ISSN 1870-4069

Evaluating Lightweight Text Classification Approaches for Arabic Texts

Dhaou Ghoul[1], Gael Lejeune[1], Lichao Zhu[2]

[1] Sorbonne Université,
STIH (Sens Texte Informatique Histoire),
France

[2] Université de Paris,
LLF (Laboratoire de Linguistique Formelle),
France

{dhaou.ghoul, gael.lejeune}@sorbonne-universite.fr,
lichao.zhu@gmail.com

Abstract. Epidemic Surveillance aims at detecting disease outbursts in the world in order to provide useful information to health authorities. Many automatic systems have been conceived to help these authorities to mine the available data with a special focus on press articles. The main goal of these systems is to select relevant articles by means of text classification. The secondary goal is to extract valuable information from these relevant texts . One of the main challenge is to handle many languages with different properties, various availability of language resources (lexicons, POS taggers…) and annotated data. In this paper we present a state of the art on Text Classification as well as Information Extraction for the Arabic language and we test different options for designing a lightweight system to process texts in written Arabic. We show that Arabic language has particular properties, making it difficult to handle properly without improving existing approaches. We propose improvements of an existing lightweight approach that would be promising for Arabic as well as more poorly endowed languages.

Keywords: Arabic texts, ligth text classification, process texts in Arabic.

1 Introduction

Available online press articles have become the main sources of information. The amount of published articles being increasing, it becomes difficult for users to get a comprehensive view of the data. This is why it is fundamental to have effective solutions in classifying the information to help web users find relevant documents in different areas. Among these areas figures the epidemic surveillance domain in which we need expertise of web-based epidemic intelligence systems that allow us to easily and automatically detect epidemic disease outbursts [6]. Epidemic surveillance consists of detection and interpretation of unstructured available information on the Internet. It aims to provide duly selected and indexed documents [22].

Text classification and information extraction are powerful techniques that help to structure data in order to help experts of many fields including epidemiology. The main goal of classification is to select relevant documents for a particular task whereas Information Extraction consists in extracting a structured representation from these documents in order to populate databases. Several approaches have been proposed for this task. Most of existing approaches are primarily designed for processing texts written in English, relying on sentence patterns [11], ontology-alike lexical resources [10] or hybrid approaches [17].

Extending the multilingual coverage therefore implies to reproduce a pipeline with language-dependent resources and processing tools. But this approach is not suitable for all languages [34] even with the help of machine learning [15, 30] pointed out the need to process poorly endowed languages or dialects without training data. This approach did not seem suitable for building a disease surveillance system for a language like Arabic. In this paper we will test DANIEL (Data Analysis for Information Extraction in any Language) [24], a lightweight approach which, according to the authors, allows handling numerous languages with a limited quantity of lexicon.

The system has evaluated 17 languages (English, French, Greek…) but has not, to the best of our knowledge, appropriately evaluated the Arabic language. The article is organized as follows: in Section 2, we present some works on Information Extraction and Text classification for the Arabic language; in Section 3, we describe a lightweight approach for text classification and information extraction; in Section 4 we analyze results and present some propositions to improve this approach.

2 Related Work

2.1 Event Extraction in Public Health

In Event Extraction task exits three main methods: rule-based, supervised and unsupervised method. In this section we present some related works from event extraction in public health. First, the rule-based systems based on pattern matching. For example, the French Animal Health Epidemic Intelligence System is based on a combined information extraction method relying on rule-based systems and data mining techniques [5].

Second, several supervised classifiers exist for detecting public health events within unstructured text. In these works, the authors used word embeddings such as TF-IDF, Word2Vec, etc. in order to capture relevant entity co-occurrences within a document [21]. Note that, this method is dominant in event extraction research even if it requires a large-scale labeled training corpus [36]. Finally, numerous unsupervised detection of events from text exist, like UPHED system[16].

This system identifies events as clusters of documents associated with labels, i.e., a set of diseases and locations describing clusters. This system achieves a precision of 60% and a recall of 71% using manually annotated real-world data. In [31], the authors showed that the problem of the size of ground-truth datasets when one wants to deal with more than a few languages even when the task is simplified as binary classification task like separating documents that contain epidemic events or not.

