KAB: Knowledge Augmented BERT2BERT Automated Questions-Answering system for Jurisprudential Legal Opinions

Saud S. Alotaibi	Amr A. Munshi	Abdullah Tarek Farag	Omar Essam Rakha	Ahmad A. Al Sallab	Majid Alotaibi
ssotaibi@uqu.edu.sa	aaamunshi@uqu.edu.sa	abdullahtarek57@gmail.com	i@omarito.me	ahmad.elsallab@gmail.com	mmgethami@uqu.edu.sa
Department of Information Systems, Umm Al-Qura University	Department of Information Systems, Umm Al-Qura University	Capiter	Faculty of Engineering, Ain Shams University	Faculty of Engineering, Cairo University	Department of Computer Engineering, Umm Al-Qura University

Summary
The jurisprudential legal rules govern the way Muslims react and interact to daily life. This creates a huge stream of questions, that require highly qualified and well-educated individuals, called Muftis. With Muslims representing almost 25% of the planet population, and the scarcity of qualified Muftis, this creates a demand supply problem calling for Automation solutions. This motivates the application of Artificial Intelligence (AI) to solve this problem, which requires a well-designed Question-Answering (QA) system to solve it. In this work, we propose a QA system, based on retrieval augmented generative transformer model for jurisprudential legal question. The main idea in the proposed architecture is the leverage of both state-of-the art transformer models, and the existing knowledge base of legal sources and question-answers. With the sensitivity of the domain in mind, due to its importance in Muslims daily lives, our design balances between exploitation of knowledge bases, and exploration provided by the generative transformer models. We collect a custom data set of 850,000 entries, that includes the question, answer, and category of the question. Our evaluation methodology is based on both quantitative and qualitative methods. We use metrics like BERTScore and METEOR to evaluate the precision and recall of the system. We also provide many qualitative results that show the quality of the generated answers, and how relevant they are to the asked questions.

Keywords:
Islamic Fatwa, Natural Language Processing, Question Answering, Transformers.

1. Introduction

Islamic Law, or Sharia, is characterized by a comprehensive set of immutable rules, which governs all aspects of Muslims lives. The Islamic jurisprudence or Fiqh is human interpretation of Sharia. Specialized and official institutions were established in several law colleges as a centralized place for issuing of fatwas for the general population. Examples are the Egyptian Dar al-Ifta, founded in 1895, and Al-Ifta is Saudi Arabia. Such schools are entitled to award certification that qualifies individuals, called Muftis, the provide answers, called Fatwa, to the Islamic legal questions. The certification and qualification process are sophisticated and takes many years, which creates scarcity in the number of Muftis. With the explosion of social media, and public websites, new channels of fatwas have emerged. While this facilitates the process of getting an answer for Muslims, however, it opens the door for many controversy or unauthentic fatwas. Also, it increases the demand and throughput of questions. With the increased demand, the number of Muftis is not matching the number of questions, from different channels. The issue is even more in the high seasons of Islam, like Ramadan or Hajj.

In this work, we aim to unleash the potential of Artificial Intelligence (AI) to deliver immediate Fatwa, an answer to a question about an Islamic Religion rule. Our focus in this work is Arabic language. We train a system based on questions and answers collected from official websites, in the same natural language they are asked. For that, we collect and release the largest dataset for that purpose. AI can power an automated Question Answering (QA), or Chatbot system, that relieves the load on the human experts.

Designing a QA system for Islamic Jurisprudential legal questions requires rigorous attention to the quality of the produced answer, because it is considered a reference for Muslim's day-to-day life decisions. Quality and relevance of the provided answer are the governing factors of such a design. While Knowledge base QA systems are easy to

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.44

design and implement, they fail to provide relevant answer, especially in cases of brand-new questions that no similar question exists in the data base. On the other hand, generative QA systems are conditioned on the asked question, but they might provide low quality answers in many cases, where they suffer the issue of "hallucination"; providing non-sense answers in some cases. Hence, we design a hybrid system, which considers prior knowledge bases, represented in repeated previous questions, questions categories and Islamic Jurisprudential reference books and sources, and at the same time, leverages the power of generative models to provide relevant answers to the asked question. The main components of our Knowledge augmented generative QA system are: 1) A knowledge base system that retrieves the relevant meta data of the question, such as: the nearest answer from the FAQ database, and the question category and 2) the generative model, which is based on encoder-decoder state-of-the art (SoTA) BERT model.

For our evaluation, we collect the largest Islamic Fatwa QA dataset, from online web sites, with 850,000 questions, answers, and question categories. This data set is used to nurture our knowledge base and train the generative encoder-decoder model. We leverage the power of state-of-the art embeddings to build the similarity match between the asked question and our knowledge base. On the generative model side, we evaluate the two main directions of building a sequence-to-sequence generative model: 1) using recurrent models like LSTM/GRU (seq2seq) powered by attention mechanisms and 2) using state-of-the art transformer models, using BERT models (BERT2BERT), which also leverages the power of pre-trained language models. We provide a full ablation of the effect of each stage in our system, with its effect on the overall performance, comparing many design alternatives in terms of preprocessing and model choice.

The rest of the paper is organized as follows: we first review the relevant literature sources to our work, to establish the necessary background that we build upon in the methodology section. Then we provide the details of our system, and the methodology of building each component. Following we provide the details of our experimental setup, and the details of our ablation studies, dataset, data preprocessing, and the main results. Finally, we conclude by discussing the main findings and the potential future directions to expand this work.

1.1 Background and related work

Chatbots and QA systems Taxonomy: A Chatbot can be thought of a high-level state-machine on top of an underlying QA engine. Chatbots can be classified according to different criteria:

Open-domain vs. Closed-domain Chatbots: Open-domain are more of conversational bots, with generic dialog flow. Closed-domain are Task-Oriented specific to an application domain.

