IMAN

المؤتمر الدولي للتطبيقات الإسلامية في علوم الحاسوب وتقنياته – إيمان 2019 (27 – 28 كانون الأول / ديسمبر 2019)
International Conference on Islamic Applications in Computer Science and Technologies - IMAN 2019 (27-28 December 2019)

Non-authentic Hadith Corpus: Design and Methodology

Taghreed Tarmom[1], Eric Atwell[2], Mohammad Alsalka[3]

School of Computing, University of Leeds, Leeds
[1]sctat@leeds.ac.uk, [2]e.s.atwell@leeds.ac.uk, [3]M.A.Alsalka@leeds.ac.uk

ABSTRACT
The primary religious text of Islam is the Quran. The Hadith—the second source—refers to any action, saying, order or silent approval of the holy prophet Muhammad that has been delivered through a chain of narrators. Each Hadith has an Isnad—the chain of narrators—and a Matan—the act of the Prophet Muhammad. In contrast to the Quran, some Hadiths, which have been handed down over the centuries, have been corrupted by narrators who were not competent in transferring them. These have been classified by Hadith scholars as a non-authentic Hadith (NAH).

To evaluate different classifiers regarding the automatic classification of Arabic Hadith, it was necessary to build Arabic Hadith corpora that contained samples of authentic and non-authentic Hadith, which were used for training models and testing. This paper aimed to create a new NAH corpus which consists of 452,624 words from six different Hadith books. The subsequent aim is to annotate this corpus to determine some Hadith features such as the Isnad, the Matan and the Hadith authenticity and to provide a ground truth.

Keywords: **Hadith Corpus, Non-authentic Hadith, Arabic, Natural Language Processing, Corpus Linguistics.**

1. Introduction

Corpus linguistics is the study of language through the collection and analysis of textual data, a corpus (plural, 'corpora'). Such data can consist of continuous text from books and websites or collections of quotations. Corpora have been compiled for different reasons and purposes. Some existing corpora were specifically designed for linguistic research, such as the prosody, grammar and discourse patterns of the language (Kennedy, 1998). Other corpora are used for natural language processing (NLP) research, for example in training and testing materials (Alkahtani & Teahan, 2015). Most NLP research for the Arabic language is focused on Modern Standard Arabic (Tarmom et al., 2018), leaving a shortage of research in classical Arabic such as the Hadith.

To evaluate different classifiers with respect to the automatic classification of the Hadith's authenticity, it was necessary to build an Arabic Hadith corpus that contained samples of authentic Hadith and non-authentic Hadith (NAH), which were used for training models and testing. We already have access to authentic Hadiths (a corpus of major books that are widely accepted as authentic, which was built by Altammami et al., 2019). Therefore, for a balanced corpus of both positive and negative examples, we need to collect an Arabic NAH corpus. In addition, most Arabic Hadith corpora concentrate on authentic Hadiths, but there is a shortage of NAH corpora. This was the main motivation for building an NAH corpus. Also, the existing Arabic corpora are quite expensive and/or are of poor quality (Alkahtani, 2015). Hence, there was a need to create a free Arabic Hadith corpus.

The NAH corpus consists of seven different corpora. Each corpus represents a Hadith book. These books can be been found on the Hadith websites such *as islamweb.net* and *almeshkat.net*. Some of them have both Hadiths (authentic and NAH), while others only contain NAH.

2. Methodology

Extracting data manually is time-consuming; therefore, we have built an application with Python to automatically extract data from the *islamweb.net* website. This application allows data collection from different pages and places them in a csv file. It also divides each Hadith by Matan, Hadith Type, authenticity, URL and the book's name. However, several books have been downloaded from *almeshkat.net* website as Word files and converted to csv files.