The same authors presented in [32] a dataset and a baseline evaluation for multilingual epidemic event extraction at sentence level. In this work, authors experiment with deep learning models based on a bidirectional LSTM (BiLSTM)[25] that use character and word representations and the multilingual BERT (BERT-multilingual-cased[1] and BERT-multilingual-uncased[2]) pre-trained language models for token sequential classification. BERT-multilingual-uncased achieves the F1, recall and precision scores with 80.99%, 79.77% and 82.25% respectively, on the dataset comprising relevant and irrelevant examples. Since Event Extraction needs both text classification and information extraction, we have to see these two tasks can be tackled for a language like Arabic.

2.2 Text Classification for Arabic Texts

Text classification consists of assigning unknown documents into predefined classes. The process of text classification is usually summarized in the following steps [29]:

1. Document pre-processing, i.e. tokenisation, stop-word removal, and stemming or lemmatisation,
2. Document modelling, i.e. representing a document in an appropriate form so that it can be processed by a machine learning algorithm,
3. Feature selection and projection,
4. Transforming features into classification rules,
5. Quality indicators and evaluation methods.

Regarding step 4, there are three approaches to text classification: rule-based approaches, machine learning approaches and hybrid approaches [28]. Rule-based approach organizes text into classes according to a set of handcrafted linguistic rules. If one wants to classify newspaper articles into two classes (*economy* and *health*), the easiest way would be to define two lists of tokens (usually words) that are the most discriminant for each class.

Then, when a new text needs to be classified, its class should be identified by comparing the tokens of the text to the list of the most representative tokens of each class. Here, the rules are made to select relevant tokens and to assign appropriate weights to them. For instance, in emotion detection, lexicons can be used to compute a probability to which the text belongs to a particular class [33].

These approaches can be quite simple to implement but, on the other hand, rule-based approaches have some disadvantages. These approaches require some knowledge of the domain and creating efficient language rules is expensive. Unlike rule-based approaches, machine learning approaches take advantage of past observations, rather than explicit expert knowledge, using pre-labeled examples as training data. For this purpose, the amount of training data should be sufficiently high.

[1] https://huggingface.co/bert-base-multilingual-cased
[2] https://huggingface.co/bert-base-multilingual-uncased

Table 1. Some examples of studies performing Arabic texts classification.

Reference	Corpus (# docs)	Classes	Algorithm	Accuracy	Year
Syiam et al. [35]	News (1132)	6	Rocchio	98%	2006
Duwairi [12]	Magazine/News (1000)	10	Naïve Bayes	95%	2007
Mesleh et al. [27]	News (1545)	9	SVM	98%	2008
Bawaneh et al. [7]	Unknown (242)	6	KNN	84%	2008
Ababneh et al. [1]	News (5121)	7	Cosine	95%	2014
Amina et al. [8]	News (6005)	9	SVM/Naïve Bayes	80%/ 70%	2017
Zinah et al. [2]	News (16757)	5	Master-slaves	88%	2018

The first step is to transform the text into an appropriate representation (usually vectors). One of the most frequently used approaches is bag of words, where a vector represents the frequency of a word in a predefined dictionary of words. Then, a learning algorithm (Naive Bayes, Support Vector Machine, KNN-neighbors, Deep Learning, ...) is applied, which takes a labeled learning corpus as input to create a classification model. With the appropriate amount of training data, text classification via machine learning exhibits more accurate results than rule-based approaches.

The strength of these two approaches can be combined by a process called hybridization [7]. Expert knowledge and machine learning methods are used to find a symbiosis between the simplicity of the linguistic rules and the efficiency of machine learning. Although much work has been devoted to the classification of available texts in English, Chinese and other common languages (Spanish, French …), few works have studied the classification of texts in Arabic. We will present here some of the most interesting works on Arabic texts classification.

These studies use different datasets with different algorithms but we unified the metrics to evaluate the performance of each approach presented in Table 1. A comparative description inspired by [3] and [13] is given in Table 1. We can see that most of existing works rely on machine learning approaches. As mentioned earlier, the text classification process comprises three main steps: pre-processing, classification and evaluation. As Arabic is a morphologically rich language, the pre-processing phase is crucial but we lack efficient pre-processing tools (contrary to an isolating language like English).

