A STATISTICAL MODEL TO ENHANCE TOPIC DETECTION OF ARABIC TEXT

Emad Aloqayli[1], Yaseen Alquran[2], Dana Hazaimeh[3], Mahmood Qudah[4]

[1]*Software Engineering Department, Faculty of Science and Information Technology, Jadara University*
[2]*Software Engineering Department, Faculty of Science and Information Technology, Jadara University*
[3]*Software Engineering Department, Faculty of Science and Information Technology, Jadara University*
[4]*Software Engineering Department, Faculty of Science and Information Technology, Jadara University*

Abstract

The exponential growth of the available Arabic documents online increased the need for techniques that help in classifying and processing these documents. The nature of the Arabic language and the presence of noisy information in the documents' contents make it difficult to guarantee accurate results when performing different processes, like topic detection and classification. In this paper, a statistical model is proposed to enhance the ability of the topic detection of Arabic text. The model is evaluated using a document set. The results are measured using precision, recall, and accuracy and the preliminary indications show that the model is able to provide promising results.

Keywords: *Arabic Language, Classification, Information Retrieval, Topic Detection*

----------***----------

1. INTRODUCTION

The rapid increase of computer technology integration in the daily life aspects, accompanied with an exponential growth of the electronic contents increased the interest in IR techniques that are able to process these contents accurately and efficiently. For example, on Facebook there are 510,000 comments posted every minute [6].

The online contents are scaling in the amount of documents available online, the diversity of the contents of these documents, and the average size of the documents. Topic Detection is the process of identifying the relation between the contents of a document and a certain set of topics. Based on previous knowledge, the topics of the new documents can be detected, however, the accumulated knowledge lacks the advances of the knowledge that can be found on the new documents.

In Arabic language, polysemy is common for single words. For example, the word spelled as (معلم) could mean a teacher or a statue, in text it is the context that identify the exact meaning of the word. In written text, Diacritic is used to differentiate the mean of the word to identify its exact meaning. Hence, مُعلم means teacher and مَعلم means statue. Generally, in online Arabic documents diacritic is not used, as it is time consuming, and the reader identifies the exact meaning of the word from the context.

Polysemy is considered as one of the noise sources to the topic detection process. When considering large scale document collections, the influence of the polysemy increases the challenge of the topic detection process in terms of its accuracy.

However, each term that appears in the document has a relation with other terms that appear in the same document and the document's topic. The extent to which this relation is strong or weak can be measured by the appearance of other terms in the document and the number of occurrences of these terms.

In this paper, a statistical model is developed based on the relation between the terms within the document and the topic of the documents in a document collection.

The rest of the paper is structured as follows. The related work is provided in section 2. The proposed model is introduced in section 3. The evaluation of the model is provided in section 4. The paper is concluded in section 5.

2. RELATED WORK

The Quranic Arabic Corpus is an annotated linguistic resource that is built by addressing the Arabic grammar, syntax and morphology for each word in the Holy Quran [5]. The corpus provides three levels of analysis: morphological annotation, a syntactic tree bank and a semantic ontology. The corpus is well-developed, however, it is designed for the Quran, and it is not possible to measure its suitability for other domains.

Automatic summarization technique can be used to enhance Arabic topic detection [3]. The use of automatic summarization to reduce the noisy information within the document and enhance the topic detection process is addressed in [3]. The researchers argued that the proposed technique significantly enhances topic detection and increases the performance of topic detection process.

Topic detection for Arabic unvowelized documents addressed in [4]. The proposed technique is implemented using adapted TF-IDF and Jaccard indicator. The study addressed the impact of stems on the topic detection process. To enhance the topic detection the researchers developed Multi-Word Terms (MWTs) extraction prototype to generate MWTs vocabularies. The study incorporate the MWTs in the topic detection process to enhance the accuract of the topic detection process.

Generally, news agencies receive their articles as a continuous stream. The stream contains articles as fragments that might or might not contain any keywords that indicate the topic of the article. The study in [1] proposed a novelty detection system for news streams with Arabic contents. The system is supposed to perform the detection process automatically.

