOI:10.1016/j.jksuci.2014.04.003.

Abuleil, S., & Evens M. (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.

Abuleil, S., & Evens, M., (2004). “Extracting Names from Arabic Text for Question-Answering Systems”. In Proceedings of RIAO'2004, pp. 638–647, France. 2004.

Aggarwal, C., & Zhai, C. (2012). “Mining text data”. Springer Science & Business Media”, 533 pages book. ISBN: 978-1-4614-3222-7

Alguliev, R.M., Aliguliyev, R.M., & Mehdiyev, C.A., (2011). “Sentence selection for generic document summarization using an adaptive differential evolution algorithm”. Swarm and Evolutionary Computation. 1(4): pp. 213-222.

Al-Kabi, M. (2013). “Towards improving Khoja rule-based Arabic stemmer”. In Proceedings of Applied Electrical Engineering and Computing Technologies (AEECT) Conference, Amman-Jordan, pp. 1-6.

Al-Kabi M. N., Kazakzeh S. A., Abu Ata B. M., Al-Rababah S. A., & Alsmadi I. M. (2014). “A Novel Root Based Arabic Stemmer”. Journal of King Saud University (Computer Information Sciences). Volume 27, Issue 2, April 2015, pp. 94-103, ISSN 1319-1578, http://dx.doi.org/10.1016/j.jksuci.2014.04.001.

Allan, J., Gupta, R., & Khandelwal, V., (2001). “Topic models for summarizing novelty”. In Proc. ARDA Workshop on Language Modeling and Information Retrieval. Pittsburgh, Pennsylvania, USA.

Al-Omari, A., & 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-Sarhan, H., Al-Shalabi, R., & Kanaan, G. (2003). “New approach for extracting Arabic roots”. In Proceedings of the 2003 Arab Conference on Information Technology (ACIT 2003), pp. 42-59. Egypt.

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

Apache Hadoop (2015). Welcome to Apache Hadoop! [Cited January, 2015], from http://hadoop.apache.org/

Apache Mahout, (2015). Latent Dirichlet Allocation. [Cited January, 2015], from https://mahout.apache.org/users/clustering/latent-dirichlet-allocation.html

Apache Mahout, (2015). K-Means clustering – basics. [Cited January, 2015], from https://mahout.apache.org/users/clustering/k-means-clustering.html.

Apache Software Foundation, (2013). “Classic Arabic Analyzer”. [Cited 03/28/2015]. Available from: http://lucene.apache.org/core/4_6_0/analyzers-common/org/apache/lucene/analysis/ar/ArabicAnalyzer.html

Benajiba, Y., (2009). “Arabic named entity recognition”, Ph.D. dissertation. Universidad Politécnica de Valencia. Valencia, Spain. http://users.dsic.upv.es/~prosso/resources/BenajibaPhD.pdf

Bender, E.M., (2013). “Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax”. Synthesis Lectures on Human Language Technologies, 6(3): 1-184. Morgan & Claypool Publishers, California (USA).

Bird, S., Klein, E., & Loper, E., (2009). “Natural language processing with Python”: O'Reilly Media, Inc. California (USA).

Blei, D. M., (2012). “Probabilistic topic models”. Communications of the ACM 55, no. 4 (2012), pp. 77-84. DOI:10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). “Latent Dirichlet allocation”. Journal of Machine Learning Research, 3, pp. 993-1022.

Brahmi, A., Ech-Cherif, A., & Benyettou, A., (2012). “Arabic texts analysis for topic modeling evaluation”. Journal Information Retrieval. 15(1): 33-53. DOI:10.1007/s10791-011-9171-y

Brank, J., Grobelnik, M., & Mladenić, D., (2005). “A survey of ontology evaluation techniques”. In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD). October 17, 2005, Ljubljana, Slovenia.

Brants, T., Chen, F., & Farahat, A., (2002). “Arabic document topic analysis”. LREC-2002 workshop on Arabic language resources and evaluation, Las Palmas, Spain. 2002.

Buck Institute for Education, (2015). Why Project Based Learning (PBL)? [Cited January, 2015], from http://bie.org/

Chambers, N., & Jurafsky, D., (2011). “Template-based information extraction without the templates”. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol 1, Stroudsburg, PA, USA.

Darwish, K., (2013). “Named Entity Recognition using Cross-lingual Resources: Arabic as an Example”. In Proceedings of the 51st Annual Meeting of the Association of Computational Linguistics (ACL), pp. 1558-1567. Sofia, Bulgaria, August 4-9, 2013.

Douzidia, F.S., & Lapalme, G. (2004). “Lakhas, an Arabic summarization system”. In Proceedings of Document Understanding Conference, DUC2004.

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

El-Haj, M.O., & Hammo, B.H. (2008). “Evaluation of query-based Arabic text summarization system”. In Proceedings of International Conference in Natural Language Processing and Knowledge Engineering. NLP-KE'08. Beijing, China.

El-Halees, A. (2006). “Mining Arabic association rules for text classification”. In Proceedings of the First International Conference on Mathematical Sciences, pp. 157-167.

ELISQ. Electronic Library Institute SeerQ. [Cited 10/15/2014]; available from: http://elisq.qu.edu.qa

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 Application”s. Poster Session of the 3rd International Semantic Web Conference, ISWC. Pp. 640-654.

Fox, E. A., Akbar, M., Abdelhamid, S. H. E. M., Elsherbiny, N. I., Farag, M. M. G., Jin, F., Leidig, J. P., & Neppali, S. T. (2014). “Digital Libraries”. Section 3, Ch. 18 in Computing Handbook, Third ed., vol. 2, Chapman & Hall/CRC Press, Taylor and Francis Group.

Fox, E. A., & Leidig, J. P. (2014). “Digital Library Applications: CBIR, Education, Social Networks, eScience/Simulation, and GIS”. Morgan & Claypool Publishers, San Francisco.

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

GitHub, Hyunjong Lee, (2014). “LatentDirichletAllocation- Implementation of Latent Dirichlet Allocation in C# (CSharp)”. Feb. 7, 2014. [Cited on 02/15/2015]. https://github.com/hyunjong-lee/LatentDirichletAllocation

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). ”Summarizing text documents: sentence selection and evaluation metrics”. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. Berkeley, USA.

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

Goncalves, M. A., Fox, E. A., & Watson, L. T. (2014). “Towards a Digital Library Theory: A Formal Digital Library Ontology”. International Journal Digital Libraries. 8(2): 91-114. DOI: 10.1007/s00799-008-0033-1.

Google Code, (2014). “Stop-Words Project- collection of stopwords in 29 languages”. Feb. 24, 2014. [Cited 03/28/2015]. Available from: https://code.google.com/p/stop-words. (This stopword list has been used for research purposes only)

Griffiths, T., & Steyvers, M., (2004). “Finding scientific topics”. Proceedings of the National Academy of Sciences 101, no.1 (2004): pp. 5228-5235. DOI:10.1073/pnas.0307752101.

Habash, N., (2010). “Introduction to Arabic natural language processing”. Synthesis Lectures on Human Language Technologies 3, no. 1 (2010): 1-187. Morgan & Claypool Publisher, California. USA.

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

Hofmann, T., (1999). “Probabilistic latent semantic analysis”. In Proceedings of the fifteenth conference on uncertainty in artificial intelligence, pp. 289-296. Stockholm Sweden, July 30 - August 1, 1999.

Hovy, E., & Lin, C.Y. (1998). “Automated text summarization and the SUMMARIST system”. TIPSTER TEXT PROGRAM PHASE III: In Proceedings of a workshop held at Baltimore, Maryland: October 13-15, 1998, p. 197-214. Association for Computational Linguistics (ACL).

Internet Archive, Heritrix, Internet Archive Web Crawler. 9 June 2011. [Cited 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 Proceedings of Innovations in Information Technology, IIT, pp. 312-316.

Kanaan, G., Al-Shalabi, R., Al-Zamil, M., & Saifan, 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., Al-Shalabi, R., Ghwanmeh S., & Al-Ma'adeed H. (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.

Kanaan, G., Al-Shalabi, R., & Sawalha, M. (2003). “Full automatic Arabic text tagging system”. In Proceedings of the International Conference on Information Technology and Natural Sciences, Amman/Jordan, pp. 258-267.

Kanaan, G., Hammouri, A., Al-Shalabi, R., & Swalha, M., (2009). “A new question answering system for the Arabic language”. American Journal of Applied Sciences. 6(4): p. 797.

Khoja, S., & Garside, R., (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. p. 79

Kim, S.N., Medelyan, O., Kan, M.Y., & Baldwin, T., (2010). “Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles”. In Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics. 15-16 July 2010. Uppsala, Sweden.

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

Larkey, L.S., Feng, F., Connell, M., & Lavrenko, V., (2004). “Language-specific models in multilingual topic tracking”. In Proceedings of SIGIR 2004, pp. 402-409. Sheffield, UK.

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. Vol. 91, pp. 312-318

Li, A., (2013). “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/digitalstrategies/222187/how-taxonomies-help-news-organizations-understand-and-categorize-their-content./

Library of Congress, (2015). Library of Congress Subject Headings. Washington, D.C. [Cited June, 2015], from http://id.loc.gov/authorities/subjects.html

LingPipe, a toolkit for processing text using computational linguistics. Alias-i. LingPipe 4.1.0. (2008). [Cited 02/15/2015]. http://alias-i.com/lingpipe.

Liu, S., Zhou, M., Pan, S., Qian, W., Cai, W., & Lian, X., (2009). “Interactive topic-based visual text summarization and analysis”, in Proceedings of the 18th ACM conference on Information and knowledge management. ACM: Hong Kong, China, pp. 543-552.

Ma, L., Goharian, N., Chowdhury, A., & Chung, M., (2003). “Extracting unstructured data from template generated web documents”. In Proceedings of the 12th international conference on Information and knowledge management. New Orleans, LA, USA. November 3-8,

Manning, C. D., Raghavan, P., & Schutze, H., (2008). “Text classification and Naïve Bayes”, Chapter 13, pages 234-265, in Introduction to Information Retrieval, Cambridge, Vol. 1: Cambridge University Press.

Mark, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I., (2009). “The WEKA Data Mining Software: SIGKDD Explorations”, Volume 11, Issue 1. http://www.cs.waikato.ac.nz/ml/weka

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

Miller, G. A., (1956). “The magical number seven, plus or minus two: Some limits on our capacity for processing information”. Psychological Review 63 (2): 81–97. DOI: 10.1037/h0043158

Murthy, U., Fox, E., Ramakrishnan, N., Kavanaugh, A., Sheetz, S., Shoemaker, D., & Srinivasan, V., (2009). “Building an Ontology for Crisis, Tragedy, and Recovery”. NKOS Workshop, ECDL 2009, 1 Oct. 2009, Corfu, Greece.

Nationsonline, (2015). “Most widely spoken Languages in the World”. [Cited 03/28/2015]. Available from: http://www.nationsonline.org/oneworld/most_spoken_languages.htm

Neto, J.L., Freitas, A.A., & Kaestner, C.A., (2002). “Automatic Text Summarization Using a Machine Learning Approach”. In Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence, p. 205-215. Springer Berlin Heidelberg.

