

# Learning Meters of Arabic Poems with Recurrent Neural Networks

A step forward for language understanding and synthesis

Mostafa A. Mahmoud

A Thesis Presented to the Faculty  
of Information Technology and Computer Science  
Nile University  
In Partial Fulfillment  
of the Requirements for the Degree  
of Master of Science

Supervised By,  
Prof. Samhaa El-Beltagy  
Assoc. Prof. Waleed A. Yousef

# Table of Contents I

- 1 Introduction
- 2 Literature Review
- 3 Datasets Design
- 4 Network Architecture and Training
- 5 Experiments and Results
- 6 Discussions
- 7 Appendix

# Introduction

# But ... What is poetry?

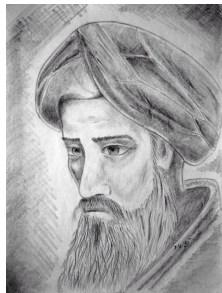
## Definition

**Poetry** is a piece of writing or speaking, which **MUST** follow specific **Patterns**.

## Example

وَلَمْ تَسْتَحْيِ فَاصْنَعْ مَا تَشَاءُ			إِذَا لَمْ تَخْشَ عَاقِبَةَ اللَّيَالِي		
تَشَاءُوْ	يَ فَصْنَعْ مَا	وَلَمْ تَسْتَحْ	لَيَالِي	شَ عَاقِبَةَ لْ	إِذَا لَمْ تَخْ
0/0//	0/0/0//	0/0/0//	0/0//	0/0/0//	0/0/0//
مُفَاعَلْ	مُفَاعَلَتُنْ	مُفَاعَلَتُنْ	مُفَاعَلْ	مُفَاعَلَتُنْ	مُفَاعَلَتُنْ

*Al-Farahidi* (718 – 786 CE) analyzed the Arabic poetry, then discovered the **Patterns**, which are the succession of consonants and vowels.



**Figure:** *Al-Farahidi*

figure taken from <https://goo.gl/ZJySa8>.

- A **poem** is a set of verses having the same meter and rhyme.

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of ِ, ُ, َ which represent the following short vowels /a/, /u/, /i/.

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of َ, ُ, ِ which represent the following short vowels /a/, /u/, /i/.
- **Consonants** carry ْ.



- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of ِ, ُ, َ which represent the following short vowels /a/, /u/, /i/.
- **Consonants** carry ْ.
- **Shadaa** indicates the letter is doubled ّ.

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of َ, ُ, ِ which represent the following short vowels /a/, /u/, /i/.
- **Consonants** carry ّ.
- **Shadaa** indicates the letter is doubled ّ.
- **Tanweenis** *harakah* and *Noon* letter with consonant at the end of the word. It sounds /n/.

# العروض Arabic Prosody

- A **foot** (*tafa'ilah* التفعيلة) is an **ordered** sequence of vowels and consonants.
- **Meter** البحر: is an **ordered** sequence of **feet**.

Meter Name	Meter feet combination
<i>al-Wafeer</i>	مفاعلتن مفاعلتن مفاعلتن مفاعلتن مفاعلتن
<i>al-Taweel</i>	فعولن مفاعيلن فعولن مفاعيلن
:	:
<i>al-Moktadib</i>	مفعولات مستفعلن مستفعلن
<i>al-Modar'e</i>	مفاعيلن فاعلاتن مفاعيلن

Feet	Scansion
فعولن	0/0//
فاعِلن	0//0/
مُسْتَفْعِلن	0//0/0/
مَفَاعِيلن	0/0/0//
مَفْعُولَات	0//0///
فَاعِلَاتن	0/0//0/
مَفَاعِلَتْن	0///0//
مُتَفَاعِلن	0//0///

إِذَا لَمْ تَخْشَ عَاقِبَةَ اللَّيَالِي      وَلَمْ تَسْتَحْيِ فَاصْنَعْ مَا تَشَاءُ  
 إِذَا لَمْ تَخْشَ      شَ عَاقِبَةَ لُ      لَيَالِي      وَلَمْ تَسْتَحْ      يَ فَصْنَعْ مَا      تَشَاءُ  
 0/0/0//      0/0/0//      0/0/0//      0/0//      0/0/0//      0/0/0//  
 مَفَاعِلَتْن      مَفَاعِلَتْن      مَفَاعِلَتْن      مَفَاعِل      مَفَاعِلَتْن      مَفَاعِل

