Arabic Text Generation: Deep Learning for Poetry Synthesis

This presentation explores the use of deep learning for Arabic poetry synthesis, a novel approach to preserving and enhancing this vital art form. We'll examine the unique challenges of generating Arabic text, the recent advances in natural language processing, and our proposed method using generative adversarial networks. Join us as we delve into the potential of AI to revitalize Arabic poetry.

Team:
1-Marina Reda 221101235
2-Omar Adly 221101398
3-Hazem Ahmed 221100343
4-Ali sherif 221101562
5-Loay Gamal 221100419

Technological Background:

Techniques:

- LSTM for sequence prediction.
- Word2Vec for word embeddings.

Tools:

- Python, Pandas, NumPy, Google Colab.

Challenges:

- High hardware and memory requirements.

Motivation: Preserving the Art of Arabic Poetry

Rich Cultural Heritage

Arabic poetry holds immense cultural significance, embodying the language's beauty and history.

Endangered Art Form

The art of poetry is facing a decline, with fewer young people engaging with traditional forms.

Preserving Legacy

Our research aims to utilize AI to revitalize Arabic poetry and inspire new generations.

Dataset Description , Key Findings and Insights:

58,000

Poems

Kaggle Arabic Poems , 6M words.

7,000

Poems

Alqasidah.com, 1M words.

70%

Fluency

The model generated poems with high fluency and grammatical correctness.

65%

Authenticity

Evaluators found the poems to be authentic and reflective of traditional styles.