Oard, D., & Gey, F.C. (2002). “The TREC-2002 Arabic/English CLIR track”. In TREC2002 notebook, pp. 81-93. Gaithersburg, Maryland, USA. http://trec.nist.gov/pubs/trec11/papers/OVERVIEW.gey.ps.gz

OCLC, (2015). Dewey Services: “Organize your materials with the world’s most widely used library classification system”. Dublin, Ohio. [Cited June, 2015], from https://www.oclc.org/dewey.en.html

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

Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J., (2007). “Topic discovery based on text mining techniques”. Information Processing & Management. 43(3): 752-768.

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

Raghavan, A., (2005). “Schema Mapper: A Visualization Tool for Incremental Semi-automatic Mapping-based Integration of Heterogeneous Collections into Archaeological Digital Libraries: The ETANA-DL Case Study”, May 2005, MS thesis, http://scholar.lib.vt.edu/theses/available/etd-05182005-114155/

Raghavan, A., Vemuri, N. S., Shen, R., Goncalves, M. A., Fan, W., & Fox, E. A., (2005). “Incremental, Semi-automatic, Mapping-Based Integration of Heterogeneous Collections into Archaeological Digital Libraries: Megiddo Case Study”. In Proceedings ECDL2005, Vienna, Sept. 18-23, 2005, 139-150, http://dx.doi.org/10.1007/11551362_13, http://fox.cs.vt.edu/talks/2005/20050919ECDLmegiddo.ppt

Richardson, R., (2007). “Using Concept Maps as a Tool for Cross-Language Relevance Determination”. PhD Dissertation in the Computer Science Department at Virginia Tech. http://scholar.lib.vt.edu/theses/available/etd-07022007-184525/

Saad, M. (2011). “Arabic Text Classification”. Lap Lambert Academic Publishing. 172 pages. ISBN-10: 3844319573.

Savoy, J. Universite de Neuchatel (UniNE), (2015). “IR multilingual Resources-Arabic Stopword List”. [Cited 03/28/2015]. Switzerland. Available from: http://members.unine.ch/jacques.savoy/clef/index.html. (This stopword list has been used for research purposes only)

Schlesinger, J.D., O’Leary, D.P., & Conroy J.M., (2008). “Arabic/English multi-document summarization with CLASSY- the past and the future”. In Computational Linguistics and Intelligent Text Processing. Editor: A. Gelbukh, Springer Berlin Heidelberg, p. 568-581.

Shaalan, K., & Raza, H., (2009). “NERA: Named entity recognition for Arabic”. Journal of the American Society for Information Science and Technology. 60(8): 1652-1663.

SLA (Special Library Association), (2015). SLA Taxonomy Division. [Cited June, 2015], from http://taxonomy.sla.org/

Smucker, M. D., Allan, J., & Carterette, B., (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.

SoruceForge, Al-Khalil Morphological System, (2011). “Al-Khalil Arabic Stemmer”. Feb. 21, 2011. [Cited 02/15/2015]. http://alkhalil.sourceforge.net/

Srinivasan, V., & Angara, P., (2014). “Classification”. Chapter 4, pages 89-103, in Edward A. Fox and Ricardo da Silva Torres, editors. Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security. Morgan & Claypool Publishers, San Francisco, March 2014, 205 pages, ISBN paperback 9781627050302, http://dx.doi.org/10.2200/S00566ED1V01Y201401ICR033

Stanford Natural Language Processing Group (2015). Stanford Named Entity Recognizer (NER). [Cited January, 2015], from http://nlp.stanford.edu/software/CRF-NER.shtml

The Center for Computational Learning Systems, Columbia University. Yasine Benajiba, (2010). “ANERCorp”. [Cited 02/15/2015]. http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

Tunkelang, D., (2009). “Faceted Search”. Morgan & Claypool Publishers, San Francisco, ISBN paperback 9781598299991, http://dx.doi.org/10.2200/S00190ED1V01Y200904ICR005

Universite de Neuchatel (UniNE), Jacques Savoy, (2015). “IR multilingual Resources-Arabic Stopword List”. [Cited 02/15/2015]. Switzerland. http://members.unine.ch/jacques.savoy/clef/index.html. [This stopword list has been used for research purposes only]

Uschold, M., & King, M., (1995). “Towards a methodology for building ontologies. In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with”. The 1995 International Joint Conference on AI (IJCAI-95). Palais de Congres Montreal, Quebec, Canada. August 20-25, 1995.

Wei, X., & Croft, W. B., (2006). “LDA-based document models for ad-hoc retrieval”, in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM: Seattle, Washington, USA, pp. 178-185.

Wikipedia, (2015). “List of languages by number of native speakers”. 27 March 2015. [Cited 03/28/2015]. Available from: http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

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

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

Yang, S., Chung, H., Lin, X., Lee, S., Chen, L., Andrew Wood, Kavanaugh, A. L., Sheetz, S. D., Shoemaker, D. J, & Fox, E. A., (2013). “PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management Through the Lens of Social Media”. Proceedings of the 10th International ISCRAM Conference. Baden-Baden, Germany, May 12-15, 2013.

Yang, S., & Magdy, M., (2014). “Ontologies”. Chapter 3, pages 63-88, in Edward A. Fox and Ricardo da Silva Torres, editors. Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security. Morgan & Claypool Publishers, San Francisco, March 2014, 205 pages, ISBN paperback 9781627050302, http://dx.doi.org/10.2200/S00566ED1V01Y201401ICR033

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

Appendix A: IRB for the Arabic NER Baseline Corpus and Evaluation Experiment

The Approval Letter

VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: February 10, 2015

TO: Edward Fox, Tarek Ghaze Kan'an, Souleiman Ibrahim Ayoub, Julia Freeman

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: Arabic Named Entitiy Recognizer

IRB NUMBER: **15-101**

Effective February 10, 2015, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the New Application request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.110 category(ies) 2**
Protocol Approval Date:	**February 10, 2015**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Date*	OSP Number	Sponsor	Grant Comparison Conducted?
02/10/2015	11119512	Qatar University	Not required (Not federally funded)

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

VirginiaTech

Institutional Review Board
Research Protocol

Once complete, upload this form as a Word document to the IRB Protocol Management System: https://secure.research.vt.edu/irb

Section 1: General Information

1.1 DO ANY OF THE INVESTIGATORS OF THIS PROJECT HAVE A REPORTABLE CONFLICT OF INTEREST? (http://www.irb.vt.edu/pages/researchers.htm#conflict)

☒ **No**
☐ **Yes,** explain:

1.2 WILL THIS RESEARCH INVOLVE COLLABORATION WITH ANOTHER INSTITUTION?

☐ **No,** go to question 1.3
☒ **Yes,** answer questions within table →

IF YES
Provide the name of the institution [for institutions located overseas, please also provide name of country]: **Amman Arab University, Amman-Jordan**
Indicate the status of this research project with the other institution's IRB: ☐ Pending approval ☐ Approved ☒ Other institution does not have a human subject protections review board ☐ Other, explain:
Will the collaborating institution(s) be engaged in the research? (http://www.hhs.gov/ohrp/policy/engage08.html) ☒ No ☐ Yes
Will Virginia Tech's IRB review all human subject research activities involved with this project? ☐ No, provide the name of the primary institution: ☒ Yes *Note: primary institution = primary recipient of the grant or main coordinating center*

1.3 IS THIS RESEARCH SPONSORED OR SEEKING SPONSORED FUNDS?

☐ **No,** go to question 1.4
☒ **Yes,** answer questions within table →

IF YES
Provide the name of the sponsor [if NIH, specify department]: QNRF
Is this project receiving federal funds? ☒ No ☐ Yes **If yes,**

1

Does the grant application, OSP proposal, or "statement of work" related to this project include activities involving human subjects that are <u>not</u> covered within this IRB application?

- ☐ No, all human subject activities are covered in this IRB application
- ☐ Yes, however these activities will be covered in future VT IRB applications, these activities include:
- ☐ Yes, however these activities have been covered in past VT IRB applications, the IRB number(s) are as follows:
- ☐ Yes, however these activities have been or will be reviewed by another institution's IRB, the name of this institution is as follows:
- ☐ Other, explain:

Is Virginia Tech the primary awardee or the coordinating center of this grant?

- ☐ No, provide the name of the primary institution:
- ☐ Yes

1.4 DOES THIS STUDY INVOLVE CONFIDENTIAL OR PROPRIETARY INFORMATION (OTHER THAN HUMAN SUBJECT CONFIDENTIAL INFORMATION), OR INFORMATION RESTRICTED FOR NATIONAL SECURITY OR OTHER REASONS BY A U.S. GOVERNMENT AGENCY?

For example – government / industry proprietary or confidential trade secret information

- ☒ **No**
- ☐ **Yes,** describe:

1.5 DOES THIS STUDY INVOLVE SHIPPING ANY TANGIBLE ITEM, BIOLOGICAL OR SELECT AGENT OUTSIDE THE U.S?

- ☒ **No**
- ☐ **Yes**

Section 2: Justification

2.1 DESCRIBE THE BACKGROUND, PURPOSE, AND ANTICIPATED FINDINGS OF THIS STUDY:

Extracting Named Entities from textual data is a very important field in the Natural Language Processing area of research. Machine learning is also an important filed of Machine Learning. Both of the previous fields are important in the artificial Intelligence / big data research area. Extracting the Named Entities and categorizing Arabic text are each a challenge because of the complexity and the nature of the Arabic language. In our ELISQ project, we are trying to build a digital library (DL) for the State of Qatar. The content of this DL will be in both the Arabic and English languages. Part of this work will be summarizing Arabic news articles for some of the news collections. Toward getting that, we need to extract the Named Entities from those articles, and assign the category of each article as well.

Toward extracting the correct Named Entities, we need to train the Named Entity Recognizer (NER) on some manually built examples. Also, to compare the result between different NERs, we need to manually build a baseline corpus. No Arabic news article corpus was found to be adequate for this purpose.
Toward automatically finding the correct category of each article, first we need to train a classifier and then the classifier will do the automatic classification. To get the classifier trained, we need first to manually assign the category of each news article for a specific number of articles, in what we called a training dataset.

In this study, we investigate the importance of building a baseline corpus for Arabic news articles to use in comparing the accuracy of different Arabic NERs. Having this corpus will help this study, and other researchers in this domain, in evaluating and comparing their NER findings.

2.2 EXPLAIN WHAT THE RESEARCH TEAM PLANS TO DO WITH THE STUDY RESULTS:

For example - publish or use for dissertation

We also study Arabic text classification toward categorizing the dataset, which also will help with building better classifiers, as well as with article summarization. We intend to publish the results, use them in evaluation studies, as well as integrate them into a dissertation.

Section 3: Recruitment

3.1 DESCRIBE THE SUBJECT POOL, INCLUDING INCLUSION AND EXCLUSION CRITERIA AND NUMBER OF SUBJECTS:

Examples of inclusion/exclusion criteria - gender, age, health status, ethnicity

Maximum 15 participants will be recruited who can read and write Modern Standard Arabic.

3.2 WILL EXISTING RECORDS BE USED TO IDENTIFY AND CONTACT / RECRUIT SUBJECTS?

Examples of existing records - directories, class roster, university records, educational records

☒ **No,** go to question 3.3
☐ **Yes,** answer questions within table

IF YES
Are these records private or public? ☐ Public ☐ Private, describe the researcher's privilege to the records:
Will student, faculty, and/or staff records or contact information be requested from the University? ☐ No ☐ Yes, provide a description under Section 14 (Research Involving Existing Data) below.

3.3 DESCRIBE RECRUITMENT METHODS, INCLUDING HOW THE STUDY WILL BE ADVERTISED OR INTRODUCED TO SUBJECTS:

A study advertisement will be emailed to potential participants who understand Arabic through some graduate student mailing lists.

