Towards an Automated Islamic Fatwa System: Survey, Dataset and Benchmarks

**Amr A. Munshi¹; Wesam H. AlSabban²; Abdullah Tarek Farag³;
Omar Essam Rakha⁴; Ahmad A. Al Sallab⁵; Majid Alotaibi⁶**

¹Department of Computer Engineering, Umm Al-Qura University, Saudi Arabia

²Department of Information Systems, Umm Al-Qura University, Saudi Arabia

³Speakol

⁴Faculty of Engineering, Ain Shams University

⁵Faculty of Engineering, Cairo University

⁶Department of Computer Engineering, Umm Al-Qura University, Saudi Arabia

¹aaamunshi@uqu.edu.sa; ²whsabban@uqu.edu.sa; ³abdullahtarek57@gmail.com; ⁴i@omarito.me;

⁵ahmad.elsallab@gmail.com; ⁶mmgethami@uqu.edu.sa

DOI: 10.47760/ijcsmc.2021.v10i04.017

Abstract— Islam is the second largest and the fastest growing religion. The Islamic Law, Sharia, represents a profound component of the day-to-day lives of Muslims. This creates a lot of queries, about specific problems, that requires answers, or Fatwas. While sources of Sharia are available for anyone, it often requires a highly qualified person, the Mufti, to provide Fatwa. To get certified for Fatwa, the Mufti needs to undergo a sophisticated and long education process that starts from basic to high school. With Islam followers representing almost 25% of planet earth population, generating a lot of queries, and the sophistication of the Mufti qualification process, creating shortage in them, we have a supply-demand problem, calling for Automation solutions. This motivates the application of Artificial Intelligence (AI) to Automated Islamic Fatwa. In this work, we explore 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. We start by surveying the State-of-The Art (SoTA) of NLP, and explore the potential use-cases to solve the problems of Question answering and Text Classification in the Islamic Fatwa Automation. We present the first and major enabler component for AI application for Islamic Fatwa, the data. We build the largest dataset for Islamic Fatwa, spanning the widely used websites for Fatwa. Moreover, we present baseline systems, for Topic Classification, Topic Modelling and Retrieval-based Question-Answering, to set the direction for future research and benchmarking on our dataset. Finally, we release our dataset and baselines to the public domain, to help advance the future research in the area.

Keywords: Islamic Fatwa; Question Answering; Text Classification; Artificial Intelligence; Machine Learning; Deep Learning; Natural Language Processing.

I. 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 population in 51 countries [64]. Also, Islam is the fastest growing religion, at growth rate 1.84% (2010-2015) per year. Islam is characterized by a comprehensive set of immutable diving rules, representing the Islamic Law, or Sharia, which governs all aspects of Muslims lives. The Islamic jurisprudence or Fiqh represents the human interpretation of Sharia. 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, of which 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). Qualification of 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, who is assisted by advanced students, which results in a certification or ijaza. This tradition, even extended to other branches of education, like medicine, law, mathematics...etc. In the modern era, specialized 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 in Saudi Arabia.

With the explosion of Social Media, and public websites, new channels of fatwas have emerged. 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 the process of getting an answer for Muslims, however, it opens the door for many controversy, unauthentic fatwas.

In this work, we aim to unleash the potential of AI to deliver immediate Fatwa, an answer to a question about an Islamic Religion rule. Usually, those answers must be given by a highly qualified experts, who had years of specialized education and verified degrees from the highest religious institutes. This creates a high-demand low-supply situation. This situation is more demanding in the high seasons, like Pilgrim, Umrah and Ramadan. Which calls for the need 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, sex...etc) 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 [35][63]. That potential is not fully unleashed in Low-Language Resources (LLR) like Arabic. Some attempts have been made such as [10][15][16][24][46][53][7]. 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 at least could be clustered and routed to the proper expert ahead of time.

We start surveying the SoTA in NLP, focused on QA systems, Chatbots and Text Classification. This leads to the discussion of the potential application areas of AI to Automated Fatwa systems, and the different use-cases scenarios. We focus our contribution on this paper to the main building block that enables other applications, which is data. We collect and build 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 also the topic and date of the fatwa when applicable. This helps us perform Exploratory Data Analysis (EDA), like Unsupervised Topic Modelling and Seasonality Analysis. To set baselines for the future research on our dataset, we build baseline models for Topic Classification, and Retrieval-based system using Word Embeddings and Text Similarity matching. We release all our models and dataset to the public domain to help advance the research in the area.