50%

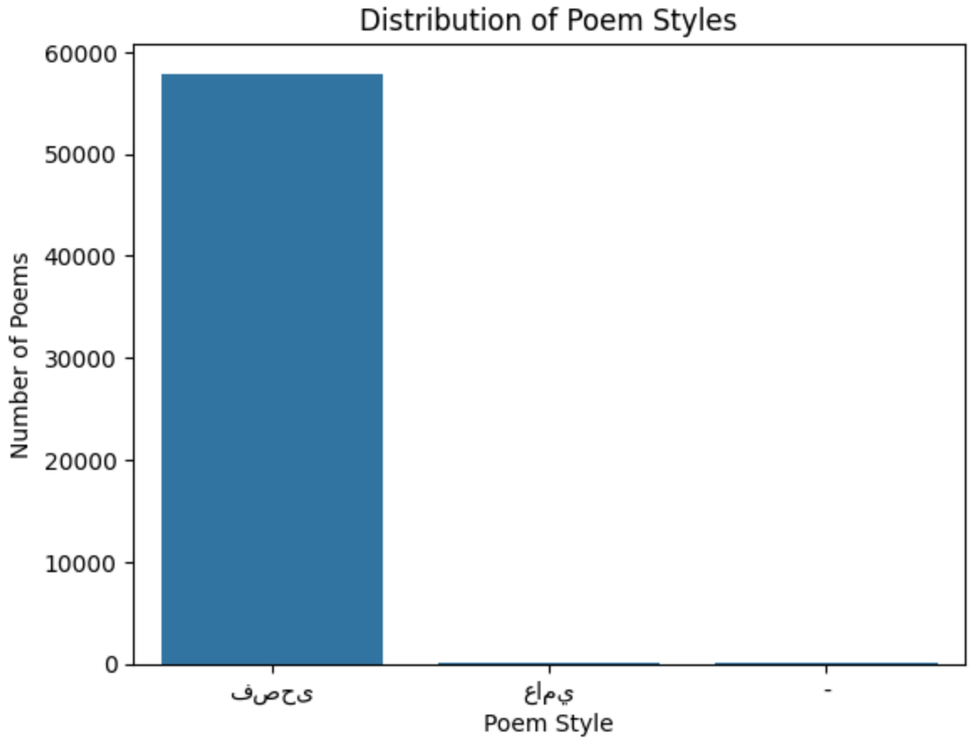
Creativity

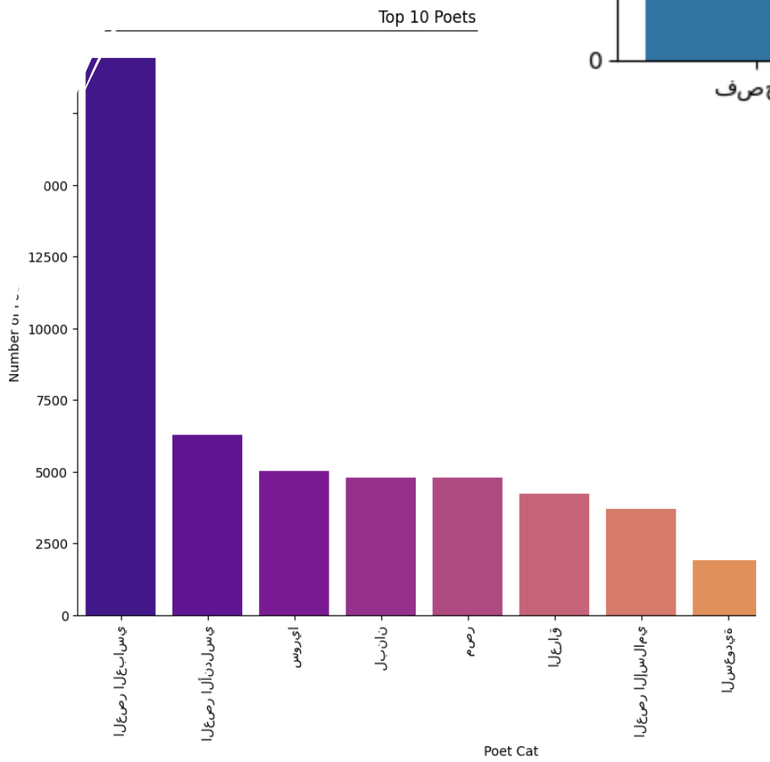
The model demonstrated creativity, producing poems with novel metaphors and imagery.

