GERMAN
INTERNATIONAL
UNIVERSITY
الجامعة الألمانية الدولية

Automatic Grading of Short Answers in Arabic Language

Presented by: Sohila Arafa - 1002329

Under supervision of: Dr. Nada Sharaf

Table of contents

01
Introduction

Motivation and the aim of the project

02
Background

An explanation for topics we need for the methodology

03
Methodology

How the project is implemented

04
Results & Analysis

Evaluating results

05
Conclusion

A summary of the project

01

Introduction

Motivation

Assessing students is an important part of the learning process.

Type of Assessment:

- MCQ
- Essay Questions
- Short Answer Questions

Manual grading is an intense process, especially for text answers.

Challenges:

- Full understanding and analysis of the answer
- Short answer grading can be inconsistent

Aim of Project

The aim of this project was to develop an approach to automate the grading process of short-answer exams in the Arabic Language.

This could be done through :

- Machine Learning (ML)
- Deep Learning (DL)
- Natural Language Processing (NLP)

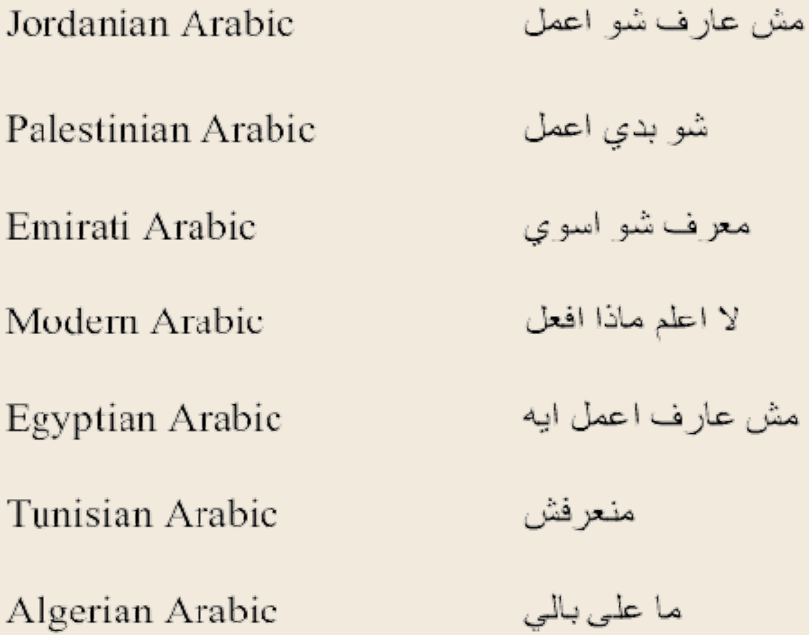
Natural Language Processing is an important part of this task. This is due to the fact that the model will not understand the natural language without it.

02

Background

Arabic Challenges

1. **Dialectal variation:** Arabic has a lot of versions. However, exams are often solved in "Standard Arabic"

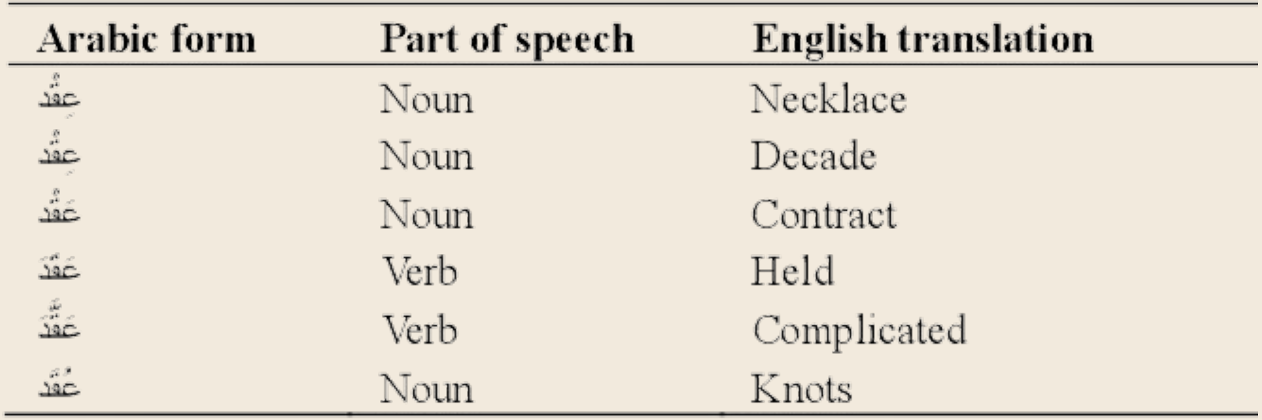


Jordanian Arabic	مش عارف شو اعمل
Palestinian Arabic	شو بدي اعمل
Emirati Arabic	معرف شو اسوي
Modern Arabic	لا اعلم ماذا افعل
Egyptian Arabic	مش عارف اعمل ايه
Tunisian Arabic	منعرفش
Algerian Arabic	ما على بالي

Arabic Challenges

2. **Complex Grammar Structure:** All verbs in Arabic have a root usually from three letters making it highly derivation.