The rest of the paper is organized as follows: first we review and discuss the literature in NLP, focused on QA, Chatbots and Text Classification. Then we discuss the possible application and research areas of AI application for Islamic Fatwa. Following, we present our dataset, with statistics and analysis of the topics and distributions of fatwas, along with seasonality analysis. Finally, we conclude with the baseline models results applied to the dataset, for QA and Topic classification, along with the suggested future directions of research in the area and main conclusions.

II. LITERATURE REVIEW

A. Chatbots 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 often called “Chit-chat bots”, and are more of conversational bots, which aims to have a flow of dialog that is generic. On the other hand, Closed-domain are Task-Oriented bots are specialized to serve certain application domain, customer service. Task-Oriented are more practical, and easier to achieve practical and satisfactory performance.

Contextualized vs Context-free Chatbots: Dialog based systems, often requires context to extend the dialog flow. Based on the context, the next answer can be given. On the other hand, Context-free bots provides 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) [14][45] transformer based [63]

The AI Fatwa system shall be developed as Task-Oriented/Closed-domain system, in a Context-Free fashion. Both Retrieval-based and Generative approaches will be evaluated, along with both Recurrent and Transformer approaches to the latter.

B. Ready-made Platforms

Among the most popular platforms for building a Chatbot are:

- Google’s Dialogflow
- Amazon Alexa
- Facebook’s Wit
- IBM’s Watson Assistant tool.

The benefits of using one of these is to obtain a quick minimum viable product (MVP), without having much previous AI knowledge. However, they need to be fine-tuned and adjusted for specific application domain, language, accent...etc.

C. State-of-The Art for Text Classifiers

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

D. 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). 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 [12], and ELMo [49].

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 [23] and ELMo [49] instead generate a representation of each word that is 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 use. 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 [14], which are based on BiLSTM models, and incorporate attention mechanisms, and the Transformer [63] 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 [35], 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. A similar approach is used in Open AI GPT [19], which is a combination of two existing ideas: transformers and unsupervised pre-training. Open AI GPT works in two stages; first train a transformer model on a very large amount of data in an unsupervised manner — using language modelling as a training signal — then we fine-tune this model on much smaller supervised datasets to help it solve specific tasks.

E. Multi-Task Learning

The field of Natural Language Processing includes dozens of tasks, among them machine translation, named-entity recognition, and entity detection. While the different NLP tasks are often trained and evaluated separately, there exists a potential advantage in combining them into one model, i.e., learning one task might be helpful in learning another task and improve its results. The idea of transfer learning enables the idea of having a common encoder (or backbone according to computer vision convention), that is shared among various tasks in jointly trained model (Multi-task learning). Hierarchical Multi-Task Learning model (HMTL) [56] provides an approach to learn different NLP tasks by training on the “simple” tasks first and using the knowledge to train on more complicated tasks. The model presents state-of-the-art performance in several tasks and an in-depth analysis of the importance of each part of the model, from different aspects of the word embeddings to the order of the tasks.

F. Potential in Arabic NLU

Arabic language is considered among the Low-NLP Resources languages, unlike English. This calls for the need of both TL and MTL to help solving this issue. Looking on the literature today, there is a wide gap in applying the above techniques to Arabic NLP tasks. Transfer learning of Word Embeddings was used in AROMA [7], 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 [1][10][24].

III. RESEARCH AREAS

G. Supervised problem formulations

Sequence-to-sequence models for Question Answering (QA)/Chatbot: Based on the collected data, an automated QA system can be built and integrated in 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 authentic answer. For that, a custom QA system needs to be built from scratch. Recent advances in Transfer Learning in NLP (see [1][10][24]) can be utilized to build a sequence-to-sequence model for a QA system. This can be further improved and enhanced by integration to a post authentication step to validate the answer quality.

Classification of Fatwa topics: This can be done in different ways:

- Classification of Fatwas areas (Pilgrim/Umrah, Financial, Prayer, Fasting...etc)
- Classification of Fatwa as Authentic/Non-authentic

Pre-trained models like BERT [23] can be used for English, and AraBERT or AraGPT [10] for Arabic, where specialized language models can be built for Fatwa.

H. Unsupervised problem formulations

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

I. Dataset building

- **Trusted sources:** These are government sources, which are reviewed and authentic. This further serves as a source of authenticity reference. Examples [2][6][22]
- **Untrusted sources:** Those are abundant. However, authenticity is not guaranteed. Such sources can provide the needed challenge to solve in the Authenticity Classification area. Examples
 - Twitter
 - Ask (ArabicASK) [11]
 - Popular sites like islamweb [38], islamway [37], binbaz [17], binothaimeen [18], and others.