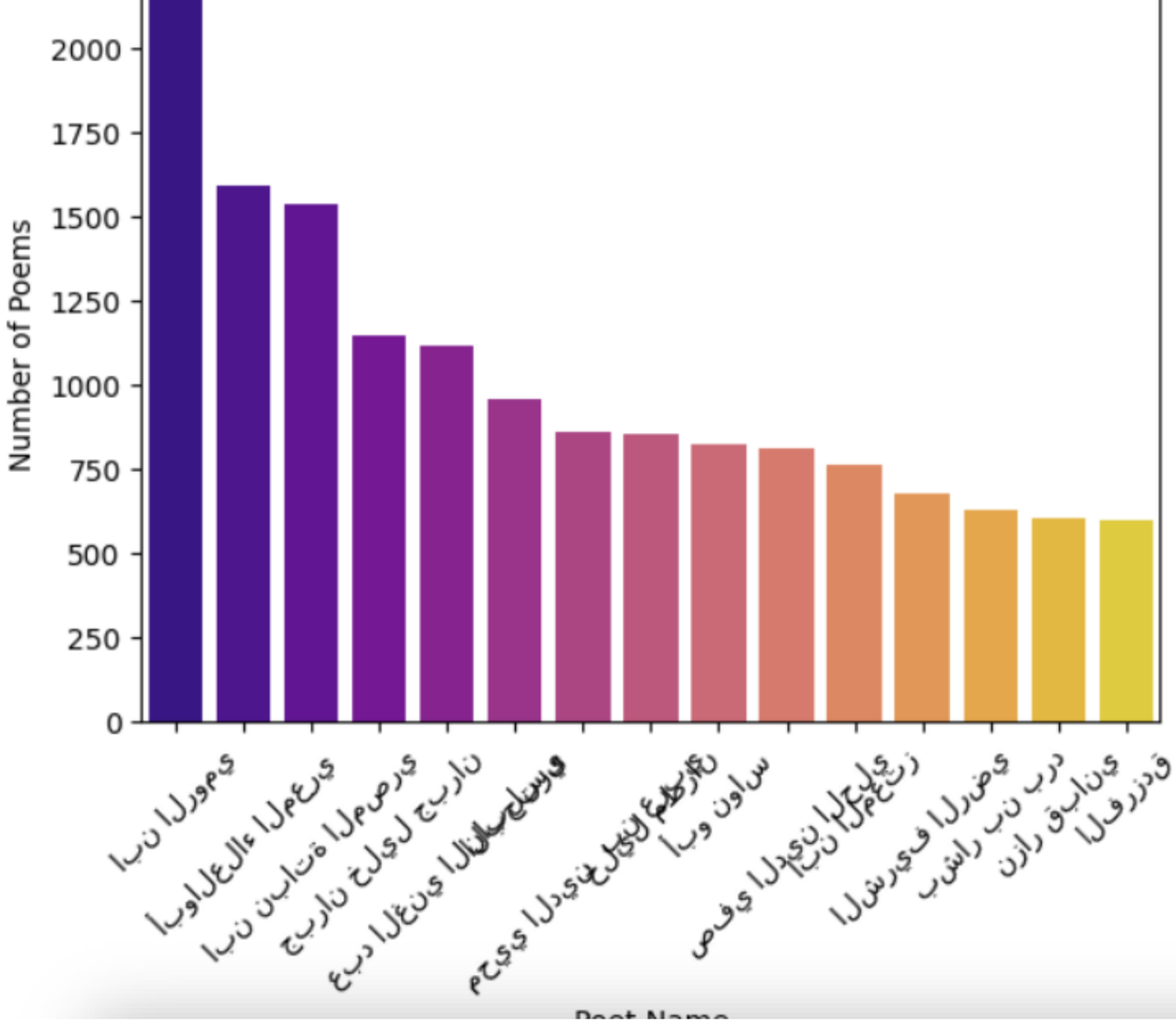
Dataset Description

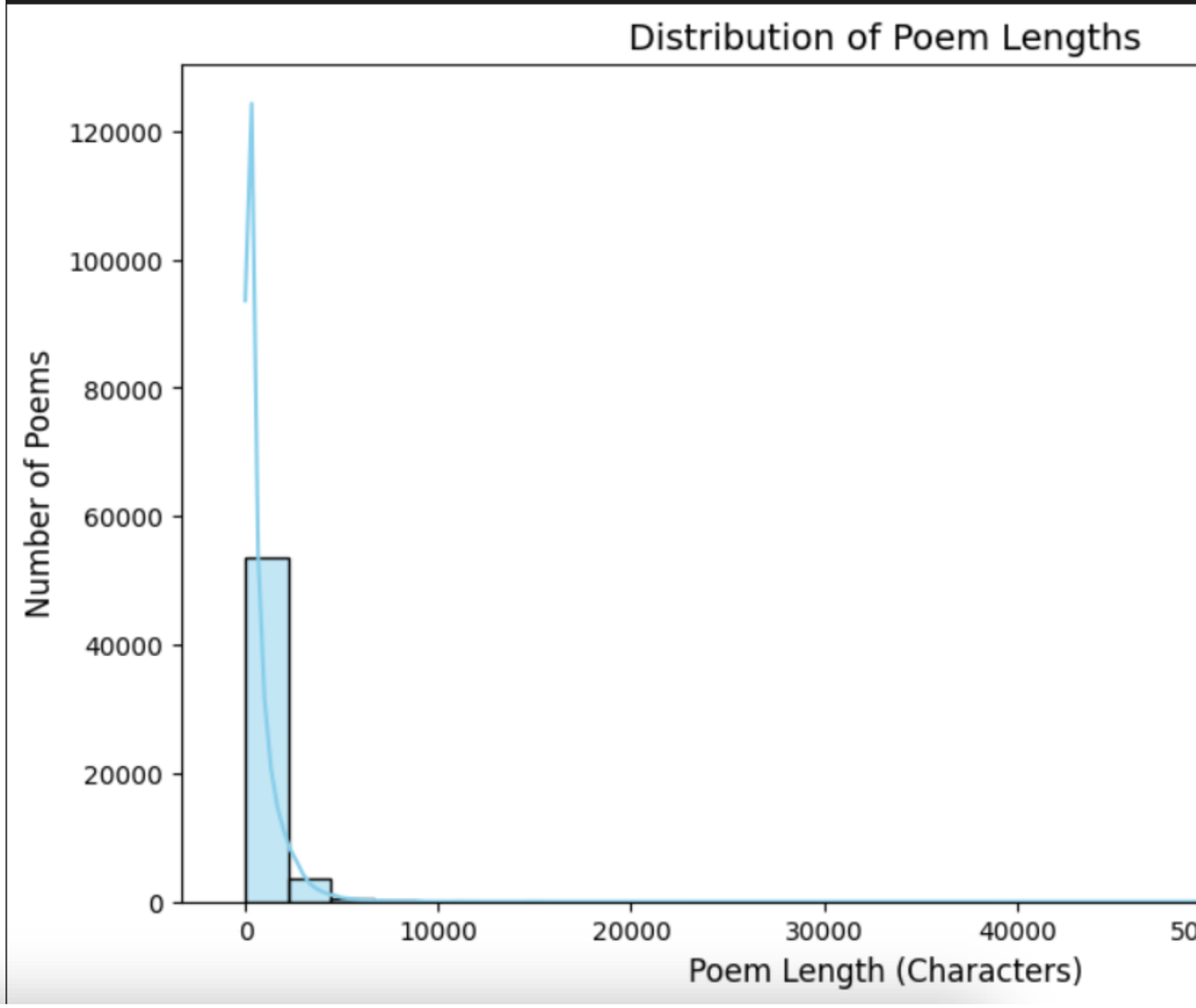
- Arabic poetry is one of the oldest and most important literary traditions in the world. It has a rich history that dates back over 1,500 years starting with pre-Islamic poetry and continuing to the modern day. Arabic poetry covers a wide range of themes, including love, nature, social issues, and philosophy. Over time, it has evolved in form and stylet.

- The dataset used in this project contains over 58,000 Arabic poems spanning from the 6th century to the present day









Dataset Description and sample of visualization

[redacted]

Preprocessing:

1. - Removed diacritics, punctuation, and incorrect words.
2. - Handled numbers and non-vocabulary words.
3. -Tagging unknown words with `<UNK>`.