Retrieval-based vs Generative: Retrieval-based systems are built using stored data base of Frequently Asked Questions (FAQ). With any given question, its text is matched based on some criteria, like cosine similarity for example, and the closet matched FAQ answer is retrieved. Generative systems generate brand new answer to the asked question, based on the understanding of the text. They follow the encoder-decoder design pattern, known as sequence-to-sequence (seq2seq). The question text is first encoded into an Embedding space, and then passed to the decoder to generate the answer. Such systems are further classifier into Recurrent based (LSTM or GRU) [1][2] or transformer based [3].

Islamic Fatwa Chatbots and QA systems: Some attempts have been made in the literature to build an automated QA or chatbot for Islamic Fatwa. Most of those are focused on knowledge and linguistic knowledge to match the asked question to the database. In [4], a retrieval based system is built using keywords matching with the NLTK text processing tool. The dataset used is focused on the questions related to the holy Quran. While keywords matching approach might work in a specific source like the holy Quran, it might fail in the general questions like the ones asked on social media and in the natural language with different accents. In our work, we build a more generic retrieval-based QA system using word embeddings matching, instead of keywords matching. This helps encoding the semantics of the question rather than exact keywords matching. Also, we rely on more general sources of questions-answers from online websites, which covers a more practical use-case. Following a similar path of string-matching similarity, [5] uses Fuzzy string matching to extract questions similarity scores, based on Quran and Hadith sources. In [6], a QA system is built from Hadith corpus. To overcome the issue of exact keyword matching, the authors resort to graph-based ranking methods to generate semantic and syntactic similarity measures, which requires an expensive language resource like Arabic WordNet (AWN). The use of graph-based method raises a question about the scalability of the system to natural language used on social media, and differences in accents. On contrary, our system is language-resources free, and is scalable via retraining on new data.

Transfer Learning in NLP: One of the biggest challenges in natural language processing (NLP) is the shortage of training data. Because NLP is a diversified field with many distinct tasks, most task-specific datasets contain only a few thousand or a few hundred thousand human-labelled training examples. However, modern deep learning-based

NLP models see benefits from much larger amounts of data, improving when trained on millions, or billions, of annotated training examples. To help close this gap in data, researchers have developed a variety of techniques for training general purpose language representation models using the enormous amount of unannotated text on the web (known as pre-training). A basic form of transfer learning has been applied in NLP in the past few years, in the form of learning useful word representations; known as "Word Embeddings". Word Embeddings have seen advances recently being applied in FastText from FaceBook [7], and ELMo [8].

Pre-trained representations can either be context-free or contextual, and contextual representations can further be unidirectional or bidirectional. Context-free models such as word2vec or GloVe generate a single word embedding representation for each word in the vocabulary. For example, the word "bank" would have the same context-free representation in "bank account" and "bank of the river." Contextual models, like BERT [9] and ELMo [8] instead generate a representation of each word that is based on the other words in the sentence.

Transfer Learning in Arabic NLP: Arabic language is considered among the Low-NLP Resources languages, unlike English. Looking on the literature today, there is a wide gap in applying the above techniques to Arabic NLP tasks. Transfer leaning of Word Embeddings was used in AROMA [10], using learnt embeddings from QALB dataset, to perform sentiment classification task. There is a high potential in applying the SOTA discussed above in the tasks of Arabic Opinion Mining (OMA) and Emotion Recognition. More recently, different pre-trained models for Arabic are released, like AraBERT and AraGPT [11][12][13].

Sequence-to-sequence generative models: Going beyond word representations, some new models appeared that focus on transfer learning on more useful architectures. Specifically, the model of encoder-decoder architecture started to take over in the field of Neural Machine Translation (NMT), like in seq2seq [1], which are based on BiLSTM models, and incorporate attention mechanisms, and the Transformer [3], which is fully based on attention gates, without any recurrent layers. Moreover, the learnt representations in that encoder, can be transferred to other tasks, like in ULMFiT [14], where a model is trained on large corpus for Neural Language Models (NLM), and then the backbone of the model is re-used to initialize a sentiment classification model on IMDB movie reviews. In BERT, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), and others.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms; an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. BERT builds upon recent work in pre-training contextual representations — including Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, and ULMFit. However, unlike these previous models, BERT is the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. Recently, BERT was used in a full encoder-decoder architecture, called BERT2BERT [15].

Knowledge augmented encoder-decoder models: Following the same architecture, other works exploited external knowledge, in addition to just the question text. In Empathetic BERT2BERT [16], where an emotion classifier is used to provide extra signal to guide the generative decoder. In RAG [17], a large scale knowledge augmentation is used, matching the question embedding to the nearest knowledge sources (like Wikipedia) embeddings. While this direction is a generic one, it does not guarantee the narrow guidance of the generator to the most relevant answer. Moreover, it requires huge training resources. Our work takes a more focused path, where we use relevant meta data to the asked question: nearest FAQ in our database, and the question category. We consider adding Islamic Jurisprudential Legal references in a similar fashion as RAG in future work.

2. Methodology

Designing a QA system for Islamic Jurisprudential legal questions requires rigorous attention to the quality of the produced answer, because it is considered a reference for Muslim's day-to-day life decisions. Hence, system design cannot be based on pure open-domain (chit-chat) architecture. On the other hand, an answer to a fatwa question is not always standard or routine answer as in closed-domain systems. So, the design must consider a balance between both design paradigms. Another design aspect of design, is to consider prior knowledge bases, represented in repeated previous questions, questions categories and Islamic Jurisprudential reference books and sources. While Fatwa QA systems cannot be considered as conversational context systems, however, such knowledge sources can be considered as a source of context that the produced answer must respect.

2.1 Knowledge-Augmented BERT2BERT (KAB)

Having those guidelines in mind, we designed the QA system as shown in Fig.1. The architecture considers both a generative and retrieval-based designs. The input question text (Q) is first processed, to clean and prepare it for next stages. Then it is passed to the Knowledge Data Base, which

contains a data base of historical question, answers, and category (QX, AX, CX) collected from online web sources of official Islamic Jurisprudential legal websites (to be described in detail in the dataset section). The question is matched against the data base to retrieve meta information about it. The most important information is the closest possible answer AX according to the historical records. This can be thought of matching a question to a list of FAQs (Frequently-Asked-Questions). Other meta information can also be retrieved, such as the question category class CX (Financial, Social, Legal…etc). The class of the historical questions is already an available information in our collected dataset. While it is not part of the current work, but the Knowledge data base can be further extended to include major Islamic Jurisprudential references, that the muftis use to give an answer. This is considered in scope of future work.

