Pertanika J. Sci. & Technol. 30 (2): 1135 - 1156 (2022)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Automated Islamic Jurisprudential Legal Opinions Generation Using Artificial Intelligence

Amr Abdullah Munshi[1], Wesam Hasan AlSabban[2], Abdullah Tarek Farag[3], Omar Essam Rakha[4], Ahmad Al Sallab[5]* and Majid Alotaibi[1]

[1]Department of Computer Engineering, Umm Al-Qura University, Makkah, Saudi Arabia
[2]Department of Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia
[3]Giza Cairo, Cairo, Egypt
[4]Faculty of Engineering, Ain Shams University, Cairo, Egypt
[5]Faculty of Engineering, Cairo University, Cairo, Egypt

ABSTRACT

Islam is the second-largest and fastest-growing religion. The Islamic Law, Sharia, represents a profound component of the day-to-day lives of Muslims. While sources of Sharia are available for anyone, it often requires a highly qualified person, the Mufti, to provide Fatwa. With Islam followers representing almost 25% of the planet earth population, generating many queries, and the sophistication of the Mufti qualification process, creating a shortage in them, we have a supply-demand problem, calling for Automation solutions. This scenario motivates the application of Artificial Intelligence (AI) to Automated Islamic Fatwa in a scalable way that can adapt to various sources like social media. In this work, the potential of AI, Machine Learning, and Deep Learning, with technologies like Natural Language Processing (NLP), paving the way to help the Automation of Islam Fatwa are explored. The work started by surveying the State-of-The-Art (SoTA) of NLP and exploring the potential use-cases to solve the problems of Question answering and Text Classification in the Islamic Fatwa Automation. The first and major enabler component for AI application for Islamic Fatwa, the data were presented by building the largest dataset for Islamic Fatwa, spanning the widely used websites for Fatwa. Moreover, the baseline systems for Topic Classification, Topic Modeling, and Retrieval-based Question-Answering are presented to set the future research and

ARTICLE INFO

Article history:
Received: 20 May 2021
Accepted: 15 September 2021
Published: 14 March 2022

DOI: https://doi.org/10.47836/pjst.30.2.16

E-mail addresses:
aaamunshi@uqu.edu.sa (Amr Abdullah Munshi)
whsabban@uqu.edu.sa (Wesam Hasan AlSabban)
abdullahtarek57@gmail.com (Abdullah Tarek Farag)
i@omarito.me (Omar Essam Rakha)
ahmad.elsallab@gmail.com (Ahmad Al Sallab)
mmgethami@uqu.edu.sa (Majid Alotaibi)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

benchmark on the dataset. Finally, the dataset is released and baselines to the public domain to help advance future research in the area.

Keywords: Artificial intelligence, Islamic fatwa, machine learning, natural language processing, question answering, text classification

INTRODUCTION

Islam is the third Abrahamic Religion, and the second-largest religion, with over 1.9 billion followers: Muslims, representing 24.9% of the earth's population in 51 countries (https://en.wikipedia.org/wiki/Islam). Also, Islam is the fastest-growing religion, at a growth rate of 1.84% (2010–2015) per year. Islam is characterized by a comprehensive set of immutable diving rules, representing the Islamic Law, or Sharia, that governs all aspects of Muslims' lives. The Islamic jurisprudence or Fiqh represents the human interpretation of Sharia. The traditional theory of Islamic jurisprudence recognizes four sources of Sharia: the Quran, sunnah (authentic hadith), qiyas (analogical reasoning), and ijma (juridical consensus). Different legal schools , also called madhhabs, the most prominent are Hanafi, Maliki, Shafi school, Hanbali, and Jafari. Classical jurisprudence was elaborated by private religious scholars, largely through legal opinions (fatwas) issued by qualified jurists (muftis). Qualifying a Mufti until certification to issue fatwas is a sophisticated learning process. Traditionally, Islamic law was taught in private circles in Mosques by a reputable Sheikh, assisted by advanced students, which results in a certification or ijaza. This tradition even extended to other branches of education, like medicine, law, and mathematics. In the modern era, specialized institutions were established in several law colleges as a centralized place for issuing fatwas for the general population. Examples are the Egyptian Dar al-Ifta, founded in 1895, and Al-Ifta in Saudi Arabia.

With the explosion of social media and public websites, new channels of fatwas have emerged. A typical scenario is that a person sends his/her question, questions are gathered and distributed among muftis, and answers are posted back, in private or public. Such channels can either be official websites, reviewed by an authority or unofficial. While this facilitates getting an answer for Muslims, it opens the door for many controversy or unauthentic fatwas.

This work aims to unleash the potential of AI to deliver immediate Fatwa, an answer to a question about an Islamic Religion rule. The main focus in this work is the Arabic language. While most existing works focus on generating answers from Quran and Hadith sources, almost none is focused on questions and answers in the social media channels. In this work, a system based on questions is trained, and answers are collected from official websites in the same natural language they are asked. For that, the work collects and releases the largest dataset for that purpose. Usually, those answers must be given by highly qualified experts who had years of specialized education and verified degrees from the highest religious institutes. It creates a high-demand low-supply situation. This situation is demanding in the high seasons, like Pilgrim, Umrah, and Ramadan, which calls for automation.

AI can answer many of these needs. At the goal, an automated Question Answering (QA)/Chatbot system can relieve the load of the human experts. However, even high gains can be achieved by simpler solutions, like categorization and routing of the question to a specialized expert, recommending an immediate answer by matching the user and the question to a database of previous answers-users who might have similar questions and/or similar background (origin, language, ethnicity, and sex) and verifying the authenticity of an answer/Fatwa.

Artificial intelligence (AI) is changing the shape of the world. In the last few years, a lot of potential has been there for applied AI in Natural Language Processing (NLP). Following the Computer Vision (CV) field, NLP has reached the so-called "Image-Net moment" with the introduction of Transfer Learning and Transformers (Devlin et al., 2018; Vaswani et al., 2017; Howard & Ruder, 2018). That potential is not fully unleashed in Low-Language Resources (LLR) like Arabic. Some attempts have been made, such as Antoun et al. (2020), Djandji et al. (2020), Nada et al. (2020), Al Sallab et al. (2014), Rashwan et al. (2015), Al Sallab et al. (2015a), Al Sallab et al. (2015b), Baly et al. (2016), Al-Sallab et al. (2017), Magooda et al. (2016), and Baly et al. (2017). One important application of AI in NLP is the area of Personal Assistants, Chatbots, and Question Answering (QA) systems, where AI delivers State-of-The-Art (SoTA) performance. Such applications are vital to domains where human experts can be overwhelmed by the high traffic of requests/questions, especially when the questions are repetitive or could be clustered and routed to the proper expert ahead of time.

