Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1462–1471
Marseille, 20-25 June 2022

Challenging the Transformer-based models with a Classical Arabic dataset: Quran and Hadith

Shatha Altammami[1,2], **Eric Atwell**[2]
King Saud University[1], University of Leeds[2]
Saudi Arabia, UK
Shaltammami@ksu.edu.sa[1]
{Scshal, E.S.Atwell}@leeds.ac.uk[2]

Abstract

Transformer-based models showed near-perfect results on several downstream tasks. However, their performance on classical Arabic texts is largely unexplored. To fill this gap, we evaluate monolingual, bilingual, and multilingual state-of-the-art models to detect relatedness between the Quran (Muslim holy book) and the Hadith (Prophet Muhammed teachings), which are complex classical Arabic texts with underlying meanings that require deep human understanding. To do this, we carefully built a dataset of Quran-verse and Hadith-teaching pairs by consulting sources of reputable religious experts. This study presents the methodology of creating the dataset, which we make available on our repository, and discusses the models' performance that calls for the imminent need to explore avenues for improving the quality of these models to capture the semantics in such complex, low-resource texts.

Keywords: Hadith, Quran, dataset, semantic similarity.

1. Introduction

Recently, state-of-the-art (SOTA) transformer-based models showed unprecedented performance that surpasses human scores on common benchmarks like SICK (Marelli et al., 2014), which created a hype into believing such models could 'understand' textual data (Bender and Koller, 2020). However, an emerging trend of probing these models yields striking results that suggest models are not able to generalize, but rather are more likely to memorize (Elangovan et al., 2021). This is due to the common practice of generating training and testing data using random split, which inadvertently leads to data leakage from the training set to the testing set reaching 70% of instance overlap (Lewis et al., 2020). Moreover, existing benchmark datasets possess a low readability index which does not reflect real-world complex data (Chandrasekaran and Mago, 2020). This is aggravated for classical Arabic (CA) since it is considered low-resource in terms of available datasets and inherently challenging for natural language processing (NLP) tasks (Habash, 2010).

Equipped with these observations, we evaluate the performance of SOTA models on a binary classification task of identifying whether two pieces of CA text convey the same underlying message, which is a special case of semantic similarity. However, before undergoing the experiments, we introduce the datasets that address the two aforementioned limitations: low-readability and data leakage.

To address the low-readability index limitation, we created a new CA dataset of related religious texts from the Quran (Muslim holy book) and the Hadith (Prophet Muhammed teachings). It comprises embedded meanings which require reasoning and deep human understanding. Moreover, these texts contain complex syntactic and rhetorical features including, verbal idioms, sarcasm, hyperbole, rhetorical questions just to name a few (Abdul-Raof, 2013).

To mitigate the limitation of data-leakage, we fine-tune the models on our extended version of the Qursim (Sharaf and Atwell, 2012) dataset consisting of related Quran-verse pairs. Then we test the best performing models on our more challenging dataset of Quran-Hadith pairs. Throughout this paper, we use the phrase 'Quran-Hadith pairs' where 'Hadith' refers to an instance of a Hadith teaching (Matn), and the term 'Quran' refers to a verse. In the next section, we give more background on these texts and discuss their challenges.

2. Background: Quran and Hadith

Muslims believe the Quran is God's divine words transmitted to the prophet Muhammad by the angel Gabriel over a period of 23 years. This holy book enjoins Muslims to follow the guidance of Prophet Muhammad in their laws, legislation, and moral guidance. In fact, most laws and legislation are obtained from Hadith, which is the reports of prophet Muhammed's statements and actions. Hence, many Islamic rulings (Fatwa) use Quran and Hadith together as evidence.

What makes these texts complex is the Arabic language inherent challenges. One of the main challenges is

its agglutinative property where a sentence can be expressed in one word by adding affixes and clitics that represent various parts of speech (Habash, 2010). For instance, the Quranic word فَسَيَكْفِيكَهُمْ is translated to "So will suffice you against them". Additional challenges are specific to the Quran, for example, it has its spelling convention (orthography) which is different from all variants of Arabic (Modern Standard Arabic (MSA), dialect Arabic, and CA). This is illustrated in Example 1, the Quran word ٱلْحَيَوَة is written in Arabic as الْحَيَاةُ. Another challenge is the embedded and different meaning of Quranic words from their meaning in Arabic. For example, the word ٱلْحَيَوَانُ in Arabic means "the animal" while in this verse it means "(is) the life".

Example 1:

وَمَا هَٰذِهِ ٱلْحَيَوٰةُ ٱلدُّنْيَآ إِلَّا لَهْوٌ وَلَعِبٌ وَإِنَّ
ٱلدَّارَ ٱلْءَاخِرَةَ لَهِىَ ٱلْحَيَوَانُ لَوْ كَانُوا يَعْلَمُونَ

Hence, it is clear that although the Quran and Hadith are both CA and cover the same domain of Islamic teachings, they are distinctive in structure, style, and orthography (Bashir et al., 2021). Hence, by training the models with pairs of Quran verses then testing it with another dataset of Quran-Hadith pairs, we ensure that we measure how well the models generalize to identify relatedness in CA texts instead of memorizing the training data. Before explaining our experiments, we briefly give an overview of previous research related to our work.

