Linking Quran and Hadith Topics in an Ontology using Word Embeddings and Cellfie Plugin

Ibtisam Khalaf Alshammari
University of Hafr Al Batin,
Hafr Al Batin, Saudi Arabia
University of Leeds, Leeds, UK
ml18ikfa@leeds.ac.uk

Eric Atwell
University of Leeds,
Leeds, UK
e.s.atwell@leeds.ac.uk

Mohammad Ammar Alsalka
University of Leeds,
Leeds, UK
m.a.alsalka@leeds.ac.uk

Abstract

Qur'an and Hadith are the sacred texts of the Islamic religion. Arabic Qur'an and Hadith texts have been analyzed and annotated by researchers using a variety of domains, representations, and formats to improve the accessibility of Islamic knowledge. However, the many and diverse Islamic resources raise a potential challenge in linking and integrating them. The main objective of this work is to link Qur'an and Hadith topics and integrate them with related knowledge from different Islamic resources. The proposed methodology is to use a combination of word embeddings-based BERT with the Cellfie tool to achieve more accurate and meaningful data integration. The results of using the CL-AraBERT word embedding model display efficiency performance in F1 score and accuracy metrics with 91% and 84% respectively. At the same time, the constructed ontology, RQHT, links the Qur'an and Hadith topics with their related knowledge properly and consistently.

1 Introduction

Islamic research has contributed to the generation of many different databases from various sources, including the Qur'an and Hadith, which may present opportunities and challenges for religious researchers. The Qur'an and Hadith are the sacred texts of Islamic law and are considered significant historical documents that illustrate the origins of the Islamic faith. Integrating the diverse Islamic datasets can be essential for deriving comprehensive insights and enabling data-driven decision-making, thus easily leveraging them in future studies. However, the heterogeneous nature of Islamic knowledge, involving data structure, domains, and formats, could impede effective data combining.

Natural language processing (NLP) advancements have provided promising solutions to address such challenges in Arabic religious texts. Recently, word embeddings using Bidirectional Encoder Representations from Transformers (BERT) have become increasingly popular in achieving state-of-the-art performances on several NLP downstream tasks (Devlin et al., 2019). It can efficiently understand and process human language by generating contextual embeddings and linking related Islamic data.

In addition to advances in NLP, the adoption of ontology development has increased lately in the Islamic domain. Ontology can structure and represent knowledge explicitly in a machine-readable format that may be integrated into computer-based applications and systems (Gruber, 1995). The consistent interpretation and integration of data across different Islamic datasets can be facilitated by leveraging the Cellfie Plugin tool. Therefore, the synergy between word embeddings and Cellfie Plugin techniques holds great potential for Qur'an and Hadith topics linking and integration.

This paper is organized as follows: Section 2 reviews the previous studies and related work of word embeddings and ontology design. The detailed methodology to link Qur'an and Hadith topics is described in Section 3. Then, Section 4 provides an overview of the evaluation process. Results and discussion of the experiment are detailed in Section 5, which is followed by future work and conclusion in Section 6.

2 Related Work

2.1 Word Embeddings for Detecting Arabic Qur'an and Hadith Semantic Similarities

Several research studies have employed various word embedding techniques to enhance the findings of semantic similarity in the Arabic language, especially the classical texts of the Qur'an and Hadith. Abdelghany et al. (2020) presented a study using the unsupervised learning algorithm Doc2vec to identify Hadith similarity either in Matn or

Sanad across various Hadith books. Similarly, Alshammeri et al. (2021) provides research on using Doc2vec embeddings to capture the Qur'anic verses' semantics and classify similar documents into 15 predefined classes. The proposed model achieved 68% and 56% in classification accuracy and F1 score, respectively.

Alsaleh et al. (2021) implemented AraBERTv0.2 and AraBERTv2 language models to determine the semantic relatedness of Qur'anic verses pairs from the QurSim dataset. The AraBERTv0.2 language model obtained the best result with a 92% accuracy score.