# Thesis Working Steps.

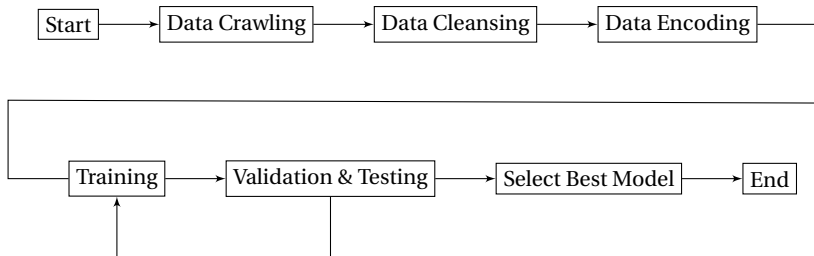


Figure: Thesis Working Steps.

# Literature Review

## Deterministic Approach

There is some literature on recognizing the meters of written Arabic poems using rule-based deterministic algorithms.

- **Abuata and Al-Omari [1]:**

## Deterministic Approach

There is some literature on recognizing the meters of written Arabic poems using rule-based deterministic algorithms.

### – **Abuata and Al-Omari [1]:**

- **Deterministic Algorithm**

- 1 Getting the input, carrying full diacritics.
- 2 Metrical scansion rules are applied to the Arud writing. 0/0/..
- 3 Grouping zero and ones to feet تنفعيلات.
- 4 A class is assigned to the input.

## Deterministic Approach

There is some literature on recognizing the meters of written Arabic poems using rule-based deterministic algorithms.

### – Abuata and Al-Omari [1]:

- Deterministic Algorithm
  - 1 Getting the input, carrying full diacritics.
  - 2 Metrical scansion rules are applied to the Arud writing. 0/0/..
  - 3 Grouping zero and ones to feet تنفعيلات.
  - 4 A class is assigned to the input.
- **Results:** 82.2% of 417 verses.



## Deterministic Approach

There is some literature on recognizing the meters of written Arabic poems using rule-based deterministic algorithms.

- **Abuata and Al-Omari [1]:**
  - Deterministic Algorithm
    - 1 Getting the input, carrying full diacritics.
    - 2 Metrical scansion rules are applied to the Arud writing. 0/0/..
    - 3 Grouping zero and ones to feet تنفعيلات.
    - 4 A class is assigned to the input.
  - **Results:** 82.2% of 417 verses.
- **Alnagdawi et al. [2],** similar approach; Context-Free Grammar; 75% correctly classed from 128.

# Machine Learning approach: Our point of departure

- Building the first intelligent model capable to learn meters from poems.

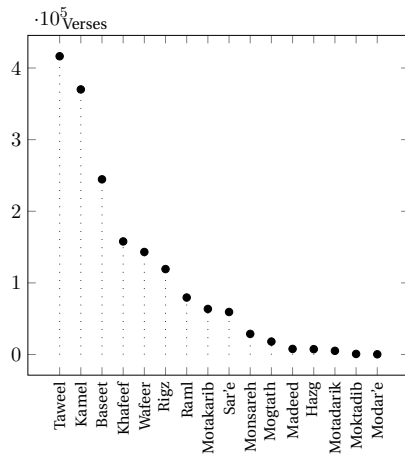
# Machine Learning approach: Our point of departure

- Building the first intelligent model capable to learn meters from poems.
- Contributing the first public available dataset of 1.7M for scientific computing.

# Datasets Design

# Dataset acquisition and cleansing

- **1,722,321** labeled data points.
- We have scrapped the Arabic datasets from الديوان [6], and الموسوعة الشعرية [5].
- Basic cleansing rules:
  - Filtering the 16 classic meters.
  - Removing unnecessary spaces.
  - Removing non-Arabic characters.
  - Factoring Shadaa and Tanween.