3.4 PROVIDE AN EXPLANATION FOR CHOOSING THIS POPULATION:

Note: the IRB must ensure that the risks and benefits of participating in a study are distributed equitably among the general population and that a specific population is not targeted because of ease of recruitment.

We will recruit graduate students who understand Arabic because user experiences, behaviors, and task performances could be different depending on the participants' academic level and familiarity with reading and writing Arabic.

Section 4: Consent Process

For more information about consent process and consent forms visit the following link: http://www.irb.vt.edu/pages/consent.htm

If feasible, researchers are advised and may be required to obtain signed consent from each participant unless obtaining signatures leads to an increase of risk (e.g., the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting in a breach of confidentiality). Signed consent is typically not required

for low risk questionnaires (consent is implied) unless audio/video recording or an in-person interview is involved. If researchers will not be obtaining signed consent, participants must, in most cases, be supplied with consent information in a different format (e.g., in recruitment document, at the beginning of survey instrument, read to participant over the phone, information sheet physically or verbally provided to participant).

4.1 CHECK ALL OF THE FOLLOWING THAT APPLY TO THIS STUDY'S CONSENT PROCESS:

☐ Verbal consent will be obtained from participants
☐ Written/signed consent will be obtained from participants
☒ Consent will be implied from the return of completed questionnaire. Note: The IRB recommends providing consent information in a recruitment document or at the beginning of the questionnaire (if the study only involves implied consent, skip to Section 5 below)
☐ Other, describe:

4.2 PROVIDE A GENERAL DESCRIPTION OF THE PROCESS THE RESEARCH TEAM WILL USE TO OBTAIN AND MAINTAIN INFORMED CONSENT:

4.3 WHO, FROM THE RESEARCH TEAM, WILL BE OVERSEEING THE PROCESS AND OBTAINING CONSENT FROM SUBJECTS?

4.4 WHERE WILL THE CONSENT PROCESS TAKE PLACE?

4.5 DURING WHAT POINT IN THE STUDY PROCESS WILL CONSENTING OCCUR?
Note: unless waived by the IRB, participants must be consented before completing any study procedure, including screening questionnaires.

4.6 IF APPLICABLE, DESCRIBE HOW THE RESEARCHERS WILL GIVE SUBJECTS AMPLE TIME TO REVIEW THE CONSENT DOCUMENT BEFORE SIGNING:
Note: typically applicable for complex studies, studies involving more than one session, or studies involving more of a risk to subjects.

☐ Not applicable

Section 5: Procedures

5.1 PROVIDE A STEP-BY-STEP THOROUGH EXPLANATION OF ALL STUDY PROCEDURES EXPECTED FROM STUDY PARTICIPANTS, INCLUDING TIME COMMITMENT & LOCATION:

1.(5 min) Brief explanation of the experiment
2.(2 min) Brief example shown to the participants
3. (3 min) Brief explanation on how we collected the data and how we will provide it to participants
4.(5 min) Participants will spend this time to read a news article
5. (5 min) Participants will spend this time to identify the article's Named Entities and category
6. Repeat steps 4 and 5 for each article until they are done

The experiment will take approximately 120-150 minutes. It will take place wherever the participants want to do this.

5.2 DESCRIBE HOW DATA WILL BE COLLECTED AND RECORDED:

Each participant will attach a list of named entities and category that he/she finds in the news article to the end of each article and save everything in a text file format

5.3 DOES THE PROJECT INVOLVE ONLINE RESEARCH ACTIVITES (INCLUDES ENROLLMENT, RECRUITMENT, SURVEYS)?

View the "Policy for Online Research Data Collection Activities Involving Human Subjects" at http://www.irb.vt.edu/documents/onlinepolicy.pdf

☒ **No,** go to question 6.1
☐ **Yes,** answer questions within table

IF YES

Identify the service / program that will be used:

- ☐ www.survey.vt.edu, go to question 6.1
- ☐ Blackboard, go to question 6.1
- ☐ Center for Survey Research, go to question 6.1
- ☐ Other

IF OTHER:

Name of service / program:
URL:
This service is…

- ☐ Included on the list found at: http://www.irb.vt.edu/pages/validated.htm
- ☐ Approved by VT IT Security
- ☐ An external service with proper SSL or similar encryption (https://) on the login (if applicable) and all other data collection pages.
- ☐ None of the above (note: only permissible if this is a collaborative project in which VT individuals are only responsible for data analysis, consulting, or recruitment)

Section 6: Risks and Benefits

6.1 WHAT ARE THE POTENTIAL RISKS (E.G., EMOTIONAL, PHYSICAL, SOCIAL, LEGAL, ECONOMIC, OR DIGNITY) TO STUDY PARTICIPANTS?

Since the particiants will deal with electronic versions of news articles that been collected from online resources, we think that there are no potiential risks for this study, other than fatigue from reading and analyzing what they read

6.2 EXPLAIN THE STUDY'S EFFORTS TO REDUCE POTENTIAL RISKS TO SUBJECTS:

We will allow the participants to split up the work over several sessions, to reduce fatigue

6.3 WHAT ARE THE DIRECT OR INDIRECT ANTICIPATED BENEFITS TO STUDY PARTICIPANTS AND/OR SOCIETY?

The study will help with building a judged corpus for Arabic news articles and that will help the researcher community interested in this field of study. Also it will help build a classifier training dataset.

Section 7: Full Board Assessment

7.1 DOES THE RESEARCH INVOLVE MICROWAVES/X-RAYS, OR GENERAL ANESTHESIA OR SEDATION?

☒ **No**
☐ **Yes**

7.2 DO RESEARCH ACTIVITIES INVOLVE PRISONERS, PREGNANT WOMEN, FETUSES, HUMAN IN VITRO FERTILIZATION, OR MENTALLY DISABLED PERSONS?

☒ **No,** go to question 7.3
☐ **Yes,** answer questions within table →

IF YES
This research involves:
☐ Prisoners
☐ Pregnant women ☐ Fetuses ☐ Human in vitro fertilization
☐ Mentally disabled persons

7.3 DOES THIS STUDY INVOLVE MORE THAN MINIMAL RISK TO STUDY PARTICIPANTS?

Minimal risk means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily activities or during the performance of routine physical or psychological examinations or tests. Examples of research involving greater than minimal risk include collecting data about abuse or illegal activities. Note: if the project qualifies for Exempt review (http://www.irb.vt.edu/pages/categories.htm), it will not need to go to the Full Board.

☒ **No**
☐ **Yes**

IF YOU ANSWERED "YES" TO ***ANY ONE*** OF THE ABOVE QUESTIONS, 7.1, 7.2, OR 7.3, THE BOARD MAY REVIEW THE PROJECT'S APPLICATION MATERIALS AT ITS MONTHLY MEETING. VIEW THE FOLLOWING LINK FOR DEADLINES AND ADDITIONAL INFORMATION: http://www.irb.vt.edu/pages/deadlines.htm

Section 8: Confidentiality / Anonymity

For more information about confidentiality and anonymity visit the following link: http://www.irb.vt.edu/pages/confidentiality.htm

8.1 WILL PERSONALLY IDENTIFYING STUDY RESULTS OR DATA BE RELEASED TO ANYONE OUTSIDE OF THE RESEARCH TEAM?

For example – to the funding agency or outside data analyst, or participants identified in publications with individual consent

☒ **No**
☐ **Yes,** to whom will identifying data be released?

8.2 WILL ANY STUDY FILES CONTAIN PARTICIPANT IDENTIFYING INFORMATION (E.G., NAME, CONTACT INFORMATION, VIDEO/AUDIO RECORDINGS)?

Note: if collecting signatures on a consent form, select "Yes."

☒ **No,** go to question 8.3
☐ **Yes,** answer questions within table →

IF YES
Describe if/how the study will utilize study codes:
If applicable, where will the key [i.e., linked code and identifying information document (for instance, John Doe = study ID 001)] **be stored and who will have access?** *Note: the key should be stored separately from subjects' completed data documents and accessibility should be limited.*
The IRB strongly suggests and may require that all data documents (e.g., questionnaire responses, interview responses, etc.) do not include or request identifying information (e.g., name, contact information, etc.) from participants. If you need to link subjects' identifying information to subjects' data documents, use a study ID/code on all data documents.

8.3 WHERE WILL DATA BE STORED?

Examples of data - questionnaire, interview responses, downloaded online survey data, observation recordings, biological samples

They will be stored in an external hard disk that will be locked in a cabinet in co-investigator Tarek Kanan's office

8.4 WHO WILL HAVE ACCESS TO STUDY DATA?

Only the investigator and co-investigators

8.5 DESCRIBE THE PLANS FOR RETAINING OR DESTROYING THE STUDY DATA

The hard disk will be kept for three years in a locked cabinet. After that period, the data will be properly deleted

8.6 DOES THIS STUDY REQUEST INFORMATION FROM PARTICIPANTS REGARDING ILLEGAL BEHAVIOR?

☒ **No,** go to question 9.1
☐ **Yes,** answer questions within table

IF YES
Does the study plan to obtain a Certificate of Confidentiality? ☐ No ☐ Yes (Note: participants must be fully informed of the conditions of the Certificate of Confidentiality within the consent process and form) *For more information about Certificates of Confidentiality, visit the following link:* *http://www.irb.vt.edu/pages/coc.htm*

Section 9: Compensation

For more information about compensating subjects, visit the following link: http://www.irb.vt.edu/pages/compensation.htm

9.1 WILL SUBJECTS BE COMPENSATED FOR THEIR PARTICIPATION?

☒ **No,** go to question 10.1

☐ **Yes,** answer questions within table

IF YES
What is the amount of compensation?
Will compensation be prorated? ☐ Yes, please describe: ☐ No, explain why and clarify whether subjects will receive full compensation if they withdraw from the study? *Unless justified by the researcher, compensation should be prorated based on duration of study participation. Payment must not be contingent upon completion of study procedures. In other words, even if the subject decides to withdraw from the study, he/she should be compensated, at least partially, based on what study procedures he/she has completed.*

Section 10: Audio / Video Recording

For more information about audio/video recording participants, visit the following link: http://www.irb.vt.edu/pages/recordings.htm

10.1 WILL YOUR STUDY INVOLVE VIDEO AND/OR AUDIO RECORDING?

☒ **No,** go to question 11.1
☐ **Yes,** answer questions within table

IF YES
This project involves: ☐ Audio recordings only ☐ Video recordings only ☐ Both video and audio recordings
Provide compelling justification for the use of audio/video recording:
How will data within the recordings be retrieved / transcribed?
How and where will recordings (e.g., tapes, digital data, data backups) be stored to ensure security?
Who will have access to the recordings?
Who will transcribe the recordings?
When will the recordings be erased / destroyed?

Section 11: Research Involving Students

11.1 DOES THIS PROJECT INCLUDE STUDENTS AS PARTICIPANTS?

☐ **No,** go to question 12.1
☒ **Yes,** answer questions within table

IF YES

Does this study involve conducting research with students of the researcher?

☒ No

☐ Yes, describe safeguards the study will implement to protect against coercion or undue influence for participation:

Note: if it is feasible to use students from a class of students not under the instruction of the researcher, the IRB recommends and may require doing so.

Will the study need to access student records (e.g., SAT, GPA, or GRE scores)?