Another study by (Alshammeri et al., 2022) combined a pre-trained AraBERT language model and Siamese transformer-based architecture to detect semantic similarity in the Qur'an text. The suggested approach was performed efficiently by achieving a 95% F1 score on the Qur'anic dataset.

2.2 Ontology Development

Many studies have focused on building Arabic ontologies covering different aspects of the Qur'an and Hadith knowledge. The popular Qur'an ontology was created by Hakkoum and Raghay (2015) is QuranOntology. Hybrid methods, such as Protégé-OWL and Jena TDB with Fuseki server, were used to develop the Qur'an ontology. A semantic-based search engine was established to support SPARQL queries in this ontology. A framework was proposed by (Alshammari et al., 2022) for linking and mapping Qur'anic ontologies. It combined RDF (resources description framework) Mapping Language (RML), Cellfie Plugin, and SDM-RDFizer to integrate the morphological annotations and syntactic analyses into Qur'an chapters, verses, and words in Quranicontology. The integrated onology was evaluated by a SPARQL server.

Al-Sanasleh and Hammo (2017) presented a prophetic ontology based on data derived from the Qur'an, Hadith, and explanatory book resources. They used the METHONTOLOGY approach, proposed by Fernández-López et al. (1997), to build their ontology from scratch. Likewise, Fairouz et al. (2020) developed an ontology based on Arabic Hadith texts using the METHONTOLOGY methodology. Their ontology's scope focused on the essential concepts mentioned in the EL-Bukhari book, especially the Knowledge related to Wudhu2, pillars of Islam, and worship.

Alsalhee and Abdullah (2022) conducted a study to construct an ontology that represents the Qur'anic stories. The object role modeling (ORM) and MappingMaster domain-specific language techniques were applied to build conceptual structure and convert Excel sheets to an OWL format. This ontology was evaluated using the SPARQL query language.

A study was conducted by (Altammami et al., 2021) to investigate the appropriateness of utilizing a Qur'an ontology as a foundation for linking Qur'an and Hadith ontology. A corpus-based evaluation approach was employed to evaluate the existing Qur'anic ontologies. The result of this study is that the QuranOntology by Hakkoum and Raghay (2015) can be used as a starting point for an Islamic ontology.

3 Methodology

3.1 Data Collection

This section presents an overview of gathering and extracting Qur'an and Hadith data from four datasets: LK-Hadith-Corpus[1], HT_Topics, QH_Dataset[2], and QuranOntology (Hakkoum and Raghay, 2015; Hakkoum and Raghay, 2016). These datasets include different data types, such as well-structured formats, semi-structured, and data lacking organisation known as unstructured.

3.1.1 LK-Hadith-Corpus

The Leeds University and King Saud University (LK) Hadith corpus is a well-structured bilingual Arabic-English Islamic Hadith. It was generated automatically using a Hadith segmentation tool to link each Hadith into its two main components, Isnad and Matn. This corpus contains 39,038 annotated Ahadith elicited from the six canonical Hadith books, including names and numbers of each book, chapter, and section. The LK-Hadith corpus also has over 10 million tokens. The total number of Hadith chapters is 238, excluding duplication (Altammami et al., 2020).

3.1.2 HT_Topics Dataset

The Hadith_Teaching_Topics dataset comprises more than 33 thousand Arabic-Matn documents and their related topics. The Arabic-Matn documents were extracted from LK-Hadith-Corpus (Altammami et al., 2020). The incomplete and empty

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

[2] https://github.com/ShathaTm/Quran_Hadith_Datasets/blob/main/QH_Dataset.csv

Hadith-teaching records were eliminated, resulting in a final dataset containing 33,169 Arabic-Matn documents. The Arabic-Matn documents have been analysed using a comparison of the BERTopic technique with different Arabic transformer-based language models and topic representations to detect hidden topics from each document. ArabicBERT model achieved the best results and generated 220 topics. Consequently, each Hadith-teaching was semantically classified into a specific related topic.