While the retrieved closest answer by itself can be considered the output of the system, however, it is not a high-quality answer. Unless the exact question, or a very near one exists in our database, the retrieved answer will be highly irrelevant to the question being asked. Also, it is not possible to classify the intent of the question, and generate a standard answer based on each intent like in conversational chat-bots, because we do not have a conversation context, and again the question-answer pair are far from being standard. Therefore, we add a generative Encoder-Decoder model stage. Generative Encoder-Decoder models by themselves can be very noisy and might hallucinate. So, we condition the answer not only on the input question text, but also on the context provided by the retrieval system (closest answer, question category…etc). Thye details of each system are provided in the next sections.

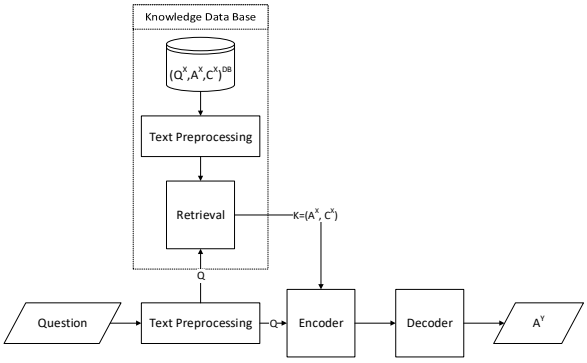

Fig.1 Knowledge Augmented Encoder-Decoder architecture

2.2 Knowledge and meta-data retrieval

Text Preprocessing. The aim of this first stage is to clean and vectorize the text into numerical indices, to reduce the noise in the text. Based on the resulting corpus of text, a vocabulary table V can be built, out of the unique tokens that will result. The more efficient the cleaning process, the smaller the vocabulary size is. A large vocabulary size will affect the model choice and size later, and hence we want to keep as small and efficient as possible. On the other hand, a small vocabulary size, might result in many Out-Of-Vocabulary (OOV) indices in the vectorization process. The next step is to vectorize the cleaned text into numerical tokens indices, $w_i \in 1,2,\dots V$ is the index of the word, selected from a vocabulary range $|V|$. The length of the sequence of tokens is padded with zeros to a maximum of N tokens.

Word Embedding Look-up Table (LUT). $E \in R^{V \times d}$, where d is the embedding dimension and V is the vocabulary size. The entries of this table are the words representations to be learnt, and thus they represent the learnable parameters of this block. Further, they can be pre-trained and fine-tuned. The result of the loop up operation is an embedding vector $e_i \in R^d$. For mathematical convenience, the look-up operation is usually done as a dot product operation, which enables an end-to-end graph that can be trained using gradient descent. In this case the word indices are converted into One-Hot-Encoded (OHE) vectors, $\hat{e} \in R^V$, which is sparse vector that has all zeros, except at the index of the of e_i. Now the embedding vector can be obtained as a simple dot product $e_i = \hat{e} \odot E$.
The word embeddings block is parametrized by the word vectors $E \in R^{V \times d}$, which are initialized randomly, and fine-tuned as part of the model optimization using gradient descent methods. It is also possible to use pre-trained embeddings tables, and fine-tune them, instead of random initializations. For that, we used two options of pre-trained embeddings: 1) Aravec [18] and 2) Fasttext [7].

Closest answer retrieval. In this approach the question is matched to the filtered subset of historic questions related to the topic. The system is shown in Fig.**2**. The Question similarity matching shall be done based on Cosine similarity. The whole model can then be end-to-end based on similarity loss objective.

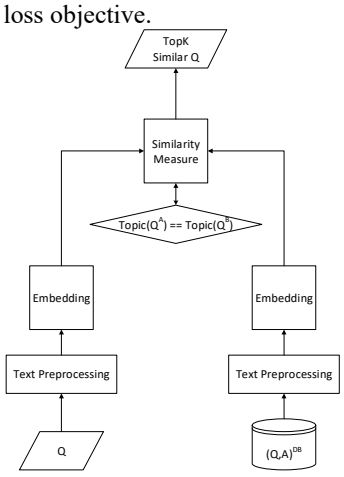

Fig.2 Closest FAQ answer retrieval architecture

Question category retrieval. This stage is to retrieve the question category to be used as meta-data in the knowledge augmentation part. We follow a hierarchical approach as in , where we have K-categories for the questions to choose from, hence we formulate the problem as a multi-class classification problem. The input question is tokenized and cleaned as described in the preprocessing pipeline, then the word embeddings are obtained for each token. The embedded tokens vectors are then aggregated using question sentence embedding, to provide question features to the softmax classification layer. The question embedding aggregates all the token embeddings vectors into one representation. We used an Arabic pre-trained transformer model, AraBERT [19] for question tokens embedding, which is a BERT transformer that was trained on Arabic data with a tweaked tokenizer that is specific for the Arabic words and Arabic word compounds.

2.3 Generative Encoder-Decoder Transformer

In our work, we employ the encoder-decoder design pattern. Mainly we follow the transformer architecture. However, we also evaluate the recurrent sequence-to-sequence models in the experiments section. The overall architecture is shown in Fig.3. The input to the encoder is the aggregated Question (Q), and the retrieved meta-data from the retrieval system (K), hence it is referred to Q+K. This is simply a concatenation of the embedding vectors of the question Q, and the knowledge K. The details of this aggregation is given in the experiments section. The aggregated input vector is first processed, tokenized and embedded into M tokens (q1...qM). Each Encoder block is based on multi-head self-attention mechanism, producing a transformed vector for the M tokens embeddings. This process is repeated over the number of layers of the encoder, producing the Question Embedding tokens (e1...eM). In the same manner, the decoder block consumes the following inputs: 1) the encoder question embeddings, 2) the ground truth answer AX, which is process, tokenized and embedded into (a1..aN) and 3) the previously generated answer tokens (b1..,bN) in an auto-regressive decoding fashion. The Multi-head attention block does not consider only the previous layer outputs, but also the encoder answer embeddings. Finally, the output modules produces the highest possible token, according to the generated probability, to generate the answer AY.