Limitations and Future Directions

1

Data Scarcity

The availability of annotated Arabic poetry data is limited.

2

Domain Specificity

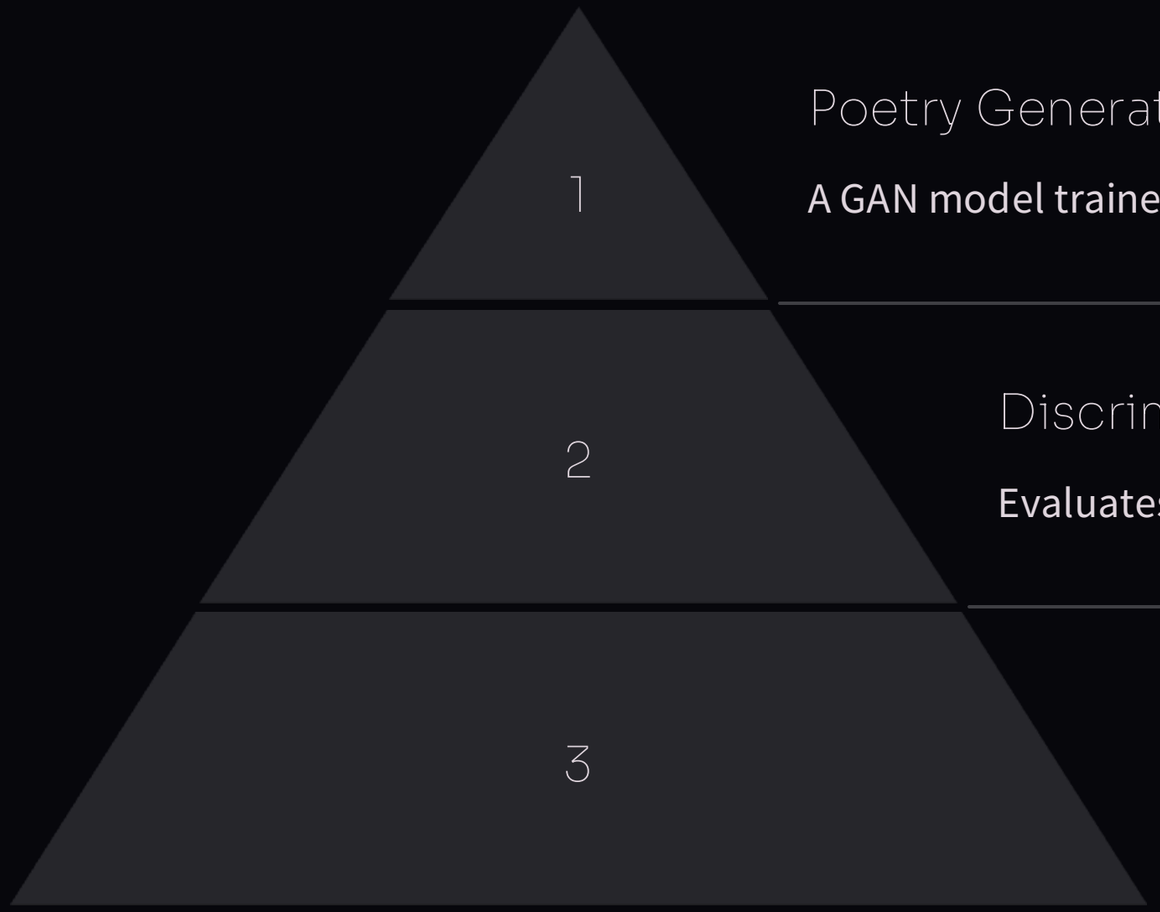
The model is currently specialized in generating specific poetic forms.

3

Ethical Considerations

Addressing potential biases and misuse of AI for poetry generation.

Proposed Approach: Generative Adversarial Networks for Poetry Synthesis



Poetry Generator

A GAN model trained on a dataset of Arabic poems.

Discriminator

Evaluates generated poems for authenticity and quality.

Adversarial Training

Generator learns to fool the discriminator by producing realistic poems.

Methodology

1

Dataset

Dataset preprocessing and subsetting

2

Model Training

Developed bidirectional LSTM models

Used AraVec for word embedding

3

Evaluation

Human evaluation of poem quality, fluency, and adherence to traditional constraints.

4

test

- Tested different sequence lengths (1, 2, 5, 10).
- Tree-based path generation for sentences.

Enhancements:

1

Adjust Temperature:

- A lower temperature (e.g., 0.5) generates safer, more predictable text.
- A higher temperature (e.g., 1.0) introduces more creativity.