3.1.3 QH_Dataset

Qur'an_Hadith_Dataset is a collection of 310 semantic relatedness pairs of Arabic Qur'an-verse and Hadith-teaching, classified as related (1) and non_related (0) pairs. The source used to build this dataset is a Fatwas website for AbdulAziz ibn Baz (a reputable Islamic scholar). The authors collected Fataws, including a Qur'an-verse and Hadith-teaching, to extract related or non-related Qur'an-verse and Hadith-teaching pairs. The limitation of this dataset is its absence of crucial data, such as the chapter names and numbers for both Qur'an-verse and Hadith-teaching. (Altammami and Atwell, 2022).

3.1.4 QuranOntology

A structured Qur'anic ontology was developed to represent and link Qur'anic knowledge from various datasets in both Arabic and English languages (Hakkoum and Raghay, 2015). It includes Qur'anic metadata and text from the Tanzil website, Qur'an descriptions from Tafsir AlJalalayn and Al-Muyasser books, Qur'anic concepts and topics discussed in Tafsir Ibn Kathir. It integrated the Semantic Qur'anic from QurSim (Sharaf and Atwell, 2012b) and the most significant Qur'anic annotations in the Qur'ana dataset (Sharaf and Atwell, 2012a).

3.2 Data Extraction and Pre-Processing

Data extraction is considered an essential step in word embeddings because of its importance in retrieving relevant information from various sources and other data repositories. To integrate relevant text data for Qur'an-verse and Hadith-teaching in QH_Dataset, we applied the following steps:

- The initial step is to extract Qur'an topics from QuranOntology corresponding to each Qur'an verse and related information such as verse numbers and chapters.
- Then, Hadith topics are meticulously extracted for each corresponding Hadith teaching from the HT_Topics dataset, and the associated details from LK-Hadith-Courpus.
- Finally, the Qur'an-Hadith-Topics (QHT) dataset is created and comprehensively combined 310 Arabic Qur'an-verse and Hadith-teaching pairs with their chapter numbers and names, Qur'an-Hadith topics, and labelled as 155 related pairs (1), and 155 non-related (0).

At the same time, preparing data is crucial for many NLP applications to enhance data quality. This process includes cleaning the raw data by identifying, correcting, and eliminating unnecessary data, such as missing and irrelevant elements. After that, the data is segmented into tokens to ensure more efficient and accurate analysis for the intended use. The topics of Qur'an-verse and Hadith-teaching pairs are then divided into 80% training and 20% testing sets.

3.3 Word Embeddings

Word embedding is an NLP technique. It represents words as dense vectors to capture semantic and syntactic relationships between texts. We employed BERT (Bidirectional Encoder Representations from Transformers) as a state-of-the-art word embedding model. BERT leverages pre-trained transformer-based language models by capturing contextual word embeddings to enhance language understanding in various NLP tasks, specifically semantic similarity.

This study uses BERT with multiple pre-trained transformer-based models to generate numerical vector representations of Qur'an-Hadith topics. The Arabic pre-trained models include: AraBERTv2 (Antoun et al., 2020), ArabicBERT (Safaya et al., 2020), CAMeL-BERT (Inoue et al., 2021), CL-AraBERT (Malhas and Elsayed, 2022), and ARBERT and MARBERT (Abdul-Mageed et al., 2021). We configured the batch size to 8, the number of epochs to 5, and the learning rate to 1e-4 for all the proposed BERT models, as a result, these settings produced the best performance in terms of accuracy and F1 score metrics.

3.4 Semantic Similarity

The semantic similarity concept has a vital significance in NLP and computational linguistics. It can support machines' understanding and process-

Models	F1 Score	Accuracy
AraBERTv2 (Antoun et al., 2020)	0.8571	0.7556
ArabicBERT (Safaya et al., 2020)	0.8981	0.8220
CAMeL-BERT (Inoue et al., 2021)	**0.9024**	0.8222
CL-AraBERT (Malhas and Elsayed, 2022)	**0.9136**	**0.8444**
ARBERT (Abdul-Mageed et al., 2021)	0.8815	0.8133
MARBERT (Abdul-Mageed et al., 2021)	**0.9085**	0.8371

Table 1: The results of the Arabic pre-trained language models.

ing of human language by computing the semantic relatedness between pairs of word embedding (Rahutomo et al., 2012). Cosine similarity, one of the standard semantic relatedness metrics, is calculated to quantify similarity and ranges from -1 (entirely dissimilar) to 1 (similar).