For Arabic texts, the pre-processing, besides tokenization and lemmatization, involves normalization of some Arabic letters. For Arabic, as well as for many other languages, there is an important gap compared to English regarding the quality of natural language processing applications. For this reason, there is a stronger interest in the scientific community towards a better treatment of both morphologically rich languages [26] and poorly endowed languages [18, 23] proposed Daniel, a lightweight approach for Classification and Information Extraction that shows convincing results for a bunch of morphologically rich languages (Greek, Polish and Russian) and remains competitive for rather isolate languages like Chinese or English.

The rationale of their approach is to avoid classical pre-processing steps, leaving out the problems of tokenization and lemmatization, and to take advantage of text type properties. Their approach seems to be limited to news and has also been applied to Arabic and a set of languages like German, Spanish or Vietnamese but without proper evaluation.

Table 2. Precision, Recall and F-Measure of some state-of-the-art Arabic Named Entity Recognition Systems.

System	Entity	Precision	Recall	F-measure	Method	Year
TAGARAB	Number	82.8	97.0	97.3	Rule based	1998
	Time	91.0	80.7	85.5		
	Location	94.5	85.3	89.7		
	Person	86.2	76.2	80.9		
Mesfar	Number	97.0	94.0	95.5	Rule based	2007
	Time	97.0	95.0	96.0		
	Location	82.0	71.0	76.0		
	Person	92.0	79.0	85.0		
PNAES	Person	93.0	86.0	89.0	Rule based	2009
ANERsys	Location	93.03	86.67	89.74	Machine learning	2008
	Person	80.41	67.42	73.35		
	Misc	71.0	54.0	61.47		
	Organisation	84.23	53.94	65.76		
Abdulhamid/Darwish	Location	93.0	83.0	88.0	Machine learning	2010
	Person	90.0	75.0	81.0		
	Organisation	84.0	64.0	73.0		
Oudah/Shaalan	Location	x	x	90.0	Hybrid approach	2012
	Person	x	x	94.0		
	Organisation	x	x	88.0		

2.3 Information Extraction for Arabic Texts

Information Extraction goes back up to the early days of natural language processing in the 1970s. In general, Information Extraction (IE) aims to acquire knowledge from a text. Two basic information extraction tasks are named entity recognition and relationship extraction. The extraction task is carried out thanks to the filling of predefined forms. This model describes a set of entities, the relationships between them and the events involving these entities [19]. For example, a template(form) for a disease should specify fields such as: ”name of disease”, ”place of disease”, ”number of victims”.

Information extraction methods can be classified into three categories: linguistic methods, statistical methods (machine learning approaches) and hybrid methods. The linguistic methods are based on a syntactic study of text. On the other hand, statistical methods make it possible to extract information without prioritizing linguistic analysis. These methods are the most used in the processing of natural language. The hybrid methods consist of combining linguistic and statistical methods [14].

“*IE in [Arabic] poses many problems because of the morphological and graphic changesin this language: polysemy, irregular and inflected derived forms, various spellings of certain words, various writings of certain combination of characters, short(diacritics) and long vowels, most of the Arabic words contain affixes*” [20].

The Named Entity Recognition(NER) is a sub-problem of Information Extraction (IE). In the Arabic language, several systems have been created on the recognition of named entities[4]. In general, the development of an information extraction system goes through three steps: (i) identify text fragments containing information, (ii) define the structure of information representation and (iii) develop the rules to identify the information and complete the proposed form.

As mentioned previously, named entity recognition is a sub-task of information extraction. Named entities are textual elements allowing particularly relevant access to document content, that is why identifying and categorizing them is a key issue for the automatic understanding of texts. The following is a non-exhaustive list of NER tools that have been used in the Arabic NER literature.

It should be noted that the list of references provided here is not set to be exhaustive. But, to conclude, we can say that systems based on hybrid approaches have shown good performance in the recognition of Arabic named entities. But adapting these approaches for a particular task involving specific entities, like epidemic surveillance and the identification of disease names, will be costly in terms of expert time for rules creation, text annotation

3 Experiment the Lightweight Daniel Approach

For the experiments presented here, we used the code provided by the authors of Daniel[3]. In this section, we will present the dataset we built for this experiment and the results obtained on Arabic texts. Then, we will discuss what we have learned from these experiments and in the last part of this section we will propose some improvements to this approach and confront it to datasets in other languages.