Arabic language can be classified into three levels:

- **Native**: which is the old language that was common hundreds of years ago.
- **Modern native**: which is the official language that is used by the governments.
- **Localized**: which is the language that is localized to each geographical area. Generally, this is the spoken language, however, it is currently being used in non official documents like social media.

3. PROPOSED MODEL

Basically, the model is based on linking the terms to the topics based on the number of occurrences of each term in the document collection in a specific topic. The model is trained on a set of documents that is classified based on their topics. Then the model is evaluated on a document collection that is extracted from a news website with no classification.

3.1 Proposed Methodology

The proposed model develops a list of Single Word Terms (SWTs) from a document collection with identified topics. The document collection is acquired from Albosala news website (*www.albosala.com*). The acquired document collection has more than 5000 documents, distributed as shown in table 1.

Table -1: Distribution of documents based on topic

Topic	Number of documents
Technology	1500
Economy	1700
Politics	1200
Medical	600

The document collection is classified into 4 document sets based on the topics.

The SWT lists contains all terms that appeared in any document in the document collection and each SWT is associated with a weight.

The weight of each term is determined using a normalized equation. The equation considers the number of occurrences of the terms and the number of terms in the list, as show in equation 1.

$$SWT_{weight} = \frac{No}{Nt} \quad \ldots\ldots\ldots (1)$$

where:
No: is the number of occurrences of the SWT in all the documents in the document set.
Nt: is the number of SWTs in the SWT list of the same topic the SWT is part of.

SWTs that have stronger relation with the topic is expected to appear more frequently in the document set. However, the number of documents in the document collection and the average size of the documents affects the number of occurrences of each term and the number of terms in the list.

Generally, the website contents is written in modern native Arabic, with no dialects. The researchers assume that the contents has no spelling errors as these documents are reviewed by specialized persons. Accordingly, during the preprocessing stage no spell check on the documents' contents is performed, and every word in the document after the preprocessing stage is considered as a valid SWTs.

The documents contents is preprocessed before adding the SWTs to the list. The preprocessing is used to filter the contents of the documents. The preprocessing process filters the stop words.

Besides that, the words in the documents are normalized by removing the dialects to have a single spelling of all the variations of the same word. For example , in Arabic it is possible to write the word ((اعلى as ((إعلى and ((أعلى. In this case , ((إعلى mean "get higher" and ((أعلى means "higher".

Based on that, SWTs lists represent a statistical relation between each SWT and its topic. Using higher number of documents to train the model it is expected that SWTs with lower relation to the topic will be ruled out. Accordingly, it is expected to enhance the topic detection by utilizing the SWTs lists.

3.2 Proposed Process

The development of the SWTs lists is performed during a training stage. During the training stage, the documents within the document collection are grouped into sets based on the topic, then the list of SWTs is extracted from the documents' sets. After that, the weight of each SWT is determined and added to the list.

Generally, the number of occurrences of an SWT in the document set is related to its link with the topic , for example, ((حكومة, which means government, is expected to appear more frequently in politics documents than other topics.

The proposed model is processed as follows:

Preprocessing Stage: each document is preprocessed before it is used in the training stage. The preprocessing includes:

- All letters that do not belong to Arabic language are removed.
- All digits are removed.
- All punctuation marks are removed.
- All diacritic are removed. In this step, the words are normalized according to the common Arabic word normalization steps discussed in [2].
- All special characters are removed.

Processing Stage: during this stage the SWTs are built and for each SWT the number of occurrences is identified.
Postprocessing Stage: in this stage the model contents is finalized by calculating the weight of each SWT.
The process is performed on three sets of documents, these sets represent politics, economy, and technology topics.

One remark on the extracted document collection is that it covers a short period of time, hence, certain set of SWTs are more frequent than others.

4. EVALUATION

As a result of the training session, the number of SWTs identified for each topic is relatively high, as shown in table 2.