The researchers started surveying the SoTA in NLP, focused on QA systems, Chatbots, and Text Classification. It leads to the discussion of the potential application areas of AI to Automated Fatwa systems and the different use-cases scenarios. They focus their contribution to this paper on the main building block that enables other applications, which is data. The work collects and builds the largest dataset of Islamic Fatwas from a diversity of the most popular Fatwa websites, official and non-official, spanning different geographical locations, accents, and backgrounds. The dataset includes the queries and answers and the topic and date of the fatwa when applicable. It helps the researchers to perform Exploratory Data Analysis (EDA), like Unsupervised Topic Modeling and Seasonality Analysis. To set baselines for future research on the dataset, they build baseline models for Topic Classification and Retrieval-based systems using Word Embeddings and Text Similarity matching. They release all models and datasets to the public domain to help advance the research in the area.

The main contributions of this work can be summarized as follows: 1) Architecture design of an Automated Islam Jurisprudential Legal Opinions Generation. 2) Collection and annotation of the largest dataset for Islam Jurisprudential Legal Opinions, with over 850,000 questions, answers, and topics, and releasing it to the public domain to help advance the state-of-the-art. 3) Baseline results for QA and Topics classification on the released dataset, using the state-of-the-art NLP models, following recurrent based and transformer approaches. 4) Such baselines are released to the public domain to help advance the state-of-the-art on the released dataset.

The rest of the paper is organized as follows: first, reviewing and discussing the literature in NLP, focusing on QA, Chatbots, and Text Classification; then discussing the possible application and research areas of AI application for Islamic Fatwa; next, presenting the dataset, with statistics and analysis of the topics and distributions of fatwas, along with seasonality analysis. Finally, concluding the baseline models results applied to the dataset, for QA and Topic classification, along with the suggested future research directions in the area and main conclusions.

Literature Review

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

Open-Domain vs. Closed-Domain Chatbots. Open-domain is often called “Chit-chat bots,” and are more conversational bots, which aim to have a flow of dialog that is generic. On the other hand, Closed-domain is Task-Oriented bots specialized in serving certain application domains and customer service. Task-Oriented is more practical and makes it easier to achieve practical and satisfactory performance.

Contextualized vs. Context-Free Chatbots. Dialog-based systems often require context to extend the dialog flow. Based on the context, the next answer can be given. On the other hand, Context-free bots provide an immediate answer to the question, and the flow restarts again. Contextualized bots are common in customer service in the IT domain, where a problem debugging tree exists beforehand, and the bot must go through the different possible root causes with the customer.

Retrieval-Based vs. Generative. This taxonomy is more concerned with the way the underlying QA system is developed. In Retrieval-based systems, the question text is matched to all the questions in the database, using a certain similarity match, like simple dot-product, Mahalanobis-distance, or cosine-similarity. On the other hand, Generative systems 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) (Bahdanau et al., 2014; Luong et al., 2015) or transformer-based (Vaswani et al., 2017).

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. Hamoud and Atwell (2016) built a retrieval-based system using keywords matching with the NLTK text processing tool. The dataset used is focused on the questions related to the holy Quran. While the 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. The researchers build a more generic retrieval-based QA system using word embeddings matching instead of keyword matching in the researchers' work. It helps encode the semantics of the question rather than exact keywords matching. Also, the work relies 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, Sihotang et al. (2020) use Fuzzy string matching to extract questions similarity scores based on Quran and Hadith sources. In Abdi et al. (2020), a QA system is built from the 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 require an expensive language resource like Arabic WordNet (AWN). The use of the graph-based method raises a question about the scalability of the system to the natural language used on social media and differences in accents. On the contrary, the system is free language resources and is scalable via retraining on new data.

State-of-the-Art in Text Classification. Deep Learning (DL) models are known for their hunger for data, which is usually a bottleneck for getting good results. One of the most effective techniques to overcome such limitations is Transfer Learning (TL), which enables learning from small data sets. At the same time, the learned representations can be reused among different tasks. A shared representation among different tasks gives rise to a new area called Multi-Task Learning (MTL). The shared representation can improve the performance over the different tasks and reduce the inference time needed by sharing common parameters.

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-labeled training examples. However, modern deep learning-based NLP models benefit 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). The pre-trained model can then be fine-tuned on small-data NLP tasks like question answering and sentiment analysis, resulting in substantial accuracy improvements compared to training on these datasets from scratch.

In the field of computer vision, researchers have repeatedly shown the value of transfer learning—pre-training a neural network model on a known task, for instance, ImageNet, and then performing fine-tuning—using the trained neural network as the basis of a new purpose-specific model. In recent years, researchers have been showing that a similar technique can be useful in many natural language tasks.

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 (Athiwaratkun et al., 2018) and ELMo (Peters et al., 2018).

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 (Devlin et al., 2018) and ELMo (Peters et al., 2018), instead generate a representation of each word based on the other words in the sentence. For example, in the sentence "I accessed the bank account," a unidirectional contextual model would represent "bank" based on "I accessed the" but not "account." However, BERT represents "bank" using both its previous and next context — "I accessed the ... account" — starting from the very bottom of a deep neural network, making it deeply bidirectional. ELMo learns contextual representations; the representation for each word depends on the entire context in which it is used. Moreover, it works at the character level, which reduces the Out-Of-Vocabulary (OOV).

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 (Bahdanau et al., 2014), which are based on BiLSTM models, and incorporate attention mechanisms, and the transformer (Vaswani et al., 2017), which is fully based on attention gates, without any recurrent layers. Moreover, the learned representations in that encoder can be transferred to other tasks, like in ULMFiT (Howard & Ruder, 2018), where a model is trained on a 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 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.

A similar approach is used in Open AI GPT (Brown et al., 2020), which is a combination of two existing ideas: transformers and unsupervised pre-training. Open AI GPT works in two stages; the first train a transformer model on a very large amount of data in an unsupervised manner—using language modeling as a training signal—then this model is fine-tuned on much smaller supervised datasets to help it solve specific tasks.

Transfer Learning in Arabic. The Arabic language is considered among the Low-NLP Resources languages, unlike English. It calls for the need of both TL and MTL to help solve this issue. Today there is a wide gap in the literature in applying the above techniques to Arabic NLP tasks. Transfer learning of Word Embeddings was used in AROMA. Al-Sallab et al. (2017) used learned embeddings from the QALB dataset to perform the 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 have been released, like AraBERT and AraGPT (Antoun et al., 2020; Djandji et al., 2020; Nada et al., 2020).

METHODOLOGY

Supervised Learning Problem Formulation

Sequence-to-Sequence Models for Question Answering (QA)/Chatbot. Based on the collected data, an automated QA system can be built and integrated into a Chatbot to answer Fatwa questions. Chatbots are either open-domain or task-specific. Open-domain systems are more designed for chatting purposes rather than the authentic answer. For that, a custom QA system needs to be built from scratch. Recent advances in Transfer Learning in NLP can be utilized to build a sequence-to-sequence model for a QA system. It can be further improved and enhanced by integration to a post-authentication step to validate the answer quality.