Some sources could be more popular than others. Some might be noisier in terms of non-qualified responses, which poses more challenge. Data can be annotated for various aspects:

- Topic
- Authenticity
- Quality (classes could be discrete as classification, or continuous as regression, which is another research question to be answered).

Custom annotation can be done using Crowd sourcing tools like: Amazon MechanicalTurk, CrowdFlower...etc. Annotation quality is a wide topic that will be addresses by the owner of this area:

- Label by multiple users
- Majority voting or other aggregation technique
- How to ensure the annotators are trained? Question can be designed to ensure that.
- Semi-automatic annotation and Active Learning to enlarge/augment the annotated data

Based on the collected datasets, the problem can be formulated as Data Science problem, and run Exploratory Data Analysis (EDA) to extract some insights like:

- Cluster topics across Time/Seasons
- Cluster topics across Regions
- Cluster topics across Language

IV. DATASET

The details of the dataset collection are shown in Table 1, of around 850K Fatwas. As discussed before, we crawl the popular websites of Islamic Fatwa, being official, like Al-Ifta-SA, Dar-al-ifta-EG and Al-ifta-JO, or non-official like islamway, islamweb,...etc. Those websites span different countries and geographical locations, accents, and backgrounds. We crawl for Question/Answer, Topic and Date. The topics and dates are not applicable or present for some websites. For Arabic AskFM, we extend the one in [11] to include 604K fatwas, by crawling the full website. A special type of QA is found in islamonline [39], where we treat the articles titles

It's better to reduce the scope to the top-k (k=5) topics, and compare to their corresponding word clouds.

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

Figure 3 Top 5 Topics/Subtopics for dar-al-ifta dataset

مجتمع وأسرة/الميراث/	
عبادات/الصلاة/	
عبادات/الزكاة/	
عبادات/الحج والعمرة/

Figure 4 Per subtopic Word Cloud for the top 5 topic/subtopic categories in dar-al-ifta dataset

L. 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 ISRISemmer for Arabic.

M. Unsupervised Topic Modelling

We use LDA as a first topic modelling attempt. LDA topic modelling works on the words co-occurrence matrix. It tries to find the latent factors that clusters sentences together. It uses Bag-of-words representation (BoW).

Generic Topic Modeling:

Figure 5 Identified topics for k=4 for all the topics in dar-al-ifta dataset

We use stemming. So الزكاة becomes زكاة. الصلاة becomes صلاة. Comparing the identified topics to the word clouds:

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

Top k=5 Topics Modeling:

Figure 6 Identified topics for k=5 for the top 5 topics/subtopics in dar-al-ifta dataset

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

Lowest categories of the top 5 seem not modelled by LDA: 'عبادات/الحج والعمرة', 'آداب وأخلاق'

In terms of topic modelling, stemming seems to perform the same as no stemming: general modelling do not perform well, while focused on top k topics give good results for the top 3, while the next 2 are not modelled.

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

N. Seasonality Analysis of Fatwa Volume Per Topic

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

Figure 7 Combined Seasonality trends of Hajj and Fasting

V. BENCHMARKS AND BASELINES

O. Topics Classifier Models

For the classifier models, we evaluate three families: the Bag-of-Words (BoW) and Sequence models: recurrent based or transformer based 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, we also evaluate a **BoW vectors**, where we use an Embedding layer for each input word. This requires padding the input sentence to a maximum of 250 words. The vocabulary size is also 2000 words and the embedding shape is 300. Embeddings layers for BoW vectors model are trained from scratch; i.e. initialized with random weights sampled from a normal distribution. For Recurrent Networks Topics Classifier

We evaluate both LSTM and GRU models, with 1 and 2 layers. For Transformer Networks Topics Classifier, we used a hugging face classification library to load AraBERT-base model and train on the dataset. We also used AraBERT's pre-process function that cleans the text and put it in the structure that AraBERT's tokenizer can read.

Baseline	Accuracy
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%

Table 2 Topics Classifiers Baseline Accuracy Results

P. Retrieval-based Question-Answering System

We evaluate the retrieval-based QA system as described in Figure 8. We used fasttext as the Embedding layer, for encoding the questions and the input. Then we use cosine similarity to compare and get the top k similar questions. The way fasttext works is 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 average them. We evaluate the QA system based on the retrieval accuracy, comparing the retrieved answer index against the true index. In order to test the similarity matching, we drop random word from the query question, before matching it against the database, which results in 96.4% retrieval accuracy. While this is a high accuracy, however, it reflects the retrieval performance from the known set of questions. In the true scenario, we might have unseen questions that do not match to the database questions, which raises the need to

have a generative QA system that generates the answer based on the summarized question state, not based on retrieval matching. We leave this for future works to tackle.