In both the encoder and decoder blocks, we follow the same pre-training and architecture design of BERT models. BERT is originally designed for text classification, so the encoder-decoder version of it is called BERT2BERT [15], which is available on hugging faces. We leverage the state-of-the art pre-trained models Arabic, AraBERT [19], for both the encoder and decoder

3. Experiments and Results

3.1 Dataset

The details of the dataset collection are shown in Table 1, of around 850,000 Fatwas (questions and answers). We crawl the popular websites of Islamic Fatwa, being official, like Al-Ifta-SA [20], Dar-al-ifta-EG [21] and Al-ifta-JO [22], or non-official like islamway [23], islamweb [24],...etc. Those websites span different countries and geographical locations, accents, and backgrounds. We crawl for question, answer, topic and date. For Arabic AskFM, we extend the one in [25] to include 604,000 fatwas, by crawling the full website. A special type of QA is found in islamonline [26], where we treat the articles titles as questions, and the bodies as answers, since they form the basic and frequently asked questions in Islamic Fatwa.

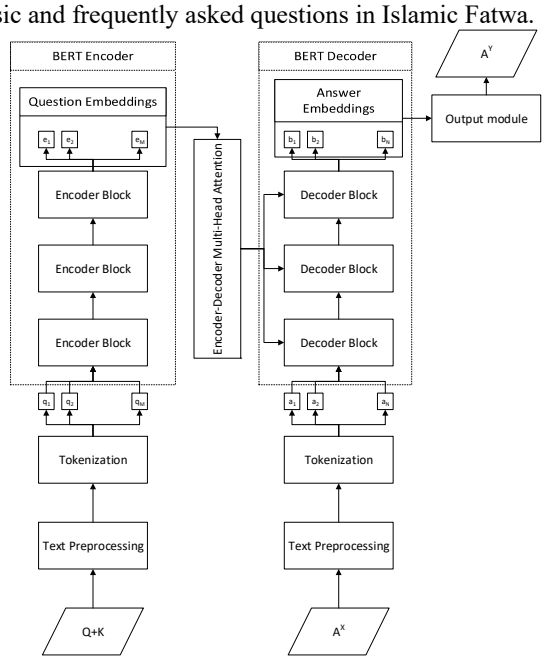

Fig.3 KAB: Knowledge Augmented BERT2BERT transformer architecture

Table 1 Dataset information, statistics, and sources

Dataset	***Question/Answers***	***Topics***	***Dates***
Al-ifta-SA [20], Dar-al-ifta-EG [21]	40,161	Yes	Yes
AskFM [25]	604,184	N/A	N/A
Islamweb [24]	126,000	Yes	Yes
Islamway [23]	15,060	N/A	Yes
Islamonline [26]	3,100	Yes	N/A
binbaz [27]	28,226	Yes	N/A
binothaimeen [28]	2,157	Yes	N/A
Islamqa [29]	30,780	Yes	Yes

3.2 Metrics

METEOR [30]. Like BLEU 1, METEOR score is based on unigram matching between the machine produced sequence and human-produced reference sequence. METEOR computes a score for this matching using a combination of unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the machine sequence are in relation to the reference. METEOR is evaluated by measuring the Pearson R correlation between the metric scores and human judgments of translation quality.

BERTScore [31]. As in other metrics, BERTScore computes a similarity score for each token in the candidate answer with each token in the ground truth answer. Hence, it is provided in terms of precision, recall and F1 measures. However, instead of exact matches, BERTScore computes token similarity using contextual embeddings. This helps to quantify the quality and relevance of the answer semantics rather than exact word match. Token embeddings are calculated using pre-trained BERT.

3.3 Text normalization and cleaning

The first step is to clean and normalize the text. An important factor in this process is to reduce the variability and noise in the text, such that, only the important tokens are kept. The following pipeline was applied: 1) Special and non-Arabic characters removal. 2) Arabic Diacritics removal. 3) Punctuation removal. 4) Numbers removal. 5) Stop words removal (using NLTK Arabic set). 6) Stemming using ISRIStemmer for Arabic.

After the cleaning process the questions and answers were pre-processed using the AraBERT preprocessor. This was necessary to do to use the Arabert weights. The pre-processor was very as it divided the words into parts that made sense in the Arabic language. After that the preprocessed text was tokenized using the AraBERT tokenizer and fed through the models.

3.4 Results and Discussions

We compare the following models: Recurrent-based sequence-to-sequence model (LSTM), Transformer-based sequence-to-sequence model (BERT2BERT) and Knowledge-Augmented BERT2BERT (KAB). Results are shown in Table 2.

Table 2 Evaluation of Encoder-Decoder seq2seq different models' setups

Model	BERT Score			METEOR
	Precision	Recall	F1	
LSTM based seq2seq	0.38	0.35	0.36	0.0005
BERT2BERT	0.57	0.37	0.44	0.036
Knowledge-Augmented BERT2BERT (KAB)	0.6	0.4	0.48	0.037

In terms of the answer quality, LSTM seq2seq model predicted the same answer for almost all the questions. The answer was a sentence said as an introduction to the answers. It was "الحمد لله والصلاة والسلام على رسول الله". The model collapsed to a mode, where the mostly repeated part of the answer is always generated. This phenomenon is not reflected in the evaluation scores since they mostly evaluate the presence of common n-grams between the ground truth and predictions. However, we can see a big drop in METEOR score. For BERT2BERT, both the encoder and decoder were initialized with the pre-trained AraBERT models. Encoder max length of 126 characters and a decoder length of 256 characters. Results show clear improvement in all scores due to the power of the transformer-based architecture, in addition to the power of transfer learning from relevant Arabic pre-trained weights in AraBERT.