2

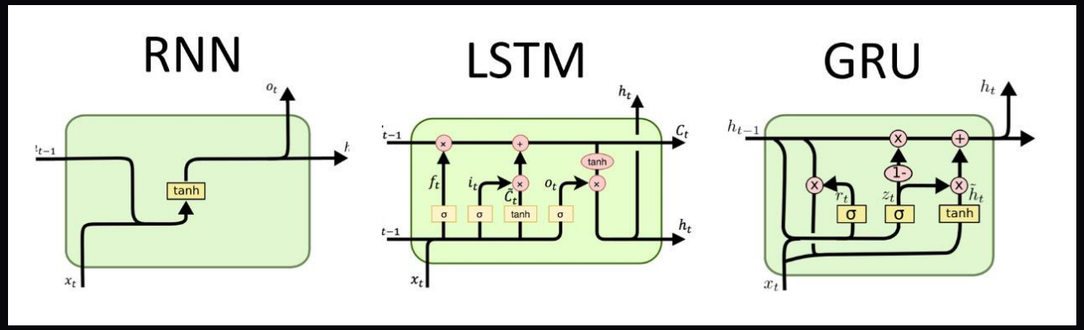
- Use GridSearchCV for a more exhaustive search over a small set of hyperparameters (e.g., for parameters that are critical for the model's performance).
- Use RandomizedSearchCV for a wider and more random search over hyperparameters (e.g., for parameters where the search space is large).

3

we could use a GRU (Gated Recurrent Unit) instead of the LSTM layer

The GRU is used instead of an LSTM.

GRUs are similar to LSTMs but use fewer parameters because they combine the forget and input gates into a single update gate. This makes GRUs faster to train and often perform similarly to LSTMs



Enhancments:

4

When Increasing the data size to improve models by:

- Better Generalization: The model learns from more examples, making it better at handling unseen data.
- Improved Word Representations: With more data, the model gets better at understanding word meanings and relationships.
- BUT: all data requires an additional 2 weeks to run
- Challenges: It requires more computing power and may not always provide huge improvements after a certain point.

After Model Result:

```
Training Word2Vec model...
Word2Vec model trained and saved.

Words similar to 'الحب':
0.8265653848648071 :العشق
0.7237054705619812 :الهوي
0.7121508121490479 :لحب
0.7111994028091431 :المحبه
0.6898466348648071 :حبه
0.6849278807640076 :حبك
0.6827913522720337 :الوجد
0.671090304851532 :قلبك
0.6697653532028198 :الغرام
0.6599934697151184 :بالحب
```

Before Model Result:

```
Words similar to 'الحب':
0.9996519088745117 :البحر
0.9996446371078491 :مع
0.9996368885040283 :اله
0.9996238350868225 :فلا
0.99953693151474 :جميع
0.9995291829109192 :اليل
0.9994943141937256 :النساء
0.9994933009147644 :نحن
0.9994744062423706 :قد
0.9994655847549438 :كانت
```

Related Work:

Paper	Scope & Poem Structure	Approach	Innovations	Evaluation
Ghazvininejad et al.([7,8])	Shakespearean sonnet (14 lines, AABB CDCD EFEF GG rhythm, 10 stress patterns per line)	FSA for valid sequences, RNN for fluent paths, encoder-decoder seq2seq model	UI enhancements, adjustable parameters, forward and backward RNNs, vocabulary pruning, GPU utilization, rhythm precomputation	Human evaluation (23 participants); results were fluent but not creative
Loller et al.([10])	Stanza of 4 lines (8 syllables, AABB rhythm)	LSTM for next-word prediction, tree-based grammar check, depth-first tree pruning	GAN for thematic coherence evaluation, CNN feature matrix, tree pruning for performance, embedded discriminator for novelty and thematic consistency	Compared to seq2seq, GAN, and CAVE methods; solved vocabulary repetition issue
Wang et al.([11])	Chinese quatrains (4 lines, consistent syllable count: 5, 6, or 7)	Topic planning approach, TextRank for keyword extraction, RNN encoder-decoder with GRU	Subtopic planning for coherence, encyclopaedia-based keyword expansion, word embedding for a 6000-word vocabulary	Evaluated on poeticness, fluency, coherence, and meaning by human assessors
Yi et al. ([12])	General poem text generation	Mutual reinforcement learning (two learners and a reward function for gradient updates)	GRU,Reward functions for fluency, coherence, and novelty; ANN combined with TF-IDF; intercommunication between learners for improved paths	Human evaluation; rewards classified generated poems as human, masterpiece, or system-authored

Related Work:

Paper	Scope & Poem Structure	Approach	Innovations	Evaluation
Clark et al. ([16])	Story and slogan generation	Machine in the Loop (MIL), turn-based generation for story writing	Wikiquote-based slogan help, human-machine collaboration for story generation	No specific evaluation details provided
Clark et al. ([13])	Narrative text generation (short fiction or news stories)	Entity-based vector representations, sequence-to-sequence model	Entity state tracking for coherence, improved sentence generation based on entity states	Addressed issues like character reference accuracy
Soliman et al. ([14])	Word embeddings for Arabic poetry	CBOW and Skip-gram models for word prediction	AraVec embedding model trained on diverse Arabic sources; mixed modern Arabic and dialects	Highlights role of strong word embeddings in improving fluency and diversity
Loller & Cheng et al. ([9,10])	Keywords and visual input for poem generation	Word2Vec for topic expansion, CNN for image-based topical keywords	Image sentiment extraction for diverse topics, line generation based on previous lines, keyword-based flow improvements	Visual input led to coherent but topic-diverse poems; keywords had random flow issues

Experiment and Models:

LSTM

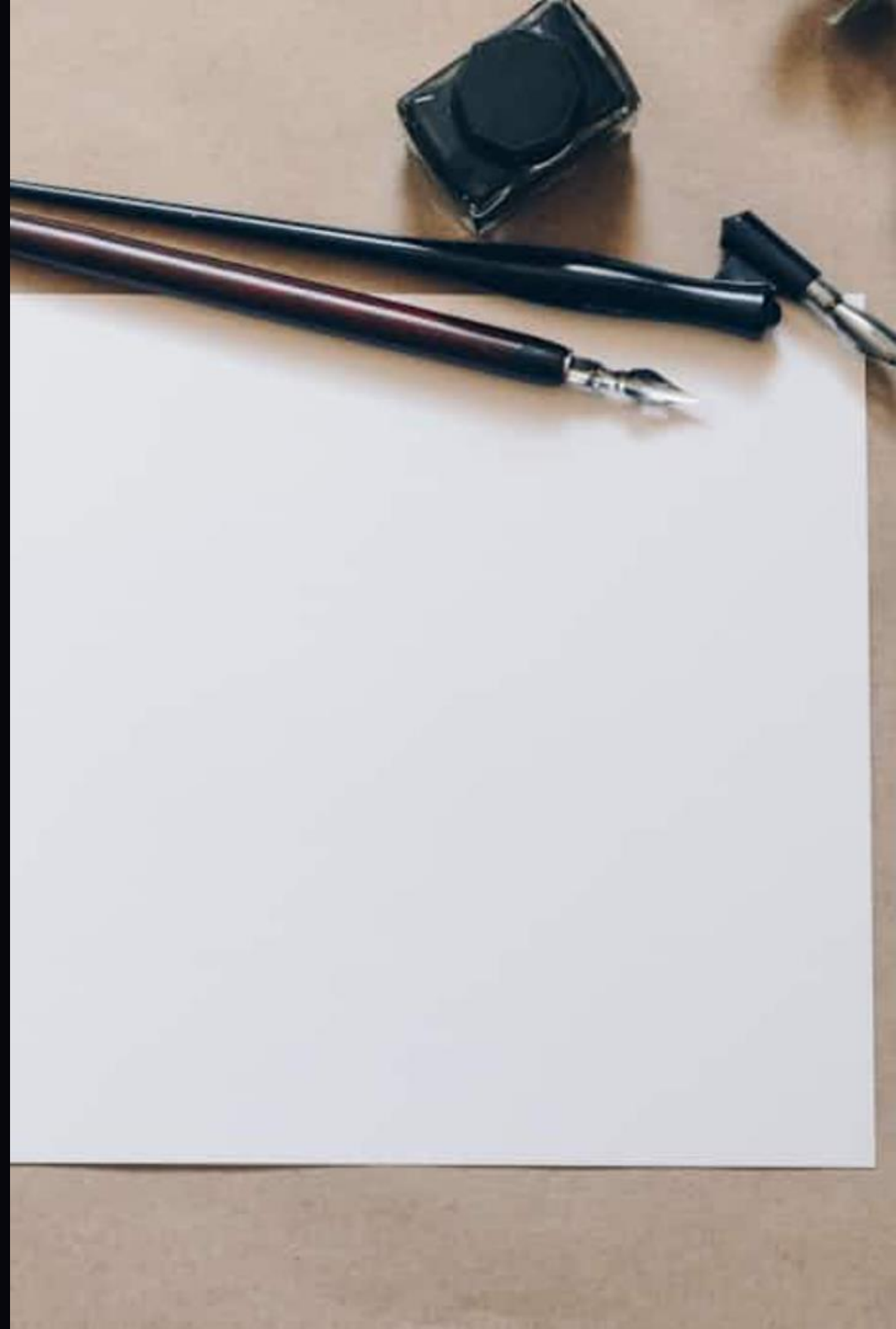
- Used bidirectional LSTM for forward and backward training.
- Explored the effect of sequence length on