Consequently, a cosine similarity matrix is generated for the pairs of Qur'an-Hadith-Topics numerical representations to capture the nuanced semantic relatedness between them. Then, the results of the cosine similarity are compared with the QH_Dataset labels to assess the performance of the best model semantic similarity.

3.5 Ontology Creation with Cellfie Plugin

The Cellfie Plugin[3] is a powerful Protégé tool developed to automatically enhance knowledge management and ontology engineering. It has the ability to convert structured data in spreadsheets (in formats like CSV or Excel) into web ontology language (OWL) format. It maps the spreadsheet data into ontology entities such as classes, individuals, and properties by using a rule-based mapping approach. Therefore, the outcomes of the related topics from the semantic similarity process can be linked with the Qur'an and Hadith data through an ontology, namely Related Qur'an and Hadith Topics (RQHT).

To build the RQHT ontology, we initially built the conceptual model for the QHT dataset and defined the ontology classes, data and object properties. The Cellfie plugin was then used to import the QHT spreadsheet and set mapping rules to ensure that each column is mapped to the appropriate classes or properties, and each row is accurately mapped to the corresponding individuals in the RQHT ontology. Figure 1 shows an example of the topics related to a Qur'an-verse and Hadith-teaching pair and integrated with Qur'an and Hadith knowledge.

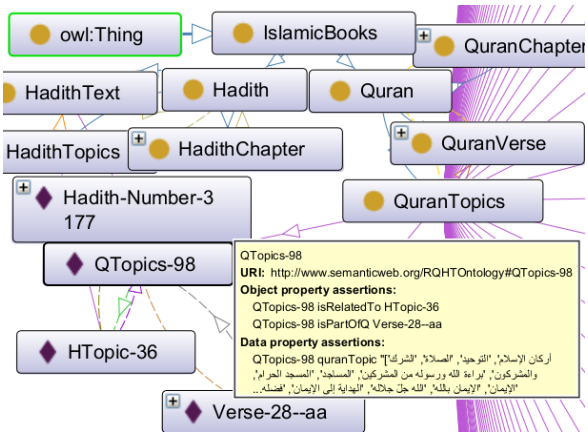

Figure 1: An example of RQHT ontology visualisation.

4 Evaluation

This section describes several methods to evaluate the semantic similarity of the word embeddings' performance on the test set, including F1 score and accuracy metrics, classifying pairs as similar or dissimilar with a threshold of 0.5. The threshold can classify topic pairs as similar if the similarity score is greater than 0.5 and dissimilar if the score is lower. These metrics comprehensively understand how effectively our method enhances and classifies relevant connections.

The F1 score provides a harmonic of precision (the accuracy of positive predictions) and recall (the true positive predictions), ranging from 0 for the worst performance to 1 for the best. Its formula is defined as follows:

$$F1\ score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

While the accuracy measures the percentage of correctly classified instances (true positives and true negatives) out of the total number of instances, it can be calculated using the following formula:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

[3] https://github.com/protegeproject/cellfie-plugin

Quranic Topics	Hadith Topics	Label	Prediction
['الإيمان', 'المؤمنون', ' صفات المؤمنين', ' ما أعده الله لهم', ' وعده إياهم', ' الإنسان والعلاقات الأخلاقية', ' الأخلاق الحميدة', ' العفة', ' غض البصر وحفظ الفرج']	[الباءة ', 'فليتزوج', 'للفرج', 'للبصر', 'وأحصن', 'أغض', 'الشباب', 'بالصوم', 'وجاء', 'استطاع']	1	0.9717
['أركان الإسلام', 'التوحيد', 'توحيد الله تعالي', 'أوامره', 'صفات الله تعالى', 'التواب', 'العمل', 'العمل الطالح', 'العمل الأثم', 'في القول', 'الغيبة ', 'الإنسان والعلاقات الأخلاقية', 'الأخلاق الذميمة', 'التجسس', 'سوء الظن', 'الغيبة', 'الفضول', 'الإنسان والعلاقات الاجتماعية', 'المجتمع', 'الإخاء']	['الشوارب', 'اللحى', 'وأعفوا', 'أحفوا', 'خالفوا', 'الكبش', 'الكفن', 'الحلة', 'الأقرن', 'وأحفوا']	0	0.3211

Figure 2: Examples of comparing the semantic similarity scores between QH_Dataset labels and CL-AraBERT model embeddings results.