Finally, in Knowledge-Augmented BERT2BERT (KAB), when BERT2BERT is supported with knowledge from the closed answer, and the question category, we see an improvement in the quantitative score. However, the quantitative score improvement does not fully reflect the quality of the generated answers. The main reason is that quantitative scores are mostly based on words and n-grams matching, without considering the quality and relevance of the answers. For that, we provide qualitative results to show the true effect of knowledge augmentation in Table 3. We can see some issues with BERT2BERT answers: they are irrelevant in some cases. Moreover, the generated answer is just a "hallucination", repeating keywords from the true answer. Those issues are not found in the KAB model.

Effect of forced start token. We noticed from the previous trial that on very rare cases where the model predicted something else other than the sentence presented above the answer began with "الجواب" (translated to the answer). So, we decided to add this word at the beginning of every answer and see what will happen. However, qualitative results of this model were not good. Although the BERT scores and METEOR score improved noticeably as in Fig.4, a good chunk of the questions got the same answer, for approximately 70% the questions. The repeated answer was " الجواب الحمد لله والصلاة والسلام على رسول الله ".

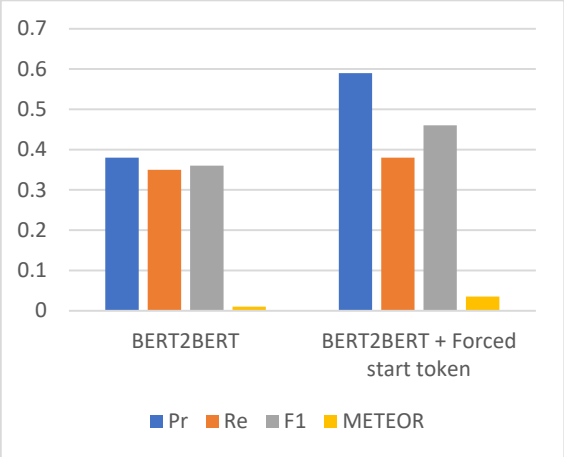

Fig.4 Bar chart for the effect of forced start token on BERT2BERT model

Effect of common sentences removal. We removed the sentences that repeated in the answers. We found that more than 80% (>75th Percentile) of the answers contained certain "introductory" sentences, that Muftis tend to include in answers. They might be in the start, middle, end. We removed every sentence in the dataset that occurred more than 300 times in the dataset. This made our answers shorter and more to the point.

The model started providing different answers to each question and moreover some of the answers made sense and most of the answers were giving an answer about the same subject as the question. Notice that the BERT scores and METEOR score did not improve much from basic BERT2BERT, because answers were inflated by predicting sentences that were exactly in the answers but provided no qualitative value. In fact, Pr, Re and F1 decreased due removing the repeated sentences (which are highly repeated and gives false high BERT score, but less quality results. Overall, METEOR score was slightly improved as in Fig.5. However, we see much improved quality of the generated answer as shown in the examples in Table 3.

Effect of closest answer augmentation. For this model we wanted to further improve the quality of the answers by concatenating the top K relevant documents to the question (in our case K=1 for memory constraints). This helps the model to better answer the questions by having references. Quantitative scores improved as shown in Fig.6, and also the qualitative answers are highly improved as shown in Table 3 It can extract information from while generating the data. We got the closest question to each question in the database using AraBERT pooled document embedding. Then we took the closest answer of the question and concatenated it to the question at hand. Now each input has a question and an answer in one sentence separated by an @ symbol. The input size is 256 and the maximum length of the question is 126 and the rest is for the closest answer. This model also removes the repeated sentences.

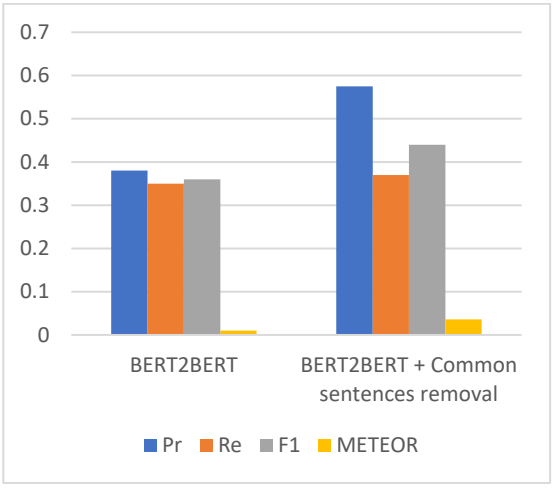

Fig.5 Bar chart for the effect of common sentences removal on BERT2BERT model

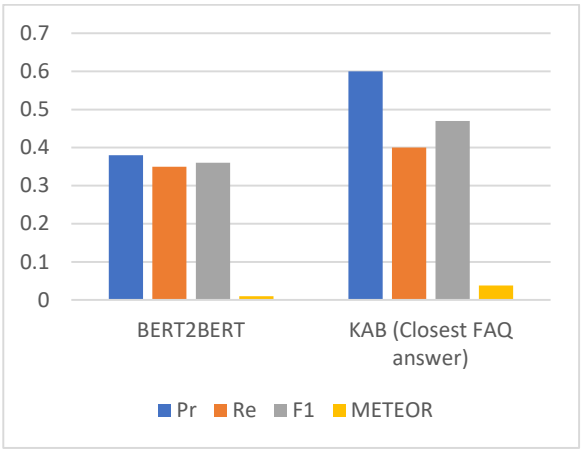

Fig.6 Bar chart for the effect of closet FAQ answer augmentation on BERT2BERT model

Effect of question category augmentation. Another knowledge signal that can help the generated answer is the category of the question being asked. Most Fatwa websites categorizes the questions into sections, which we use as a label for the question category while collecting our dataset. Using this information, we train a hierarchal transformer-based classifier using AraBERT pre-trained weights. In the run time, the classifier is queried for the question category, and the class label is embedded and appended to the knowledge augmented model, along with the closest FAQ answer. This improves the scores as shown in Fig.7.