☒ No

☐ Yes

11.2 DOES THIS PROJECT INCLUDE ELEMENTARY, JUNIOR, OR HIGH SCHOOL STUDENTS?

☒ **No,** go to question 11.3

☐ **Yes,** answer questions within table →

IF YES

Will study procedures be completed during school hours?

☐ No

☐ Yes

If yes,

Students not included in the study may view other students' involvement with the research during school time as unfair. Address this issue and how the study will reduce this outcome:

Missing out on regular class time or seeing other students participate may influence a student's decision to participate. Address how the study will reduce this outcome:

Is the school's approval letter(s) attached to this submission?

☐ Yes

☐ No, project involves Montgomery County Public Schools (MCPS)

☐ No, explain why:

You will need to obtain school approval (if involving MCPS, click here: http://www.irb.vt.edu/pages/mcps.htm). Approval is typically granted by the superintendent, principal, and classroom teacher (in that order). Approval by an individual teacher is insufficient. School approval, in the form of a letter or a memorandum should accompany the approval request to the IRB.

11.3 DOES THIS PROJECT INCLUDE COLLEGE STUDENTS?

☐ **No,** go to question 12.1

☒ **Yes,** answer questions within table →

IF YES

Some college students might be minors. Indicate whether these minors will be included in the research or actively excluded:

☐ Included

☒ Actively excluded, describe how the study will ensure that minors will not be included: **The participants will be graduate students; we will make sure that they will not be minors.**

Will extra credit be offered to subjects?

☒ No
☐ Yes

If yes,

What will be offered to subjects as an equal alternative to receiving extra credit without participating in this study?

Include a description of the extra credit (e.g., amount) to be provided within question 9.1 ("IF YES" table)

Section 12: Research Involving Minors

12.1 DOES THIS PROJECT INVOLVE MINORS (UNDER THE AGE OF 18 IN VIRGINIA)?

Note: age constituting a minor may differ in other States.

☒ **No,** go to question 13.1
☐ **Yes,** answer questions within table

IF YES

Does the project reasonably pose a risk of reports of current threats of abuse and/or suicide?

☐ No
☐ Yes, thoroughly explain how the study will react to such reports:

Note: subjects and parents must be fully informed of the fact that researchers must report threats of suicide or suspected/reported abuse to the appropriate authorities within the Confidentiality section of the Consent, Assent, and/or Permission documents.

Are you requesting a waiver of parental permission (i.e., parent uninformed of child's involvement)?

☐ No, **both** parents/guardians will provide their permission, if possible.
☐ No, **only one** parent/guardian will provide permission.
☐ Yes, describe below how your research meets **all** of the following criteria (A-D):
Criteria A - The research involves no more than minimal risk to the subjects:
Criteria B - The waiver will not adversely affect the rights and welfare of the subjects:
Criteria C - The research could not practicably be carried out without the waiver:
Criteria D - (Optional) Parents will be provided with additional pertinent information after participation:

Is it possible that minor research participants will reach the legal age of consent (18 in Virginia) while enrolled in this study?

☐ No
☐ Yes, will the investigators seek and obtain the legally effective informed consent (in place of the minors' previously provided assent and parents' permission) for the now-adult subjects for any ongoing interactions with the subjects, or analysis of subjects' data? If yes, explain how:

For more information about minors reaching legal age during enrollment, visit the following link:
http://www.irb.vt.edu/pages/assent.htm

*The procedure for obtaining assent from minors and permission from the minor's guardian(s) must be described in **Section 4** (Consent Process) of this form.*

Section 13: Research Involving Deception

For more information about involving deception in research and for assistance with developing your debriefing form, visit our website at http://www.irb.vt.edu/pages/deception.htm

13.1 DOES THIS PROJECT INVOLVE DECEPTION?

☒ **No,** go to question 14.1
☐ **Yes,** answer questions within table

IF YES
Describe the deception:
Why is the use of deception necessary for this project?
Describe the debriefing process:
Provide an explanation of how the study meets all the following criteria (A-D) for an alteration of consent: Criteria A - The research involves no more than minimal risk to the subjects: Criteria B - The alteration will not adversely affect the rights and welfare of the subjects: Criteria C - The research could not practicably be carried out without the alteration: Criteria D - (Optional) Subjects will be provided with additional pertinent information after participation (i.e., debriefing for studies involving deception): *By nature, studies involving deception cannot provide subjects with a complete description of the study during the consent process; therefore, the IRB must allow (by granting an alteration of consent) a consent process which does not include, or which alters, some or all of the elements of informed consent.*
The IRB requests that the researcher use the title "Information Sheet" instead of "Consent Form" on the document used to obtain subjects' signatures to participate in the research. This will adequately reflect the fact that the subject cannot fully consent to the research without the researcher fully disclosing the true intent of the research.

Section 14: Research Involving Existing Data

14.1 WILL THIS PROJECT INVOLVE THE COLLECTION OR STUDY/ANALYSIS OF EXISTING DATA DOCUMENTS, RECORDS, PATHOLOGICAL SPECIMENS, OR DIAGNOSTIC SPECIMENS?

Please note: it is not considered existing data if a researcher transfers to Virginia Tech from another institution and will be conducting data analysis of an on-going study.

☐ **No,** you are finished with the application
☒ **Yes,** answer questions within table

IF YES
From where does the existing data originate? Public online resources
Provide a detailed description of the existing data that will be collected or studied/analyzed: it's an extracted collection of text from online news articles

Is the source of the data public?

☐ No, continue with the next question

☒ Yes, you are finished with this application

Will any individual associated with this project (internal or external) have access to or be provided with existing data containing information which would enable the identification of subjects:

- **Directly** (e.g., by name, phone number, address, email address, social security number, student ID number), or
- **Indirectly through study codes** even if the researcher or research team does not have access to the master list linking study codes to identifiable information such as name, student ID number, etc or
- **Indirectly through the use of information that could reasonably be used in combination to identify an individual** (e.g., demographics)

☐ No, collected/analyzed data will be completely de-identified

☐ Yes,

If yes,

Research will not qualify for exempt review; therefore, if feasible, written consent must be obtained from individuals whose data will be collected / analyzed, unless this requirement is waived by the IRB.

Will written/signed or verbal consent be obtained from participants prior to the analysis of collected data? -select one-

This research protocol represents a contract between all research personnel associated with the project, the University, and federal government; therefore, must be followed accordingly and kept current.

Proposed modifications must be approved by the IRB prior to implementation except where necessary to eliminate apparent immediate hazards to the human subjects.

Do not begin human subjects activities until you receive an IRB approval letter via email.

It is the Principal Investigator's responsibility to ensure all members of the research team who interact with research subjects, or collect or handle human subjects data have completed human subjects protection training prior to interacting with subjects, or handling or collecting the data.

---------END---------

The Recruitment Materials and Announcement Email

We are conducting a study trying to collect named entities as well as category information from Arabic news articles toward building a baseline corpus. This corpus will be used to evaluate Arabic NERs and text classifiers. We need participants to partake in our experiment. The duties of the participants will include:

1. Reading news articles
2. Identify named entities in each article
3. Recording the named entities
4. Identifying (from a small set) the right category(ies) for each article, and recording that information

With the results of the study we plan to use the recorded information to help us evaluate the accuracy or precision of Arabic NERs and text classifiers.

If you are a graduate student who can read and write Modern Standard Arabic and are over the age of 18, and would like to participate, please contact

Tarek Kanan: tarekk@vt.edu

He will respond with instructions on how to participate. Thank you!

The Consent and Instruction Form

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Informed Consent for Participants in Research Projects Involving Human Subjects

Title of Project
Extracting Named Entities and Categories from Arabic News Articles.

Investigators
Dr. Edward A. Fox, Tarek Kanan, Soulieman Ayoub, Julia Freedom

I. Purpose of this Research/Project
We would like to test our new Arabic Named Entity Recognizer tool and train a Classifer. To do that, and because we could not find a good Arabic news article corpus to use for our NER and classification experiments we decided to build a manually judged corpus. Participants will make the needed judgments, so a new research (training) corpus can be built from the aggregation of their submissions.

II. Procedures
All participants will be asked to read every news article that will be assigned to them, and manually identify the named entities and the categories from that article. In short, the task is to take each file provided, read it, add to the bottom three lists of named entities and a category, save the expanded file, and then submit all the expanded files back to the investigators, along with this document.

III. Risks
There are no more than minimal risks involved. If a participant becomes fatigued, they may pause their work, and continue in one or more subsequent sessions and days, after having an adequate rest. The entire effort should take no more than 3 hours. If that amount of time is not available, participants can submit as many files as they have time to complete.

IV. Benefits
We will acknowledge the participants as a group for their help in any publication we will develop related to this work.

V. Extent of Anonymity and Confidentiality
All data collected in this study will be anonymized. We will not collect any kind of participants' information

VI. Compensation
Participants will not receive any compensation for their help.

VII. Freedom to Withdraw
At any point in the study, participants may withdraw from the study.

VIII. Subject's Responsibilities
I agree to participate in this study for free. I will extract the named entities and categories for each news article to the best of my knowledge.

IX. Instructions

1. Please read the attached ExampleArabic.pdf file. Your goal is to prepare a file like this for each of a set of news articles.
2. You need to read each news article and extract all of the Person, Organization, and Location entities, plus the category.
3. For the category section, please assign up to your knowledge the category/ies that describe(s) each news article. Please choose the category from this list: Sports, Economics, Politics, Social Issues, Art & Culture, and Miscellaneous.
4. Please assign at least one category to each news article. In general, the category will be the main idea(s) the news article talking about; please see ExampleArabic.pdf for an illustration.
5. At the bottom of each file please write the words Person, Organization, Location, and Category, each followed by a colon ":", each on a separate line
6. For each of the four added words at the bottom of each file, please follow it with all of the correct entities and a category of the type that you find in the article, and separate them with commas
7. Please use the 'Save As" option to overwrite the original file
8. Make sure you don't use a new name for a file and that the encoding is Unicode-8

X. Subject's Permission
After reading this Consent Form, and possibly after an email exchange with the co-investigators. I have had all my questions answered. I confirm that I am not a minor. By submitting the requested files, with the requested additions, and returning this document with those files, I hereby acknowledge the above and give my voluntary consent.

A Cover Letter

This file "CoverLetter-IRB.docx" is only for IRB explanations and will not be used with the proposed study.

Dear IRB staff,

We have attached two example files as support materials for this study; the first one is "ExampleArabic.pdf" and the second one is "ExampleEnglish.pdf".

Since the study will be over Arabic news articles, the "ExampleArabic.pdf" is the one we will use and send to participants to follow.

The reason we created and attached the "ExampleEnglish.pdf" file as a support material is to help IRB understand the study.

Please let us know if you have any questions.