Al-Kabi et al. (2014) noted that removing diacritics enhanced the classification result. Hence, we used the pyarabic-master library to remove diacritics from each Hadith. The pyarabic-master library is an Arabic text-processing library for Python which supplies useful functions to manipulate Arabic text. For example, it can tokenize text into words, remove tashkeel 'diacritics' and filter out non-Arabic words and so on (Zerrouki, 2010). Figure 1 shows a Hadith with full diacritics, while Figure 2 shows the same Hadith after removing all diacritics.

" مَنْ كَذَبَ عَلَيَّ مُتَعَمِّدًا فَلْيَتَبَوَّأْ مَقْعَدَهُ مِنَ النَّارِ "

Figure 1: Hadith with full diacritics.

" من كذب علي متعمدا فليتبوأ مقعده من النار "

Figure 2: Hadith after removing diacritics.

Each Hadith has been manually checked to verify that it was labelled correctly. To do that, each Hadith was read for the following reasons:

1. To verify that the Isnad and the Matan have been correctly separated. As mentioned above, the Isnad is the chain of narrators; however, we do not know the specific number of narrators. While some Hadiths have nine narrators, some have more, and some have fewer. Therefore, separating the Isnad and the Matan by the numbers of narrators is not useful. We have noticed that the Matan is written between parentheses; thus, we used this as a rule in the separating process in our application. We then manually compared the original Hadith with the Isnad and the Matan that had been separated to correct any errors.

   We believe that this process will be more accurate if we use a compression-based segmenter to solve it. However, this segmenter needs corpora for training models and testing; thus, we have built this corpus.

2. To verify the authenticity of each Hadith. The author's comment at the end of each Hadith must be read. Our application searches for the expressions that the author used to describe the Hadith's authenticity. For example, if the author wrote '**حَدِيثٌ صَحِيحٌ**', then this Hadith will be authentic (see Figure 3). Figure 3 shows a sample of our application code to label the authenticity of each Hadith. During this stage, we also read each Hadith to see if it was labelled accurately, and we corrected it if it was wrong.

3. To manually add a topic label for each Hadith.

These steps took approximately one month (~80 hours) per corpus to complete. We spent one hour checking nine Hadiths, and we worked four hours daily to check approximately 36 Hadiths per day. Figure 4 shows the workflow diagram for building each corpus that was produced for this research.

```
if 'حديث صحيح'in hadith:
  degree = 'صحيح'
  authenticity = 'authentic'
elif 'حديث موقوف' in hadith and 'حديث صحيح'in hadith:
    degree =  'صحيح'
    authenticity = 'authentic'
elif 'حديث حسن' in hadith:
    degree = 'حسن'
    authenticity = 'authentic'
```

Figure 3: Sample of our application code to label the authenticity of each Hadith.

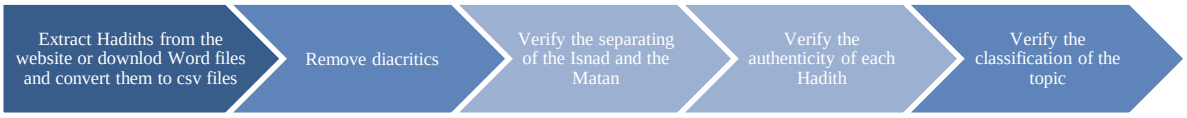

Figure 4: A workflow diagram for building each corpus that was produced for this research.

3. Non-authentic Hadith Corpus Content

The first book that was used to create our corpus was found on *islamweb.net*. It was 'الأباطيل والمناكير والمشاهير للجورقاني', which was written by 'عبد الرحمن بن عمر الجورقاني', who died in 1148. It has 732 Hadiths, including both authentic and non-authentic examples. The author added his comments after each Hadith to describe the authenticity.