Pure retrieval-based system. To complete our analysis, we compare KAB to the pure retrieval-based model, which

just retrieves the answer of the closed FAQ in the database. Which this approach is simple, but it suffers two main practical issues: 1) the asked question might be a new one, not related to any of the saved questions in the data base, in which case the returned answer will be weird and completely irrelevant to the asked question, leading a naïve system. 2) in terms of qualitative scores, we might get misleading precision, recall values, since the retrieved answer will contain a lot of common keywords, however, we see a clear bad performance in the qualitative results. We provide some of those results in Table 4. While the retrieved closest answer is apparently like the asked question, in terms of keyword, and even the category of the question, however, it is completely irrelevant to the true answer or the asked question.

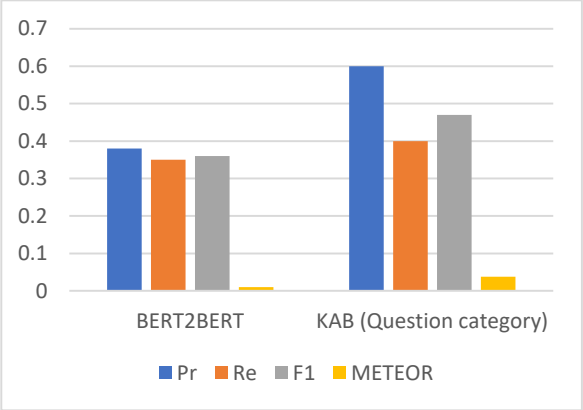

Fig.7 Bar chart for the effect of question category augmentation on BERT2BERT model

Table 3 Qualitative results of sample questions and answers for BERT2BERT and KAB models

Question	**Ground truth Answer**	**BERT2BERT answer**	**Knowledge-Augmented BERT2BERT (KAB) answer**
هل تتفاوت درجات الثواب للأعمال الصالحة هل هذه الدرجات متساوية من ناحية الثواب التي هي القرآن والقراءة والقتال والصدقات	لا الثواب يختلف على حسب نية العبد وإخلاصه وصدقه واستكماله العمل يختلف هذا يصلي وهذا يصلي وبين صلاتيهما أعظم مما بين المشرق والمغرب هذا صلاته قد أقبل عليها واعتنى بها وخشع فيها لله وكملها وهذا عنده نقص فيها فبينهما فرق وهكذا صيامهم وهكذا حجهم يختلفون	العمل الصالح هي الأعمال الصالحة التي 'تقرب العبد إلى الله تعالى	أما بعد فالثواب المترتب على الأعمال الصالحة لا 'يتفاوت بحسب العمل
شخص وجد مبلغا من المال في مكينة ' الصرافة فلا يعلم أن يذهب بها الشيخ لكن موضوع من الخارج أم في جوف المكينة السائل شخص صرف ونسي النقود في المكينة الشيخ لكن من خارج أم في جوف 'المكينة السائل لا من خارج	ا اذهب للراجحي وقل له في اليوم الفلاني صرف إنسان ونسي هذه الدراهم انظر ما يمكن أن يخرج من عند الصراف إلا مقيدا عنده فمثلا إذا كان مقيدا عنده مثلا عشرة يسألونه أنت يا فلان أخذت مائة هل فقدت منها شيئا ويعلمون بهذا السائل إذا لم أجد الشخص الشيخ على كل حال أنت اذهب إليهم وأخبراهم بالواقع إذا لم تحصل على نتيجة فانتظر لك كم يوم لعله يجيء وإذا يئست منه تصدق بها عليه	الصرافة التي في الصرافة لها حكم اللقطة فإن كان'	الواجب على الشخص المذكور أن يبحث عن صاحب المال وأن يسأله عنه
'هل يجوز نداء الشهر وسؤاله مثال أيها الشهر ترفق فدمع الفراق تدفق وفؤاد المحبين تمزق عسانا بركب المقبولين نلحق وعسانا من النيران نعتق فلا تفارقنا إلا ورضى المولى تحقق'	الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد فمثل هذا النداء والخطاب للشهر لا حرج فيه وهو أسلوب سائغ تتابع عليه المسلمون دون نكير المهم أن يكون مقصوده مباحا وكلماته صحيحة لا تتضمن مخالفة شرعية كما في المثال المذكورواللهأعلم	'ال الشهرة والشهرة والشهرة والشهرة والشهر'	'أما بعد فلا حرج في نداء الشهر القمري بالشهر القمري'
'إذا حلق أحد الناس شعره أو خففه يقول له بعض الآخرين نعيما ويقول الذي حلق رأسه أو خففه الله ينعم عليك فهل هذا القول جائز أم لا افتونا بارك الله فيكم'	الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد فلم يثبت عن النبي صلى الله عليه وسلم أنه كان يقول لمن حلق شعر رأسه أو قصره شيئا والظاهر أنه لا حرج في القول المذكور في السؤال لكن ينبغي ألا يتخذ عادة خشية أن يعتقد الجاهل أنه سنة والله أعلم	'الة الشعر والشعر والشعر والشعر والشعر والشعر'	'أما بعد فلا حرج في قول الشعر المذكور
'هل صيام الثلاثة أيام الأولى من كل شهر عربي بدعة أم سنة وكذلك النصف من شعبان'	الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعدفصوم ثلاثة أيام من كل شهر سنة والأفضل أن تكون الثالث عشر والرابع عشر والخامس عشر كما بيناه في الفتوى رقم وأما تخصيص صوم النصف من شعبان فليس بسنة لعدم صحة الدليل بذلك كما في الفتوى رقم والله أعلم	الصيام الثلاثة الاولى الثلاثة الاولى من شعبان هي الصيام'	'أما بعد فإن الصيام في شهر شعبان سنة وليس بواجب
'هل يجوز استخدام أدوات التجميل التي تحتوي على كحول مثل كريم الجسم'	الحمد لله أدوات التجميل المحتوية على كحول الأحوط عدم استعمالها للخلاف المعروف في نجاسة المسكر وإذا اضطر الإنسان إليها وكانت نسبة الكحول يسيرة فلا بأس إن شاء الله	'آلات التجميل التي تحتوي على كحول أو نجسة لا يجوز استعمالها	'آلات التجميل التي تحتوي على كحول أو نجسة لا يجوز استعمالها