Classification of Fatwa Topic. It can be done in different ways: Classification of Fatwas areas (Pilgrim/Umrah, Financial, Prayer, and Fasting) and Classification of Fatwa as Authentic/Non-authentic. Pre-trained models like BERT (Devlin et al., 2018) can be used for English, and AraBERT or AraGPT (Antoun et al., 2020) for Arabic, where specialized language models can be built for Fatwa.

Unsupervised Learning Problem Formulation

Fatwa text can be modeled based on the topics using topic modeling techniques such as Latent Dirichlet Analysis (LDA). It can be used to annotate the collected dataset as well. Also, Active Learning can be utilized for data annotation. Active learning is concerned with semi-supervised labeling of samples through an iterative selection of which samples to present to the annotator, based on the learner performance in the previous iteration. It can be used for custom dataset building. Ready platforms like AWS GroundTruth can be used.

Dataset Building

The work relies on data collection via web scrapping of public websites. This approach offers free annotation for the question-answer pairs, where both already exist in a structured way on the web page. Also, the topic class of the question is sometimes present. The website's sources are listed in Table 1. Such sources can be classified into trusted sources: These are government sources, which are reviewed and authentic. It further serves as a source of authenticity reference, for example, https://www.dar-alifta.org/ar/Default.aspx?sec=fatwa&1&Home=1, https://www.alifta.gov.sa, https://aliftaa.jo/ and untrusted sources which are abundant. However, authenticity is not guaranteed. Such sources can provide the needed challenge to solve in the Authenticity Classification area. Examples: Ask and ArabicASK (Essam, 2017), islamweb (https://www.islamweb.net/ar/), islamway (https://ar.islamway.net/fatawa/source/), binbaz (https://binbaz.org.sa/fatwas/kind/1), binothaimeen (https://binothaimeen.net/site), and others.

System Architecture

The overall architecture is shown in Figure 1, following the described phases and use case sequence diagram. First, the data as [Question, Topic (intent), Answer] is collected from different data sources. This data will train the topic classifier and the QA engines (described as Topic Expert). The Question router takes care of routing the question, either to the human expert in the early/manual mode or to the QA Expert engine according to the classified topic by the topic classifier. In the following sections, more details are provided for each component.

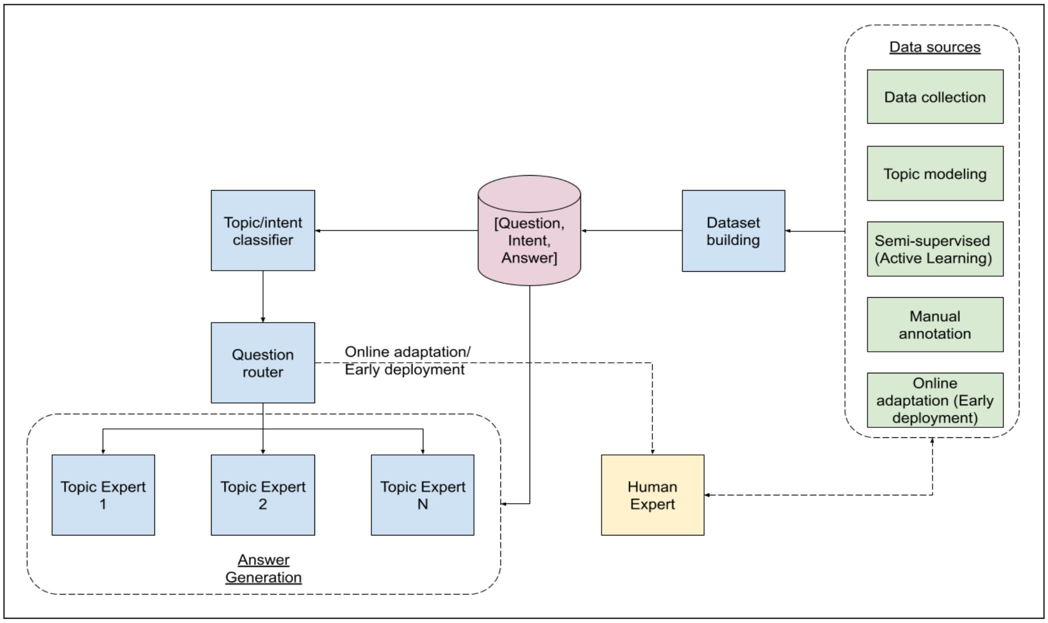


Figure 1. Overall architecture

Online Adaptation. During operation, especially in early deployment, a human expert should be involved as part of the assessor of the results produced to avoid misleading/ vague interpretation of the answer.

Topic Classification. This component is concerned with classifying the user intent/ topic to narrow down the answer space. The plan is to develop this module based on two approaches and evaluate them: Recurrent based (LSTM/GRU) and Transformer based (BERT, AraBERT, AraGPT2).

The data used to train the classifier will depend on the available data source. It will enable the development of different versions according to the phase of available data, which will be detailed in the project plan.

Bag-of-Word Models

All bag of words models had the same model architecture and had a vocabulary size of 2000. The architecture consisted of 2 Dense layers with 1000 nodes followed by 512 node layers and then the classification layer. Dropouts were applied after every Dense Layer to reduce overfitting. The bag of words models used the following text features: binary, count, frequency, and TFIDF.

Moreover, the BoW vectors are also evaluated, where an Embedding layer for each input word is used. It requires padding the input sentence to a maximum of 250 words. The vocabulary size is 2000 words, and the embedding shape is 300. Embeddings layers for the BoW vectors model are trained from scratch, i.e., initialized with random weights sampled from a normal distribution.

Recurrent Based Models

Each word is first passed on an Embedding block, providing a vector per word. The separate word vectors can be aggregated using a recurrent neural network (RNN), which acts as a state machine that sequentially processes the token vector, resulting in a hidden state that is updated with each token in the sequence. The final hidden state can then be used to represent the whole sequence. A single layer RNN is parametrized by input weights U, hidden transition weights V and output weights O, as shown in Figure 2. It is possible to stack more such layers straightforwardly using the representations of the hidden states. LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Chung et al., 2014) recurrent models were used.

