ARABIC POEMS GENERATION USING LSTM, MARKOV-LSTM AND PRE-TRAINED GPT-2 MODELS

Asmaa Hakami, Raneem Alqarni, Mahila Almutairi and Areej Alhothali

Department of Computer Science,
King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

Nowadays, artificial intelligence applications are increasingly integrated into every aspect of our lives. One of the newest applications in artificial intelligence and natural language is text generation, which has received considerable attention in recent years due to the advancements in deep learning and language modeling techniques. Text generation has been investigated in different domains to generate essays and books. Writing poetry is a highly complex intellectual process for humans that requires creativity and high linguistic capability. Several researchers have examined automatic poem generation using deep learning techniques, but only a few attempts have looked into Arabic poetry. Attempts to evaluate the generated pomes coherence in terms of meaning and themes still require further investigation. In this paper, we examined character-based LSTM, Markov-LSTM, and pre-trained GPT-2 models in generating Arabic praise poems. The results of all models were evaluated using BLEU scores and human evaluation. The results of both BLEU scores and human evaluation show that the Markov-LSTM has outperformed both LSTM and GPT-2, where the character-based LSTM model gave the lowest yields in terms of meaning due to its tendency to create unknown words.

KEYWORDS

Arabic Poems, Markov, GPT-2, Deep Neural Networks, & Natural Language Processing.

1. INTRODUCTION

The developments of artificial intelligence have made it possible to compare the capabilities of machines with human abilities, such as the ability to generate texts of various forms [1]. The developments of artificial intelligence have made it possible to compare the capabilities of machines with human abilities, such as the ability to generate texts of various forms [1]. One of these texts is poetry, which is artistic literature that uses aesthetic and rhythmic language style to convey meanings or evoke emotions that affect the person who reads or hears [2]. Poetry can be used to express a specific feeling, situation, or scene or describe qualities of a character or place. It is one of the essential aspects of language in the world. Moreover, it is important to introduce the history and culture of the people, especially the Arab community. Their history, customs, and social principles are held over in this art. It also indicates the strength and durability of their language [3].

Poetry generation is one of the most interesting yet challenging Natural Language Processing (NLP) tasks. Several researchers seek to build models to generate textual data in different domains. One of the domains that were recently examined is poem generation. Several attempts were made in the NLP to generate poems in different languages, but many challenges were

David C. Wyld et al. (Eds): CMLA, CIoT, CSEIT, NCS, NLPD, SPM - 2021
pp. 139-147, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111512

encountered due to the meaning of the generated poem being unclear and understandable [4] or the structure of the poem being so chaotic and not thematic. Furthermore, other researchers focused on a specific type of poem or the style of specific writers [5] (see Section 2 for more details).

This work aims to take up this challenge and develop three different models to generate Arabic praise poems. LSTM, Markov-LSTM, and Pre-trained GPT-2 were chosen in this research due to their promising performance in other text generation tasks. We also evaluate the performance of the models against each other. This paper is organized as follows: Section 2 gives details about related work. Section 3 presents the used dataset. Section 4 describes the methodology for Arabic poem generation, which includes the pre-processing, and the proposed approaches to be compared. Section 5 presents the used evaluation criteria and analysis of the obtained results. Finally, Section 6 shows conclusions and future work.

2. RELATED WORK

Several studies that looked into generating poems and stories were considered in this research to benefit from the previous experiences in the same field. A number of these studies are represented in this section.

Authors in [4] built a model to generate coherent Chinese poetry in meaning with a flexible clear description of the topic for which the poem was created. The model was evaluated on three poetry domains which are quatrain, iambic, and chinoiserie lyric. Working Memory Model has been used as a method to create a poetry line by taking the previous line into account. The previous line is stored in local memory to be combined with the following line. A Topic Trace (TT) mechanism has been created to record the topics in a more explicit way. The poetry experts compared the results of this model with other approaches. The model received higher scores which indicates that the model generated poems with better quality and cohesion. Moreover, this model can create different types of poetry. The mechanism of TT helped increasing performance; however, there is still a gap between the generated poetry and human poetry.