3. Related Work

In this section we discuss research, mostly in the digital humanities, that aim to utilize advancements in NLP to identify similarities in sacred scriptures.Then we enumerate the existing Arabic benchmarks for text similarity and relatedness tasks.

3.1. Semantic similarity

There have been several attempts to use NLP techniques to find the semantic similarity and relatedness among religious text, ranging from within the same book (Saeed et al., 2020) to different religious scriptures (Verma, 2017; Varghese and Punithavalli, 2019; Peuriekeu et al., 2021; Qahl, 2014). These studies also differ in the scope, where some measure the semantic similarity at the corpus level (Qahl, 2014) or at the verse level (Alshammeri et al., 2020; Alsaleh et al., 2021).

The glaring weakness of the studies which compare different religious books (e.g., the Quran and the Bible) is language, in which the translations are used instead of the original texts. Ideally, such study should use the texts in their original languages (e.g. CA, Hebrew, Greek) to keep their true meaning which could be lost in the translations. This is due to the inherent biases or misinterpretations by the translators. For example, the Quran's meaning is translated into more than 60 English translations, each one rectifying deficiencies in the previous versions (Kidwai, 1987). Hence, multilingual models with zero-shot learning could be the answer to such studies.

One of the studies that focused on the Quran and is more related to our work is (Alshammeri et al., 2020). They trained a doc2vec model (Le and Mikolov, 2014) on the Quran corpus to obtain embeddings for each verse. Then they calculated the cosine similarity between the Quran pairs. To verify their results, they examined whether the pair of verses with high cosine similarity falls within the same concept in a Quran ontology (Abbas, 2009). Another study on the Quran presented by (Alsaleh et al., 2021) utilizes AraBERT (Antoun et al., 2020), a transformer-based model, to identify semantic similarity between Quran verses. Their findings show promising results. Contrarily, our experiments explore if such SOTA models fine-tuned on the Quran pairs dataset can generalize to perform as well when presented with the Quran-Hadith pairs dataset. Hence, our work is not comparable to the previous research, because they used the same dataset for both training and testing. In the following section, we compare the available dataset and discuss the feasibility of incorporating it in our experiments.

3.2. Arabic datasets

Recently many Arabic resources have been published[1] for different NLP downstream tasks (Alyafeai et al., 2021) including question answering (Mozannar et al., 2019), offensive language detection (Mubarak et al., 2020) and more datasets for sentiment analysis, machine translation, and topic classification. However, few are dedicated to semantic similarity task, and only one is in CA. Table 1 shows the existing Arabic datasets for the semantic similarity and relatedness tasks.

,

Dataset	Arabic Variant	num of pairs	Year
Qursim	CA	7,679	2012
Q2Q	MSA	15,712	2019
STS	MSA	1104	2017

Table 1: The available Arabic semantic similarity and relatedness dataset.

Qursim (Sharaf and Atwell, 2012) is a semantic relatedness dataset consisting of 7,679 related Quran-verse pairs. It is extracted from the well-known Quran commentary of Ibn Kathir, an early Islamic scholar who died in 1373. His methodology is clearly stated in the introduction of his book, where each verse is discussed and explained by referring to other verses in the Quran that contain more details or explain other aspects of the same topic. This is the only available

[1] https://arbml.github.io/masader/

CA semantic similarity dataset we are aware of.

The second dataset is dedicated to a Semantic Question Similarity shared task that was conducted during the workshop NLP Solutions for Under Resourced Languages (NSURL) (Seelawi et al., 2019). The dataset consists of three fields, question1, question2, and label. The label is 1 if both questions have a similar answer, or 0 otherwise. The third available dataset is used at the SemEval-2017 STS shared task (Cer et al., 2017). It consists of Arabic pairs machine translated from English then checked by human translators. The label of each pair is in a range from 0 to 5 with five being the most similar.

The last two datasets are not comparable to our dataset of related Quran and Hadith texts. This is because finding the semantic relatedness between Quran and Hadith is different from the standard STS task. The aim is detecting similarity in the underlying meaning and message conveyed by these scared texts, which is more complex.

4. Datasets

In this section, we explain the methodology of creating the Quran-Hadith dataset, which is the main contribution of this paper. Then we describe the process of extending the Qursim (Sharaf and Atwell, 2012) to produce the datasets used in the training/fine-tuning and validation phase across the various models. These datasets are available on our repository[2]. We hope it will be useful for the wider NLP community, particularly those working on under-resourced languages since there is an increased interest in applying NLP tools on such texts like religious scripture but challenges are still unresolved (Bounhas, 2019; Bashir et al., 2021).

4.1. Testing Dataset: Quran-Hadith (QH) Pairs

To build this dataset, we follow five steps: 1) Selecting the sources where a reputable Islamic scholar explicitly stated the relatedness. 2) collect the text, then extract the Quran and the Hadith as a pair. 3) ensure the whole Quran verse or Hadith teaching is obtained by cross-referencing the original corpora. 4) create samples of non-related pairs. 5) process the dataset to remove punctuation and diacritics. The following paragraphs discuss these steps.