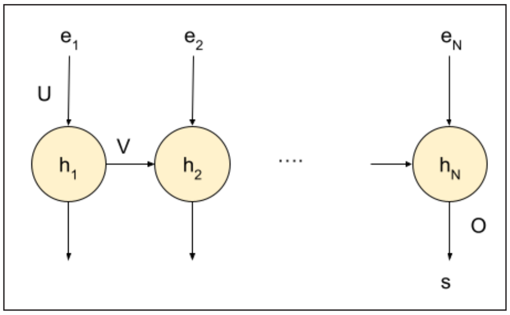


Figure 2. Single-layer recurrent neural network

This model was implemented in three different flavors, using an embedding layer to embed the words from scratch—two pre-trained Arabic word embedding models, AraVec and Fasttext, with 300 embedding dimensions. The single-layer LSTM is used, and then layers are added. The first is another LSTM layer, and the second is a dense layer that was put between the LSTM and the output layer, resulting in 100 dimensions sentence embedding. The GRU model is the same network as the LSTM network except that the LSTM was replaced with a GRU layer, with 1 and 2 layers. It also had 100 dimensions.

Transformer Based Models

In this section, attention-based sentence embedding models are used. It started with BERT (Devlin et al., 2018). BERT is based on the full attention mechanism introduced in Vaswani et al. (2017) for sequence-to-sequence models, which is the encoder-decoder architecture. The encoder is based on a full self-attention mechanism, resulting in an embedding vector for each input token of the input sentence. However, only one vector for the whole sentence is needed for sentence representation. BERT introduces a new CLS token at the input that has learnable embeddings to work this out. The output token vectors can then be ignored, except the first one, which corresponds to the CLS token and holds a representation for the whole sentence, as shown in Figure 3. The whole encoder is pre-trained on different upstream tasks, like Next Sentence Prediction (NSP) and Masked Language Modeling (MLM), and can then be fine-tuned to any downstream task, like sentence classification.

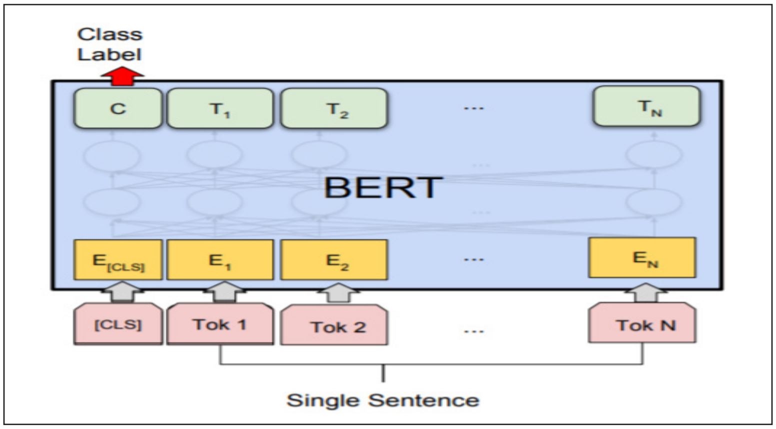


Figure 3. BERT for single sentence classification (Devlin et al., 2018)

Topic Router. This component routes the question to the expert system based on the classified topic/intent. According to the mode of operation, the question might be routed to a human expert in the manual/early mode.

Answer Generation (Topic Experts). Based on the intent of the user, a specialized QA system (Topic Expert) is triggered. Retrieval-based answer generation: in this approach, the question is matched to the filtered subset of historic questions related to the topic. The system is shown in Figure 4.

The Question similarity matching shall be done based on some similarity metrics (L1, L2, Euclidean distance, Mahalanobis distance, and Cosine similarity). It can also be a learnable objective system, as shown in Figure 4. The dataset to train such a model can be formed by clustering the questions per topic and embedding the questions data using a Neural Network model. The learning signal is simply if the topics of the two questions match or not. The whole model can then be end-to-end based on similarity loss objectives.

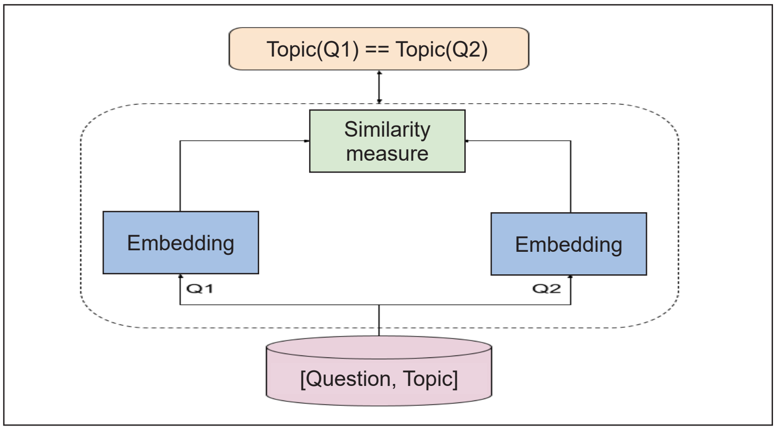


Figure 4. Similarity-based Question matching model

RESULTS AND DISCUSSIONS

Dataset

The details of the dataset collection are shown in Table **1**, of around 850K Fatwas. As discussed before, the popular websites of Islamic Fatwa are crawled, being official, like Al-Ifta-SA, Dar-al-ifta-EG, and Al-ifta-JO, or non-official like islamway and islamweb. Those websites span different countries and geographical locations, accents, and backgrounds. The work crawls for Question/Answer, Topic, and Date. The topics and dates are not applicable or present for some websites. For Arabic AskFM, the work extends the one in (Essam, 2017) to include 604K fatwas by crawling the full website. A special type of QA is found in islamonline, where the article titles are treated as Questions, and the bodies as answers, since they form the basic and frequently asked questions in Islamic Fatwa.

Table 1
Dataset information, statistics, and sources

Dataset	Question/Answers	Topics	Dates
Al-ifta-SA (https://www.alifta.gov.sa) and Dar-al-ifta-EG (https://www.dar-alifta.org/ar/Default.aspx?sec=fatwa&1&Home=1)	3,450	Yes	Yes
AskFM (Essam, 2017)	604,184	N/A	N/A
Islamweb (https://www.islamweb.net/ar/)	126,000	Yes	Yes
Islamway (https://ar.islamway.net/fatawa/source/)	15,060	N/A	Yes
Islamonline (Islamonline, n.d.)	3,100	Yes	N/A
binbaz (https://binbaz.org.sa/fatwas/kind/1)	28,226	Yes	N/A
binothaimeen (https://binothaimeen.net/site)	2,157	Yes	N/A
AlFawzan (https://www.alfawzan.af.org.sa)	2,000	N/A	Yes
Islamqa (https://islamqa.info/)	30,780	Yes	Yes
Fatwapedia (http://fatawapedia.com/)	34,661	Yes	N/A

Topics Analysis

Traditional jurisprudence distinguishes two principal branches of law, ibadat (rituals) and muamalat (social relations); each can be further subdivided into more subtopics (Figure 5). Another plane of categorization is by the action mandated, which falls in one of five categories: mandatory, recommended, neutral, abhorred, and prohibited.