Talafha and Rekabdar's [6] work was the first to propose an Arabic poem generation model using deep learning algorithms. The model contains two parts: The first part is a Bi-directional Gated Recurrent Unit (Bi-GRU) to generate the first line. The second part is a modified Bi-GRU encoder-decode. The proposed approach uses a hybrid model that combines a TextRank algorithm and a word embedding technique to extract keywords. FastText approach used to build the word-level embedded model. The dataset contains 80,506 verses from 20,106 Arabic poems that expressed love and religion. Quantitative evaluation using BLEU scores shows that the proposed model outperforms other deep learning approaches. On the other hand, the results of qualitative evaluation by humans show that the proposed model gives higher scores in terms of Coherence, Meaning, and Poeticness compared to other approaches in the same area.

Another study [5] developed a model called "poet without emotions" using deep learning algorithms for generating Arabic poems simulating the poems of the poet Nizar Qabbani.The model has built using Long Short-Term Memory (LSTM) networks that are a modified Recurrent Neural Network (RNN) version, making it more convenient to remember memory data. The model trained on 10,000 verses of Nizar Qabbani poems that are text sequences with the same length. So, the most generated poetic text by this model contains one verse of poetry. The accuracy of the generated poetic text reached 93%, which is a fairly good result. As Gharbi [5] states that, "it is an acceptable result if we take into account the simplicity of the used structure, the training and data formatting processes, in addition to the volume of the training text, compared to traditional text generation methods.

While [7] proposed a story generation model called “Story Scrambler” using RNN and LSTM. The model has given a sequence of input, which is considered a window. After that, the model had to predict the following word using the SoftMax activation function then the window updated. There were two types of input the first one is two stories with different content, and the second input is two stories with the same content but different narration. The model tested on various numbers of RNN layers, batch size, and input sequence length. It obtained minimal train loss of 0.01 when the size of RNN was 512 with three layers, the batch size was 100, and the input sequence length was 50. Moreover, it was discovered through experiments when increasing the number of layers beyond three and the batch size beyond 100 results in overfitting. The model evaluated by humans, and the accuracy of it was 63%.

Astigarraga et al. [8] proposed a model to generate poetry in the Basque language using two Markov chains. Poetry is one of the exhibitions of traditional Basque culture, especially in events and competitions. The model generates poems in the style of an existing author in less than a minute. Two different datasets were used to train the model. The first one was the Txirrita dataset, which has 2127 verses of poetry by a famous Txirrita. The second one was the Mixed dataset, which has 18913 lines compilation of sentences collected from Basque newspaper and poetry sung. Besides, some linguistic tools have been used to generate verse, such as rhyme search to find words that rhyme with the given word. Also, Latent Semantic Analysis (LSA) method has been used to measure the semantic relationship between pairs of words and sentences. The poem result will consist of four lines, each has thirteen syllables long, and all of them sharing a rhyme. The evaluation metric was only 2-gram because the system goal was to produce a poem somewhat different from the dataset. However, human evaluation is needed to assess poems. Therefore, four peoples familiar with the poems evaluated the generated poems. Each of them analyzed twenty poems, ten from each dataset. Their impression was positive, stating that they were well-formed poems, although not of human-produced quality. But they also found that the internal coherence of the whole poem was pretty poor. Furthermore, they stated that poems created from the Txirrita dataset seemed more natural and closer to the style of the Basque poems compared with the Mixed dataset.

From the previous studies, we found that most of the researches that are concerned with building a model to generate poems is limited to both English and Chinese languages. There is a lack of research that generates Arabic poems. Also, none of the previous studies were specialized in generating praise poems in the Arabic language.

3. DATASET

The dataset in this research was collected from praise poems written by different Muslim poets who were born and lived during different eras and in several countries, such as AL-Arjani, who lived in the Andalusian era, Ibn Al-Khayyat, who lived in the Mamluk era. The dataset consisted of 34,466 verses that have been collected manually from the AL-diwan website [9], which includes a large number of poems in various fields such as Ghazal poetry. There were also many different fields/aspects of praise poetry, including praising people such as Prophet Muhammad peace be upon him, tribes such as Quraysh. Also, countries such as Iraq and Egypt. This is an example of a poem that praise Prophet Muhammad peace be upon him:

“أَجِد مَدحَ خَيرِ الخَلقِ ذاتاً وَجَودَةً
وَجِد عَن سِوى ما سَنَّهُ لَكَ حَيدَةً
وَأَنشِد هَوىً فيهِ اِكتَفي وَمَوَدَّةً
مَدَحتُ رَسولَ اللَهِ بَدأً وَعَودَةً
وَمِقدارُهُ في البَدءِ وَالعَودِ أَعظَم”.