Many thanks,

Tarek Kanan, Co-Investigator

tarekk@vt.edu

An Arabic Example

الدوحه – انور الخطيب:
تحقيق الامن والاستقرار واعاده بناء
مصر من جديد على اساس المساواه
والعداله الاجتماعيه.. مهام ملحه ارتاى عدد
من الكتاب والمحللين السياسيين في قطر
ان على الدكتور محمد مرسي رئيس مصر
المنتخب ان يركز عليها في بدايات عهده
الجديد.. مؤكدين ان نجاح الدكتور مرسي
كاول رئيس مصري منتخب من الشعب
يؤسس لبناء مصر الجديده وسينعكس
ايجاب ا على باقي الدول العربيه. ونوهوا
بما وصفوه ب »الحمل الثقيل« الذي ورثه
مرسي، الذي لن يتمكن دون تكاتف كل
القوى السياسيه في مصر، من بناء مصر
الجديده على قواعد جديده تضمن لجميع
المواطنين المساواه في الفرص وتحقيق
العداله الاجتماعيه واعاده عجله التنميه الى
الدوران في البلاد.
فمن جانبه راى الدكتور عبد الحميد
الانصاري عميد كليه الشريعه والقانون
السابق في جامعه قطر ان المهمه الاساس
للرئيس الجديد تحقيق الاصلاح وتنفيذ
الوعود التي قطعها خلال حملته الانتخابيه..
مضيفا ان تحقيق الاستقرار في مصر
سينعكس ايجاب ا على الوضع المصري
الداخلي وبالضروره ان ذلك سينعكس على
الدول العربيه.
واكد الدكتور الانصاري ان انتخاب
الرئيس الجديد تم بصوره ديمقراطيه وان
صندوق الاقتراع هو من حسم هويه الفائز
وان على جميع المصريين ان يتقبلوا هذه
النتيجه ويتقبلوا نتيجه صندوق الاقتراع ..
فالفائز اصبح رئيسا لجميع المصريين مهما
كانت انتماءاتهم السياسيه، وفي المقابل
فان على الرئيس الفائز والذي كان مرشحا
لجماعه الاخوان المسلمين ولحزب الحريه
والعداله ان يمارس مهامه كرئيس لمصر
وليس رئيسا لحزب والا يفرق بين ابناء
الوطن حسب انتماءاتهم السياسيه او الدينيه.
وكرر التاكيد على ضروره ان يبادر
الرئيس الجديد بتحقيق الاصلاح الداخلي
والتركيز على قضيه الاقتصاد الذي وصل
في مصر الى مرحله الحضيض وان يعمل
على دفع المصريين لمزيد من الانتاج
والعمل لانقاذ اقتصاد البلاد المتهاوي. كما
راى ضروره ان يقوم الرئيس المصري
الجديد بطمانه الجميع والتاكيد على حرص
مصر في العهد الجديد على اقامه علاقات
طيبه مع جميع الدول على قاعده الاحترام
المتبادل.
من جهتها عبرت الكاتبه الدكتوره موزه

طرحت خلال العقود الماضيه شعار
»الاسلام هو الحل« وهم وصلوا الى سده
الحكم في مصر الان وهي مطالبه بتحويل
هذا الشعار الى برامج اجتماعيه واقتصاديه
وسياسيه ليرى الشعب المصري نتائجه على
ارض الواقع.
ودعا ال اسحاق الرئيس المصري الجديد
الى التركيز على القضايا الداخليه المصريه..
مؤكدا ان قوه مصر وقوه الشعب المصري
ستنعكس ايجابا بالضروره على الوضع
العربي باكمله.
واعتبر الدكتور ربيعه الكواري ان تجربه
الانتخابات الرئاسيه في مصر كانت ناجحه
بدليل فوز مرشح الاخوان المسلمين.. معتبرا
ان الشعب المصري اختار الشخص المناسب
في المكان المناسب. واكد من خلال هذا
الاختبار انه لا يمكن ان ينحاز الا للثوره.
وقال الدكتور ال كواري ان التغيير
والتجديد مطلوب وان على الرئيس الجديد
ان يسعى في هذه المرحله الى تحقيق مبدا
العداله الاجتماعيه بين المواطنين والقضاء
على مشاكل الفقر والبطاله واعاده الحياه
للاقتصاد المصري وان ينفذ البرنامج الذي
انتخبه الشعب المصري على اساسه.. مضيفا
ان قضايا التعليم والصحه والعمل وبسط
الامن والاستقرار هي القضايا الملحه التي
يجب ان يعمل عليها الرئيس الجديد.
من جهته دعا الدكتور عيسى مطر
الاستشاري في مؤسسه حمد الطبيه الرئيس
الجديد الى ان يكون ملتصقا بهموم ومشاكل
الناس وان ينزل الى الشارع ليسمع مطالبهم
وقضاياهم.. معبرا عن الامل ان يكون في
نجاح الرئيس الجديد خير لمصر وللامه
العربيه والاسلاميه.
فوز مرسي يؤسس لبناء مصر الجديده
دعوا الرئيس المصري لتحقيق الامن والاستقرار .. مواطنون:
ق وه مصر ست ن عكس اي جاب ا على ال عال م العربي
د. عبد الحميد الانصاريعيسى ال اسحاقد. موزه المالكيابراهيم الابراهيم

Person: انور ، الخطيب ، محمد ، مرسي ، عبد ، الحميد ، الانصاري ، موزه ، المالكي ، ابراهيم ، عيسى ، اسحاق ، ربيعه ، الكواري ، عيسى ، مطر

Organization: كليه الشريعه والقانون ، جامعه قطر ، حزب الحريه والعدال ، العالم العربي ، جماعه الاخوان المسلمين ، مؤسسه حمد الطبيه

Location: الدوحه , قطر , مصر

Category: سياسة
* فقط اختر احد التصنيفات ذات الصلة من القائمة { رياضة، اقتصاد ، فن

{ سياسة , قضايا اجتماعية , ثقافة

An English Example

6 February 2015
AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, “has all the qualities” required to take the helm of the club, Ancelotti told a news conference. “I enjoy Zidane’s work, he’s doing very well,” Ancelotti said.
After a difficult start of the season, Castilla are top of Spain’s third tier league. “He’s doing very well in his first year in charge. He’s taken Castilla to first place and he needs to keep up the good work.
“It’s pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid,” said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao’s reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Person: Zinedine, Zidane, Carlo, Ancelotti, Martin, Odegaard

Organization: Real, Madrid, Castilla, Athletic, Bilbao

Location: Spain, Madrid, Bilbao, Norway, Europe

Category: Sports

Appendix B: IRB for the Arabic LDA and Summary Evaluation Experiments

The Approval Letter

VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: February 16, 2015

TO: Edward Fox, Tarek Ghaze Kan'an, Souleiman Ibrahim Ayoub, Julia Freeman

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: Evaluating topical relevance and template summary quality for Arabic news articles

IRB NUMBER: **15-161**

Effective February 13, 2015, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the New Application request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.110 category(ies) 2**
Protocol Approval Date:	**February 13, 2015**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

VirginiaTech

Institutional Review Board
Research Protocol

Once complete, upload this form as a Word document to the IRB Protocol Management System: https://secure.research.vt.edu/irb

Section 1: General Information

1.1 DO ANY OF THE INVESTIGATORS OF THIS PROJECT HAVE A REPORTABLE CONFLICT OF INTEREST? (http://www.irb.vt.edu/pages/researchers.htm#conflict)

☒ **No**
☐ **Yes,** explain:

1.2 WILL THIS RESEARCH INVOLVE COLLABORATION WITH ANOTHER INSTITUTION?

☐ **No,** go to question 1.3
☒ **Yes,** answer questions within table →

IF YES
Provide the name of the institution [for institutions located overseas, please also provide name of country]: **Amman Arab University, Amman-Jordan**
Indicate the status of this research project with the other institution's IRB: ☐ Pending approval ☐ Approved ☒ Other institution does not have a human subject protections review board ☐ Other, explain:
Will the collaborating institution(s) be engaged in the research? (http://www.hhs.gov/ohrp/policy/engage08.html) ☒ No ☐ Yes
Will Virginia Tech's IRB review all human subject research activities involved with this project? ☐ No, provide the name of the primary institution: ☒ Yes *Note: primary institution = primary recipient of the grant or main coordinating center*

1.3 IS THIS RESEARCH SPONSORED OR SEEKING SPONSORED FUNDS?

☐ **No,** go to question 1.4
☒ **Yes,** answer questions within table →

IF YES
Provide the name of the sponsor [if NIH, specify department]: QNRF
Is this project receiving federal funds? ☒ No ☐ Yes **If yes,**

Does the grant application, OSP proposal, or "statement of work" related to this project include activities involving human subjects that are <u>not</u> covered within this IRB application?

- ☐ No, all human subject activities are covered in this IRB application
- ☐ Yes, however these activities will be covered in future VT IRB applications, these activities include:
- ☐ Yes, however these activities have been covered in past VT IRB applications, the IRB number(s) are as follows:
- ☐ Yes, however these activities have been or will be reviewed by another institution's IRB, the name of this institution is as follows:
- ☐ Other, explain:

Is Virginia Tech the primary awardee or the coordinating center of this grant?

- ☐ No, provide the name of the primary institution:
- ☐ Yes

1.4 DOES THIS STUDY INVOLVE CONFIDENTIAL OR PROPRIETARY INFORMATION (OTHER THAN HUMAN SUBJECT CONFIDENTIAL INFORMATION), OR INFORMATION RESTRICTED FOR NATIONAL SECURITY OR OTHER REASONS BY A U.S. GOVERNMENT AGENCY?

For example – government / industry proprietary or confidential trade secret information

☒ **No**
☐ **Yes,** describe:

1.5 DOES THIS STUDY INVOLVE SHIPPING ANY TANGIBLE ITEM, BIOLOGICAL OR SELECT AGENT OUTSIDE THE U.S?

☒ **No**
☐ **Yes**

Section 2: Justification

2.1 DESCRIBE THE BACKGROUND, PURPOSE, AND ANTICIPATED FINDINGS OF THIS STUDY:

Find the topics, and generating summaries, from textual data, is a very important field in the Natural Language Processing (NLP) area of research. NLP is important in the artificial Intelligence / big data research area. Evaluating the generated topic and summary from Arabic text is a challenge because of the complexity and the nature of the Arabic language. In our ELISQ project, we are trying to build a digital library (DL) for the State of Qatar. The content of this DL will be in both the Arabic and English languages. Part of this work will be summarizing Arabic news articles for some of the news collections. Toward getting that, we need to generate or extract the topics for those articles, and to fill in a template to serve as a summary for each article.

Toward extracting the correct article topic and summary, we created a tool that will help us find out each article's main topic and then we used this topic combined with other attributes to fill in a summary template. To evaluate our topic extractor tool and our final filled in template summary, we need to manually ask people to read each article in our testing corpus, and rate the quality of the topic and the summary, using Likert scale from 0-10.

In this study, we are testing the importance of our topic extraction tool and our automatic filled in template summaries for Arabic news articles to use in evaluating the accuracy of our results.

2.2 EXPLAIN WHAT THE RESEARCH TEAM PLANS TO DO WITH THE STUDY RESULTS:

For example - publish or use for dissertation

We intend to publish the results, use them in evaluation studies, as well as integrate them into a dissertation.

Section 3: Recruitment

3.1 DESCRIBE THE SUBJECT POOL, INCLUDING INCLUSION AND EXCLUSION CRITERIA AND NUMBER OF SUBJECTS:

Examples of inclusion/exclusion criteria - gender, age, health status, ethnicity

Maximum 15 participants will be recruited who can read and write Modern Standard Arabic.

3.2 WILL EXISTING RECORDS BE USED TO IDENTIFY AND CONTACT / RECRUIT SUBJECTS?

Examples of existing records - directories, class roster, university records, educational records

☒ **No,** go to question 3.3
☐ **Yes,** answer questions within table

IF YES
Are these records private or public? ☐ Public ☐ Private, describe the researcher's privilege to the records:
Will student, faculty, and/or staff records or contact information be requested from the University? ☐ No ☐ Yes, provide a description under Section 14 (Research Involving Existing Data) below.