Overall, word cloud for all topics is shown in Figure 6. It is better to reduce the scope to the top-k (k=5) topics (Figure 7) and compare them to their corresponding word clouds (Figure 8).

Text Cleaning

The following pipeline was applied:

- Special and non-Arabic Characters Removal.
- Arabic Diacritics Removal.
- Punctuations Removal.
- Numbers Removal.
- Stop Words Removal (using NLTK Arabic set).
- Stemming using ISRIStemmer for Arabic.

Unsupervised Topic Modeling

The work use LDA as a first topic modeling attempt. LDA topic modeling works on the words co-occurrence matrix. It tries to find the latent factors that cluster the sentences together. It uses Bag-of-Words representation (BoW). LDA topics for k=4 topics are shown in Figure 9, and k=5 in Figure 10.

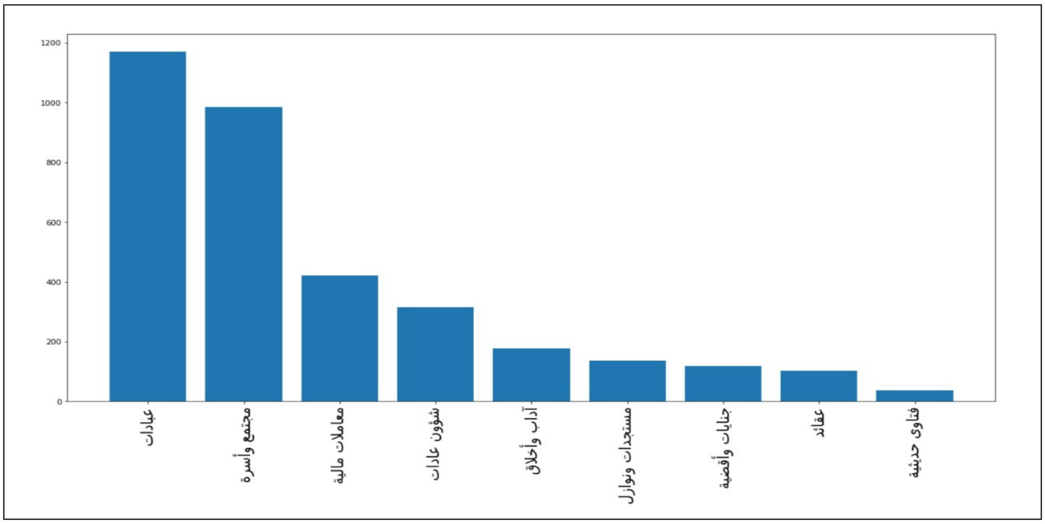


Figure 5. Distribution of fatawa per topic for dar-al-ifta dataset

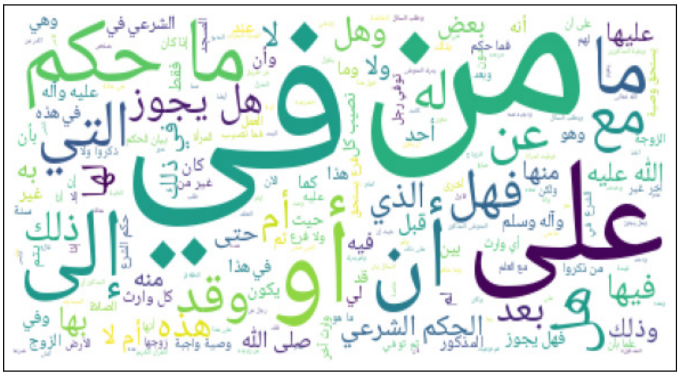
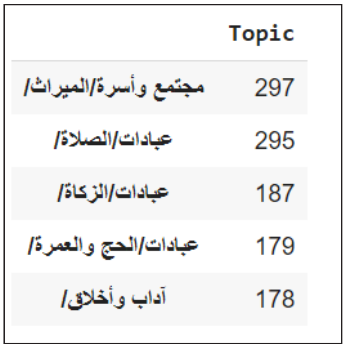


Figure 6. Word cloud for dar-al-ifta dataset



	Topic
مجتمع وأسرة/الميراث/	297
عبادات/الصلاة/	295
عبادات/الزكاة/	187
عبادات/الحج والعمرة/	179
آداب وأخلاق/	178

Figure 7. Top 5 Topics/Subtopics for dar-al-ifta dataset

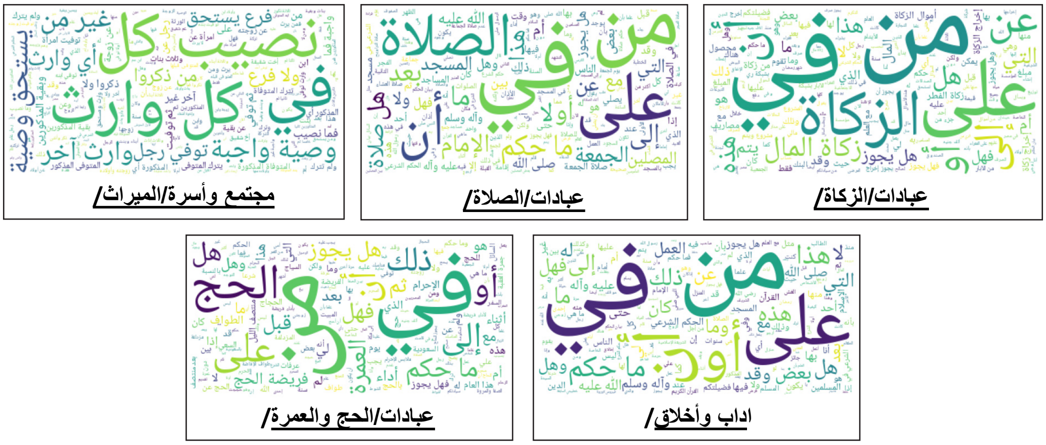


Figure 8. Per subtopic word cloud for the top 5 topic/subtopic categories in dar-al-ifta dataset