Figure 8 Similarity-based Question matching model.

Query	Answer / Fatwa
س: بسبب العلاج ومواعيد مش هفطر في رمضان غير أقل من 3 ساعات وباقي اليوم صايمة هو انا كدة ينفع مصومش؟	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 كتاب "الخلافة" حال دون وإياكم أجمعين 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

Table 3 Sample questions and retrieved answers

VI. CONCLUSION

In this paper we present the first steps towards building an Automated Fatwa System using AI and Deep Learning NLP methods. We survey the State-of-Art methods and provide use-case scenario design for a system that performs topic/intent classification and Question-Answer retrieval. This leads to the discussion of dataset collection, where we present the largest dataset of Islamic Fatwas. For this dataset, we performed unsupervised topic modelling, seasonality analysis, along with baseline models for both topic classification and QA. We evaluate our baselines in various aspects like: the effect of sequence modelling, the effect of pre-trained embeddings and language models. Also, we provide baselines for the widely used models in NLP in literature. Finally, we release all our models, 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 the project number 20-UQU-IF-P3-001.

REFERENCES

- [1] Abu Nada, A. M., Alajrami, E., Al-Saqqa, A. A. & Abu-Naser, S. S., 2020. Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach.
- [2] AiIftaSA, n.d. alifta. <https://www.alifta.gov.sa>.
- [3] Al Sallab, A. A. et al., n.d. Towards Deep Learning Models for Sentiment Analysis in Arabic.
- [4] Al Sallab, A. et al., 2015. *Deep learning models for sentiment analysis in Arabic*. s.l., s.n., p. 9–17.
- [5] Al Sallab, A., Rashwan, M., Raafat, H. & Rafea, A., 2014. *Automatic Arabic diacritics restoration based on deep nets*. s.l., s.n., p. 65–72.
- [6] AlIftaJO, n.d. alifta-jo. <https://alifta-jo/>.
- [7] AlFawzan, <https://alfawzan.af.org.sa/ar>
- [8] Al-Sallab, A. et al., 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)*, Volume 16, p. 1–20.
- [9] Alzantot, M. et al., 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.
- [10] Antoun, W., Baly, F. & Hajj, H., 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- [11] AskFM98k, n.d. askfm98k. <https://omarito.me/arabic-askfm-dataset/>.
- [12] Athiwaratkun, B., Wilson, A. G. & Anandkumar, A., 2018. Probabilistic fasttext for multi-sense word embeddings. *arXiv preprint arXiv:1806.02901*.
- [13] Badaro, G. et al., 2019. A survey of opinion mining in Arabic: a comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Volume 18, p. 1–52.
- [14] Bahdanau, D., Cho, K. & Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- [15] Baly, R. et al., 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*. s.l., s.n., p. 603–610.
- [16] Baly, R. et al., 2016. A meta-framework for modeling the human reading process in sentiment analysis. *ACM Transactions on Information Systems (TOIS)*, Volume 35, p. 1–21.
- [17] Binbaz, n.d. binbaz. <https://binbaz.org.sa/fatwas/kind/1>.
- [18] Binothaimeen, n.d. binothaimeen. <https://binothaimeen.net/site>.
- [19] Brown, T. B. et al., 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- [20] Carlini, N. & Wagner, D., 2017. *Towards evaluating the robustness of neural networks*. s.l., s.n., p. 39–57.
- [21] Chung, J., Gulcehre, C., Cho, K. & Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- [22] DarAlIftaEG, n.d. Dar-al-ifta. <https://www.dar-alifta.org/ar/Default.aspx?sec=fatwa&l&Home=1>.
- [23] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [24] Djandji, M., Baly, F., Hajj, H. & others, 2020. *Multi-Task Learning using AraBert for Offensive Language Detection*. s.l., s.n., p. 97–101.
- [25] Evgeniou, T. & Pontil, M., 2004. *Regularized multi-task learning*. s.l., s.n., p. 109–117.
- [26] Fallahzadeh, R. & Ghasemzadeh, H., 2017. *Personalization without user interruption: Boosting activity recognition in new subjects using unlabeled data*. s.l., s.n., p. 293–302.
- [27] Fatawapedia, n.d. fatawapedia. <http://fatawapedia.com/>.
- [28] Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y., 2016. *Deep learning*. s.l.:MIT press Cambridge.
- [29] Goodfellow, I. J., Shlens, J. & Szegedy, C., 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- [30] Grosse, K. et al., 2017. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*.