3.3 DESCRIBE RECRUITMENT METHODS, INCLUDING HOW THE STUDY WILL BE ADVERTISED OR INTRODUCED TO SUBJECTS:

A study advertisement will be emailed to potential participants who understand Arabic through some graduate student mailing lists.

3.4 PROVIDE AN EXPLANATION FOR CHOOSING THIS POPULATION:

Note: the IRB must ensure that the risks and benefits of participating in a study are distributed equitably among the general population and that a specific population is not targeted because of ease of recruitment.

We will recruit graduate students who understand Arabic because user experience, behavior, and task performance could be different depending on the participants' academic level and familiarity with reading and writing Arabic.

Section 4: Consent Process

For more information about consent process and consent forms visit the following link: http://www.irb.vt.edu/pages/consent.htm

If feasible, researchers are advised and may be required to obtain signed consent from each participant unless obtaining signatures leads to an increase of risk (e.g., the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting in a breach of confidentiality). Signed consent is typically not required for low risk questionnaires (consent is implied) unless audio/video recording or an in-person interview is involved. If researchers will not be obtaining signed consent, participants must, in most cases, be supplied with consent information in a different format

(e.g., in recruitment document, at the beginning of survey instrument, read to participant over the phone, information sheet physically or verbally provided to participant).

4.1 CHECK ALL OF THE FOLLOWING THAT APPLY TO THIS STUDY'S CONSENT PROCESS:

☐ Verbal consent will be obtained from participants
☐ Written/signed consent will be obtained from participants
☒ Consent will be implied from the return of completed questionnaire. Note: The IRB recommends providing consent information in a recruitment document or at the beginning of the questionnaire (if the study only involves implied consent, skip to Section 5 below)
☐ Other, describe:

4.2 PROVIDE A GENERAL DESCRIPTION OF THE PROCESS THE RESEARCH TEAM WILL USE TO OBTAIN AND MAINTAIN INFORMED CONSENT:

4.3 WHO, FROM THE RESEARCH TEAM, WILL BE OVERSEEING THE PROCESS AND OBTAINING CONSENT FROM SUBJECTS?

4.4 WHERE WILL THE CONSENT PROCESS TAKE PLACE?

4.5 DURING WHAT POINT IN THE STUDY PROCESS WILL CONSENTING OCCUR?
Note: unless waived by the IRB, participants must be consented before completing any study procedure, including screening questionnaires.

4.6 IF APPLICABLE, DESCRIBE HOW THE RESEARCHERS WILL GIVE SUBJECTS AMPLE TIME TO REVIEW THE CONSENT DOCUMENT BEFORE SIGNING:
Note: typically applicable for complex studies, studies involving more than one session, or studies involving more of a risk to subjects.

☐ Not applicable

Section 5: Procedures

5.1 PROVIDE A STEP-BY-STEP THOROUGH EXPLANATION OF ALL STUDY PROCEDURES EXPECTED FROM STUDY PARTICIPANTS, INCLUDING TIME COMMITMENT & LOCATION:

1.(5 min) Brief explanation of the experiment
2. (3 min) Brief explanation on how we collected the data and how we will provide it to participants
3. (5 min) Participants will spend this time to read a news article with its corresponding topic and summary.
4. (5 min) Participants will spend this time to evaluate the relevance of the topic and the quality of the summary using a Likert scale (rating each of the topic relevance and summary quality on a scale of 0-10. 0 indicates a topic without relevance, or an extremely low quality summary, while 10 indicates a totally relevant topic and an extremly high quality summary.
5. Repeat steps 3 and 4 for each article until they done

The experiment will take approximately 100-120 minutes. It will take place wherever the participants want to do this.

5.2 DESCRIBE HOW DATA WILL BE COLLECTED AND RECORDED:

Each participant will attach the evaluation for the topic and the summary to the end of each article and save everything in a text file format.

5.3 DOES THE PROJECT INVOLVE ONLINE RESEARCH ACTIVITES (INCLUDES ENROLLMENT, RECRUITMENT, SURVEYS)?

View the "Policy for Online Research Data Collection Activities Involving Human Subjects" at http://www.irb.vt.edu/documents/onlinepolicy.pdf

☒ **No,** go to question 6.1
☐ **Yes,** answer questions within table

IF YES

Identify the service / program that will be used:

- ☐ www.survey.vt.edu, go to question 6.1
- ☐ Blackboard, go to question 6.1
- ☐ Center for Survey Research, go to question 6.1
- ☐ Other

IF OTHER:

Name of service / program:
URL:
This service is...

- ☐ Included on the list found at: http://www.irb.vt.edu/pages/validated.htm
- ☐ Approved by VT IT Security
- ☐ An external service with proper SSL or similar encryption (https://) on the login (if applicable) and all other data collection pages.
- ☐ None of the above (note: only permissible if this is a collaborative project in which VT individuals are only responsible for data analysis, consulting, or recruitment)

Section 6: Risks and Benefits

6.1 WHAT ARE THE POTENTIAL RISKS (E.G., EMOTIONAL, PHYSICAL, SOCIAL, LEGAL, ECONOMIC, OR DIGNITY) TO STUDY PARTICIPANTS?

Since the particiants will deal with electronic versions of news articles that been collected from online resources, we think that there are no potiential risks for this study, other than fatigue from reading and analyzing what they read.

6.2 EXPLAIN THE STUDY'S EFFORTS TO REDUCE POTENTIAL RISKS TO SUBJECTS:

We will allow the participants to split up the work over several sessions, to reduce fatigue.

6.3 WHAT ARE THE DIRECT OR INDIRECT ANTICIPATED BENEFITS TO STUDY PARTICIPANTS AND/OR SOCIETY?

The study will help with evaluating topic extracting using LDA results and text summaries generated by filling out templates, and that will help the researcher community interested in this field of study.

Section 7: Full Board Assessment

7.1 DOES THE RESEARCH INVOLVE MICROWAVES/X-RAYS, OR GENERAL ANESTHESIA OR SEDATION?

☒ **No**
☐ **Yes**

7.2 DO RESEARCH ACTIVITIES INVOLVE PRISONERS, PREGNANT WOMEN, FETUSES, HUMAN IN VITRO FERTILIZATION, OR MENTALLY DISABLED PERSONS?

☒ **No,** go to question 7.3
☐ **Yes,** answer questions within table →

IF YES
This research involves: ☐ Prisoners ☐ Pregnant women ☐ Fetuses ☐ Human in vitro fertilization ☐ Mentally disabled persons

7.3 DOES THIS STUDY INVOLVE MORE THAN MINIMAL RISK TO STUDY PARTICIPANTS?

Minimal risk means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily activities or during the performance of routine physical or psychological examinations or tests. Examples of research involving greater than minimal risk include collecting data about abuse or illegal activities. Note: if the project qualifies for Exempt review (http://www.irb.vt.edu/pages/categories.htm), it will not need to go to the Full Board.

☒ **No**
☐ **Yes**

IF YOU ANSWERED "YES" TO ***ANY ONE*** OF THE ABOVE QUESTIONS, 7.1, 7.2, OR 7.3, THE BOARD MAY REVIEW THE PROJECT'S APPLICATION MATERIALS AT ITS MONTHLY MEETING. VIEW THE FOLLOWING LINK FOR DEADLINES AND ADDITIONAL INFORMATION: http://www.irb.vt.edu/pages/deadlines.htm

Section 8: Confidentiality / Anonymity

For more information about confidentiality and anonymity visit the following link: http://www.irb.vt.edu/pages/confidentiality.htm

8.1 WILL PERSONALLY IDENTIFYING STUDY RESULTS OR DATA BE RELEASED TO ANYONE OUTSIDE OF THE RESEARCH TEAM?

For example – to the funding agency or outside data analyst, or participants identified in publications with individual consent

☒ **No**
☐ **Yes,** to whom will identifying data be released?

8.2 WILL ANY STUDY FILES CONTAIN PARTICIPANT IDENTIFYING INFORMATION (E.G., NAME, CONTACT INFORMATION, VIDEO/AUDIO RECORDINGS)?

Note: if collecting signatures on a consent form, select "Yes."

☒ **No,** go to question 8.3
☐ **Yes,** answer questions within table →

IF YES

Describe if/how the study will utilize study codes:

If applicable, where will the key [i.e., linked code and identifying information document (for instance, John Doe = study ID 001)] **be stored and who will have access?**

Note: the key should be stored separately from subjects' completed data documents and accessibility should be limited.

The IRB strongly suggests and may require that all data documents (e.g., questionnaire responses, interview responses, etc.) do not include or request identifying information (e.g., name, contact information, etc.) from participants. If you need to link subjects' identifying information to subjects' data documents, use a study ID/code on all data documents.

8.3 WHERE WILL DATA BE STORED?

Examples of data - questionnaire, interview responses, downloaded online survey data, observation recordings, biological samples

They will be stored in an external hard disk that will be locked in a cabinet in co-investigator Tarek Kanan's office.

8.4 WHO WILL HAVE ACCESS TO STUDY DATA?

Only the investigator and co-investigators.

8.5 DESCRIBE THE PLANS FOR RETAINING OR DESTROYING THE STUDY DATA

The hard disk will be kept for three years in a locked cabinet. After that period, the data will be properly deleted.

8.6 DOES THIS STUDY REQUEST INFORMATION FROM PARTICIPANTS REGARDING ILLEGAL BEHAVIOR?

☒ **No,** go to question 9.1
☐ **Yes,** answer questions within table

IF YES

Does the study plan to obtain a Certificate of Confidentiality?

☐ No
☐ Yes (Note: participants must be fully informed of the conditions of the Certificate of Confidentiality within the consent process and form)

For more information about Certificates of Confidentiality, visit the following link:
http://www.irb.vt.edu/pages/coc.htm

Section 9: Compensation

For more information about compensating subjects, visit the following link: http://www.irb.vt.edu/pages/compensation.htm

9.1 WILL SUBJECTS BE COMPENSATED FOR THEIR PARTICIPATION?

☒ **No,** go to question 10.1
☐ **Yes,** answer questions within table

IF YES
What is the amount of compensation?
Will compensation be prorated? ☐ Yes, please describe: ☐ No, explain why and clarify whether subjects will receive full compensation if they withdraw from the study? *Unless justified by the researcher, compensation should be prorated based on duration of study participation. Payment must not be contingent upon completion of study procedures. In other words, even if the subject decides to withdraw from the study, he/she should be compensated, at least partially, based on what study procedures he/she has completed.*

Section 10: Audio / Video Recording

For more information about audio/video recording participants, visit the following link: http://www.irb.vt.edu/pages/recordings.htm

10.1 WILL YOUR STUDY INVOLVE VIDEO AND/OR AUDIO RECORDING?

☒ **No,** go to question 11.1
☐ **Yes,** answer questions within table

IF YES
This project involves: ☐ Audio recordings only ☐ Video recordings only ☐ Both video and audio recordings
Provide compelling justification for the use of audio/video recording:
How will data within the recordings be retrieved / transcribed?
How and where will recordings (e.g., tapes, digital data, data backups) be stored to ensure security?
Who will have access to the recordings?
Who will transcribe the recordings?
When will the recordings be erased / destroyed?

Section 11: Research Involving Students

11.1 DOES THIS PROJECT INCLUDE STUDENTS AS PARTICIPANTS?

☐ **No,** go to question 12.1
☒ **Yes,** answer questions within table

IF YES

Does this study involve conducting research with students of the researcher?