In the SWT lists, SWTs have their weights distributed over a wide range, SWTs that are common to appear in different topics, and have relatively a low weight is expected to introduce noise to the topic detection process.

Accordingly, a cutoff level is used to determine if the SWT to appear in the final list. The first factor used is the number of occurrences of the SWT.

Based on the distribution of the results the number is set to 10, that is if the SWT has less than 10 number of occurrences in the document set, then it is removed from the SWT list, table 2 shows the final number of SWTs that is listed.

Table -2: SWT lists

Topic	Number of SWTs	Number of SWTs after cutoff
Technology	45,000	3,750
Economy	27,000	2,800
Politics	21,500	2,350

Applying the cutoff has two advantages:

- Shortness the SWTs list, thus, only SWTs that are actually relating to the topic appear in the list.
- Compacting the SWT lists, for those SWTs that have actual relation to the topics is expected to enhance the topic detection accuracy and efficiency.

The proposed model is evaluated using two methodologies:

- Using a special set of documents that have contents designed for the evaluation purposes. The documents are selected such that it have SWTs that already appeared in the SWT lists. However, the contents of the documents are not altered, it is used as extracted from the website.
- Using a random set of documents. Documents in this set is selected randomly and its contents had not been checked before the evaluation process.

4.1 Specially Selected Documents

The documents in this set is selected such that it reflects three different schemes:
• **Biased Scheme:** the terms that appear in these documents is biased toward one of the topics, that is the document has SWTs that represents a certain topic.
• **Balanced Scheme:** the set of documents in this scheme is randomly selected from different topics, such that the documents have contents that already identified in the SWT lists.
• **Unidentified Scheme:** the documents selected for this scheme have some terms that had not been identified in any of the developed SWT lists.

Fig 1 shows the precision results of the specially selected documents. The results indicate that as the number of retrieved documents increases the precision decrease. Fig 2 shows the recall results of the specially selected documents. The results indicate that as the number of retrieved documents increases the recall increase.

Figure 3 shows the results of the f-measure for the specially selected documents. The results indicate that as the number of documents increase, the harmony between the precision and recall increase, which is a positive indication.