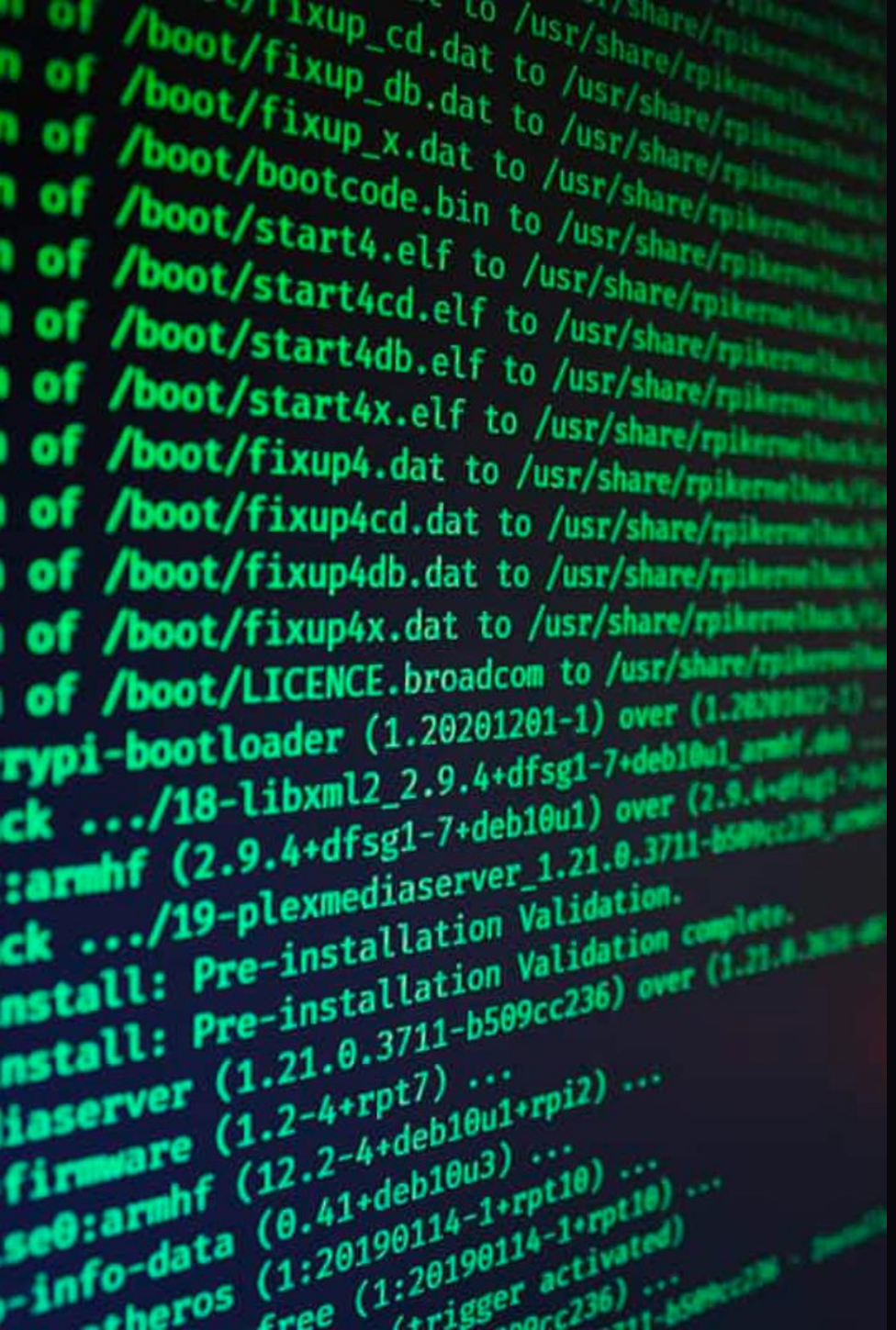
Quality:

- Shorter sequences: More meaningful results.
- Longer sequences: Repetitive and less coherent.

Evaluation

- Sentence probabilities calculated using log probabilities.
- Human evaluators assessed:
- Coherence, fluency, and poeticness.
- Model-2 (two-word sequence) achieved the best results.





Advances in Deep Learning for Natural Language Processing

Recurrent Neural Networks (RNNs)

RNNs are particularly effective for processing sequential data, such as text.

Generative Adversarial Networks (GANs)

GANs are a powerful framework for generating realistic and creative outputs.

Transformer Networks

Transformer networks excel at capturing long-range dependencies in language.