Table 4 Qualitative samples comparing KAB vs. Retrieval based systems

Question	Ground truth answer	Matched FAQ question	Retrieval based answer	KAB answer
هل تتفاوت درجات الثواب للأعمال الصالحة هل هذه الدرجات متساوية من ناحية الثواب التي هي القرآن والقراءة والقتال والصدقات	لا الثواب يختلف على حسب نية العبد وإخلاصه وصدقه واستكماله العمل يختلف هذا يصلي وهذا يصلي وبين صلاتيهما أعظم مما بين المشرق والمغرب هذا صلاته قد أقبل عليها واعتنى بها وخشع فيها لله وكملها وهذا عنده نقص فيها فبينهما فرق وهكذا صيامهم وهكذا حجهم يختلفون	ما حكم الذكر الجماعي بعد الصلاة على وتيرة واحدة كما يفعله البعض وهل السنة الجهر بالذكر أو الإسرار ؟	سنة الجهر بالذكر عقب الصلوات الخمس وعقب صلاة الجمعة بعد التسليم لما ثبت في الصحيحين عن ابن عباس رضي الله عنهما «أن رفع الصوت بالذكر حين ينصرف الناس من المكتوبة كان على عهد النبي صلى الله عليه وسلم قال ابن عباس «كنت أعلم إذا انصرفوا بذلك إذا سمعته» أما كونه جماعيا بحيث يتحرى كل واحد نطق الآخر من أوله إلى آخره وتقليده في ذلك فهذا لا أصل له بل هو بدعة وإنما المشروع أن يذكروا الله جميعا بغير قصد لتلاقي الأصوات بدءا ونهاية	أما بعد فالثواب المترتب على الأعمال الصالحة لا يتفاوت بحسب العمل'
هل يجوز إسقاط الحمل إذا لم يبلغ أشهر وذلك لوجود أطفال صغار والصغرى تبلغ عاما ونصفاوستكون الولادة في غير وطننا'	الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعدفإن إسقاط الحمل محرم شرعا فهو من الإفساد في الأرض وإهلاك النسل قال الله تعالى وإذا تولى سعى في الأرض ليفسد فيها ويهلك الحرث والنسل والله لا يحب الفساد البقرةويشتد التحريم وتعظم الجريمة إذا وصل الحمل إلى أربعة أشهر وهي المرحلة التي تنفخ فيها الروح فيكون من قتل النفس التي حرم الله إلا بالحقويكون التحريم أشد إذا كان الباعث على ذلك خشية الإملاق كما كان أهل الجاهلية يفعلون وقد نهى الله تعالى عن ذلك ووصفه بالخطأ الكبير فقال الله تعالى ولا تقتلوا أولادكم خشية إملاق نحن نرزقهم وإياكم إن قتلهم كان خطئا كبيرا الإسراءولمزيد من الفائدة نرجو الاطلاع على الفتوى رقم والله أعلم	هل يجوز للمرأة المطلقة ولديها أولاد جعل وليها ابنها في عقد الزواج لأن أباها لم يحب أن يتدخل في موضوع الزواج لبعض التحفظات غير الشرعية مثل لأن الزوج معدد الزوجات؟	الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد فلا يحق للولي أن يمتنع عن تزويج المرأة لمجرد كون من أراد أن يتقدم للزواج منها متزوجا من قبل فليس هذا بمسوغ شرعا لعدم تزويجه إياها فإن رفض تزويجها لغير عذر كان عاضلا لها فتنتقل الولاية إلى الولي الأبعد أو إلى القاضي في قول بعض الفقهاء والذي يلي الأب في الترتيب هو الابن ومن الفقهاء من قدم الابن على الأب كما هو مبين بالفتوى رقم فننصح المرأة بأن تجتهد في محاولة إقناع وليها بتزويجها منه ولتشفع إليه بمن يرجى أن يكون قوله مقبولا عنده فإن اقتنع فالحمد لله وإلا فلترفع أمره إلى القاضي الشرعي لينظر في الأمر فإن ثبت عنده العضل زوجها أو أمر وليها الأبعد بتزويجهاوالله أعلم	أما بعد فلا يجوز إسقاط الجنين في أي مرحلة من مراحل الجنين ولو كان'

4. Conclusion

In this work, we design and evaluate a large-scale Automatic Islam Jurisprudential Legal Question-Answering system. The system leverages both retrieval-based knowledge-based systems, and generative-based sequence-to-sequence machine learning models. Our design leverages the state-of-the art BERT2BERT model, pretrained using AraBERT transformer model for Arabic language. Our experimental setup is supported with the largest available QA dataset for Islam Jurisprudential Legal Question-Answers, which is collected from trusted online sources to match day-to-day questions language. We also designed rigorous algorithms to clean and process the questions, to detect and remove repeated introductions in the answers. Our evaluation framework involved standard metrics like METEOR and BERT Score to evaluate the retrieval and precision abilities of the system. Our results show superior quantitative results, both qualitatively and quantitively. We evaluated different setups for the generative encoder-decoder models, like LSTM/GRU and transformer-based systems, where we find that AraBERT with BERT2BERT encoder-decoder model is the best choice. Future work includes: 1) extension to other languages than only Arabic, and 2) Including reference Automatic Islam Jurisprudential Legal reference books as source of knowledge.

Acknowledgments

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding their research work through the project number 20-UQU-IF-P3-001.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv Prepr. arXiv1409.0473, 2014.

[2] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” Aug. 2015, Accessed: Aug. 09, 2018. [Online]. Available: http://arxiv.org/abs/1508.04025.

[3] A. Vaswani et al., “Attention is all you need,” arXiv Prepr. arXiv1706.03762, 2017.

[4] B. Hamoud and E. Atwell, “Quran question and answer corpus

for data mining with WEKA," in 2016 Conference of Basic Sciences and Engineering Studies (SGCAC), 2016, pp. 211–216.