Metrics	
Axioms	2497
Class count	9
Object property count	4
Data property count	6
Individual count	463

Table 2: The RQHT ontology metrics.

Regarding the RQHT ontology, it is evaluated based on various criteria: correctness, completeness, and consistency. HermiT tool is the first freely available OWL Reasoner based on a novel "hypertableau" calculus (Glimm et al., 2014).
It is employed in this work to assess the consistency of the ontology and detect subsumption links between classes.

5 Results

This section provides an analysis of the performance of the proposed methodology for linking Qur'an and Hadith topics semantically. Table 1 shows the performance of the mentioned pretrained Arabic language models. CAMeL-BERT, CL-AraBERT, and MARBERT models performed quite similar findings in F1 score, but CL-AraBERT is the best among them with 0.9136 F1 score and 0.8444 accuracy score. The CL-AraBERT has successfully categorised the binary classification of Qur'an and Hadith topic pairs.

Figure 2 illustrates examples of the results obtained from the semantic similarity scores based on

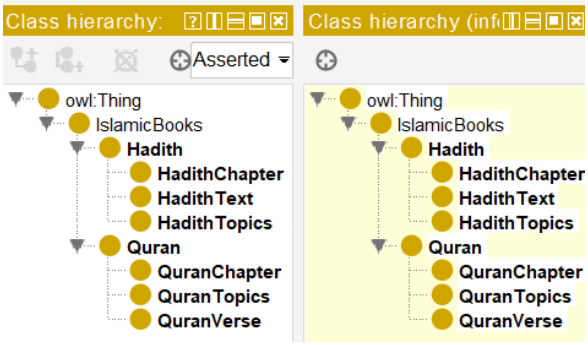

Figure 3: An example of the inferred class hierarchy using the HermiT reasoner.

CL-AraBERT model embeddings. The predicted scores compared to the labels of the QH_Dataset. As previously mentioned, the Qur'an and Hadith topics with semantic similarity scores exceeding the threshold of 0.5 are considered related. Subsequently, the results show no significant difference and explain the effectiveness of the model embedding in capturing the semantic relatedness of the Qur'an and Hadith topics.

Concerning RQHT ontology, Table 2 presents some of the ontology information. Also, Figure 1 illustrates the visualization of our ontology. It explains the related Qur'an and Hadith topics with their related knowledge, such as verse and Hadith numbers. Our experiment provides a consistent mapping between the classes, their instances and related topics by using HermiT reasoner. For example, the HermiT reasoner detected the inferred class hierarchy consistently, as shown in Figure 3.

6 Conclusion

In this paper, we described the experimental phases of linking Qur'an and Hadith topics and their related knowledge from various datasets. To the best of our knowledge, there is no work on linking Qur'an and Hadith topics and integrating them to related knowledge from many Qur'an and Hadith resources using word embeddings and the Cellfie Plugin tool.

This research consists of extracting the semantic similarity scores of Qur'an and Hadith topics and then comparing them to the labels of Qur'an-verse and Hadith-teaching pairs. Then, the Cellfie tool was employed to build the RQHT ontology by converting structured data in spreadsheets into OWL formats. The results of the conducted experiment have shown that the semantic similarity scores using the CL-AraBERT model are similar to

the labels and obtained an F1 score of 0.91. Meanwhile, the ontology correctly links the Qur'an and Hadith topics with their related data.