4. METHODOLOGY

As shown in Figure 1, the followed methodology to generate praise poems consists of four phases. In this section, these phases are discussed in more detail.

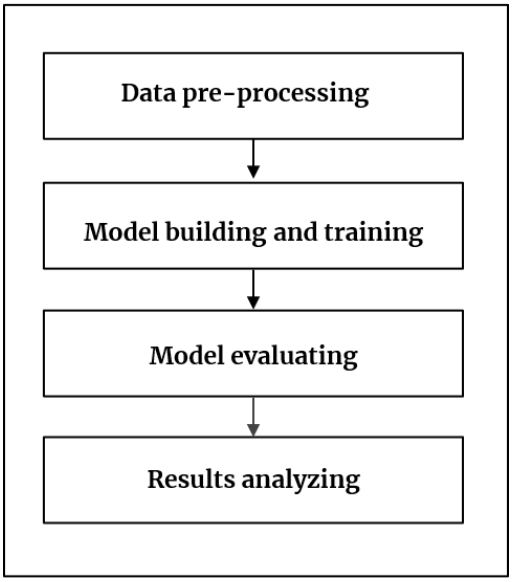

Figure 1. The phases of the methodology.

4.1. Data pre-processing

In this phase, we performed dataset pre-processing by removing or modifying data that is incorrect, incomplete, irrelevant, or duplicate. We removed unrelated characters or symbols such as punctuation marks, line space, $, #, brackets, parentheses, and other unrelated characters from the dataset. Table 1 represents an example from the input before and after processing.

Table 1. Example of the pre-processing stage

Before processing	After processing
!غير الرياحِ التي في التيهِ تنزلقُ	غير الرياحِ التي في التيهِ تنزلقُ
هُنا مدائحُ (حَسَّانٍ) على شفتي	هُنا مدائحُ حَسَّانٍ على شفتي
مَنْ ذا يُطَبِّبُ في الإنسانِ جوهرَهُ؟	مَنْ ذا يُطَبِّبُ في الإنسانِ جوهرَهُ

4.2. Proposed Methods

In this work, three approaches for text generation (praise poems generation in particular) are proposed, which are the character-based LSTM model, Markov-LSTM model, and GPT-2 pre-trained model. In this section, the models are discussed in more detail.

4.3. Character-based LSTM model

The first proposed model is the character-based LSTM model. It is implemented by using the Keras library [1]and composed of three layers. The first layer is the embedding layer that takes integers indices (which stand for specific characters) and turns them into dense vectors of 256 dimensions. Before inputting the data to this layer, we map all existing characters in the dataset to a numerical representation. The second layer is an LSTM layer with 1024 units. LSTM is a type of RNN that is capable of handling long-term dependency and vanishing gradient problem. RNN models can be useful to model time series data or sequential data such as natural language text [10]. The third layer is a dense layer with size outputs equals to vocab size. This model takes the input as a sequence of characters, and the length of this sequence is 200 characters, and tries to predict the next character at each time step. For example, if the input is the sequence shown in Figure 2 (A), and in this case, the model expects the letter "ر" and the output to be as shown in Figure 2 (B). The problem can be regarded as a classification issue at this point. The model at this time step will predict the class of the next character based on the previous LSTM state and the current input. Therefore, the categorical cross-entropy loss function is used, and it is employed across the last dimension of the predictions. This model is compiled with Adam optimizer. The string length of the input through the training is 200 characters, but the model can be run on start strings of any length. An example of the generated verses by using this model is shown and analysed in Section 5.

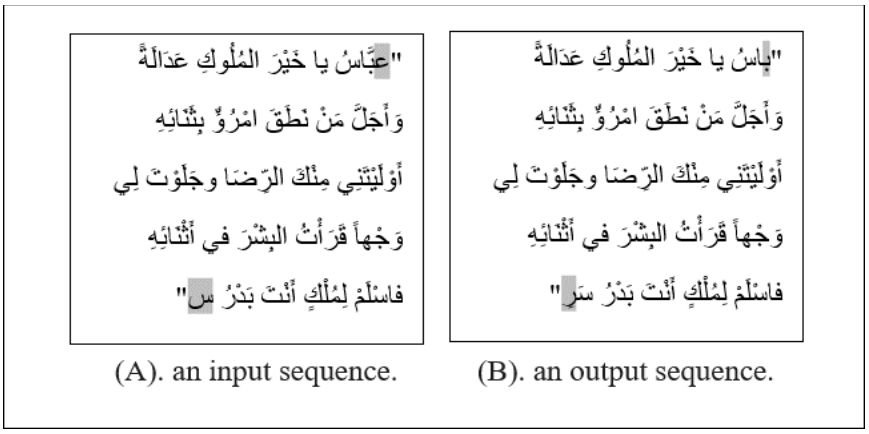

Figure 2. An example of the input-output of character-based LSTM model.