After we finished annotating this corpus, we found out that all Hadith books which contain authentic and NAH have been removed from the *islamweb.net* website for an audit process. Therefore, we moved to the *almeshkat.net* website. Several books have been downloaded as Word files and converted to csv files. These books are 'مائة حديث ضعيف وموضوع منتشرة بين اللآلئ المصنوعة في الأحاديث', 'اللآلئ المصنوعة في الأحاديث الموضوعة الجزء الأول', 'الخطباء والوعاظ الجد الحثيث في بيان ما ليس بحديث', 'الأحاديث الضعيفة في كتاب رياض الصالحين', 'الموضوعة الجزء الثاني and 'الفوائد المجموعة في الأحاديث الموضوعة'. The columns in Figure 5 show our corpus contents. The *No.* refers to the book number in this corpus, then the *Corpus Name.* In the *Corpus Name* field, *N* refers to a non-authentic word, while _1 and _2 means that this book has two parts (_1 is part one and _2 is part two). This is followed by *Book's Title* and *Author*. Some

books have Isnad, Matan and comments, while others just have Matan and comments, so we added *Book's Contents* to clarify these contents. Then we added *Book's Description*, then *No. of words* and finally, the *URL*.

No.	Corpus Name	Book's Title	Author	Book's Contents	Book's Description	No. of words	URL
1	N1	الأباطيل والمناك	أبو عبد الله الهم	Isnad, Matan, Cc	ألفه الحافظ الجورقانى	121,080	http://www.i
2	N2	مائة حديث ضع	إحسان العتيبى	Matan, Commer	تحذير من رواية الأحاديـ	2,898	https://www.
3	N3_1	اللآلئ المصنوعة	جلال الدين الس	Isnad, Matan, Cc	كتاب يبحث فى علم الحـ	15,421	https://www.
4	N3_2	اللآلئ المصنوعة	جلال الدين الس	Isnad, Matan, Cc	كتاب يبحث فى علم الحـ	151,382	https://www.
5	N4	الأحاديث الضعيـ	إحسان العتيبى	Isnad, Matan, Cc	الحمد لله والصلاة والسلا	5,675	https://www.
6	N5	الجد الحثيث فى	أحمد بن عبد الـ	Matan, Commer	-	16,382	https://www.
7	N6	الفوائد المجموع	الإمام محمد بن	Matan, Commer	قال الشوكانى رحمه الله	139,786	https://www.

Figure 5: The NAH corpus contents.

4. Annotating process and some challenges

At this time, the annotating process has been completed for three books. Thus, each Hadith in these three corpora has seven primary features or attributes (see the columns in Figure 6). The first feature is the *Full Hadith*, which contains all the Hadiths as they appear in the book. Figure 7 shows how the author wrote each Hadith; the Hadith number (red square) is followed by the Hadith type (blue square). After that, we show that he wrote the Isnad (between the blue square and the black square), followed by the Matan which was written between parentheses (black square). Finally, at the end of Hadith, the author describes the authenticity of each Hadith (green square). Some other features for each corpus include the *Isnad*, the *Matan*, the *Author's Comments,* the *Hadith Type* and *Authenticity*.

No.	Full Hadit	Isnad	Matan	Author's (	Degree	Authentic	Topic
1	رقم الحديث: 1	فقال : فيما أخبر	من كذب على م	أخرجه الإمام أبو	صحيح	authentic	لايوجد
2	رقم الحديث: 2	أخبرنا أبو الفرج	لا تكذبوا على فـ	أخرجه الإمام أبو	صحيح	authentic	لايوجد
3	رقم الحديث: 3	أخبرنا أبو طاهر	بلغوا عنى ولو آ	فارتكب هذه الـ	مرفوع	non-authenti	لايوجد
4	رقم الحديث: 4	فقد أخبرنا حمزة	أقبل رجل ، فلما	وفيه الدلالة على	صحيح	authentic	لايوجد
5	رقم الحديث: 5	أخبرنا محمد بن	عن الرجل لا يحـ		مقطوع		لايوجد
6	رقم الحديث: 6	أخبرنى أبو بكر ع	كذاب والله ، لو		مقطوع		لايوجد
7	رقم الحديث: 7	أخبرنا محمد بن	إذا علم الرجل مـ		مقطوع		لايوجد
8	رقم الحديث: 8	إأخبرنا عبد الملك	يا أبا عبد الله ،	فظهر بهذا الإجم	مقطوع		لايوجد
9	رقم الحديث: 9	فقد أخبرنا حمزة	لأن أعرف علة حـ		مقطوع		لايوجد
10	رقم الحديث: 10	أخبرنا عبد الملك	من لم يعرف حد		مقطوع		لايوجد
11	رقم الحديث: 11	أخبرنا أبو بكر اله	إن العالم إذا لم	فمما يعرف به ص	مقطوع		لايوجد