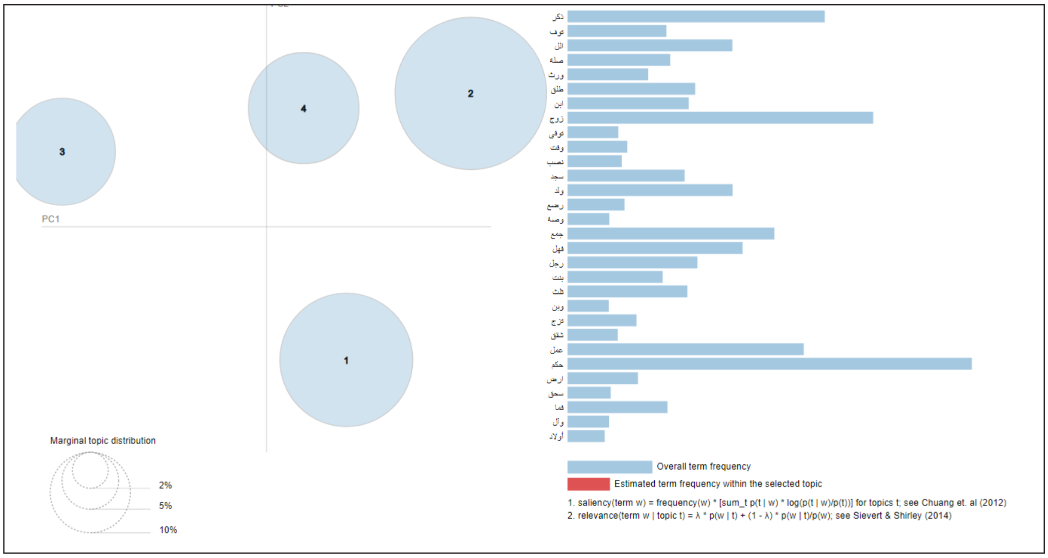


Figure 9. Identified topics for k=4 for all the topics in dar-al-ifta dataset

Stemming is used. So الزكاة becomes زكة. الصلاة becomes صلة. Comparing the identified topics to the word clouds:

- 2 seem about الزكاة+الصلاة
- 3 seem about الميراث
- 1 seems about 'عبادات/الحج والعمرة', '/آداب وأخلاق/'

Comparing the word clouds per category to the modeled keywords (hover over the circles):

- 2 seem about الزكاة
- 4 and 5 seem about الميراث
- 3 seems about الصلاة

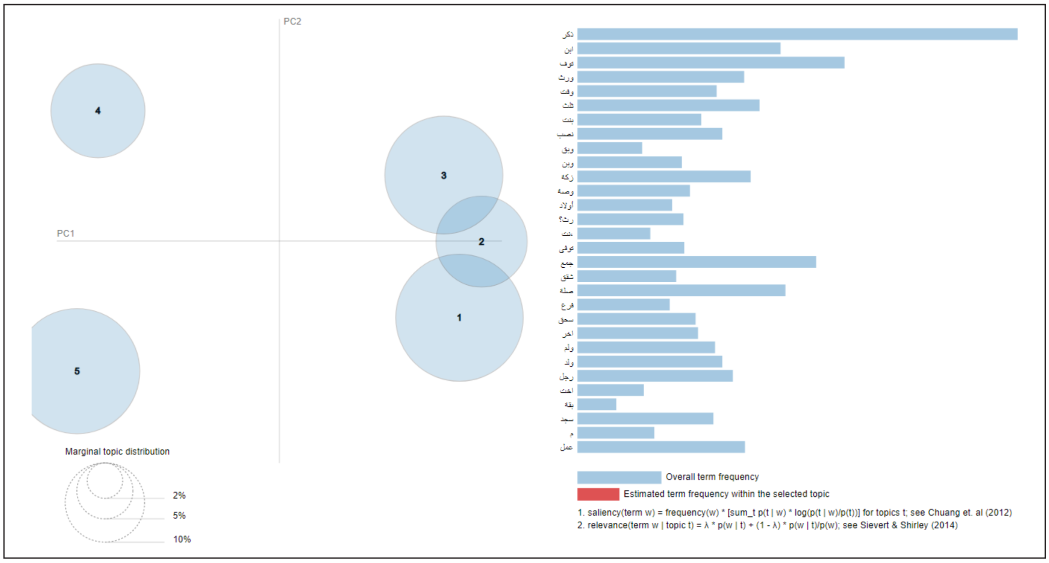


Figure 10. Identified topics for k=5 for the top 5 topics/subtopics in dar-al-ifta dataset

The lowest categories of the top 5 seem not modeled by LDA: 'عبادات/الحج والعمرة', '/آداب وأخلاق/'

In terms of topic modeling, stemming seems to perform the same as no stemming: general modeling does not perform well, while focused on top k topics give good results for the top three, while the next two are not modeled.

The work also evaluates a supervised BoW model with TF-IDF features to benchmark the unsupervised topic model. Promising 85% test accuracy shows a good correlation signal for the top 5 topics/subtopics. In terms of the BoW model, stemming performs much better than no stemming. TF-IDF also seems the best text feature to use with this simple BoW model.

Seasonality Analysis

The Hijri month trend for syam questions is shown in Figure 11. The x-axis is Hijri month. The y-axis is the number of questions. The graph clearly shows that the number of fasting-related questions increases a lot around Ramadan, month 9. Moreover, this is quite logical. Another logical trend is for "hajj," where the peak is around 10-11-12 (hajj months peaking at ذي الحجة 12th month), and down around months 1-2 (محرم و صفر). Also, the combined trends show a high volume of fasting questions season.

Baselines

Topics Classification. For the classifier models, the work evaluates three families: the Bag-of-Words (BoW) and Sequence models: recurrent-based or transformer-based models. The

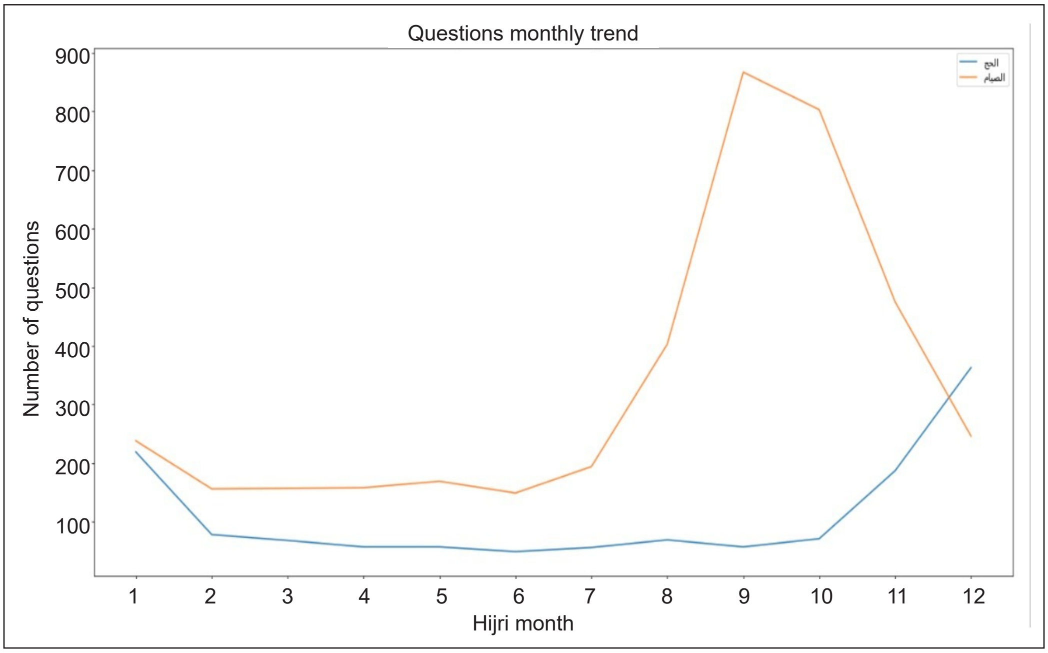


Figure 11. Combined Seasonality trends of Hajj and Fasting

reason behind this evaluation is two folds: 1) A public dataset provide a comprehensive baseline that includes the state-of-the-art classification models and 2) To compare the performance of models that are capable of modeling context, like recurrent and transformer methods, to other models that exist in the literature, like the popular BoW. The aim is to show that the complexity of the dataset calls for the need for context-aware models.