4.4. Markov-LSTM Model

The second model in this study generates Arabic poems using the Markov-LSTM model. Markov models were used in several fields. One of them is in text generation [11] and shows promising results in short text generation. Markov-LSTM depends on the probability of the next word based on the current [11]. This model uses Markovify functions to generate verses based on every word probability. So, if the current word is "وإني", and the probability of the word "لما" to come after it is 20%, where the word " حين" has the probability of 25%, Then the word "حين" will be chosen and so on. The first word in each verse was chosen randomly. Any verses result from the Markovify functions that contain the word "الله" (Allah) was removed to avoid inappropriate use of this word. Furthermore, the result verses were encoded and used as input to the LSTM model. The LSTM model is composed of four layers to generate a poem. LSTM has long-term memory,

[1]https://keras.io/

so that it is better to predict the properties of the next verse, such as rhyme scheme. So, the appropriate new verse of the poem can be selected each time by the LSTM method from all previous verses generated by Markovify. An example of the generated verses by using this model is shown and analysed in Section 5.

4.5. Pre-trained GPT-2 Model

The third model focuses on fine-tuning a pre-trained GPT-2 model. GPT-2 is a large-scale unsupervised language model produced by Open-AI. It was trained on 40GB of Internet text to predict the next word. Due to the concern of Open-AI about malicious technical applications, Open-AI is not releasing the trained model. Instead, Open-AI launches a much smaller model as an experiment in responsible disclosure [12]. There are three versions of GPT-2 are released, which are the small version (124 Million Parameters), the medium version (355 Million Parameters), and the large version (774 Million Parameters). Moreover, larger models are more knowledgeable, but these models take a longer time for fine-tuning and text generation [12]. For this work, the small version of the GPT-2 model is fine-tuned on the dataset to generate praise poems. Because our dataset with a size of 2.01 BM and that less than the minimum recommended size (10 MB) to use the medium version of GPT-2. We fine-tune this model by using the gpt-2-simple package that is a Python package, and it wraps existing model fine-tuning. Also, it makes text generation easier and allowing for prefixes to force the text to start with a given phrase. The input data to this model is a single text file as the model requires. We use the Finetune function to fine-tune the pre-trained GPT-2 model on our datasets. We set the parameters for the Finetune function as following: The steps parameter was set to 2000. The restore From parameter set to "fresh"; to begin training from the base GPT-2. The learning rate parameter for the training is set to1e-4 by default. Also, we use the Generate function to generate text after fine-tuning this model on our datasets. This function has an important parameter which is the temperature. The higher the temperature, the syntactically incorrect and the unique the text. We set it to 0.7 as the minimum recommended value (recommended to keep the temperature value between 0.7 and 1.0) [13]. An example of the generated verses by using this model is shown and analysed in Section 5.

5. RESULT AND DISCUSSION

In this section, the results and evaluation of the three models that were used to generate Arabic poems will be represented.

5.1. Quantitative Evaluation

The BLEU scores were used to evaluate models, which indicate the identity of the generated text and the reference text. It has many ways of measuring, such as 1-gram, 2-gram, 3-gram, and 4-gram. Each gram represents the number of words that will be taken from both texts and compared to each other. The value of BLEU ranges from 0 to 1. The higher the BLEU value, the higher the similarity between the generated sentence and the reference sentence. The BLEU scores are calculated by the following equation [14]:

$$BLEU = \mathrm{BP}.\, exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5.1}$$

Table 2 represents a sample of the outputs for all models. The verses of the Character-based LSTM model were inconsistent and had no meaning because the model predicted a letter by letter. It creates new words that are not present in the dataset or the Arabic language dictionary.