Diacritics	With Shadda	Without shadaa	With tanween	Without tanween
Shape	دّ	دّ	دّ	دّ + دّ

# Data Representation

- Diacritics are standalone characters.

- $\text{len مرَّحَبًا} \neq \text{len مرحبا}$
- We have represented the letter and its diacritic as a **single character**.

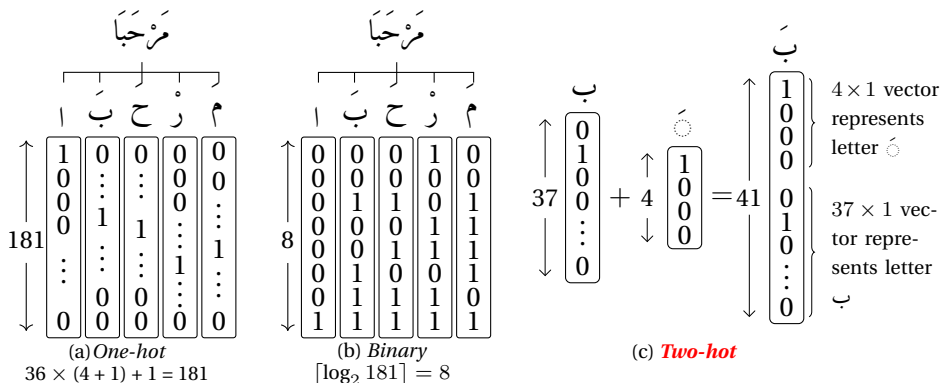


Figure: Different encoding mechanisms

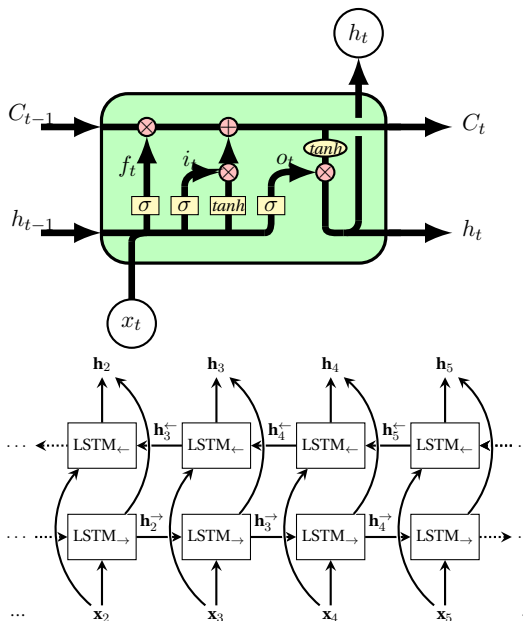
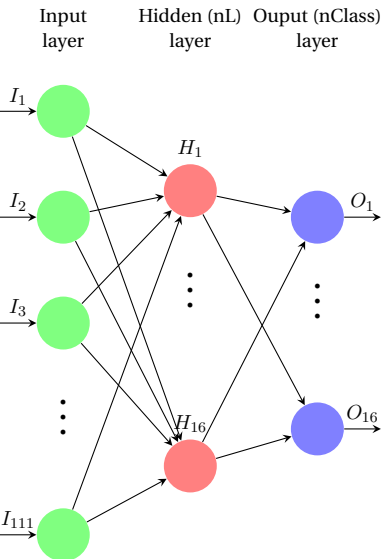
# Network Architecture and Training

# Which Network?

- **Pattern** is a sequence of characters.
- Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.
- In theory, RNNs are capable of handling long-term dependencies. However, in practice they do not, due to the **exploding gradient problem**
- LSTMs was designed to solve the long-term dependency problem using internal memory gates.



# Neural Networks Architectures



## Experiments and Results

- **Dataset Configurations** ( $3 \times 2 \times 2$ ):

- Encoding technique (3): BinE, OneE, TwoE.
- Diacritics (2): 0D, 1D.
- Trimming (2): 0T, 1T.

- **Network Configurations** ( $2 \times 2 \times 2 \times 2$ ):

- Loss functions (2): *Weighted* or *Non-Weighted* (**1, 0**) respectively.
- The number of layers (2): nL.
- The number of cell units (2): nU.
- Cell type (2): LSTM, Bi-LSTM.