All bag of words models had the same model architecture and had a vocabulary size of 2000. The architecture consisted of 2 Dense layers with 1000 nodes followed by 512 node layers and then the classification layer. Dropouts were applied after every Dense Layer to reduce overfitting. The bag of words models used the following text features: binary, count, frequency, and TFIDF.

Moreover, the work also evaluates BoW vectors, where an Embedding layer is used for each input word. It requires padding the input sentence to a maximum of 250 words. The vocabulary size is 2000 words, and the embedding shape is 300. Embeddings layers for the BoW vectors model are trained from scratch, i.e., initialized with random weights sampled from a normal distribution. For Recurrent Networks Topics Classifier, both LSTM and GRU models are evaluated, with 1 and 2 layers. For Transformer Networks Topics Classifier, a hugging face classification library was used to load the AraBERT-base model and train on the dataset. AraBERT's pre-process function is also used to clean the text and put it in the structure that AraBERT's tokenizer can read. The precision, recall, and F_1 measures for each classifier in Table 2 are also reported for completeness.

Table 2
Topics classifiers baseline results

Model	Accuracy (%)	Precision (%)	Recall (%)	F_1 (%)
BoW-Binary	53.3	47	41	42
BoW-Count	53.4	48	39	41
BoW-Frequency	51	39	30	30
BoW-TF-IDF	**53.5**	**44**	**42**	**43**
BoW Vectors	47	36	34	34
1-Layer LSTM	54	44	38	39
1-Layer GRU	**55**	**45**	**40**	**42**
2-Layer LSTM	53	46	36	38
2-Layer GRU	54	45	38	43
AraBERT	**70**	**59**	**57**	**56**
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
BoW-Binary Features	53.3			
BoW-Count	53.4			
BoW-Frequency	51			
BoW-TF-IDF	**53.5**			
BoW Vectors	47			
1-Layer LSTM	52			
1-Layer GRU	53			
2-Layer LSTM	50			
2-Layer GRU	**56**			
AraBERT	**70**			

Effect of Sentence Embedding

Bag-of-words Models. On the other hand, BoW lacks the advantage of context; hence sequence models, like recurrent and transformers models, outperform them, by 2–8%, excluding the effect of pre-trained word vectors. However, for text classification, keywords of vocabulary could be more critical. Hence, there is no huge gap in the performance.

Recurrent Models. For recurrent-based models, GRU layers outperform LSTM layers by 4–7%. Adding extra layers for both options does not seem to have an advantage.

Transformer-Based Models. Transformer-based models were introduced as a replacement to recurrent models. In BERT, the idea of pre-trained language models was coupled with the full attention bi-directional transformers. The dependence on a pre-trained language model raises the question of the efficiency of off-the-shelf models for multi-lingual text classification. Thus, language-specific, pre-trained models were introduced in AraBERT and AraGPT-2. The results show a clear advantage of Arabic-specific language models, with a 20–26% advantage over the generic model, fine-tuned on the task. AraBERT is outperforming AraGPT-2 by 6%.

Effect of Transfer Learning Pre-Trained Embeddings

Transfer learning in NLP can be viewed at the word level, with pre-trained word Embeddings or language models, like BERT and GPT. Thus, a higher 9–16% advantage over the BoW models was seen when pre-trained embeddings with recurrent models or pre-trained language models with transformers. The effect of pre-trained language models in AraBERT is more dominant, giving the top score of 70%.

Effect of Text Features

While the differences are small, we can see an advantage for the frequency and TF-IDF features, probably because of the normalization effect they introduce to the features. Following is the binary features model, which also keeps 0/1 hard-normalized features. Finally, the least performer is the count model, which does not perform any normalization, giving a false advantage to the frequent words. The counting model is not far from the other since rigorous text cleaning removes irrelevant words. The BoW vectors model is the least performer because it starts from randomly initialized Embeddings.

Retrieval-based Question-Answering System. The work evaluates the retrieval-based QA system as described in Figure 4. Fasttext is used as the Embedding layer for encoding the questions and the input. Then cosine similarity is compared and get the top k similar questions. Fasttext works by treating each word as a bag of character ngrams (from 3 to 6 in practice). Each word vector is represented by summing the vectors of its character ngrams plus a specific word vector for the word itself. The sentence is represented by normalizing each word vector in the question with its L2 norm and then averaging them. Sample questions and generated answers are shown in Table 4. The QA system is evaluated based on the retrieval accuracy, comparing the retrieved answer index against the true index. In order to test the similarity matching, the random word is dropped from the query question before matching it against the database, which resulted in 96.4% retrieval accuracy. However, while this is highly accurate, it reflects the retrieval performance from the known set of questions. In the true scenario, unseen questions might not match the database questions, which requires a generative QA system that generates the answer based on the summarized question state, not based on retrieval matching. This work leaves this for future works to tackle.

In addition, precision, recall, and F_1 measures per topic are reported in addition to their macro and weighted average in Table 3.

Table 3
Retrieval QA system precision, recall, and F_1

Topic	Precision (%)	Recall (%)	F_1 (%)
العقيدة الإسلامية	28.41	30.69	29.51
الفضائل والتراجم	10.89	10.49	10.69
القرآن الكريم	46.90	35.81	40.61
اللباس والزينة	27.59	25.40	26.45
طب وإعلام وقضايا معاصرة	25.64	31.63	28.32
فقه الأسرة المسلمة	53.14	68.03	59.67
فقه الجنايات	19.51	15.38	17.20
فقه العبادات	77.49	72.07	74.68
فقه المعاملات	64.09	66.71	65.37
فقه المواريث	73.04	77.93	75.40
فكر وسياسة وفن	13.68	10.92	12.14
الأذكار والأدعية	16.33	12.50	14.16
Macro Average	29.11	27.92	28.25
Weighted Average	53.51	53.47	53.21