Also, we noticed that sometimes the generated text reproduces the exact verses in the dataset. The Markov-LSTM model verses have common characteristics with the previous model, except that sometimes it changes one word when taking a verse from the dataset. The generated verse is better than the Character-based LSTM model because it does not create new unknown words. The verse of the Pre-trained GPT-2 model was close to the Markov-LSTM model in terms of relational coherence and logic.

Table 2. Models' output

Model name	**The generated verses in Arabic**	**The generated verses in English**
Character-based LSTM	وَأحسَنُ مِنكَ لَم تَرَ قَطُّ عَيني وَأعظَمُ غُرَّةٍ يُعازِلُ في حَقٍّ اِتَّخَذتَ بِها وَفي العُراقِ إِذا أَضَلَّتَ حَبّاً مَعفونَفَينِ بِهِ مَكانَ عَلى أَيّامِن الجَنزِلبِسِ وَهوَ سُوري ذَريعٌ كَما غَدَّ الغَمامُ عَلى الدُجى	My eyes did not see better than you (seed text), nor did they see greater than your face. And in Iraq if she astray because the huge love (metaphor). He is dismissing in the right way that you take it. They are chaste by it (word of chaste written with misspelling). Place on Ayamen (Ayamen is the plural of word of right in Arabic), jeans, and it is Syrian (these unrelated words to each other). It is fast as the spread of a thin white cloud over the darkness of the night.
Markov-LSTM	إذا قيلَ هَل سارَ فَاقلِقني أنشدت مدحي فيك من فنون المعاني عَجائِبَ لا تَنفَكُّ عَينَي إلى جدث بعد الحياة اذا التذكار أحياني ولا تعتذر غير ما يَعْني أَنتَ اِلّا في مَقامِ الشُكرِ يا رَبِّ اشْفَعْني أوْتِرْنِي	If it is said he walked, worry me I sang praise in you from the arts of meanings Wonders do not open my eyes To a grave after life, if the souvenir revives me And do not apologize except for what is meant You are only in the place of thanks giving, O Lord, preemptive me or see me.
Pre-trained GPT-2	وَأحسَنُ مِنكَ لَم تَرَ قَطُّ عَيني وَلَم أَجِر قَومٍ يَخافُ الذُبابَ بِها وَنَزَّهُ مِن حَيثُ لانتِزارُ أَخرَفُ فَكَم بُعدٍ غَيرِ القَولِ فيما خُلِقوني ُتَوارَى أَنَّهُ عَنِ الأَعمارِ وَحَلَّه وَما إِذا حَلَّ قَلبَ المُقَلَّدِ في كَرَمٍ	And better than you, my eyes did not see(seed text) And I did not reward a people that feared flies for it And he pure from where Lantizar is senile How much a dimension of unspoken from what they created me He hid as from the ages and resolved it And whether the heart is resolved of the imitator in the vineyard

Table 3 shows the average BLEU-1 scores of all models. From the Character-based LSTM model and Pre-trained GPT-2 model scores, they noticed that several new words were created not from the dataset since the ratio of 1-gram is not high. The Markov-LSTM model had a high score on the 1-gram which means the model does not generate new words that are not in the dataset, but it used some consecutive words as in the dataset.

Table 3. Models' BLEU Score

GRAM	CHARACTER BASED LSTM	MARKOV-LSTM	PRE-TRAINED GPT-2
1-GRAM	0.552026	0.760297	0.560736

5.2. Qualitative Evaluation

The BLEU scores do not evaluate the quality of generated poems, such as meaning and coherence. So, it is required to get a human evaluation in this field of research. Besides, human evaluation is hard because they have different opinions and tastes in poems. In this work, three poems were chosen randomly from each model to evaluate them by two experts. The evaluation was based on four criteria which are: meaning, coherence, rhyme, and rhythm. Each criterion takes a score from zero to five (zero is the worst). Table 4 shows the result of human evolution. As shown in the result, the Markov-LSTM model got higher scores in terms of meaning, coherence, and rhyme compared with the other two models.