## Total Experiments Configurations

Dataset Conf. (12)  $\times$  Network Conf. (16) = 192 Experiment.

# Overall Accuracy!

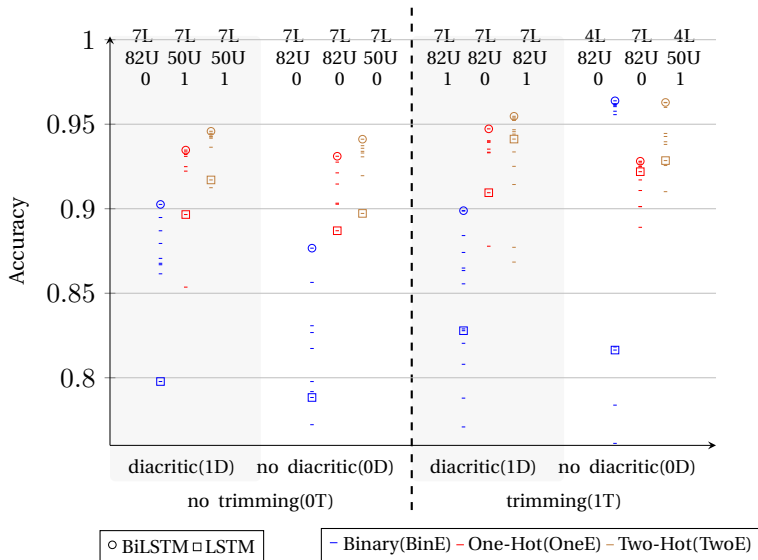
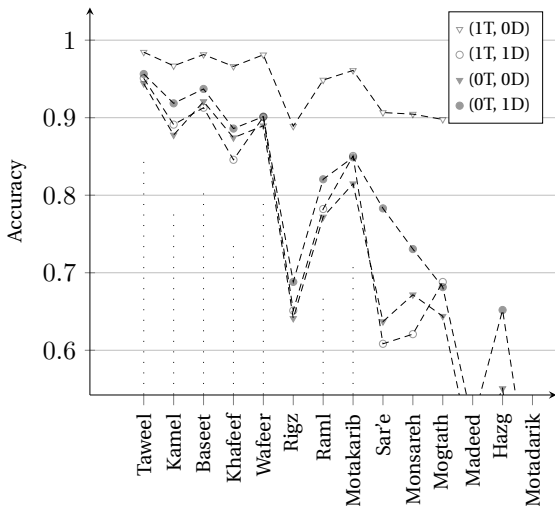


Figure: Overall accuracy of the 192 experiments

# Detailed Analysis for Overall Accuracy winner!

Ref.	Accuracy	Test Size
[2]	75%	128
[1]	82.2%	417
DNN	96.38%	150,000

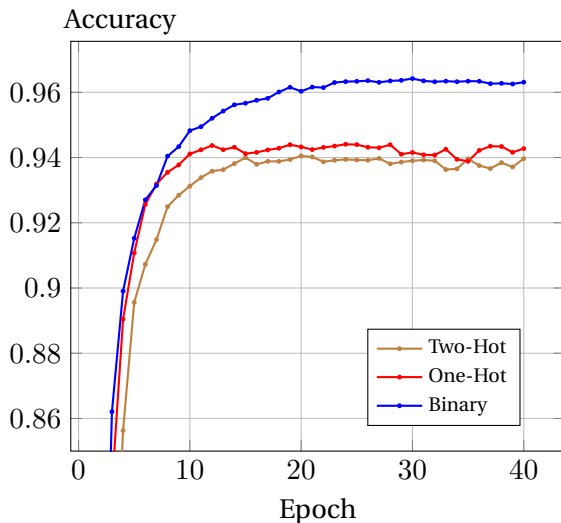
**Table:** Overall accuracy of this article compared to literature.



**Figure:** The per-class accuracy score of the best four models.

# Discussions

# Encoding effect



**Figure:** Encoding effect on Learning rate with the best model (1T, 0D, 4L, 82U, 0W, BinE) and when using the two other encodings instead of BinE.