3.1 Getting Dataset and Resources

Since the Daniel code is designed to work with structured press articles, we managed to build a corpus of press articles in Arabic (Table 3). The method itself relies on detecting repeated character strings to perform both classification and Information Extraction. For epidemic surveillance, the authors assume that a simple list of disease names obtained from Wikipedia pages is sufficient. The rationale is that in press articles, journalists use most common words in order to ensure that the information is conveyed properly. If scientific names are used, it is only additional to these common words since the target of the press articles is mainly composed by regular speakers rather than specialists.

With this resource, the system provides a binary classification stating if a given press article is related to epidemics or not. An article is tagged as relevant if a substring S of a disease name D is found in salient positions (title, first paragraph and last paragraph) and if length$(S)/$length$(D) >= \theta$. θ is a threshold that can be manually tuned for each language, but the authors report that $\theta = 0.8$ provides good results in various languages. This heuristic is also used to identify the disease named entities of the document.

The definition of a relevant document may need to be better defined. It is not completely clear on how the frontier between a relevant and an irrelevant article is drawn

[3] https://github.com/rundimeco/daniel

Table 3. Statistics for the Arabic dataset, French dataset and multilingual dataset.

	ar Corpus	fr Corpus	multi Corpus
Documents	41,432	2,733	2,129
Paragraphs	$488 * 10^3$	$12 * 10^3$	$25 * 10^3$
Avg. paragraphs	$11.8(\pm 23)$	$4.5(\pm 2.3)$	$12(\pm 8)$
Characters	$97 * 10^6$	$11 * 10^6$	$5 * 10^6$
Avg.characters	$2,347(\pm 3010)$	$4,168(\pm 4,604)$	$2,402(\pm 23,99)$

since the guidelines used for human annotators[4] only indicate that in relevant articles "the main theme of the article is epidemics". Once the relevant document is detected, the system tries to locate the event at country level. There again, common names from Wikipedia are used.

If the name of a country is repeated, this country is supposed to be where the epidemics take place. Else, the location is the country where the article has been published. This rule is referred to as "implicit location". Therefore, when we built our dataset, we created a resource indicating for each particular press source, the country where it is published (see Table 3), the other two corpora have been used in [31].

3.2 Testing the Daniel Approach

Creating a reference dataset for epidemic surveillance was not our first goal since it is rather costly to build a dataset of sufficient size, so we only propose to evaluate the output of the system. This configuration does not allow us to evaluate recall, but at least we can evaluate precision. In our opinion, it is possible to verify the soundness of the approach for our purpose. Furthermore, previous research on this approach showed that the system usually produced worse results in precision than in recall.

It is a somewhat counter-intuitive statement, considering the lightness of the lexical resources. Together with the DANIEL code, we have provided a reference dataset of more than 2,000 annotated articles in French. We will also perform experiments with the reference dataset "corpus_daniel[5]" used in [23] which contains around 2,100 annotated documents in five languages. These two datasets have been annotated with the same guidelines. These datasets will be exploited to test our improvement proposals for the DANIEL approach.

4 Results and Discussion

In this section, we present the first results obtained on the Arabic corpus. Among the relevant documents, we annotated 50 documents in order to assess precision. 54% of them were related to epidemic events.

This was not a good result and was far from what was reported in the literature for other languages. Therefore, we wanted to find out more about this result and to get an insight of the classification errors.