☒ No

☐ Yes, describe safeguards the study will implement to protect against coercion or undue influence for participation:

Note: if it is feasible to use students from a class of students not under the instruction of the researcher, the IRB recommends and may require doing so.

Will the study need to access student records (e.g., SAT, GPA, or GRE scores)?

☒ No

☐ Yes

11.2 DOES THIS PROJECT INCLUDE ELEMENTARY, JUNIOR, OR HIGH SCHOOL STUDENTS?

☒ **No,** go to question 11.3

☐ **Yes,** answer questions within table

IF YES

Will study procedures be completed during school hours?

☐ No

☐ Yes

If yes,

Students not included in the study may view other students' involvement with the research during school time as unfair. Address this issue and how the study will reduce this outcome:

Missing out on regular class time or seeing other students participate may influence a student's decision to participate. Address how the study will reduce this outcome:

Is the school's approval letter(s) attached to this submission?

☐ Yes

☐ No, project involves Montgomery County Public Schools (MCPS)

☐ No, explain why:

You will need to obtain school approval (if involving MCPS, click here: http://www.irb.vt.edu/pages/mcps.htm). Approval is typically granted by the superintendent, principal, and classroom teacher (in that order). Approval by an individual teacher is insufficient. School approval, in the form of a letter or a memorandum should accompany the approval request to the IRB.

11.3 DOES THIS PROJECT INCLUDE COLLEGE STUDENTS?

☐ **No,** go to question 12.1

☒ **Yes,** answer questions within table

IF YES

Some college students might be minors. Indicate whether these minors will be included in the research or actively excluded:

☐ Included

☒ Actively excluded, describe how the study will ensure that minors will not be included: **The participants**

will be graduate students; we will make sure that they will not be minors.

Will extra credit be offered to subjects?

☒ No
☐ Yes

If yes,

What will be offered to subjects as an equal alternative to receiving extra credit without participating in this study?

Include a description of the extra credit (e.g., amount) to be provided within question 9.1 ("IF YES" table)

Section 12: Research Involving Minors

12.1 DOES THIS PROJECT INVOLVE MINORS (UNDER THE AGE OF 18 IN VIRGINIA)?

Note: age constituting a minor may differ in other States.

☒ **No,** go to question 13.1
☐ **Yes,** answer questions within table

IF YES

Does the project reasonably pose a risk of reports of current threats of abuse and/or suicide?

☐ No
☐ Yes, thoroughly explain how the study will react to such reports:

Note: subjects and parents must be fully informed of the fact that researchers must report threats of suicide or suspected/reported abuse to the appropriate authorities within the Confidentiality section of the Consent, Assent, and/or Permission documents.

Are you requesting a waiver of parental permission (i.e., parent uninformed of child's involvement)?

☐ No, **both** parents/guardians will provide their permission, if possible.
☐ No, **only one** parent/guardian will provide permission.
☐ Yes, describe below how your research meets **all** of the following criteria (A-D):
Criteria A - The research involves no more than minimal risk to the subjects:
Criteria B - The waiver will not adversely affect the rights and welfare of the subjects:
Criteria C - The research could not practicably be carried out without the waiver:
Criteria D - (Optional) Parents will be provided with additional pertinent information after participation:

Is it possible that minor research participants will reach the legal age of consent (18 in Virginia) while enrolled in this study?

☐ No
☐ Yes, will the investigators seek and obtain the legally effective informed consent (in place of the minors' previously provided assent and parents' permission) for the now-adult subjects for any ongoing interactions with the subjects, or analysis of subjects' data? If yes, explain how:

For more information about minors reaching legal age during enrollment, visit the following link: http://www.irb.vt.edu/pages/assent.htm

*The procedure for obtaining assent from minors and permission from the minor's guardian(s) must be described in **Section 4** (Consent Process) of this form.*

Section 13: Research Involving Deception

For more information about involving deception in research and for assistance with developing your debriefing form, visit our website at http://www.irb.vt.edu/pages/deception.htm

13.1 DOES THIS PROJECT INVOLVE DECEPTION?

☒ **No,** go to question 14.1
☐ **Yes,** answer questions within table

IF YES
Describe the deception:
Why is the use of deception necessary for this project?
Describe the debriefing process:
Provide an explanation of how the study meets all the following criteria (A-D) for an alteration of consent: Criteria A - The research involves no more than minimal risk to the subjects: Criteria B - The alteration will not adversely affect the rights and welfare of the subjects: Criteria C - The research could not practicably be carried out without the alteration: Criteria D - (Optional) Subjects will be provided with additional pertinent information after participation (i.e., debriefing for studies involving deception): *By nature, studies involving deception cannot provide subjects with a complete description of the study during the consent process; therefore, the IRB must allow (by granting an alteration of consent) a consent process which does not include, or which alters, some or all of the elements of informed consent.*
The IRB requests that the researcher use the title "Information Sheet" instead of "Consent Form" on the document used to obtain subjects' signatures to participate in the research. This will adequately reflect the fact that the subject cannot fully consent to the research without the researcher fully disclosing the true intent of the research.

Section 14: Research Involving Existing Data

14.1 WILL THIS PROJECT INVOLVE THE COLLECTION OR STUDY/ANALYSIS OF EXISTING DATA DOCUMENTS, RECORDS, PATHOLOGICAL SPECIMENS, OR DIAGNOSTIC SPECIMENS?

Please note: it is not considered existing data if a researcher transfers to Virginia Tech from another institution and will be conducting data analysis of an on-going study.

☐ **No,** you are finished with the application
☒ **Yes,** answer questions within table

IF YES
From where does the existing data originate? Public online resources
Provide a detailed description of the existing data that will be collected or studied/analyzed: it's an extracted collection of text from online news articles

Is the source of the data public?

☐ No, continue with the next question

☒ Yes, you are finished with this application

Will any individual associated with this project (internal or external) have access to or be provided with existing data containing information which would enable the identification of subjects:

- **Directly** (e.g., by name, phone number, address, email address, social security number, student ID number), or
- **Indirectly through study codes** even if the researcher or research team does not have access to the master list linking study codes to identifiable information such as name, student ID number, etc or
- **Indirectly through the use of information that could reasonably be used in combination to identify an individual** (e.g., demographics)

☐ No, collected/analyzed data will be completely de-identified

☐ Yes,

If yes,

Research will not qualify for exempt review; therefore, if feasible, written consent must be obtained from individuals whose data will be collected / analyzed, unless this requirement is waived by the IRB.

Will written/signed or verbal consent be obtained from participants prior to the analysis of collected data? -select one-

This research protocol represents a contract between all research personnel associated with the project, the University, and federal government; therefore, must be followed accordingly and kept current.

Proposed modifications must be approved by the IRB prior to implementation except where necessary to eliminate apparent immediate hazards to the human subjects.

Do not begin human subjects activities until you receive an IRB approval letter via email.

It is the Principal Investigator's responsibility to ensure all members of the research team who interact with research subjects, or collect or handle human subjects data have completed human subjects protection training prior to interacting with subjects, or handling or collecting the data.

---------END---------

The Recruitment Materials and Announcement Email

We are conducting a study to evaluate, for Arabic news articles, the results of our topic extraction tool, as well as our automatically generated summaries. This evaluation will be used to judge the quality and accuracy of our methods and results. We need participants to partake in our experiment. The duties of the participants will include:

1. Reading news articles along with our automatically generated summaries and topics;
2. Evaluating our automatically generated results, i.e., giving a number, on the scale of 0-10, rating each topic and summary; and
3. Saving those numbers so we can record the assessment of quality.

With the results of the study we plan to use the recorded information to help us evaluate the quality of our topic extraction and summarization methods.

If you are a graduate student who can read and write Modern Standard Arabic and are over the age of 18, and would like to participate, please contact

Tarek Kanan: tarekk@vt.edu

He will respond with instructions on how to participate. Thank you!

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Informed Consent for Participants in Research Projects Involving Human Subjects

Title of Project
Evaluating topical relevance and template summary quality for Arabic news articles.

Investigators
Dr. Edward A. Fox, Tarek Kanan, Soulieman Ayoub, Julia Freeman

I. Purpose of this Research/Project
We would like to test our new Arabic topic extractor tool and our summary template generator. Participants will make the needed quality judgments, so we can judge our tools and results.

II. Procedures
All participants will be asked to read every news article assigned to them, along with its corresponding topic and summary. They will rate (evaluate) the topics and summaries. In short, the task is to 1) take each file provided; 2) read it; 3) add to the bottom two numbers, one for topic relevance and one for the quality of the summary, using a Likert scale (rating); 3) save the expanded file, and then 4) submit all of the expanded files back to the investigators, along with this document. Each of the numbers will be in the range 0-10; 0 is for a totally nonrelevant topic or an extremely bad summary, while 10 is for totally relevant topic and extremely high quality summary. You can use any number between them; for example 5 will mean a topic is somewhat relevant or a summary has average quality.

III. Risks
There are no more than minimal risks involved. If a participant becomes fatigued, they may pause their work, and continue in one or more subsequent sessions and days, after having an adequate rest. The entire effort should take no more than 2 hours. If that amount of time is not available, participants can submit as many files as they have time to complete.

IV. Benefits
We will acknowledge the participants as a group for their help in any publication we will develop related to this work.

V. Extent of Anonymity and Confidentiality
All data collected in this study will be anonymized. We will not collect any kind of participants' information.

VI. Compensation
Participants will not receive any compensation for their help.

VII. Freedom to Withdraw
At any point in the study, participants may withdraw from the study.

VIII. Subject's Responsibilities
I agree to participate in this study for free. I will make a best effort attempt to evaluate and assign a number in the range 0-10 for topic relevance, and a similar number indicating the quality of a summary, for each news article.

IX. Instructions

1. You need to read each news article and its corresponding topic and summary, and then rate the topic and the summary using a Likert scale.
2. For the topic section, please assign, based on your knowledge, the number that reflects the relevance of the topic for the article. Please choose a number in the range 0-10. For example, 0 will indicate a topic that is not at all relevant to the article, while 10 means the topic is fully relevant to the article. The other numbers will indicate the degree of relevance. For example, 5 will mean average or medium relevance. The topic will be in the form of an ordered list of (e.g.,10) words.
3. For the summary section, based on your knowledge, and using a scale in the range 0-10, please assign a number that reflects the quality of the summary for the article. For example, 0 will mean an extremely low quality summary of the article, while 10 will mean an extremely high quality summary for the article. The numbers in between will reflect the level of quality, for example 5 will mean average quality. The summary will be in the form of small paragraph that contains several sentences.
4. At the bottom of each file please write the numbers you assign for the relevance of the topic, and for the quality of the summary.
5. Please use the 'Save As" option to overwrite the original file

6. Make sure you don't use a new name for a file and that the encoding is Unicode-8

X. Subject's Permission

After reading this Consent Form, and possibly after an email exchange with the co-investigators, I have had all my questions answered. I confirm that I am not a minor. By submitting the requested files, with the requested additions, and returning this document with those files, I hereby acknowledge the above and give my voluntary consent.

Appendix C: A Modified Version of the Standardized Taxonomy and a Significance Test for the Results of P-Stemmer

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 stemmers. Table C-1 shows the F1 results using the SVM classifier for P-Stemmer and Stem1, Stem2, Stem3, Stem8, and Stem10.