[5] M. T. Sihotang, I. Jaya, A. Hizriadi, and S. M. Hardi, "Answering Islamic Questions with a Chatbot using Fuzzy String-Matching Algorithm," in Journal of Physics: Conference Series, 2020, vol. 1566, no. 1, p. 12007.

[6] A. Abdi, S. Hasan, M. Arshi, S. M. Shamsuddin, and N. Idris, "A question answering system in hadith using linguistic knowledge," Comput. Speech \& Lang., vol. 60, p. 101023, 2020.

[7] B. Athiwaratkun, A. G. Wilson, and A. Anandkumar, "Probabilistic fasttext for multi-sense word embeddings," arXiv Prepr. arXiv1806.02901, 2018.

[8] M. E. Peters et al., "Deep contextualized word representations," arXiv Prepr. arXiv1802.05365, 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv Prepr. arXiv1810.04805, 2018.

[10] A. Al-sallab, R. Baly, H. Hajj, K. B. Shaban, W. El-hajj, and G. Badaro, "AROMA : A Recursive Deep Learning Model for Opinion Mining in Arabic as a Low Resource Language," vol. 16, no. 4, 2017.

[11] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," Feb. 2020, Accessed: Jul. 05, 2021. [Online]. Available: http://arxiv.org/abs/2003.00104.

[12] M. Djandji, F. Baly, H. Hajj, and others, "Multi-Task Learning using AraBert for Offensive Language Detection," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 97–101.

[13] A. M. Abu Nada, E. Alajrami, A. A. Al-Saqqa, and S. S. Abu-Naser, "Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach," 2020.

[14] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv Prepr. arXiv1801.06146, 2018.

[15] C. Chen et al., "bert2BERT: Towards Reusable Pretrained Language Models," Oct. 2021, Accessed: Jan. 31, 2022. [Online]. Available: http://arxiv.org/abs/2110.07143.

[16] T. Naous, W. Antoun, R. A. Mahmoud, and H. Hajj, "Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with Little Data," Mar. 2021, Accessed: Jan. 29, 2022. [Online]. Available: https://arxiv.org/abs/2103.04353.

[17] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, Accessed: Jan. 31, 2022. [Online]. Available: http://arxiv.org/abs/2005.11401.

[18] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," Procedia Comput. Sci., vol. 117, pp. 256–265, 2017.

[19] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv Prepr. arXiv2003.00104, 2020.

[20] AiIftaSA, "alifta," https://www.alifta.gov.sa.

[21] DarAlIftaEG, "Dar-al-ifta," https://www.dar-alifta.org/ar/Default.aspx?sec=fatwa&1&Home=1.

[22] AlIftaJO, "alifta-jo," https://aliftaa.jo/.

[23] Islamway, "islamway," https://ar.islamway.net/fatawa/source/.

[24] Islamweb, "islamweb," https://www.islamweb.net/ar/.

[25] AskFM98k, "askfm98k," https://omarito.me/arabic-askfm-dataset/.

[26] Islamonline, "islamonline," https://islamonline.net/.

[27] Binbaz, "binbaz," https://binbaz.org.sa/fatwas/kind/1.

[28] Binothaimeen, "binothaimeen," https://binothaimeen.net/site.

[29] Islamqa, "islamqa," https://islamqa.info/.

[30] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Apr. 2019, Accessed: Feb. 02, 2022. [Online]. Available: https://arxiv.org/abs/1904.09675.

Authors Biography

Saud S. Alotaibi is an associate professor of Computer Science at the Umm Al-Qura University, Makkah, Saudi Arabia. He received the Bachelor of Computer Science degree from King Abdul Aziz University, Jeddah, KSA, in 2000, the Master's degree in Computer Science from King Fahd University, Dhahran, KSA, in May 2008, the Ph.D. degrees in Computer Science from Colorado State University, Fort Collins, USA, in August 2015. From January 2009 to 2017, he worked as a Vice Deputy of IT Center at Umm Al-Qura University, Makkah, KSA. Then, he was designated as Vice COE of the Smart Campus. Currently, he is the dean of College of Computer and Information Systems at the same university. His current research interests include Emotional Intelligence, Data Mining, Natural Language Processing, Machine Learning, Deep Learning, Computer Networks, Wireless Sensor Networks and Network Security.

AMR A. MUNSHI received the B.Sc. degree in computer engineering from Umm Al-Qura University, Makkah, Saudi Arabia, in 2008, and the M.Sc. and PhD degrees in computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2014 and 2019, respectively. Currently Dr. Munshi is an Assistant Professor in the Computer Engineering department at Umm Al-Qura University, Makkah, Saudi Arabia. His research interests include data mining, smart grids and big data analytics. Dr. Munshi is a member of the Golden Key International Honor Society and currently serves as an Editor of the Alberta Academic Review journal.

Abdullah Tarek Farag received B.Sc. degree in computer science from MSA university in 2018. He then started working as a computer vision engineer for a year and a half. Then worked as a data scientist and an NLP engineer.

Omar Essam Rakha received the B.Sc. degree from the faculty of engineering, Ain shams university and has since been working as an AI research engineer with specific interest in NLP.

Ahmad A. Al Sallab received his B.Sc. from the Faculty of Engineering, Cairo University in Electronics and Communications. He acquired his M.Sc. and Ph.D. in 2009 and 2013 from Cairo University in Artificial intelligence. Ahmad has 17 years of practical experience, where he worked for reputable multi-national organizations in the industry like Intel and Valeo. He has over 40 publications and book chapters in top IEEE and ACM journals and conferences, in addition to 8 patents filed in European and German patent offices, with applications in Speech, NLP, Computer Vision and Robotics.

Majid Alotaibi received Ph.D. from The University of Queensland, Brisbane, Australia, in 2011. Currently, he is an associate professor with the Department of Computer Engineering, Umm AlQura University, Makkah, and the Co-founder of the SMarT Lab. His current research interests include mobile computing, mobile and sensor networks, wireless technologies, IoT in Healthcare, Smart Cities, ad-hoc networks, computer networks (wired/wireless), RFID, antennas and propagation, and radar. Digital transformations: Frameworks, Strategy, Enterprise Architecture, Governance, Technologies, and Data.