[4] https://daniel.greyc.fr/guidelines.pdf

[5] https://tinyurl.com/ResearchGate-DanielCorpus

أمير منطقة الرياض بالنيابة يدشن المرحلة الثانية للحملة الوطنية للتحصين ضد أمراض الحصبة

Prince Governor of Riyadh opens the second phase of the National Measles Campaign

بالنيابة الرياضمنطقة أمير عبدالعزيز بن سعد بن محمد الأمير الملكي السمو صاحب دشن واس م ٢٠١١ ديسمبر ٠٣ الموافق هـ ١٤٣٣ محرم ٠٨ الرياض
وأعرب . الإمارة بديوان ,وذلكوالنكاف الألمانية والحصبة الحصبةأمراض ضد للتحصين الوطنية للحملة الثانية المرحلة اليوم الحكم بقصر سموه مكتب في
الحملة،مثمناً تدشينه على بالنيابة الرياض منطقة أمير لسمو والتقدير الشكر عن العبدالكريم سليمان بن عدنان الدكتورالرياضبمنطقة الصحية الشؤون عام مدير
والعسكرية الحكومية والمعاهد والجامعات الكليات وطالبات طلاب تستهدف الحملة أن وبين .الرياض بمنطقة الصحية للخدمات دعم من سموه يقدمه ما
الفيروسية الأمراض أحد هو الحصبة مرض أن العبدالكريم وأوضح سنوات. ٦ إلى أشهر ٩ عمر من والأطفال عاماً ٢٤ إلى ١٩ عمر من الأشخاص وجميع
شبيهة وأعراض الحرارة درجة في بارتفاع علاماته وتبدأ ابريل إلى يناير شهر بين ما الفترة خلال انتشاره ويزداد رئيسي بشكل الأطفال تصيب التي
بالشفاء غالباً وتنتهي والاختفاء التلاشي في الأعراض تبدأ ثم العلوية والأطراف والصدر الوجه على طفح ظهور يعقبها أيام ٤ أو ٣ لمدة البرد بنزلات
لذلك المرسومة الخطة وفق المدارس على وتوزيعها الطبية الفرق تجهيز أكملت الرياض صحة أن وأكد حياته. طوال مناعة المصاب الطفل واكتساب
الخاصة الأمصال جميع تسلمت قد كانت إدارته أن إلى مشيراً الأولية، الصحية الرعاية لإدارة التابعة الصحية والقطاعات الوقائي الإشراف إدارة من
الميدانية الصحية الفرق لدعم والممرضات الممرضين من بمجموعة مشاركته خلال من الخاص الصحي بالقطاع استعانت كما الوزارة، وفرتها التي بالحملة
الطلاب تطعيم أن وبين . التطعيم سابقة عن النظر بغض كافة، والطالبات الطلاب لتطعيم المدرسية، الصحية الوحدات من المشاركة الفرق إلى بالإضافة
فرق وترسل به المحيط النطاق في الواقعة الكليات صحي مركز كل يغطي بحيث منطقة كل في الصحية المراكز من فرق عبر سيتم الكليات في والطالبات
الطلاب. جميع تطعيم من الانتهاء وحتى أسابيع خمسة ولمدة الأسبوع مدار وعلى يومياً الكليات داخل للعمل المراكز هذه من

Fig. 1. An example of True Positive: Measles in Saudi Arabia.

المؤتمرنت - صنعاء: جريمة جديدة لمليشيات المشترك

A new crime for the common militia.

قتل الطفل يحيى هارون في جريمة بشعة تنم عن وحشية مفرط ة وإرهاب يستهدف الجميع كبارا وصغارا , أقدم قناصة من أحزاب اللقاء المشترك وحلفائهم
وأذيالهم صباح اليوم على قتل الطفل يحيى جميل هارون ١٣ سنة في حي الحصبة بالعاصمة صنعاء. وقال شهود عيان أن أولئك القناصة المتمركزين في معهد
التوجيه والإرشاد صوبوا أسلحتهم باتجاه الطفل يحيى هارون وهو في طريقه لإحضار فطور لأسرته وأطلقوا عليه أعيرة نارية أصابته في الرأس وفجرت
دماغه واستشهد على الفور.
وكان الطفل الشهيد يسكن مع أسرته التي تنتمي إلى منطقة بني الحارث بمحافظة صنعاء في حارة الخرابة جوار المعهد البيطري.
وقوبلت هذه الجريمة الشنيعة باستنكار وتنديد شعبي واسع ومطالبات بسرعة القبض على قتلة هذا الطفل ومن يقف وراءهم وإحالتهم إلى أجهزة العدالة
لينالوا جزاءهم العادل والرادع.
واعتبرت منظمات المجتمع المدني وحقوق الإنسان والطفل أن هذه الجريمة لم تكن الأولى التي ترتكبها تلك العناصر الإجرامية بحق الأطفال, إذ سبق
لها أن قتلت وأصابت العديد من الأطفال الأبرياء بدم بارد في عمليات قصف لمنازل المواطنين بقذائف الهاون والبوازيك وغيرها من الأسلحة في أحياء
الحصبة والخرابة وقرية الدجاج والإدارة المحلية وبيت القحوم والمنازل المجاورة لوزارة الداخلية خلال الفترة الماضية, مشددة على سرعة قيام أجهزة الأمن
بواجبها والقبض على قتلة الأطفال والمواطنين الأبرياء.
واعتبرت تلك المنظمات أن عدم تطبيق القانون وملاحقة الجناة سيجعل أولئك المجرمين يتمادون في غيهم ويقتلون مزيدا من الأطفال الأبرياء ما داموا
بعيدا عن الملاحقة والمساءلة القانونية.