Figure 6: Features of each corpus in the NAH.

رقم الحديث 1

(حديث مرفوع) فقال : فيما أخبر أبو الفضل محمد بن طاهر بن علي المقدسي ، رضي الله عنه ، قال : أخبرنا علي بن أحمد بن البندار ، قال : حدثنا أبو طاهر محمد بن العباس المخلص ، قال : حدثنا عبد الله بن محمد بن عبد العزيز البغوي ، قال : حدثنا أبو خيثمة زهير بن حرب ، قال : حدثنا إسماعيل بن إبراهيم ، عن عبد العزيز بن صهيب ، عن أنس ، عن النبي صلى الله عليه وسلم ، قال : " من كذب علي متعمدا فليتبوأ مقعده من النار " . هذا حديث صحيح ، أخرجه الإمام أبو الحسين مسلم بن الحجاج النيسابوري في صحيحه ، عن أبي خيثمة زهير بن حرب هكذا .

Figure 7: A screenshot of the first Hadith from the first book.

In some Hadith books, Hadiths have been classified by their topics, so we added the *Topic*

feature. In the first book, we noticed that there was confusion in this classification, because in *the prayer* 'الصلاة' topic, there are some sections about *the charity* 'الزكاة'. In addition, in *the fasting* 'الصوم' topic, there are some sections about *the pilgrimage* 'الحج'. This may cause a problem if we try to use automatic topic classifications.

In Hadith books, there are different types of Hadith, such as *Maqtu`* 'مقطوع', *Mawquf* 'موقوف' and *Marfo'* 'مرفوع'. The *Maqtu`* Hadith refers to sayings, actions and explanations attributed to a man who was meet the Prophet Muhammad's friends (a successor), whether it is a narration attributed of that man or otherwise (Ibn al-Salah, 1236). The *Mawquf* Hadith describes a statement or action of the Prophet Muhammad's friends (the sahaba). The *Marfo'* Hadith refers to any action, saying or order that was done by Prophet Muhammad and has been delivered through a chain of narrators (Ibn al-Salah, 1236). All these types of Hadith could be either authentic or non-authentic. To make that determination, Hadith scholars follow certain rules. As previously stated, the author describes the authenticity of each Hadith, but some Hadiths lack comments. Therefore, we do not know their authenticity. For example, in Figure 7, the beginning of the Hadith, in the first book, lists *Marfo'*, and at the end, the author states that this is an authentic Hadith (highlighted in yellow). By contrast, in Figure 8, the author does not describe the authenticity of this Hadith.

رقم الحديث 14
(حديث مرفوع) أخبرنا حمد بن نصر بن أحمد الحافظ ، أخبرنا عبد الرحمن بن غزو بن محمد ، قال : حدثنا أحمد بن إبراهيم بن أحمد بن تركان ، أخبرنا محمد بن الحسين بن علي ، قال : حدثنا محمد بن جعفر بن علي بن أحمد بن محمد بن الأحنف بن قيس التميمي الخوارزمي ، قال : حدثنا مأمون بن أحمد السلمي ، قال : حدثنا أحمد بن عبد الله الجويباري الهروي ، قال : حدثنا سفيان بن عيينة ، عن ابن طاوس ، عن أبيه ، عن ابن عباس ، عن النبي صلى الله عليه وسلم ، قال : " الإيمان لا يزيد ولا ينقص " .