4.1.1. QH Sources

The traditional approach of building a dataset using crowd-sourcing is not possible for religious texts since it requires domain experts. Hence, we used two reliable sources of reputable Islamic scholars who explicitly mentioned the relatedness between the pairs. First, the Sahih Albukhari book incorporates a collection of Hadith-teachings organized into topics by a well-known scholar named Muhammed Albukhari who died in 870. We take advantage of the book structure to create our dataset. In many cases, section headings consist of a Quran-verse, which the scholar used to imply that it is related to the Hadiths within the section. Table 2 shows an instance of such cases. We used the LK-Hadith-Corpus [3] (Altammami et al., 2020a) since it provides a well-structured version of the canonical Hadith collections. We extracted these section-headings and their associated Hadith Matn (the actual Hadith teaching without the chain of narrators [Isnad]) as a related Quran-Hadith pair.

Chapter	كتاب البيوع
Section	يَا أَيُّها الَذينَ آمَنُوا لاَ تَأْكُلُوا الرِبا أَضْعَافًا مُضَاعَفَةً وَاتَقوا اللَه لَعَلَكمْ تُفْلِحُونَ.
Hadith	حَدَّثَنَا آدَمُ، حَدَّثَنَا ابْنُ أَبِي ذِئْبٍ، حَدَّثَنَا سَعِيدٌ الْمَقْبُرِيُّ، عَنْ أَبِي هُرَيْرَةَ، عَنِ النَّبِيِّ صلى الله عليه وسلم قَالَ: لَيَأْتِيَنَّ عَلَى النَّاسِ زَمَانٌ لاَ يُبَالِي الْمَرْءُ بِمَا أَخَذَ الْمَالَ، أَمِنْ حَلاَلٍ أَمْ مِنْ حَرَامٍ.

Table 2: Example from Albukhari book.

The second source is a website dedicated to Abdul-Aziz ibn Baz (died in 1999) who was a reputable Islamic scholar that answered religious questions on mainstream media which was later collected and archived on a website[4]. Most of these archived Fatwas (a ruling on an Islamic law given by a recognized authority) contains an answer to a specific question supported by a Quran-verse and a Hadith-teaching. However, some Fatwas contain several Quran verses and Hadith teachings to answer complex questions consisting of various topics. Hence, the relatedness is not clear if taken out of context. Therefore, we collected only the Fatwas which contain one Quran-verse and one Hadith-teaching to ensure they are related to a distinct topic. Then we extracted the Quran-verse and the Hadith-teaching as a pair.

4.1.2. QH Cross-referencing and filtration

The collected pairs from Albukhari and Binbaz were further processed to ensure the full text is included, because the scholar may mention part of the Quran-verse or the Hadith-teaching. So, we cross-referenced the extracted text to find its complete instance on Tanzil[5]

[2]https://github.com/ShathaTm/Quran_Hadith_Datasets

[3]https://github.com/ShathaTm/LK-Hadith-Corpus

[4]https://binbaz.org.sa/fatwas

[5]https://tanzil.net/

for the Quran, and the LK Hadith Corpus for Hadith-teaching (Matn). Finally, the duplicates were removed to form our 155 related pairs.

4.1.3. QH Non-related pairs

To create the negative samples of non-related pairs, we use the Binbaz website tagging feature where each Fatwa is tagged with a topic. We randomly selected a Quran-verse from a Fatwa and a Hadith-teaching from another Fatwa that has a different topic tag. This formed our balanced dataset of 310 related and not-related QH pairs, which is used in the testing phase. To train/fine-tune the models we used the Quran-Quran pair dataset which we explain next.

4.2. Training: Quran-Quran (QQ) Pairs

We used part of the Qursim dataset(Sharaf and Atwell, 2012)) mentioned in Section 3.2. This is because the authors of Qursim analysed the pairs manually to ensure their relatedness is clear regardless of Ibn Kathir's comments. They found that not all pairs showed clear relatedness out of contexts. Therefore, each pair is classified into one of three categories: Strong relation(3,079 pairs), weak relation (3,718 pairs), or no-obvious relation (883 pairs). Since this was done by one annotator, we used the pairs with strong relations only. To further process the dataset, we follow (Alsaleh et al., 2021) methodology to remove duplicates.

4.2.1. QQ Non-related pairs

To create the QQ negative samples of non-related pairs we undergo a different approach than (Alsaleh et al., 2021). Instead of randomly extracting two Quran verses and assume they are not related, we used the Quran ontology by (Hakkoum and Raghay, 2016), since it is the most comprehensive Quran ontology (Altammami et al., 2020b), to extract pairs that do not share the same ontological concepts. Fig 1 shows the algorithm of creating the negative samples.

First, two Quran verses are selected from Tanzil dataset (x, y). Then the associated Quran ontology concepts (Cx, Cy) are extracted from the Quran ontology. The extracted concepts are tokenized into words $c_1, c_2, ...c_n \in Cx$ and $c_1, c_2, ...c_n \in Cy$. After that a comparison is conducted to ensure there is no match in verse1 concepts and verse2 concepts$(Cx \cap Cy = \emptyset)$. Otherwise, these pairs are discarded and the system restarts to extract new pairs of Quran verses until it finds a pair with no intersection of concepts. Then the algorithm ensures the pair in both orders is not already in the negative sample list $(x, y) \notin (X, Y)$ and $(y, x) \notin (X, Y)$. Finally it is added to the list of negative samples $(x, y) \in (X, Y)$ and the process is repeated until the number of collected negative samples is 2,548. This formed our balanced dataset of 5,096 related and non-related QQ pairs which we shuffle and divide into 80% training(fine-tuning) and 20% validation.