Fig. 2. A False Positive example: there is a confusion because a city is named “Measles”.

We present here two examples of documents deemed relevant by DANIEL. In these two illustrations, the red colour represents the disease and the yellow colour the place of the event. The article presented in Figure 1 describes cases of measles in Riyadh (Saudi Arabia) and is, according to the guidelines mentioned earlier, a true positive. On the contrary, Figure 2 shows a False Positive: because this article speaks about a city which is called “measles”.

We can see with the True Positive example that the repetition patterns, the so-called “relevant content” algorithm, succeeds in detecting the main subject of the article. This is not surprising as the 5W rule used in this article is also known in Arabic rhetorics[6]. In the other example, the misclassification is not due to the algorithm itself, but rather to the lexical resources it exploits.

The character string extracted is a rather small one and therefore tends to be more frequent and is prone to ambiguity. In [23] the only parameter used to avoid such False Positive cases is the θ ratio between the found substring and the lexical entry of the database. In the code published online, some corrections are made to take into account different positions in the document. But neither of these two methods is suitable to resolve the problem identified here.

Tuning the ratio would not help here, since the full string is found in the document.Modifying the relevant positions would not help either, since repetitions in first paragraph and body of the article are usually quite efficient to assess the theme of the document. Another solution might be to remove this disease name from the database, but this would surely lead to an increasing number of False Negatives. The French corpus and the multilingual corpus provided by the authors show similar False Positives cases.

In the multilingual corpus we had a False Positive case regarding “Odra” (“Measles” in Polish) because Odra is also the name of a river and the name of a small city. Another example involved the French name for ”scabies” which is “gale”. The substring “gale” is not uncommon in French so that some False Positives identified in the data were due to the substrings of that particular disease name. An alternative approach may be to try linguistic pre-processing and Named Entity recognition but, it would imply a paradigm shift.

So we want to find a solution that keeps the originality and the multilinguality of the original approach. We believe that the length of the disease name is the key. The longer the disease name is, the less ambiguous it is and the more confident the system should be. Setting a minimum length threshold would not fulfil the purpose. Some disease names are short, and the length would need to be tuned according to the language. The solution we propose here is to take into account not only the θ ratio but also the length of the disease name itself to compute a confidence score.

4.1 Taking into account the length of the disease names

In this configuration, we sort the documents selected by the system with respect to the length of the disease names. We take advantage of the two annotated corpora at our disposal. We observe that False Positives come mostly from short disease names. We try different configurations, first with the longest disease names ($length >= 10$) and then with names of length 9, 8 …and so on until all the disease names are introduced.

The rationale of this experiment is to see from a ROC curve how recall and precision evolve. Figure 3 shows the results obtained on the French corpus and multilingual corpus, where we can see a property of the length of the disease names. Longer disease names lead to a better precision with a recall quite low.

Shorter disease names are introduced step by step, increasing the recall but at some point with an important cost in precision. In order to have a reasonable number of False Positives, some improvements of the algorithm need to be performed.