- [31] Gui, L. et al., 2018. Negative transfer detection in transductive transfer learning. *International Journal of Machine Learning and Cybernetics*, Volume 9, p. 185–197.
- [32] Hammerla, N. Y., Halloran, S. & Plötz, T., 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*.
- [33] Harutyunyan, H. et al., 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, Volume 6, p. 1–18.
- [34] Hong, J.-H., Ramos, J. & Dey, A. K., 2015. Toward personalized activity recognition systems with a semipopulation approach. *IEEE Transactions on Human-Machine Systems*, Volume 46, p. 101–112.
- [35] Howard, J. & Ruder, S., 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- [36] Islamqa, n.d. islamqa. <https://islamqa.info/>.
- [37] Islamway, n.d. islamway. <https://ar.islamway.net/fatawa/source/>.
- [38] Islamweb, n.d. islamweb. <https://www.islamweb.net/ar/>.
- [39] Islamonline, <https://islamonline.net/>
- [40] Kaiser, L. et al., 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.
- [41] Kuleshov, V., Thakoor, S., Lau, T. & Ermon, S., 2018. Adversarial examples for natural language classification problems.
- [42] Liu, P., Qiu, X. & Huang, X., 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.
- [43] Liu, X., He, P., Chen, W. & Gao, J., 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.
- [44] Li, Z. & Hoiem, D., 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, Volume 40, p. 2935–2947.
- [45] Luong, M.-T., Pham, H. & Manning, C. D., 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.
- [46] Magooda, A. et al., 2016. *RDL_{[T]{e}{a}m} at SemEval-2016 Task 3: RDI Unsupervised Framework for Text Ranking*. s.l., s.n., p. 822–827.
- [47] Mahmoud, R. A., Hajj, H. & Karamah, F. N., 2020. A systematic approach to multi-task learning from time-series data. *Applied Soft Computing*, Volume 96, p. 106586.
- [48] Min, W., Mott, B. W., Rowe, J. P. & Lester, J. C., 2017. *Deep LSTM-based goal recognition models for open-world digital games*. s.l., s.n.
- [49] Peters, M. E. et al., 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- [50] Pilonetto, G., Dinuzzo, F. & De Nicolao, G., 2008. Bayesian online multitask learning of Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 32, p. 193–205.
- [51] Qu, C. et al., 2019. *Learning to selectively transfer: Reinforced transfer learning for deep text matching*. s.l., s.n., p. 699–707.
- [52] Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., 2018. Improving language understanding by generative pre-training.
- [53] Rashwan, M. A. 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*, Volume 23, p. 505–516.
- [54] Roth, K., Lucchi, A., Nowozin, S. & Hofmann, T., 2018. Adversarially robust training through structured gradient regularization. *arXiv preprint arXiv:1805.08736*.
- [55] Sanh, V., Debut, L., Chaumond, J. & Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- [56] Sanh, V., Wolf, T. & Ruder, S., 2019. *A hierarchical multi-task approach for learning embeddings from semantic tasks*. s.l., s.n., p. 6949–6956.
- [57] Soliman, A.B., Eissa, K. and El-Beltagy, S.R., 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117, pp.256-265.
- [58] Sun, S. et al., 2018. Training augmentation with adversarial examples for robust speech recognition. *arXiv preprint arXiv:1806.02782*.
- [59] Sun, X., Kashima, H. & Ueda, N., 2012. Large-scale personalized human activity recognition using online multitask learning. *IEEE Transactions on Knowledge and Data Engineering*, Volume 25, p. 2551–2563.
- [60] Szegedy, C. et al., 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- [61] Tang, Z., Li, L. & Wang, D., 2016. *Multi-task recurrent model for speech and speaker recognition*. s.l., s.n., p. 1–4.
- [62] Tramèr, F. et al., 2017. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*.
- [63] Vaswani, A. et al., 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

- [64] Wikipedia, n.d. Wikipedia. <https://en.wikipedia.org/wiki/Islam>.
- [65] Yang, Z., Salakhutdinov, R. & Cohen, W. W., 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.
- [66] Yuan, X., He, P., Zhu, Q. & Li, X., 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, Volume 30, p. 2805–2824.
- [67] Zamir, A. R. et al., 2018. *Taskonomy: Disentangling task transfer learning*. s.l., s.n., p. 3712–3722.
- [68] Zhang, X. & Mahoor, M. H., 2016. Task-dependent multi-task multiple kernel learning for facial action unit detection. *Pattern Recognition*, Volume 51, p. 187–196.