Table C-1. F1-Measure Results for the P-Stemmer and the Five Larkey Stemmers

	SVM					
	P-Stemmer	**Stem1**	**Stem2**	**Stem3**	**Stem8**	**Stem10**
Art	0.918	0.915	0.912	0.912	0.921	0.920
Economy	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-rank test (Wilcoxon, 1945), a non-parametric statistical hypothesis test. It can be used when comparing two related samples, which in our case involves P-Stemmer and any one of the five Larkey stemmers. This test is very popular for information retrieval evaluation (Smucker, Allan & Carterette, 2007). Table C-2, Table C-3, Table C-4, Table C-5, and Table C-6 show the calculations and results, with the final Wilcoxon (Wcal) values. The text inside the boxes below represents the formulas used to generate the values in the tables toward calculating the final test results.

Let N be the number of pairs, sample size (5 in our case).

For i= 1, …, 5, let $X_{1,i}$ denote the five F1 measure values of one of the Larkey stemmers (Stem1, Stem2, Stem3, Stem8, or Stem10) and $X_{2,i}$ denote the five F1 measure values of P-Stemmer.

Sgn is the sign of the value (+/-).

Abs is the absolute value | |.

R_i is the rank

Our Hypotheses are:

H_0: 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.

vs.

H_1: The median difference of the F1 measure of P-Stemmer and each one of Larkey's stemmers (Stem1, Stem2, Stem3, Stem8, or Stem10) is greater than zero.

Table C-2. The Values toward Calculating W_{cal} for the Wilcoxon Signed-Rank Test for the F1 Measure Ordered by Absolute Differences (abs) between P-Stemmer and Stem1

			$X_{2,i}$ - $X_{1,i}$			
i	**$X_{2,i}$**	**$X_{1,i}$**	**sgn**	**abs**	**R_i**	**sgn * R_i**
4	0.991	0.990	1	0.001	1	1
3	0.915	0.913	1	0.002	2	2
1	0.918	0.915	1	0.003	3	3
5	0.964	0.960	1	0.004	4	4
2	0.935	0.919	1	0.016	5	5

The W_{cal} value is calculated using the formula: $\sum_i^N [sgn(X2, i\ -\ X1, i) * Ri]$

W_{cal} = [1+2+3+4+5] = 15

We tested our hypotheses with α= 0.05 ➔ $W_{\alpha=0.05,5}$ = zero

If (W_{cal} >= $W_{\alpha,N}$) ➔ 15 >= 0 then we reject H_0 (accept H_1)

For our test 15 >= zero, so we reject H_0 (accept H_1) and conclude that, using the F1 measure for evaluation, our P-Stemmer is significantly different from Stem1.

Table C-3. The Values toward Calculating W_{cal} for the Wilcoxon Signed-Rank Test for the F1 Measure Ordered by Absolute Differences (abs) between P-Stemmer and Stem2

			$X_{2,i} - X_{1,i}$			
i	**$X_{2,i}$**	**$X_{1,i}$**	**sgn**	**abs**	**R_i**	**sgn * R_i**
4	0.991	0.993	-1	0.002	1	-1
1	0.918	0.912	1	0.006	2	2
3	0.915	0.908	1	0.007	3	3
5	0.964	0.955	1	0.009	4	4
2	0.935	0.918	1	0.017	5	5

The W_{cal} value is calculated using the formula: $\sum_i^N [sgn(X2, i\ -\ X1, i) * Ri]$

W_{cal} = [-1+2+3+4+5] = 13

We tested our hypotheses with α= 0.05 ➔ $W_{\alpha=0.05,5}$ = zero

If (W_{cal} >= $W_{\alpha,N}$) ➔ 13 >= 0 then we reject H_0 (accept H_1)

For our test 13 >= zero, so we reject H_0 (accept H_1) and conclude that, using the F1 measure for evaluation, our P-Stemmer is significantly different from Stem2.

Table C-4. The Values toward Calculating W_{cal} for the Wilcoxon Signed-Rank Test for the F1 Measure Ordered by Absolute Differences (abs) between P-Stemmer and Stem3

			$X_{2,i} - X_{1,i}$			
i	**$X_{2,i}$**	**$X_{1,i}$**	**sgn**	**abs**	**R_i**	**sgn * R_i**
4	0.991	0.993	-1	0.002	1	-1
1	0.918	0.912	1	0.006	2	2
5	0.964	0.955	1	0.009	3	3
2	0.935	0.904	1	0.031	4	4
3	0.915	0.864	1	0.051	5	5

The W_{cal} value is calculated using the formula: $\sum_i^N [sgn(X2,i\ -\ X1,i) * Ri]$

W_{cal} = [-1+2+3+4+5] = 13

We tested our hypotheses with α= 0.05 ➔ $W_{\alpha=0.05,5}$ = zero

If (W_{cal} >= $W_{\alpha,N}$) ➔ 13 >= 0 then we reject H_0 (accept H_1)

For our test 13 >= zero, so we reject H_0 (accept H_1) and conclude that, using the F1 measure for evaluation, our P-Stemmer is significantly different from Stem3.

Table C-5. The Values toward Calculating W_{cal} for the Wilcoxon Signed-Rank Test for the F1 Measure Ordered by Absolute Differences (abs) between P-Stemmer and Stem8

			$X_{2,i}$ - $X_{1,i}$			
i	**$X_{2,i}$**	**$X_{1,i}$**	**sgn**	**abs**	**R_i**	**sgn * R_i**
4	0.991	0.993	-1	0.002	1.5	-1.5
5	0.964	0.962	1	0.002	1.5	1.5
1	0.918	0.921	1	0.003	3	3
2	0.935	0.910	1	0.025	4	4
3	0.915	0.889	1	0.026	5	5

The W_{cal} value is calculated using the formula: $\sum_i^N [sgn(X2,i\ -\ X1,i) * Ri]$

W_{cal} = [-1.5+1.5+3+4+5] = 12

We tested our hypotheses with α= 0.05 ➔ $W_{\alpha=0.05,5}$ = zero

If (W_{cal} >= $W_{\alpha,N}$) ➔ 12 >= 0 then we reject H_0 (accept H_1)

For our test 13 >= zero, so we reject H_0 (accept H_1) and conclude that, using the F1 measure for evaluation, our P-Stemmer is significantly different from Stem8.

Table C-6. The Values toward Calculating W_{cal} for the Wilcoxon Signed-Rank Test for the F1 Measure Ordered by Absolute Differences (abs) between P-Stemmer and Stem10

			$X_{2,i} - X_{1,i}$			
i	$X_{2,i}$	$X_{1,i}$	**sgn**	**abs**	R_i	**sgn * R_i**
4	0.991	0.992	1	0.001	1	1
1	0.918	0.920	1	0.002	2	2
5	0.964	0.959	1	0.005	3	3
3	0.915	0.896	1	0.019	4	4
2	0.935	0.900	1	0.035	5	5

The W_{cal} value is calculated using the formula: $\sum_i^N [sgn(X2, i - X1, i) * Ri]$

W_{cal} = [1+2+3+4+5] = 15

We tested our hypotheses with α= 0.05 ➔ $W_{\alpha=0.05,5}$ = zero

If (W_{cal} >= $W_{\alpha,N}$) ➔ 15 >= 0 then we reject H_0 (accept H_1)

For our test 13 >= zero, so we reject H_0 (accept H_1) and conclude that, using the F1 measure for evaluation, our P-Stemmer is significantly different from Stem10.

We used the Wilcoxon signed-ranked test to test the statistical difference between our proposed stemmer and each of Larkey's stemmers, with the P-value less than or equal to 0.05. We did the test five times and successfully rejected our null hypothesis, for each one of the five tests, which states 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 stemmers.

Modified Taxonomy

Our standardized taxonomy contains 13 different categories, five in the first level and eight in the second level; see Figure 2-8 in chapter 2. To expand our taxonomy and enable it to cover more Arabic news stories, we decided to modify the original taxonomy in Figure 2-8 by adding another category to the first level of the taxonomy only for browsing purposes. Our modified standardized taxonomy that can be used for browsing and navigating any Arabic news articles will contain 14 categories in total, of which six are in the main level, with eight in the second. The extra category added is called Miscellaneous (Misc.); this category will basically hold any news article unable to fit in any of the original five categories in the main level. With this we can guarantee that our taxonomy covers more topics like Health, Science, etc. that have not been covered by the original five categories and their sub-categories. Figure C-1 shows the modified version of the standardized taxonomy.

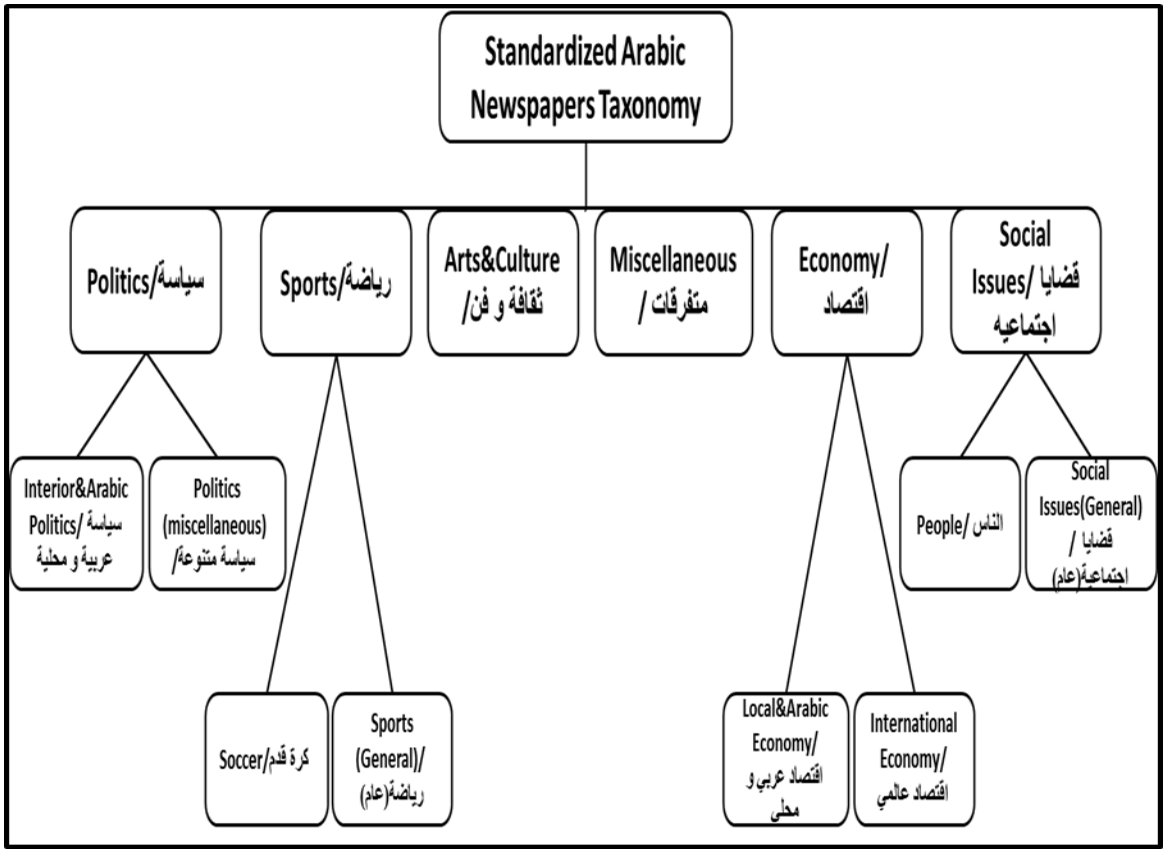

Figure C-1: Modified Version of the Standardized Taxonomy for Browsing