4.2.2. QQ augmentation

To enlarge our training data of QQ pairs, we take advantage of the fact that the meaning of the Quran is carefully translated into many languages at the verse level. Therefore, we aim to utilize zero-shot transfer learning in our experiments by fine-tuning the multilingual models on the Quran translations of 43 languages provided electronically on Tanzil. Additionally, since there are 17 different English translations on Tanzil, we fine-tune the English-Arabic cross-lingual model GigaBERT (Lan et al., 2020) using the English translations of the training dataset. Furthermore, the Tanzil project provides an electronic version of Arabic commentaries aligned at the verse level. Hence, we extracted the two available commentaries (Aljlyleen and Almuyaser) of each pair to triple our Arabic training/fine-tuning dataset.

4.3. Data Preprocessing

The Quran and Hadith contain diacritics which are important for readability. Furthermore, the Quran contains its other special symbols and signs that indicate how the verse should be read. For example, the sign on the bottom right of Fig 2 is a small ۚ that is found on the top of words indicating that a reciter of the Quran is allowed to pause at that instance. However, since the models were pre-trained on Arabic corpora without these diacritics and symbols, we removed them.

The training datasets used throughout the experiments in the next section is shown in Table 3, while for validation another set of 1,024 Arabic QQ-pairs is used.

Training dataset pairs	# of pairs
Ar Quran	4,072
Ar Quran + Tafseer	12,216
En Quran	20,360
M Quran	256,536

Table 3: The different training/fine-tuning datasets. Ar= Arabic, En= English, M= Multilingual (43 languages).

5. Experimental Setup

In this section, we discuss the different models trained, fine-tuned on the QQ dataset and tested for identifying relatedness in the QH dataset. This is not semantic similarity in a traditional way where being synonymous or directly equivalent is what we measure, but rather identifying relatedness in the underlying religious teaching. The models are given [text1, text2, label] where label is 1 to indicates their relatedness and 0 otherwise.

5.1. Evaluation metric

The models' performance is measured using the common accuracy and F1 scores. In addition, the Matthews correlation coefficient (MCC) is used, which takes the value between -1 to +1. It produces a reliable score in

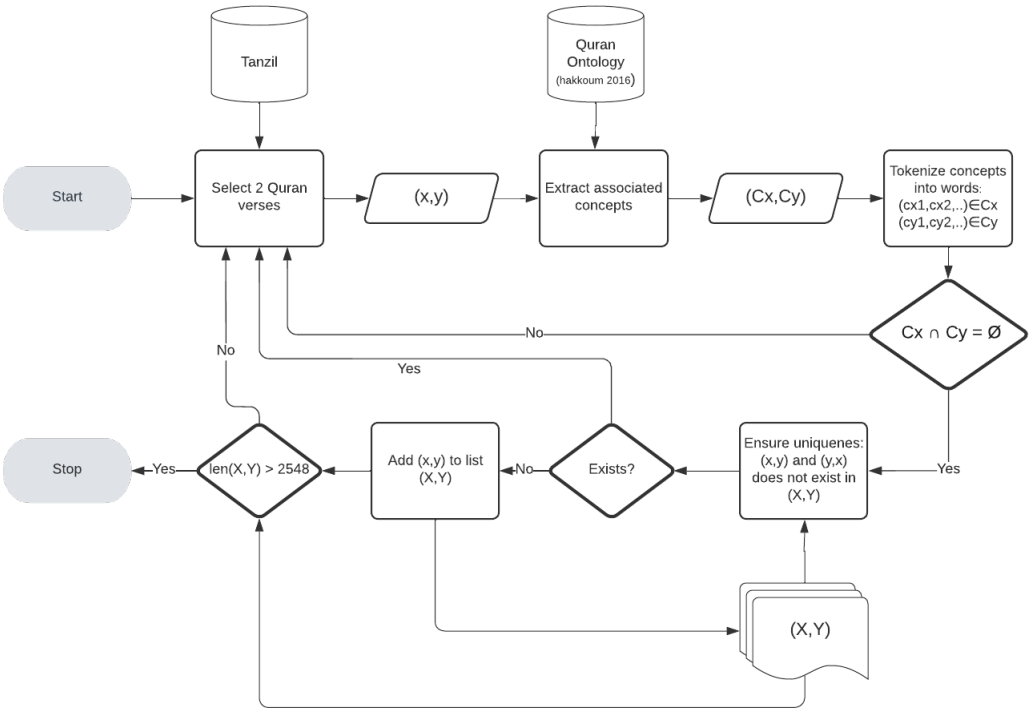

Figure 1: Algorithm of creating negative QQ samples.

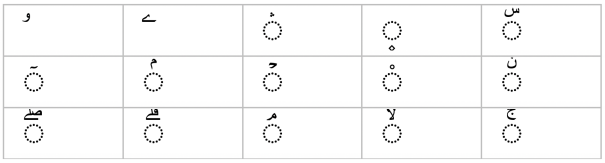

Figure 2: Some of the Quran symbols.

evaluating binary classifications, where a high score is produced only if the model preforms well on the majority of the negative instances and the majority of positive instances (Chicco and Jurman, 2020).