Table 4
Sample questions and retrieved answers

Query	Answer / Fatwa
س: بسبب العلاج ومواعيد مش هفطر فى رمضان غير أقل من 3 ساعات وباقى اليوم صايمه هو انا كدة ينفع !مصوموش؟	ج: إذا كان يتعبك الصوم فلك رخصة في الفطر في هذه الحالة
س: السلام عليكم ورحمة الله وبركاته، شيخنا الغالي ،بداية أحبك في الله الحق، وأسأل الله أن يسددك ويؤيدك بروح من عنده، وأن يبارك في علمك ووقتك وقلمك ، وأن يجعله مرشدا للمهتدين وصارما على المبدلين، سؤالي الأول ، أين بقية الحديث عن الخلافة أحكامها وأيامها؟	ج: وعليكم السلام ورحمة الله وبركاته وأحبك الله، وتقبل دعاءك الصالح، آمين وإياكم أجمعين كتاب "الخلافة" حال دون إكمال مسودات فصوله النهائية كتاب "ابن تيمية ومعركة الحرية" وبعض الكتب والفتاوى والدراسات التي تطرأ وتحتاج إلى بيان، مما يضطرني للتوقف، وهناك كتب عديدة لم أستطع الفراغ منها لهذا السبب، يسر الله إخراجها في أقرب وقت
س: طب ممكن تقترح اسم كتب عن سيرة النبي صلى الله عليه وسلم	ج: كتاب مختصر مع الدروس المستفادة:السيرة النبوية دروس وعبر لمصطفى السباعي..كتاب متوسط الحجم:الرحيق المختوم:السيرة النبوية .عرض وقائع وتحليل أحداث، لعلي الصلابي

CONCLUSION

This paper presents the first steps towards building an Automated Fatwa System using AI and Deep Learning NLP methods. The State-of-Art methods are surveyed and provide use-case scenario design for a system that performs topic/intent classification and Question-Answer retrieval. It leads to the discussion of dataset collection, where the largest dataset of Islamic Fatwas is presented. For this dataset, unsupervised topic modeling, seasonality analysis, and baseline models for topic classification and QA are performed. The baselines are evaluated in various aspects like the effect of sequence modeling, the effect of pre-trained embeddings, and language models. Also, the baselines for the widely used models in NLP in literature are provided. Finally, all the models are released, benchmarks, and data to the public domain to help advance the research in the area.

ACKNOWLEDGEMENT

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

REFERENCES

Abdi, A., Hasan, S., Arshi, M., Shamsuddin, S. M., & Idris, N. (2020). A question answering system in hadith using linguistic knowledge. *Computer Speech & Language, 60*, Article 101023. https://doi.org/10.1016/j.csl.2019.101023

Al Sallab, A. A., Baly, R., Badaro, G., Hajj, H. M., El-Hajj, W., & Shaban, K. (2015a, March 9-10). Towards deep learning models for sentiment analysis in Arabic. In *Machine Learning and Data Analytics Symposium - MLDAS 2015* (pp. 1-5). Doha, Qatar. https://doi.org/10.18653/v1/W15-3202

Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El-Hajj, W., & Shaban, K. (2015b, July 26-31). Deep learning models for sentiment analysis in Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing* (pp. 9-17). Beijing, China. https://doi.org/10.18653/v1/W15-3202

Al Sallab, A., Rashwan, M., Raafat, H., & Rafea, A. (2014, October 25). Automatic Arabic diacritics restoration based on deep nets. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 65-72). Doha, Qatar. https://doi.org/10.3115/v1/W14-3608

Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., & Badaro, G. (2017). Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16*(4), 1-20. https://doi.org/10.1145/3086575

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

Athiwaratkun, B., Wilson, A. G., & Anandkumar, A. (2018). *Probabilistic fasttext for multi-sense word embeddings*. arXiv Preprint. https://doi.org/10.18653/v1/P18-1001

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. arXiv Preprint.

Baly, R., Badaro, G., Hamdi, A., Moukalled, R., Aoun, R., El-Khoury, G., El-Sallab, A., Hajj, H., Habash, N., Shaban, K. B., & El-Hajj, W. (2017). Omam at semeval-2017 task 4: Evaluation of English state-of-the-art sentiment analysis models for Arabic and a new topic-based model. In *Proceedings of the 11th*

International Workshop on Semantic Evaluation (SEMEVAL-2017) (pp. 603-610). ACM Publishing. https://doi.org/10.18653/v1/S17-2099

Baly, R., Hobeica, R., Hajj, H., El-Hajj, W., Shaban, K. B., & Al-Sallab, A. (2016). A meta-framework for modeling the human reading process in sentiment analysis. *ACM Transactions on Information Systems (TOIS), 35*(1), 1-21. https://doi.org/10.1145/2950050

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … & Amodei, D. (2020). *Language models are few-shot learners*. arXiv Preprint.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv Preprint.

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

Djandji, M., Baly, F., Antoun, W., & Hajj, H. (2020). 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* (pp. 97-101). European Language Resource Association.

Essam, O. (2017). *Arabic AskFM dataset*. Data Science.

Hamoud, B., & Atwell, E. (2016). Quran question and answer corpus for data mining with WEKA. In *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)* (pp. 211-216). IEEE Publishing. https://doi.org/10.1109/SGCAC.2016.7458032

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification*. arXiv Preprint. https://doi.org/10.18653/v1/P18-1031

Luong, M. T., Pham, H., & Manning, C. D. (2015). *Effective approaches to attention-based neural machine translation*. arXiv Preprint. https://doi.org/10.18653/v1/D15-1166

Magooda, A., Sayed, A. M., Mahgoub, A. Y., Ahmed, H., Rashwan, M., Raafat, H., Kamal, E., & & Al Sallab, A. A. (2016). RDI_Team at SemEval-2016 Task 3: RDI unsupervised framework for text ranking. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 822-827). ACM Publishing. https://doi.org/10.18653/v1/S16-1127

Nada, A. M. A., Alajrami, E., Al-Saqqa, A. A., & Abu-Naser, S. S. (2020). Arabic text summarization using AraBERT model using extractive text summarization approach. *International Journal of Academis Information Systems Research, 4*(8), 6-9.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. arXiv Preprint. https://doi.org/10.18653/v1/N18-1202

Rashwan, M. A., Al Sallab, A. A., Raafat, H. M., & Rafea, A. (2015). Deep learning framework with confused sub-set resolution architecture for automatic Arabic diacritization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(3), 505-516. https://doi.org/10.1109/TASLP.2015.2395255

Sihotang, M. T., Jaya, I., Hizriadi, A., & Hardi, S. M. (2020). Answering Islamic questions with a chatbot using fuzzy string-matching algorithm. In *Journal of Physics: Conference Series* (Vol. 1566, No. 1, p. 012007). IOP Publishing. https://doi.org/10.1088/1742-6596/1566/1/012007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 5998-6008). Long Beach, USA.