[6] https://en.wikipedia.org/wiki/Five_Ws

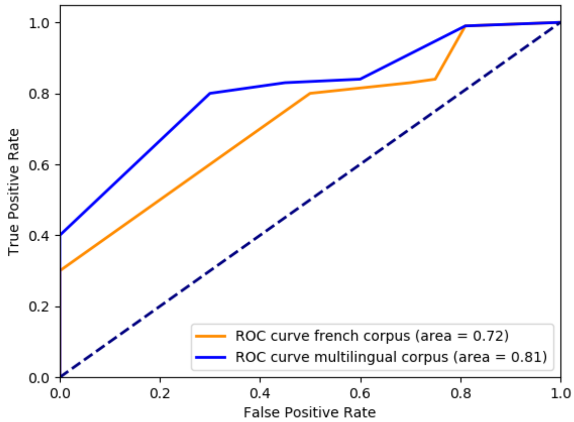

Fig. 3. ROC curve of the DANIEL system with different compositions of the lexical resource.

4.2 Discussion

We advocate that there are two ways to do this. The first one would be to use a measure to assess the potential ambiguity of the substrings found in the document. This can be done using additional lexical resources or measures like the adaptation measure [9] would surely help to improve results without setting or learning language-dependent length parameters. The other way, and more promising, would be to use long disease names to bootstrap the system by getting for each language a bunch of annotated documents with great confidence scores.

These annotated documents will then be used to learn words of the domain that are not disease names so that it would be possible to resolve ambiguities for shorter words independently of expert data. In the future, we plan to build an annotated dataset of sufficient size for Arabic to experiment with this solution, since it would help to assess how the lightweight approach for text classification and Information Extraction can be useful for Arabic texts.

References

1. Ababneh, J., Almanmomani, O., Hadi, W., El Omari, N., Alibrahim, A.: Vector space models to classify arabic text. International Journal of Computer Trends and Technology, vol. 7, pp. 219–223 (2014)
2. Abutiheen, Z. A., Aliwy, A. H., Aljanabi, K. B. S.: Arabic text classification using master-slaves technique. Journal of Physics: Conference Series, vol. 1032, no. 1, pp. 012052 (2018)
3. Al Tahrawi, M. M., Al Khatib, S. N.: Arabic text classification using polynomial networks. Journal of King Saud University - Computer and Information Sciences, vol. 27, no. 4, pp. 437–449 (2015)

4. Alruily, M., Ayesh, A., Zedan, H.: Arabic language in the context of information extraction task. In: Journal of Computational Linguistics Research. (2012)
5. Arsevska, E., Valentin, S., Rabatel, J., De Goër De Hervé, J., Falala, S., Lancelot, R., Roche, M.: Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. Public Library of Science, vol. 13, no. 8, pp. 1–25 (2018)
6. Barboza, P., Vaillant, L., Mawudeku, A., Nelson, N. P., Hartley, D. M., Madoff, L. C., Linge, J. P., Collier, N., Brownstein, J. S., Yangarber, R., et al.: Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. Public Library of Science, vol. 8, no. 3, pp. 1–9 (2013)
7. Bawaneh, M., Alkoffash, M., Alrabea, A.: Arabic text classification using K-NN and Naive Bayes. Journal of Computer Science, vol. 4 (2008)
8. Chouigui, A., Khiroun, O. B., Elayeb, B.: Ant corpus: An arabic news text collection for textual classification. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications. pp. 135–142 (2017)
9. Church, K.: Empirical estimates of adaptation: The chance of two noriega's is closer to p/2 than p 2. In: Proceedings of the 18th International Conference on Computational Linguistics. pp. 173–179 (2000)
10. Collier, N.: Towards cross-lingual alerting for bursty epidemic events. Journal of Biomedical Semantics, vol. 2, no. 5, pp. 1–11 (2011)
11. Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., Yangarber, R.: Building support tools for Russian-language information extraction. In: Habernal, I., Matoušek, V. (eds.) Proceedings of the 14th international conference on Text, Speech and Dialogue. pp. 380–387. Springer (2011)
12. Duwairi, R.: Arabic text categorization. Int. Arab J. Inf. Technol., vol. 4, pp. 125–132 (2007)
13. Elhassan, R., Ahmed, M.: Arabic text classification review. International Journal of Computer Science and Software Engineering, vol. 4, pp. 2409–4285 (2015)
14. Elsadig, M., Ahmed, A., Himmat, M.: Information extraction methods and extraction techniques in the chemical document's contents: Survey. ARPN Journal of Engineering and Applied Sciences, vol. 10, pp. 1068–1073 (2015)
15. Etzioni, O., Fader, A., Christensen, J., Soderland, S.: Open information extraction: The second generation. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 3–10 (2011)
16. Fisichella, M.: Unified approach to retrospective event detection for event- based epidemic intelligence. International Journal on Digital Libraries, pp. 1–26 (2021)
17. Freifeld, C. C., Mandl, K. D., Reis, B. Y., Brownstein, J. S.: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. Journal of the American Medical Informatics Association, vol. 15, no. 2, pp. 150–157 (2008)
18. Gales, M. J., Knill, K. M., Ragni, A., Rath, S. P.: Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In: Spoken Language Technologies for Under-Resourced Languages. pp. 16–23. International Speech Communication Association (2014)