5.2. Baseline Model

For the baseline, we use the character embeddings (Bojanowski et al., 2017) since it is one of the fundamental embeddings that have proven to be a major breakthrough in the field of semantic similarity, and because it is more useful for morphologically-rich languages that contain many rare words. Specifically, we have adapted the methodology proposed by (Nagoudi et al., 2017) which was introduced at the SemEval Task1 and produced one of the top results. Their approach enhances the embedding model by incorporating Inverse Document Frequency (idf) weighting and Part-of-Speech (POS) tagging to give more weight to highly descriptive words in a text.

To compute the idf, we used the Quran as the background corpus, in which the idf for each word is measured using Eq1, where s is the total number of sentences in the corpus (Quran verses), and ws is the number of sentences that contain the word w. While the POS tags were obtained using CAMeL Tools (Obeid et al., 2020).

$$idf(w) = log(s/ws) \quad (1)$$

To calculate the word embedding, we trained a character n-grams model, *FastText*(FT), using Genism[6] library on several corpora including Quran, Hadith, and the associated commentaries. We used vector size 300, window 5, and the minimum count of 3 to ignore rare words.

The limitation of context-free models is that they generate a single word embedding representation for each word in the vocabulary. However, since we trained the model on a specialized corpus, we assume it will capture the semantics of the religious classical Arabic words. Once we get the embedding for each word, the verse embedding is calculated as shown in Eq2, where w_k represents a word in the verse, $POSw_k$ represents the POS tag of the word w, which is used to assign the corresponding weight as proposed by (Nagoudi et al., 2017), and v_k as the word vector. Once the embedding for each verse is calculated, we measure the cosine similarity between the pairs considering those with more than 0.5 as related.

[6]https://radimrehurek.com/gensim/models/fasttext.html

$$V = \sum_{k=1}^{i} (idf(w_k)) \times POS\_weight(POSw_k) \times v_k \quad (2)$$

We considered another baseline model where several machine learning classifiers (SVM, Random Forests, Naive Bayes) were trained with the aforementioned verses embedding as the features. The best performance was obtained using the Random Forests, so we incorporate it in Table 4.

5.3. Transformer-based models

Our experiments consist of fine-tuning pre-trained transformer-based models to classify pairs as related or not. We used the Tesla K80 GPU available on Google Colab. For a given model, we experiment using the same hyper-parameter values of learning-rate 2e-5, patch-size 16, the maximum sequence length of 128 tokens, while using early stopping to avoid over-fitting.

There are several Arabic pre-trained transformer-based models, however, we are interested in those trained on CA or MSA. This is based on the recent study that showed variant proximity of pre-training data to fine-tuning data is the most significant factor (Inoue et al., 2021). One of the widely used models is AraBERT (Antoun et al., 2020) which is trained on 24GB of Arabic text in the news domain. There are several versions of this model, but we used the AraBERTv02 because previous studies showed its superior performance (Inoue et al., 2021; Alsaleh et al., 2021).

A similar model incorporated in the experiments is ArabicBERT (Safaya et al., 2020), which is trained on 95GB of text mainly from the Arabic portion of the OSCAR corpus. A recently released model is CAMeLBERT-CA (Inoue et al., 2021) trained on the 6GB OpenITI Corpus which consists of CA texts (Belinkov et al., 2019).

Since zero-shot learning showed promising results, we experiment with two multilingual modules by fine-tuning them on the English and the multilingual datasets. We used mBERT(Devlin et al., 2018) which is trained on Wikipedia dump, and XLMRoberta (Conneau et al., 2019) trained on filtered Common-Crawl data. In addition to this, we study the performance of a bilingual module named GigaBERT (Lan et al., 2020), designed for English-to-Arabic cross-lingual transfer tasks and trained on news-wire, Wikipedia, and web crawl data.

5.4. Sentence-BERT

Sentence-BERT (SBERT) is a framework introduced to enable obtaining semantically meaningful sentence embeddings that can be compared using cosine similarity (Reimers and Gurevych, 2019). According to the authors, SBERT results in better representations than BERT for tasks involving text comparison. Therefore, we incorporate their multilingual models in our experiments. We used the threshold of 0.5 and above to classify pairs as related.

6. Results

In this section, we show the models' performance on the validation dataset of QQ pairs. Followed by their performance on the testing dataset of QH pairs.

6.1. Validation

Table 4 shows the modules' performance on the 1,019 Arabic QQ pairs. Moreover, we show the training dataset types, the number of training pairs, and fine-tuning time. The table is categorized into three section based on the model's type. The first shows the performance of the baseline model using FastText embeddings and cosine similarity as explained in Section 5.2. The enhanced baselines incorporates a machine learning (ML) model trained on the two Arabic datasets of 4,072 QQ pairs and 8,144 Tafseer pairs.

The second section records the performance of the monolingual models fine-tuned on the Arabic datasets. The third section comprises the results of using zero-shot learning where the cross-lingual GigaBert model is fine-tuned on the English translations of the QQ pairs, and the other multilingual models, mBert and XLMRoberta, are fine-tuned in the first iteration on the English QQ pairs, and the second iteration on the multilingual QQ pairs. Finally, the performance of SBERT models are shown separately in Table 5 since they do not require fine-tuning.