## Encoding

- The encoding method is a transformer function  $\mathcal{T}$ , which transforms discrete input values  $X$ .
- It is assumed the network  $\eta_1$  is the most accurate network which can “decode”  $\mathcal{T}(X)$ .
- If we have another encoding function  $\mathcal{T}_2$  and we try to use the same network  $\eta_1$  for the  $\mathcal{T}_2$  as  $\eta_1(\mathcal{T}_1(X)) = (\eta_1 \cdot \mathcal{T}_1 \cdot \mathcal{T}_2^{-1})(\mathcal{T}_2(X))$ . This network may be of complicated architecture to be able to “decode” a terse or complex pattern  $\mathcal{T}_2(X)$ .



# Classifying Arabic Non-Poem Text

We found 99.2% of our 150k testing observations which correctly classified have a score range between 0.94 and 1.0.

## Arabic Article

قاد الدولي المصري محمد صلاح فريقه ليفربول للعودة إلى صدارة الدوري الإنجليزي الممتاز، بعد الفوز على ضيفه بورنموث بثلاثية نظيفة، خلال المباراة التي جمعتها مساء السبت بالجولة الـ 26 من المسابقة. ونستعرض في التقرير التالي أبرز الأرقام التي حققها صاحب الـ 26 عاماً بعد العودة للتسجيل أمام بورنموث: يعد بورنموث بوابة صلاح للعودة للتسجيل هذا الموسم في بريميرليج

source: <https://www.yallakora.com/epl/2545/News/360950/>

خلال المباراة التي جمعتها							
مساء السبت بالجولة الـ 26 من المسابقة				مساء السبت بالجولة الـ 26 من المسابقة			
خلال	مباراة	لتي	جمعتها	مساء	سبت	الجولة	الـ 26 من المسابقة
خلال	مباراة	لتي	جمعتها	مساء	سبت	الجولة	الـ 26 من المسابقة
0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//
0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//	0/0/0//
مفاعيلن	مفاعيلن	مفاعيلن	مفاعيلن	مفاعيلن	مفاعيلن	مفاعيلن	مفاعيلن

# Bibliography



## (1) Abuata, Belal and Al-Omari, Asma

A Rule-Based Algorithm for the Detection of Arud Meter in Classical Arabic Poetry  
*International Arab Journal of Information Technology*. (2017), 15.



## (2) Alnagdawi, Mohammad and Rashaideh, Hasan and Aburumman, Ala

Finding Arabic Poem Meter Using Context Free Grammar  
*J. of Commun. & Comput. Eng.* (2013), 3, 52-59.



## (3) Colah

Understanding Lstm Networks

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>, 2015.



## (4) Petar Veličković

Collection of Latex Tikz figures

<https://github.com/PetarV-/TikZ>.



## (5) الموسوعة الشعرية

Department of Culture and Tourism – Abu Dhabi

<https://poetry.dctabudhabi.ae>.



## (6) الديوان

Al-Diwan website

<https://www.aldiwan.net>.

# Questions!

Questions.

# Appendix

# RNN, Architectures

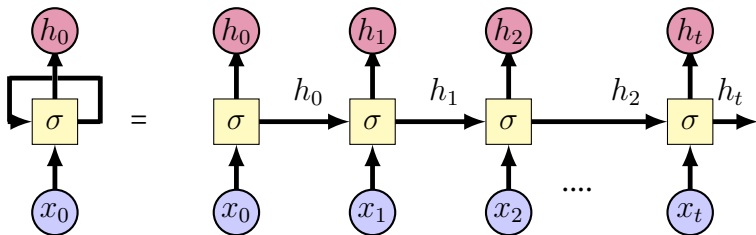


Figure: Recurrent Neural Networks Loops adapted from [3]