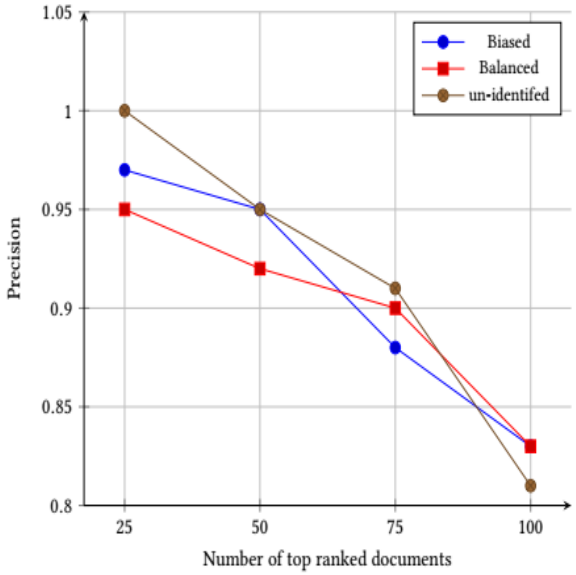

Fig -1: Precision for specially selected documents

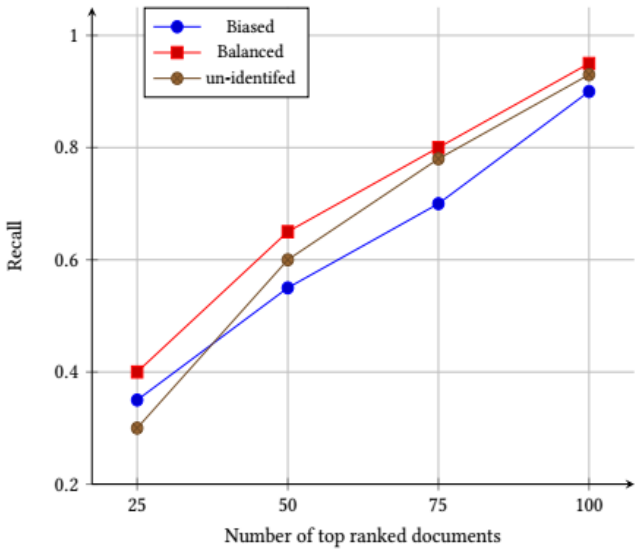

Fig -2: Recall for specially selected documents

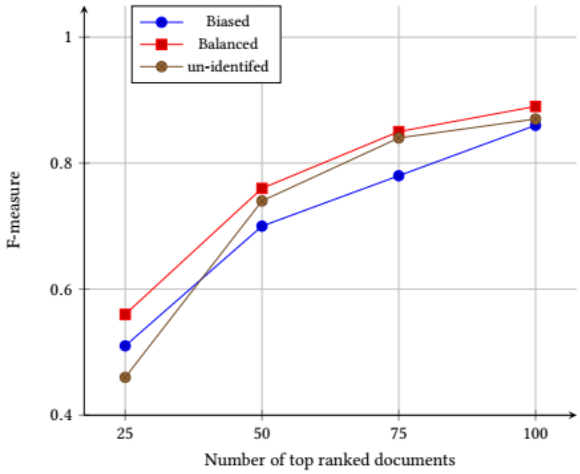

Fig -3: F-measure for specially selected documents

4.2 Randomly Selected Documents

The documents in this set is selected such that none of these documents appeared in the training session. This ensures that the evaluation of this set is not biased by the availability of the contents of these documents in the SWT lists.

The selected documents for this set has relatively higher average size and contains terms that satisfy the diversity of the topics.

Fig 4 shows the precision results of the randomly selected documents. The results indicate that as the number of retrieved documents increases the precision decrease.

Fig 5 shows the recall results of the randomly selected documents. The results indicate that as the number of retrieved documents increase the recall increase.

Fig 6 shows the results of the f-measure for the randomly selected documents. The results indicate that as the number of documents increases, the harmony between the precision and recall increases, which is a positive indication.