Background: The Unique Challenges of Arabic Text Generation

Complex Script

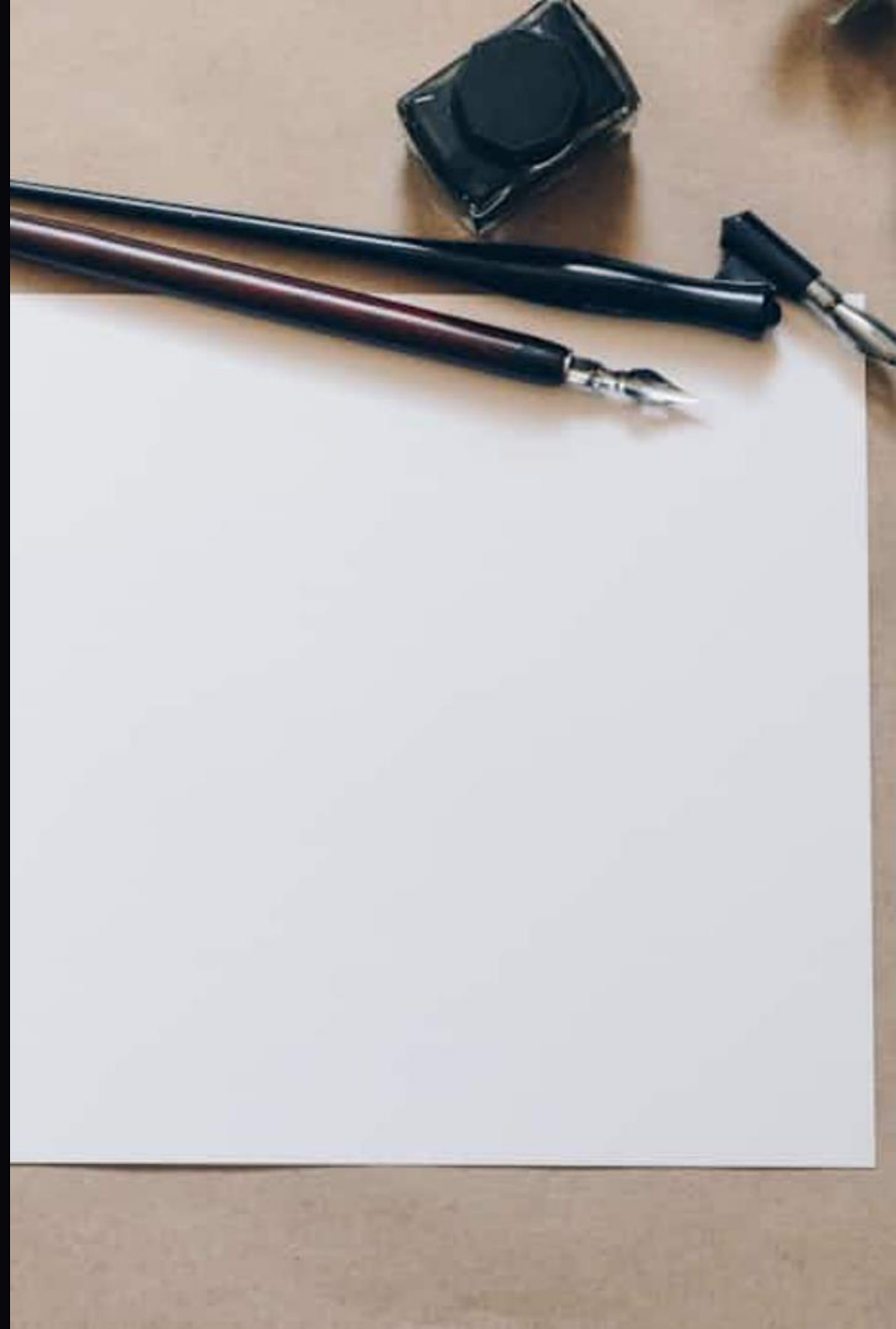
Arabic script's right-to-left direction and diacritical marks pose challenges for text processing.

Rich Semantic Nuance

Arabic poetry is characterized by intricate rhyme schemes and poetic devices.

Limited Training Data

The availability of annotated Arabic poetry datasets is limited, hindering model training.



Results:

- Generated poetry examples:

- Shorter sequences produced more coherent lines.
- Longer sequences favored frequent words.

- Evaluation:

- Model-2 outperformed others in coherence.

```
Generating text for Model_1...
Generated text for Model_1: عينيك السبات كمنجات تلمح يغضبه وعاثوا شيرين

Generating text for Model_2...
Generated text for Model_2: عينيك باعناق والفداء خراءب لكفر قتلاها صادم

Generating text for Model_5...
Generated text for Model_5: عينيك ومناي تهربين واملا مقتبس بمشيب يطلبه

Generating text for Model_10...
Generated text for Model_10: عينيك راعفا المءنث اوجهم اجواء رغيبة وتنقل
Log probability for Model_1: -64.103
Log probability for Model_2: -74.213
Log probability for Model_5: -33.622
Log probability for Model_10: -31.395
```

Conclusion:

Our research demonstrates the potential of deep learning to preserve and revitalize Arabic poetry. We invite further research to address the limitations and expand the model's capabilities. Together, we can leverage AI to safeguard this valuable cultural heritage and inspire future generations of poets.