The result demonstrates that the best performing models on the Arabic QQ pairs validation dataset are the monolingual models. Contrarily, multilingual and SBERT showed the worst performance across its different models, which highlights the impact of the pretraining data. Additionally, the baseline model does not perform as well as it does on the SemEval task. This could be attributed to the dataset of QQ pairs since they are considered complex with embedded meanings and polysemy that is hardly captured in non-contextual representations.

In Table 6, we show examples of CAMeLBERT-CA classification results. It seems that the model produced promising results even with some of the incorrectly classified pairs. For example, the first pair although not related as a sentence, it contains words that have semantic similarities like the word حميم which mean 'warm' or 'closeness [in relationships]'. We also show the transliteration of this example to indicate the possible shallow similarity that was identified by the model. The second example was classified as not-related which is appropriate because this relatedness requires knowing what the pronouns are referring to. Bear in mind that the Arabic text is

Model	Fine-tuning	# of pairs	Time	MCC	Acc.	F1
Baseline FastText (Nagoudi et al., 2017)	-	-	-	-2.8	49.2	64.1
FastText + Random Forests	Ar	4,072	0:00:65	+57.6	78.4	76.5
FastText + Random Forests	Ar,Taf	12,216	0:24:32	+55.4	77.4	75.9
AraBERT (Antoun et al., 2020)	Ar	4,072	0:01:57	+74.6	86.7	**85.4**
AraBERT	Ar, Taf	12,216	0:04:37	+81.0	90.2	**89.6**
ArabicBERT (Safaya et al., 2020)	Ar	4,072	0:01:33	+88.4	94.1	**94.0**
ArabicBERT	Ar, Taf	12,216	0:04:37	+94.1	97.0	**97.0**
CAMeLBERT-CA(Inoue et al., 2021)	Ar	4,072	0:01:38	+82.2	90.9	**90.5**
CAMeLBERT-CA	Ar, Taf	12,216	0:04:51	+86.6	93.2	**93.1**
GigaBERT (Lan et al., 2020)	En	20,360	0:07:51	+16.0	55.9	67.0
XLM-Roberta (Conneau et al., 2019)	En	20,360	0:08:44	+20.8	57.8	68.2
XLM-Roberta	M, Ar, En	260,608	1:52:40	+29.0	64.3	61.6
mBERT (Devlin et al., 2018)	En	20,360	0:08:10	-24.7	44.2	61.3
mBERT	M, Ar, En	260,608	1:52:40	+39.0	69.0	65.2

Table 4: Models performance on the QQ validation dataset shown in %. Overall best models based on F1 score are highlighted in bold. Taf = Arabic Tafseer (commentaries).

Model	MCC	Acc.	F1
paraphrase-xlm-r-m-v1	+19.6	59.8	61.4
distiluse-base-m-cased-v2	+30.5	60.4	37.8
distiluse-base-m-cased-v1	+30.3	56.8	35.3
stsb-xlm-r-m	+16.6	58.4	60.0

Table 5: Performance of SBERT multilingual models on the QQ dataset.

different from the translations which usually explicitly state the meaning embedded in the original Arabic text.

Although monolingual modules produced promising results, can its performance be generalized to other CA datasets in the same domain? To answer this question, we evaluate these monolingual models on the testing dataset of QH pairs.

6.2. Testing

We investigate how well the best-performing models generalize to the dataset of QH pairs. The results are shown in Table 8. It is clear that CAMeLBERT-CA's performance is superior, yet there is a significant drop across all the models' F1 score compared to their performance on the validation dataset of QQ pairs in Table 4.

7. Analysis and Discussion

In this section, we analyse the results of the ∼20 points drop in F1 score across the models on the QH dataset. So we extract several examples classified by CAMeLBERT-CA in Table 7 to illustrate where it did well and discuss the causes of misclassification. Example 1 shows a Hadith and Quran that consist of different words except for one keyword that occurs in different morphological forms (الإيلاء - يؤلون), but the model was able to identify the relatedness. However, many of the correctly classified pairs consist of a clear message as in Example 2. While 70% of the incorrectly classified pairs have the label '1' (related), but the model fails to identify the relatedness and predicts '0'(not-related) as the label. This could be attributed to several reasons. First, the two texts consist of different words but have the same underlying message as shown in Example 3. Second, several words in these texts are vague; hence, referring to exegesis is essential to understand it as shown in Example 4. Third, many Hadiths are a narration of a detailed incident that has a moral behind it, while the Quran states the explicit guidance as shown in Example 5.

The analysis leads to the following conclusion. For machines to capture the underlying meanings in the Quran and the Hadith, it might be more possible through a knowledge-based approach. However, many ontologies have been developed for the Quran, but none of which cover the larger scope of Hadith (Altammami et al., 2020b). Creating such resources is expensive, and utilizing deep learning models could be the alternative. Yet, the idea of the current models reaching a human-level understanding of such texts remains somewhat elusive. This is aligned with a recent study by (Chandrasekaran and Mago, 2020) who tested the transformer-based models on a more complex dataset of 50 English sentence pairs. Their performance decreases by more than 10 points compared to their performance on the common benchmarks (e.g. SICK). This is amplified for under-resourced languages as shown by our results. Hence, measuring the sensitivity of the models to the under-resourced text and their increase in complexity should be considered.