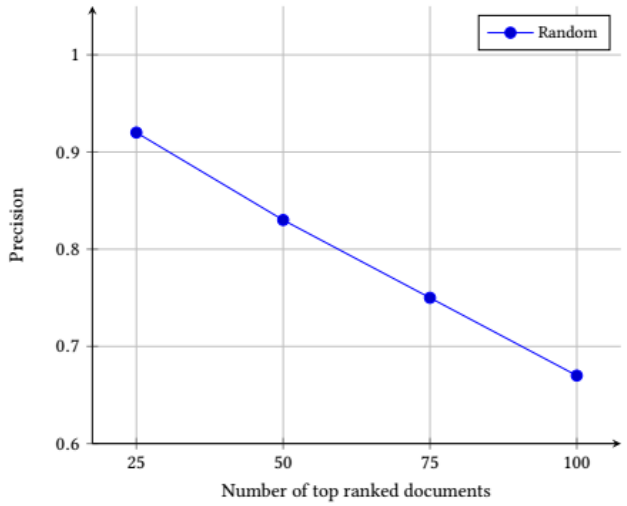

Fig -4: Precision for randomly selected documents

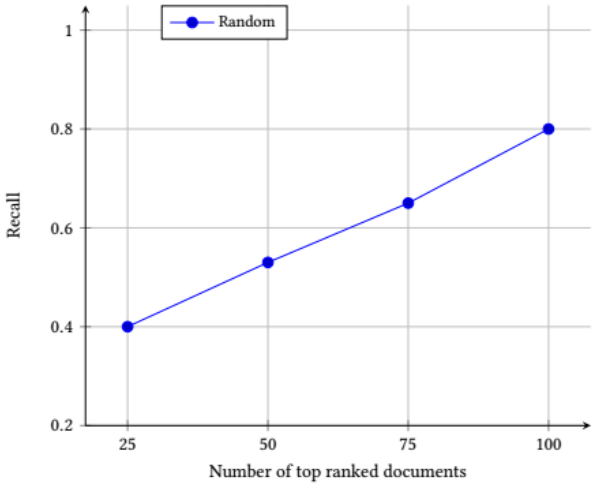

Fig -5: Recall for randomly selected documents

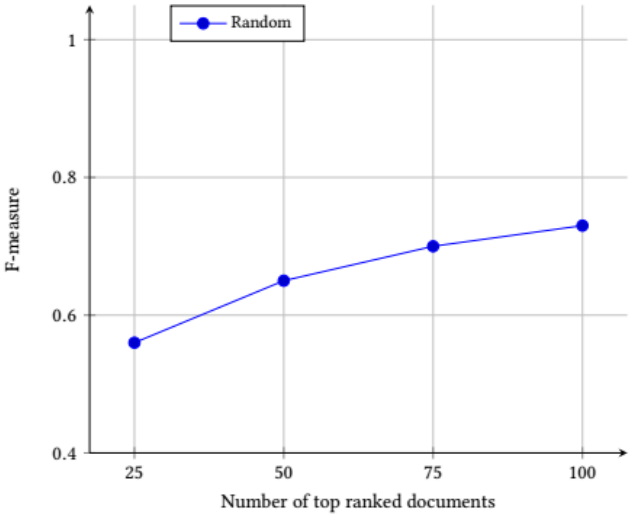

Fig 6: F-measure for randomly selected documents

4.3 Discussion of the Results

The evaluation of the proposed model indicates that it achieves high precision and recall results, and these results are harmonized as the f-measure indicates.

The developed lists reflects the relation between the SWTs and the topics. Currently, three lists are developed for politics, economy, and technology. The number of documents used to develop these list are relatively small and represents news

5. CONCLUSION

In this paper a model is proposed to enhance the topic detection process of documents with Arabic contents. The proposed model is developed based on the statistical relation between the SWTs and the documents topics. A relatively small set of document collection that covers three main topics: politics, technology, and economy are used to develop the SWTs list.

The evaluation of the proposed model shows that it achieves high accuracy results. The model can be used for topic detection of any document with Arabic contents, regardless of its contents size.

Further enhancements is being considered to further improve and evalaute the model and increase its usability for other information retrieval operations.

The development of SWTs list using a large scale document collection enable enhancing the current topic detection levels. The researchers are investigating enhancements on the lists to enable subtopic and partial topic detection.

ACKNOWLEDGEMENT

This work is supported and funded by Jadara University. The authors like to thank Dr. Naser Alodat who provided invaluable comments and notes.

REFERENCES

[1]. Mohammed Al-Kabi, Niveen Z. Halalsheh, Muhammad Dabour, and Heider A. Wahsheh. Arabic news: Topic and novelty detection. In Proceedings of the 3rd International Conference on Information and Communication Systems, ICICS '12, pages 7:1--7:5, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1327-8.

[2]. Mayy M. Al-Tahrawi and Sumaya N. Al-Khatib. Arabic text classification using polynomial networks. Journal of King Saud University - Computer and Information Sciences, 27(4):437 -- 449, 2015. ISSN 1319-1578.

[3]. R. Koulali, M. El-Haj, and A. Meziane. Arabic topic detection using automatic text summarisation. In 2013 ACS International Conference on Computer Systems and Applications (AICCSA), pages 1–4, May 2013

[4]. Rim Koulali and Abdelouafi Meziane. Topic Detection and Multi-word Terms Extraction for Arabic Unvowelized Documents, pages 614--623. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-25631-8.

[5]. LRG. The quranic arabic corpus. http://corpus.quran.com/, 2017. Accessed: 2017-03-20. Zephoria.

[6]. The top 20 valuable facebook statistics. https://zephoria.com/top-15-valuable-facebook-statistics/, 2017-04-03. Accessed: 2017-04-08