For future work, we plan to extend RQHT ontology by linking more Qur'an and Hadith data from different types of datasets, such as covering more languages and question-answering datasets. This work could provide a unification of Islamic resources and could contribute to many aspects of Islamic or religious education disciplines.

Limitations

The potential limitation of this work is the small amount of QH_Dataset, related Qur'an-verse and Hadith-teaching pairs. Thus, the experiment was conducted on topics that are related to the pairs of Qur'an-verse and Hadith-teaching.

Ethics Statement

In this study, we used datasets derived from publicly accessible data resources. Including Qur'an Ontology, QH_Dataset, and LK-Hadith-Corpus. The HT_Topics dataset is built by us and will be available in the future.

Acknowledgements

The first author would like to express her deepest gratitude for her supervisor's comments. Also, she would like to thank the University of Hafr Al Batin for sponsoring her PhD studies at the University of Leeds through the Saudi Arabia Cultural Bureau (SACB).

References

Ahmed Abdelghany, Hammam Abdelaal, Abdulrahman Kamr, and Passent Elkafrawy. 2020. Doc2vec: An approach to identify hadith similarities. *Australian Journal of Basic and Applied Sciences*, 14:46–53.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Hanan A. Al-Sanasleh and Bassam H. Hammo. 2017. Building domain ontology: Experiences in developing the prophetic ontology form quran and hadith. In *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, pages 223–228.

Abdullah Alsaleh, Eric Atwell, and Abdulrahman Altahhan. 2021. Quranic verses semantic relatedness using AraBERT. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 185–190, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Rusul Yousif Alsalhee and Abdulhussein Mohsin Abdullah. 2022. Building quranic stories ontology using mappingmaster domain-specific language. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(1):684–693.

Ibtisam Khalaf Alshammari, Eric Atwell, and Mohammad Ammar Alsalka. 2022. Automatic mapping of quranic ontologies using rml and cellfie plugin. In *Natural Language Processing and Information Systems*, pages 307–314, Cham. Springer International Publishing.

Menwa Alshammeri, Eric Atwell, and Mhd Alsalka. 2022. A siamese transformer-based architecture for detecting semantic similarity in the quran. *International Journal on Islamic Applications in Computer Science And Technology*, 9(4).

Menwa Alshammeri, Eric Atwell, and Mohammad Ammar Alsalka. 2021. Classifying verses of the quran using doc2vec. In *ACL Anthology*. ACL.

Shatha Altammami and Eric Atwell. 2022. Challenging the transformer-based models with a classical Arabic dataset: Quran and hadith. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1462–1471, Marseille, France. European Language Resources Association.

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2020. The arabic–english parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 8(2).

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2021. Towards a joint ontology of quran and hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 9(2):01–12.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Bendjamaa Fairouz, Taleb Nora, and Arari Amina Nouha. 2020. An ontological model of hadith texts. *International Journal of Advanced Computer Science and Applications*, 11(4).

Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. 1997. Methontology: From ontological art towards ontological engineering. In *AAAI Conference on Artificial Intelligence*.

Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, and Zhe Wang. 2014. Hermit: An owl 2 reasoner. *Journal of Automated Reasoning*, 53:245–269.

Thomas R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907–928.

Aimad Hakkoum and Said Raghay. 2015. Advanced search in the qur'an using semantic modeling. In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–4.

Aimad Hakkoum and Said Raghay. 2016. Semantic qa system on the qur'an. *Arabian Journal for Science and Engineering*, 41:5205–5214.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *ArXiv*, abs/2103.06678.

Rana Malhas and Tamer Elsayed. 2022. Arabic machine reading comprehension on the holy qur'an using cl-arabert. *Inf. Process. Manag.*, 59:103068.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. volume 4.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 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, Barcelona (online). International Committee for Computational Linguistics.

Abdul-Baquee Sharaf and Eric Atwell. 2012a. QurAna: Corpus of the Quran annotated with pronominal anaphora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 130–137, Istanbul, Turkey. European Language Resources Association (ELRA).

Abdul-Baquee Sharaf and Eric Atwell. 2012b. 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, Istanbul, Turkey. European Language Resources Association (ELRA).