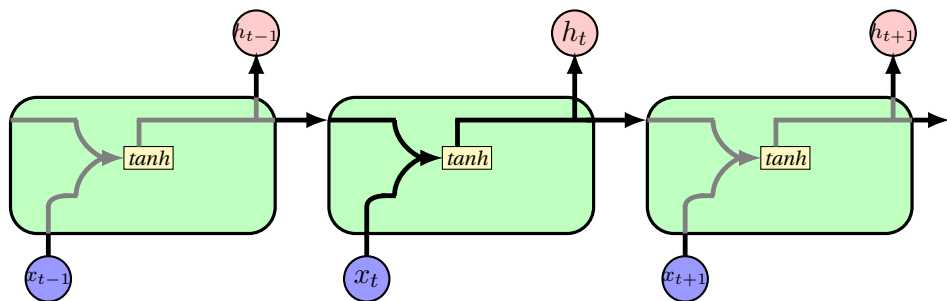


Figure: A single recurrent layer adapted from [3]

# LSTM Architectures

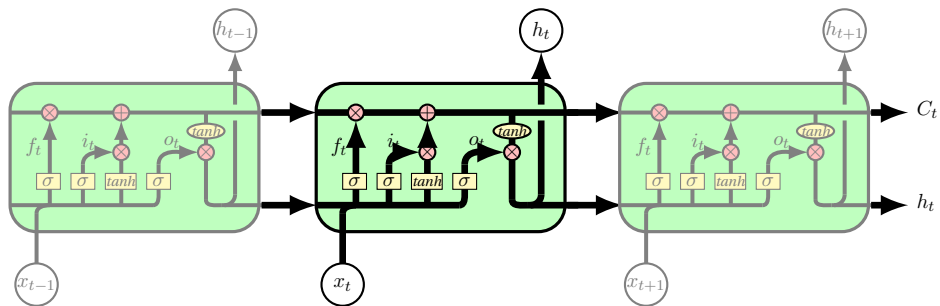


Figure: Unfold LSTM adapted from [3]

## Bi-LSTM Motivation

- *Harry* is the king, and he will travel next week.
  - The new book which makes the big sale is named *Harry* Potter.
- 
- Bi-LSTM models always outperform LSTM models.
  - This means that models cannot learn the pattern from one direction; it should be two directions together.



# LSTM Architectures

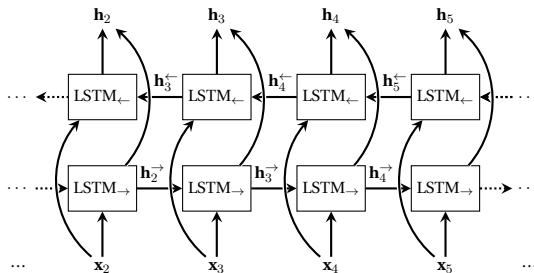


Figure: bidirectional long short-term memory [4]

# Classifying Arabic Non-Poem Text

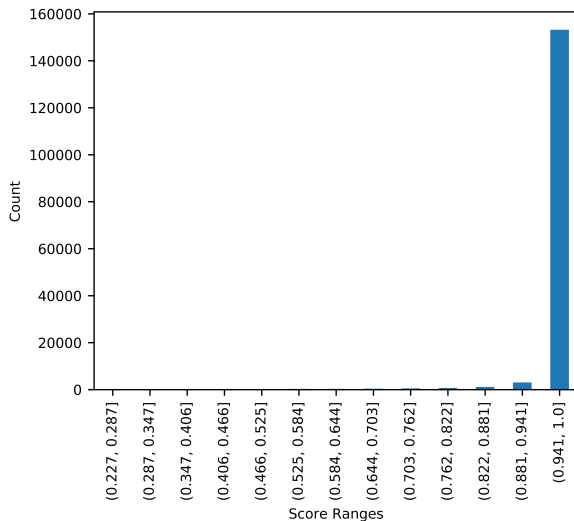


Figure: Testing data score ranges distribution.

# Does Semantic Matter?

تدفق في البطحاء بعد تبهطل  
وسار بأركان العقيش مقرنصا  
يقول: وما بال البحاط مقرمطا  
إذا أقبل البحراط طاح بهمة  
يكاد على فرط الحطيف ييقبق  
وقعق في البیداء غیر مزرکلی  
وهام بكل القارطات بشنکلی  
ویسعی دواما بین هک وهنکلی  
وإن أقرط البحطوش ناء بكلکلی  
ویضرب ما بین الهماط وکندل