Table 4. Human Evaluation

Criteria / Model name	Meaning	Coherence	Rhyme	Rhythm
Character-based LSTM	0.5	0.5	1.5	1.5
Markov-LSTM	1	0.75	2	1.5
Pre-trained GPT-2	0.5	0.5	1.5	1.5

6. CONCLUSION AND FUTURE WORK

In this paper, three models are presented for generating Arabic poems in the field of praise. Three models were built, which are Character-based LSTM, Markov-LSTM, and pre-trained GPT-2. The results of the Markov-LSTM model were better than the other two models based on the BLEU-1 score. As for the consistency of the verses, and the clarity of their meaning there is no wide difference between it and the pre-trained GPT-2 model. The generated poems still lacked some grammatical rules and logical sequences of words and their interconnection with each other. In the future, we aim to reduce the runtime and increase the number of verses in the dataset. Finally, improve the grammatical rules of the lines that the models generate.

REFERENCES

[1] Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. Journal of Service Research, 21(2), 155-172.

[2] Ibrahim Mahmud Ahmad, Aabd Alrahim (2018). Alqusu abalaghatuh fy alshier alerby alqdym [The storytelling and its rhetoric in ancient Arabic poetry]. Majalat albahth alelmy fy aladab [Journal of Scientific Research in Literature], 211-232.

[3] Abdul-Sahib, Ali (2011). Fi mafhum alshier wlghth: khasayis alnas alshaerii[On the concept of poetry and its language: characteristics of the poetic text] University of Sharjah Journal for Humanities and Social Sciences, 111(460), 1-17.

[4] Yi, X., Sun, M., Li, R., & Yang, Z. (2018, July). Chinese poetry generation with a working memory model. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (pp. 4553-4559).

[5] Gharbi, G. (2019). Poetry Without Emotion: An experiment in gener-ating Arabic poetry by using deep learning. In Artificial Intelligence Applications In The Service of The Arabic Language

(1st ed., pp. 174-186). Riyadh, KSA: Wojooh Publishing & Distribution House. Retrieved October 02, 2020, from https://kaica.org.sa/site/page/89.

[6] Talafha, S., & Rekabdar, B. (2019, January). Arabic Poem Generation with Hierarchical Recurrent Attentional Network. In 2019 IEEE 13thInternational Conference on Semantic Computing (ICSC) (pp. 316-323).IEEE.

[7] Pawade, D., Sakhapara, A., Jain, M., Jain, N., & Gada, K. (2018).Story scrambler-automatic text generation using word level rnn-lstm. International Journal of Information Technology and Computer Science(IJITCS), 10(6), 44-53.

[8] Astigarraga, A., Martínez-Otzeta, J. M., Rodriguez, I., Sierra, B., & Lazkano, E. (2017, August). Markov Text Generator for Basque Poetry. In International Conference on Text, Speech, and Dialogue (pp. 228-236). Springer, Cham.

[9] Aldiwan: mawsueat alshier alearabi [Aldiwan: An Encyclopedia of Arabic Poetry] [Online] Available at: <https://www.aldiwan.net>.

[10] Shukla, N., & Fricklas, K. (2018). Machine learning with TensorFlow. Greenwich: Manning.

[11] Szymanski G., & Ciota,Z. (2004). On-line text generation using Markov models, Proceedings of the International Conference Modern Problems of Radio Engineering, Telecommunications and Computer Science, pp. 339-341.

[12] Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better language models and their implications. OpenAI Blog https://openai. com/blog/better-language-models.

[13] Utane, N. (2020, April 17). Complete guide to build and deploy a tweet generator app into production. Retrieved December 22, 2020, from https://towardsdatascience.com/complete-guide-to-build-and-deploy-a-tweet-generator-app-into-production-5006729e583c.

[14] Zhukov, V., Golikov, E., & Kretov, M. (2017). Differentiable lower bound for expected BLEU score. arXiv preprint arXiv:1712.04708.

AUTHORS

Asmaa Hakami is a senior computer science student at King Abdul-Aziz University. Her research interest lies in the areas of machine learning, deep learning, natural language processing, and computer version.

Raneem Alqarni is a senior computer science student at King Abdul-Aziz University. Her research interest lies in the areas of machine learning, deep learning, natural language processing and computer version.

Mahila Almutairi is a computer science student at King Abdul-Aziz University. Her research interest lies in the areas of machine learning, deep learning, and natural language processing.

Areej Alhothali is an assistant professor in the faculty of computer science and information technology at King Abdul-Aziz University. She earned her master's and Ph.D. degrees in computer science (artificial intelligence) from the University of Waterloo, Canada. Her research interest lies in the areas of machine learning, deep learning, natural language processing, intelligent agent systems, affective computing, and sentiment analysis.