19. Grishman, R., Sundheim, B.: Message understanding conference 6: A brief history. In: International Conference on Computational Linguistics. vol. 96, pp. 466–471 (1996)
20. Hajjar, M., Zreik, K.: Classification of arabic information extraction methods. In: 2nd International Conference on Arabic Language (2009)
21. Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Kenneth, D. M., Mawudeku, A., Eysenbach, G., Brownstein, J. S.: Use of unstructured event-based reports for global infectious disease surveillance. Emerging Infectious Diseases, vol. 15, no. 5, pp. 689–695 (2009)
22. Lejeune, G., Lucas, N., Doucet, A.: Tentative d'approche multilingue en extraction d'information. In: JADT Journées internationales d'Analyse statistique des Données Textuelles. pp. 1259–1267 (2010)
23. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. Artificial Intelligence in Medicine, (2015)
24. Lejeune, G., Brixtel, R., Lecluze, C., Doucet, A., Lucas, N.: Daniel : Veille épidémiologique multilingue parcimonieuse. In: Proceedings of Traitement Automatique des Langues Naturelles. pp. 787–788 (2013)
25. Ma, X., Hovy, E. H.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. Computing Research Repository, (2016)
26. Marton, Y., Habash, N., Rambow, O., Alkhulani, S.: Shared task system: The CADIM arabic dependency parser (2013)
27. Mesleh, A. M., Kanaan, G.: Support vector machine text classification system: Using ant colony imization based feature subset selection. In: 2008 International Conference on Computer Engineering Systems. pp. 143–148 (2008)
28. Mesleh, A. M.: Support vector machines based arabic language text classification system: Feature selection comparative study. In: Advances in Computer and Information Sciences and Engineering, Proceedings of the 2007 International Conference on Systems, Computing Sciences and Software Engineering. Part of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering. pp. 11–16 (2007). doi: 10.1007/978-1-4020-8741-7_3
29. Mirończuk, M. M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, vol. 106, pp. 36–54 (2018)
30. Munro, R.: Processing short message communications in low-resource languages. Ph.D. thesis, Stanford University (2012)
31. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., Odeo, M.: Multilingual epidemiological text classification: A comparative study. In: International Conference on Computational Linguistics. pp. 6172–6183 (2020)
32. Mutuvi, S., Boros, E., Doucet, A., Lejeune, G., Jatowt, A., Odeo, M.: Token-level multilingual epidemic dataset for event extraction. In: Berget, G., Hall, M. M., Brenn, D., Kumpulainen, S. (eds.) Linking Theory and Practice of Digital Libraries. Springer International Publishing. pp. 55–59 (2021)
33. Recupero, D. R., Dragoni, M., Buscaldi, D., Alam, M., Cambria, E. (eds.): Proceedings of 4th Workshop on Sentic Computing, Sentiment Analysis, Opinion Mining, and Emotion Detection. Co-located with the 15th Extended Semantic Web Conference, vol. 2111 (2018)

34. Steinberger, R.: A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation, pp. 1–22 (2011)
35. Syiam, M., Tolba, M., Fayed, Z., Abdel-Wahab, M., Ghoniemy, S., Habib, M.: An intelligent system for arabic text categorization. International journal of cooperative information systems, vol. 6, no. 1, pp. 1–19 (2006)
36. Zhan, L., Jiang, X.: Survey on event extraction technology in information extraction research area. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference. pp. 2121–2126 (2019)