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]

1- Motivation

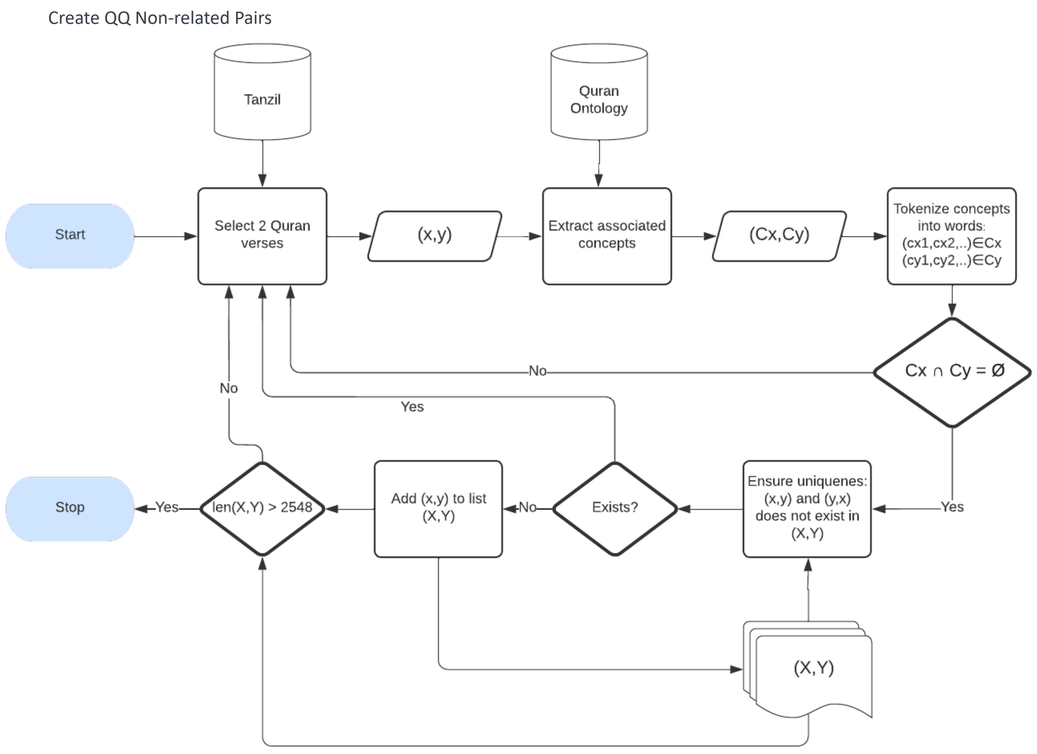
- Transformer-based models showed near-perfect results on several downstream tasks. However, their performance on Classical Arabic (CA) texts is largely unexplored.
- CA is considered low-resource in terms of available datasets and inherently challenging for natural language processing tasks.

2- Contribution

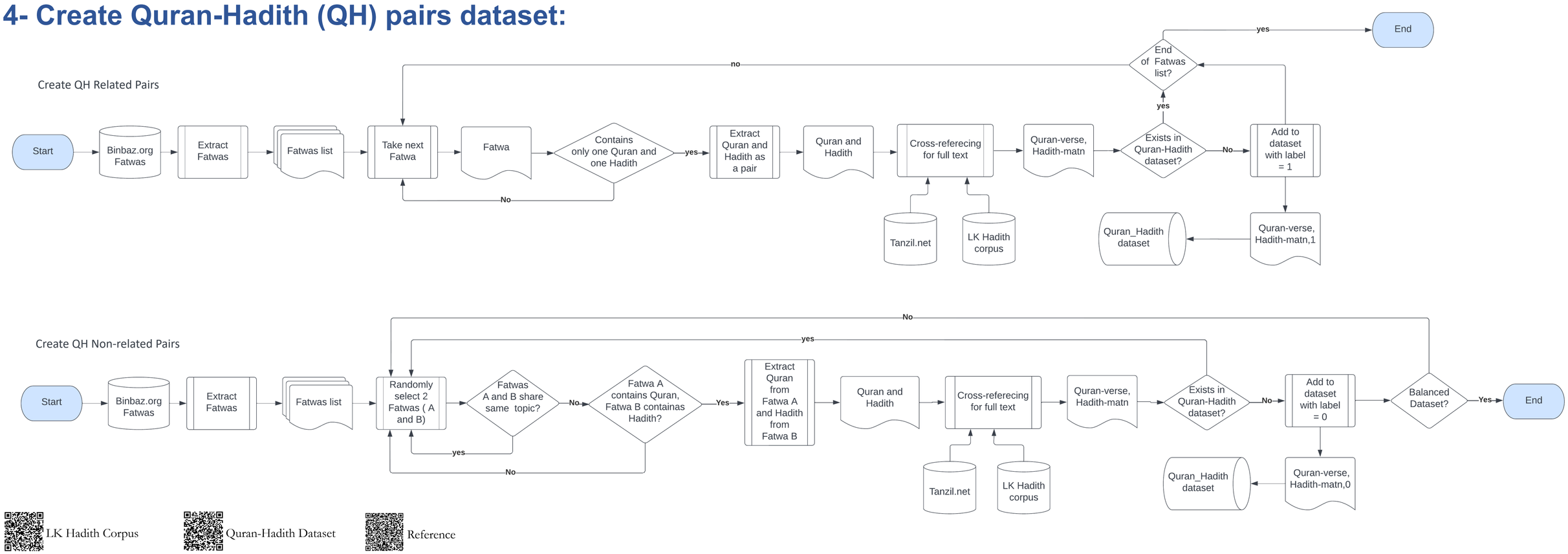
- We proposed a framework to built a CA dataset of related and non-related Quran-verse (Muslim holy book) and Hadith-teaching (Prophet Muhammed teachings) pairs by consulting sources of reputable religious experts.
- We evaluate monolingual, bilingual, and multilingual state-of-the-art Transformer-based models on our new CA dataset to detect relatedness between the Quran and the Hadith.
- The models' performance calls for the imminent need to explore avenues for improving the quality of these models to capture the semantics in such complex texts.