Figure 8: An example of a Hadith with the author's comment missing.

Table 1 shows the expressions that were used by authors to describe the authenticity of each Hadith.

Authentic	Non-authentic
صحيح	موضوع
حسن صحيح	غير صحيح
حسن	باطل
محتمل التحسين	مضطرب
قوي	منكر
جيد	كذب
رواته ثقات	ضعيف
	ليس لهذا الحديث أصل
	موقوف منكر

Table 1: Expressions used by authors to describe the authenticity of each Hadith.

There are other types of Hadith for which authors did not describe the authenticity, such as غريب (a hadith that has been transferred by only one narrator), عزيز (a hadith that has been transferred by two narrators) and مشهور (a hadith that has been transferred by three or more narrators). To verify the authenticity of these Hadiths, we must study each one and confer with Hadith scholars, which will take additional time.

The third book, which is 'اللآلئ المصنوعة في الأحاديث الموضوعة الجزء الأول', has two parts; this book went through a harder annotation process, since it is written in a very old style, and it does not have a clear structure. Thus, it had a manual annotating process, which took more time than we expected. Figure 9 shows an image of the original 'اللآلئ المصنوعة في الأحاديث الموضوعة' book pages. Figure 10 is a screenshot of the 'اللآلئ المصنوعة في الأحاديث الموضوعة' Word file that was used to build our corpus. We learned that annotating Isnad, Matan and author's comments in this book was a difficult task, because Matans in this book were not written between parentheses as in the other Hadith books, so there is no option other than to manually annotate them. Also, the Hadiths have been written one after the other without numbers.

Figure 9: A screenshot of the original 'مخطوطة كتاب اللآلئ المصنوعة في الأحاديث الموضوعة' book pages (*www.alukah.net*, 2016).

كتاب التوحيد

(الحاكم) (ج) أنبأنا إسماعيل بن محمد الشعراني أخبرت عن محمد بن شجاع الثلجي أخبرني حبان بن هلال عن حماد بن سلمة عن أبي الهزم عن أبي هريرة قال قيل يا رسول الله مم ربنا قال من ماء مرور لا من أرض ولا من سماء خلق خيلا فأجراها فعرقت فخلق نفسه من ذلك العرق.

موضوع اتهم به محمد بن شجاع ولا يضع مثل هذا مسلم قلت ولا عاقل قال الذهبي في الميزان ابن شجاع هذا كان فقيه العراق في وقته وكان حنفياً صاحب تصانيف وكان من أصحاب بشر المريسي وكان ينتقص الإمامين الشافعي وأحمد، وكان من وصيته التي كتبها عند موته ولا يعطى من ثلثي إلا من قال القرآن مخلوق، وقال ابن عدي كان يضع أحاديث في التشبيه ينسبها إلى أصحاب الحديث فيتهم بذلك منها هذا الحديث وحبان بن هلال ثقة، قال الذهبي هذا الحديث مع كونه أتى من المكذب فهو من وضع الجهمية ليذكروه في معرض الاحتجاج به، على أن نفسه اسم لشيء من مخلوقاته، فكذلك إضافة كلامه إليه من هذا القبيل إضافة ملك بل كلامه بالأولى، قال وعلى كل حال فما يعد مسلم هذا في أحاديث الصفات تعالى الله عن ذلك انتهى والله أعلم.

Figure 10: A screenshot of the 'اللآلئ المصنوعة في الأحاديث الموضوعة' Word file that was used to build our corpus.

5. Conclusion and Future work

This paper described the production of the NAH corpora in detail. The NAH consists of 452,624 words from six different Hadith books. Each corpus took approximately one month (~80 hours) to go through the annotation process. They have different types of Hadiths, with the authenticity described at the end of each by the authors. In addition, some types of Hadiths, such as غريب and مشهور, lack a description. Hence, we do not know the authenticity of these Hadiths.