8. Conclusion and future work

This work highlights that state-of-the-art models fall short when presented with complex texts from a low-resource language. We demonstrate this by testing several models on a binary classification task of identifying relatedness between the Quran and the Hadith, which are classical-Arabic religious texts with embedded meanings. To fine-tune the models, we extend the

ID	Label	Prediction	Quran Verse1	Quran Verse2
Ex1	0	1	نار حامية	ولا يسأل حميم حميما
			Transliteration: [Naarun hamiyah] It is a Fire, intensely hot. [101:11]	Transliteration: [Wa laa yas'alu hameemun hameemaa] And no friend will ask [anything of] a friend. [70:10]
Ex2	1	0	وإذ قلنا للملائكة اسجدوا لآدم فسجدوا إلا إبليس أبى واستكبر وكان من الكافرين	قال فاهبط منها فما يكون لك أن تتكبر فيها فاخرج إنك من الصاغرين
			And [mention] when We said to the angels, "Prostrate before Adam"; so they prostrated, except for Iblees. He refused and was arrogant and became of the disbelievers. [2:34]	[Allah] said, "Descend from Paradise, for it is not for you to be arrogant therein. So get out; indeed, you are of the debased. [7:13]
Ex3	1	1	كل امرئ منهم يومئذ شأن يغنيه	ولا يسأل حميم حميما
			For every man, that Day, will be a matter adequate for him. [80:37]	And no friend will ask [anything of] a friend. [70:10]

Table 6: Examples of QQ pairs classified by CAMeLBERT-CA.

ID	Label	Prediction	Quran	Hadith
Ex1	1	1	للذين يؤلون من نسائهم تربص أربعة أشهر فإن فاءوا فإن الله غفور رحيم	كان يقول في الإيلاء الذي سمى الله لا يحل لأحد بعد الأجل إلا أن يمسك بالمعروف أو يعزم بالطلاق كما أمر الله عز وجل
			For those who swear not to have sexual relations with their wives is a waiting time of four months, but if they return [to normal relations] - then indeed, Allah is Forgiving and Merciful. [2:226]	If the period of Ila expires, then the husband has either to retain his wife in a handsome manner or to divorce her as Allah has ordered.
Ex2	1	1	وإني لغفار لمن تاب وآمن وعمل صالحا ثم اهتدى	قال رسول الله صلى الله عليه وسلم التائب من الذنب كمن لا ذنب له
			But indeed, I am the Perpetual Forgiver of whoever repents and believes and does righteousness and then continues in guidance. [20:82]	He who repents of a sin is like him who has committed no sin.
Ex3	1	0	يا أيها الذين آمنوا لا تأكلوا الربا أضعافا مضاعفة واتقوا الله لعلكم تفلحون	النبي صلى الله عليه وسلم قال ليأتين على الناس زمان لا يبالي المرء بما أخذ المال أمن حلال أم من حرام
			O you who have believed, do not consume usury, doubled and multiplied, but fear Allah that you may be successful. [3:130]	Certainly a time will come when people will not bother to know from where they earned the money, by lawful means or unlawful means.
Ex4	1	0	يا أيها الذين آمنوا إنما الخمر والميسر والأنصاب والأزلام رجس من عمل الشيطان فاجتنبوه لعلكم تفلحون.	أن رسول الله صلى الله عليه وسلم قال لا سبق إلا في نصل أو خف أو حافر
			O you who have believed, indeed, intoxicants, gambling, [sacrificing on] stone alters [to other than Allah], and divining arrows are but defilement from the work of Satan, so avoid it that you may be successful. [5:90]	Wagers are allowed only for shooting arrows, or racing camels or horses.
Ex5	1	0	وما أرسلنا من قبلك إلا رجالا نوحي إليهم فاسألوا أهل الذكر إن كنتم لا تعلمون	خرجنا في سفر فأصاب رجلا منا حجر فشجه في رأسه ثم احتلم فسأل أصحابه فقال هل تجدون لي رخصة في التيمم فقالوا ما نجد لك رخصة وأنت تقدر على الماء فاغتسل فمات فلما قدمنا على النبي صلى الله عليه وسلم أخبر بذلك فقال قتلوه قتلهم الله ألا سألوا إذ لم يعلموا فإنما شفاء العي السؤال إنما كان يكفيه أن يتيمم ويعصب على رأسه خرقة ثم يمسح عليها ويغسل سائر جسده
			And We sent not before you except men to whom We revealed [Our message]. So ask the people of the message if you do not know.[16:43]	We set out on a journey. One of our people was hurt by a stone, that injured his head. He then had a sexual dream. He asked his fellow travelers: Do you find a concession for me to perform tayammum? They said: We do not find any concession for you while you can use water. He took a bath and died. When we came to the Prophet (), the incident was reported to him. He said: They killed him, may Allah kill them! Could they not ask when they did not know? The cure for ignorance is inquiry. It was enough for him to perform tayammum and to pour some drops of water or bind a bandage over the wound (the narrator Musa was doubtful); then he should have wiped over it and washed the rest of his body.

Table 7: Examples of QH pairs classified by CAMeLBERT-CA.

Model	Fine-tuning	MCC	Acc.	F1
ArabicBERT	Ar	+3.6	51.7	50.0
ArabicBERT	Ar+Taf	+32.3	66.1	65.1
CAMeLBERT-CA	Ar	+31.9	65.7	67.9
CAMeLBERT-CA	Ar+Taf	**+55.3**	**77.2**	**74.8**
AraBERT	Ar	+12.9	56.3	59.1
AraBERT	Ar+Taf	+41.4	70.6	69.6

Table 8: Models performance on the QH testing dataset.