3- Create Quran-Quran (QQ) pairs dataset:

- We used the pairs of verses annotated with a strong relation in the Qursim dataset [1].
- To create the QQ negative samples of non-related pairs, we used the Quran ontology[2] to extract random pairs that do not share the same ontological concepts.
- We used Quran translations and commentaries for data augmentation.



4- Create Quran-Hadith (QH) pairs dataset:



LK Hadith Corpus Quran-Hadith Dataset Reference

5- Experiment

- Baseline:
  Verse vector: $V = w \sum_{k=1}^{i} \big(idf(w_k)\big) \times POS\_weight(POS\ w_k) \times v_k$
  Cosine similarity > 0.5
- Word-embeddings + Machine Learning.(Random Forests)
- Transformer-based Models

6- Results: Validation on QQ dataset of 1,024 pairs

Model	Fine-tuning data	# of pairs	F1
Baseline	-	-	64.1
FastText + Random Forests	Ar	4,072	76.5
FastText + Random Forests	Ar + Tafeer	12,216	75.9
AraBERT	Ar	4,072	85.4
AraBERT	Ar + Tafeer	12,216	89.6
ArabicBERT	Ar	4,072	94.0
ArabicBERT	Ar + Tafeer	12,216	97.0
CAMeLBERT-CA	Ar	4,072	90.5
CAMeLBERT-CA	Ar + Tafeer	12,216	93.1
GigaBert	En	20,360	67.0
XLM-Roberta	En	20,360	68.2
XLM-Roberta	M, Ar, En	260,608	61.6
mBERT	En	20,360	61.3
mBERT	M, Ar, En	260,608	65.2

7- Results: Testing best models on QH dataset of 310 pairs

Model	Fine-tuning data	F1
AraBERT	Ar	56.9
AraBERT	Ar + Tafeer	69.6
ArabicBERT	Ar	50.0
ArabicBERT	Ar + Tafeer	65.1
CAMeLBERT-CA	Ar	67.9
CAMeLBERT-CA	Ar + Tafeer	74.8

Examples of CAMeLBERT-CA Results

Label	Predic.	Hadith	Quran
1	1	كان يقول في الإيلاء الذي سمى الله لا يحل لأحد بعد الأجل إلا أن يمسك بالمعروف أو يعزم بالطلاق كما أمر الله عز وجل	للذين يؤلون من نسائهم تربص أربعة أشهر فإن فاءوا فإن الله غفور رحيم
1	1	قال رسول الله صلى الله عليه وسلم التائب من الذنب كمن لا ذنب له	وإني لغفار لمن تاب وآمن وعمل صالحا ثم اهتدى
1	0	النبي صلى الله عليه وسلم قال ليأتين على الناس زمان لا يبالي المرء بما أخذ المال أمن حلال أم من حرام	يا أيها الذين آمنوا لا تأكلوا الربا أضعافا مضاعفة واتقوا الله لعلكم تفلحون
1	0	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	وما أرسلنا من قبلك إلا رجالا نوحي إليهم فاسألوا أهل الذكر إن كنتم لا تعلمون