We will continue building and annotating this corpus, to add several Hadith books. it will then be used for training models and for testing the automatic classification/segmentation of Arabic Hadith.

Acknowledgments

We would like to thank Sheik Dr Abdullah Bin Yusuf Al-Judai, President of the European Council for Fatwa and Research, for his explaining some concepts in the science of Hadith. Also, we would like to thank Dr Fares Al-Qunaieer and his group for their effort of translating 330 machine learning terms from English to Arabic; this file helped us when we translated the abstract into the Arabic language.

References

Al-Kabi, M., Wahsheh, H. and Alsmadi, I. (2014). A topical classification of Hadith Arabic text. IMAN, 2014, 2.

Alkahtani, S. (2015). Building and verifying parallel corpora between Arabic and English. PhD thesis. Computer Science, Bangor University. Available at http://e.bangor.ac.uk/6546/1/saad_alkahtani_dissertation.pdf

Alkahtani, S. and Teahan, W. (2015, December). A new parallel corpus of Arabic/English. In Proceedings of the Eighth Saudi Students Conference in the UK, January (p. 279). World Scientific.

Altammami, S., Atwell, E. and Alsalka, A., 2019. Text segmentation using n-grams to annotate Hadith corpus. In Proceedings of the 3rd Workshop on Arabic Corpus Linguistics, 31–39.

Alukah.net. (2016). *مخطوطة كتاب اللآلي المصنوعة في الأحاديث الموضوعة*. [online] Available at https://www.alukah.net/manu/files/manuscript_5520/makhtot.pdf [Accessed 21 Nov. 2019].

Ibn al-Salah. (1236). Muqaddimah Ibn al-Salah 'Introduction to the Science of Hadith'. pp. 193–195, Dar al-Ma'arif, Cairo.

Kennedy, G. (1998). An introduction to corpus linguistics. London & New York: Longman.

Tarmom, T., Teahan, W., Atwell, E. and Alsalka, M. (2018). Compression vs traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. Submitted to *Journal of Natural Language Engineering*.

Zerrouki, T. (2010). Pyarabic, an Arabic language library for Python. Available at https://pypi.python.org/pypi/pyarabic/.

الملخص العربي

في الإسلام، هنالك توافق على نصين دينيين. النص الأساسي هو القرآن الكريم والثاني هو الحديث الشريف. أما الحديث فانه يشير إلى أي عمل أو قول أو أمر أو موافقة من النبي صلى الله عليه وسلم. للحديث مكونين أساسيين هما الإسناد والذي هو عبارة عن سلسلة من الرواة الذين ساعدوا في ايصال هذا الحديث إلينا والمتن والذي يمثل قول الرسول صلى الله عليه وسلم أو فعله. على عكس القرآن، فإن بعض الأحاديث قد أُفسدت على مر القرون بواسطة رواة لم يكونوا مؤهلين في إيصالها إلينا لذلك قام علماء الحديث بتصنيف الأحاديث إلى أحاديث صحيحة وغير صحيحة.

لتقييم المصنفات المختلفة فيما يتعلق بالتصنيف التلقائي للحديث العربي، كان من الضروري بناء ذخيرة نصية للحديث العربي والتي تحتوي على عينات من الأحاديث الصحيحة والغير صحيحة، والتي يمكن استخدامها كنماذج للتدريب والاختبار. يهدف هذا البحث بالدرجة الاولى إلى تكوين ذخيرة نصية جديدة للحديث غير الصحيح والتي تحتوي على ٤٥٢،٦٢٤ كلمة من ستة كتب حديث مختلفة. الهدف الثاني هو توسيم هذه الذخيرة النصية لتحديد بعض معالم الحديث مثل الإسناد، والمتن وصحة الحديث وتوفير العلامات المرجعية.