Quran-verse pairs dataset (Sharaf and Atwell, 2012). Then we test the models on our carefully created 310 Quran-Hadith pairs dataset, which could be extended by incorporating different sources of Fatwas. Such datasets are vital to gauge and improve models' performance on low-resource languages. Hence, we aim to develop more classical Arabic datasets of different genres like poetry. This is because poetry incorporates rhetorical features that can be used to challenge SOTA models to identify satirical, romantic, or panegyric texts, which is a complex task that requires deep understanding of the language.

9. Acknowledgements

This research has been supported by a scholarship from King Saud University. We thank the anonymous reviewers for their constructive comments.

10. Bibliographical References

Abbas, N. H. (2009). Quran 'search for a concept' tool and website. Master's thesis, University of Leeds.

Abdul-Raof, H. (2013). *Qur'an translation: Discourse, texture and exegesis*. Routledge.

Alsaleh, A. N., Atwell, E., and Altahhan, A. (2021). Quranic verses semantic relatedness using arabert. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 185–190. Leeds.

Alshammeri, M., Atwell, E., and Alsalka, M. A. (2020). Quranic topic modelling using paragraph vectors. In *Proceedings of SAI Intelligent Systems Conference*, pages 218–230. Springer.

Altammami, S., Atwell, E., and Alsalka, A. (2020a). Constructing a bilingual hadith corpus using a segmentation tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3390–3398.

Altammami, S., Atwell, E., and Alsalka, A. (2020b). Towards a joint ontology of quran and hadith. *International Journal on Islamic Applications in Computer Science And Technology*.

Alyafeai, Z., Masoud, M., Ghaleb, M., and Alshaibani, M. S. (2021). Masader: Metadata sourcing for arabic text and speech data resources.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Bashir, M. H., M Azmi, A., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., and Qadir, J. (2021). Arabic natural language processing for qur'anic research: A systematic review.

Belinkov, Y., Magidow, A., Barrón-Cedeño, A., Shmidman, A., and Romanov, M. (2019). Studying the history of the arabic language: language technology and a large-scale historical corpus. *Language Resources and Evaluation*, 53(4):771–805.

Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bounhas, I. (2019). On the usage of a classical arabic corpus as a language resource: related research and key challenges. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–45.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Chandrasekaran, D. and Mago, V. (2020). Domain specific complex sentence (dscs) semantic similarity dataset. *arXiv preprint arXiv:2010.12637*.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elangovan, A., He, J., and Verspoor, K. (2021). Memorization vs. generalization: quantifying data leakage in nlp performance evaluation. *arXiv preprint arXiv:2102.01818*.

Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Hakkoum, A. and Raghay, S. (2016). Ontological approach for semantic modeling and querying. *International Journal on Islamic Applications in Computer Science and Technology*, pages 37–37.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Kidwai, A.-R. (1987). Translating the untranslatable: a survey of english translations of the quran. *The Muslim World Book Review*, 7(4):66–71.

Lan, W., Chen, Y., Xu, W., and Ritter, A. (2020). An empirical study of pre-trained transformers for arabic information extraction. *arXiv preprint arXiv:2004.14519*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Lewis, P., Stenetorp, P., and Riedel, S. (2020). Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pages 216–223. Reykjavik.

Mozannar, H., Maamary, E., El Hajal, K., and Hajj, H. (2019). Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy, aug. Association for Computational Linguistics.

Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., and Al-Khalifa, H. (2020). Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France, may. European Language Resource Association.

Nagoudi, E. M., Ferrero, J., and Schwab, D. (2017). Lim-lig at semeval-2017 task1: Enhancing the semantic similarity for arabic sentences with vectors weighting. In *Proceedings of the 11th International*

Workshop on Semantic Evaluation (SemEval-2017), pages 134–138.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May. European Language Resources Association.

Peuriekeu, Y. M., Noyum, V. D., Feudjio, C., Goktug, A., and Fokoue, E. (2021). A text mining discovery of similarities and dissimilarities among sacred scriptures. *arXiv preprint arXiv:2102.04421*.

Qahl, S. H. M. (2014). *An automatic similarity detection engine between sacred texts using text mining and similarity measures*. Rochester Institute of Technology.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saeed, S., Haider, S., and Rajput, Q. (2020). On finding similar verses from the holy quran using word embeddings. In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pages 1–6. IEEE.

Safaya, A., Abdullatif, M., and Yuret, D. (2020). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

Seelawi, H., Mustafa, A., Al-Bataineh, H., Farhan, W., and Al-Natsheh, H. T. (2019). Nsurl-2019 task 8: Semantic question similarity in arabic. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 1–8.

Sharaf, A.-B. and Atwell, E. (2012). Qursim: A corpus for evaluation of relatedness in short texts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2295–2302.

Varghese, N. and Punithavalli, M. (2019). Lexical and semantic analysis of sacred texts using machine learning and natural language processing. *International Journal of Scientific & Technology Research*, 8(12):3133–3140.

Verma, M. (2017). Lexical analysis of religious texts using text mining and machine learning tools. *International Journal of Computer Applications*, 168(8):39–45.