3. **Lexical Ambiguity:** one word could have multiple meanings.



Arabic form	Part of speech	English translation
عِقْد	Noun	Necklace
عِقْد	Noun	Decade
عَقْد	Noun	Contract
عَقَدَ	Verb	Held
عَقَّدَ	Verb	Complicated
عُقَد	Noun	Knots

Code Switching

Code-switching is alternating between two or more languages or language varieties in a single conversation.

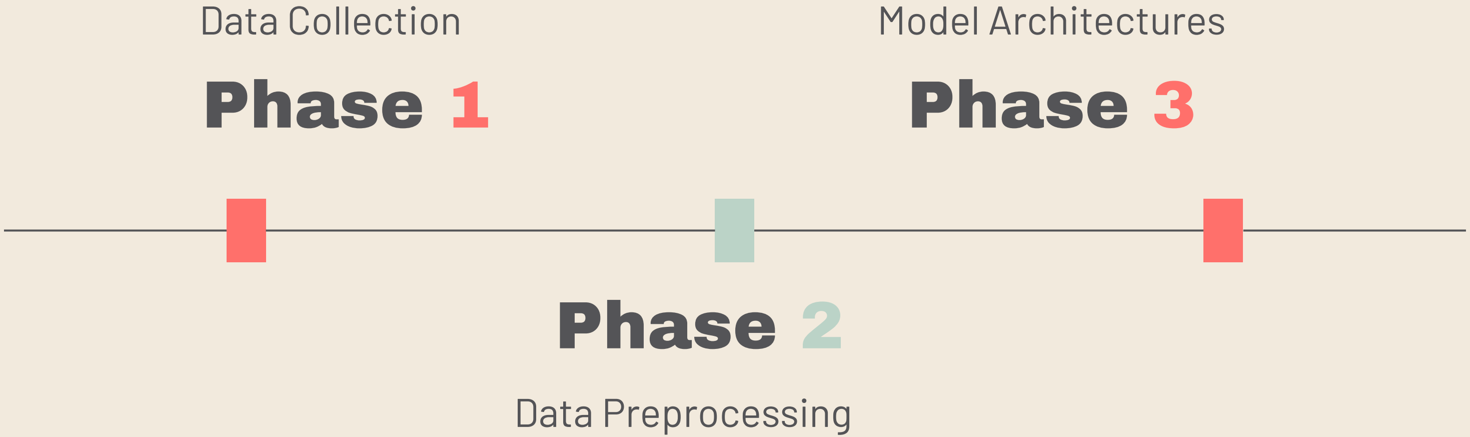
An example of code-switching in our community could be seen on ``Social Media". It is easier to many people to express themselves by
alternating between English and Arabic.

Want to help me make un castillo?
¡Sí! Let's make a castle grande!

03

Methodology

Methodology



Data Collection

Phase 1

Phase 2

Data Preprocessing

Model Architectures

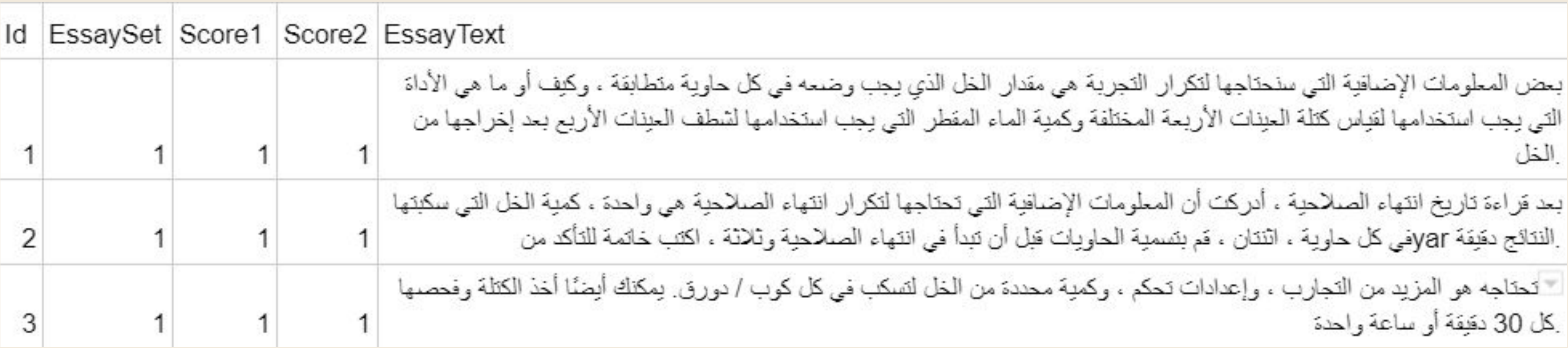
Phase 3

Data set

The automated-student assessment prize - short answer scoring dataset by the Hewlett Foundation was used.

The dataset consists of 10 questions and 17043 rows.

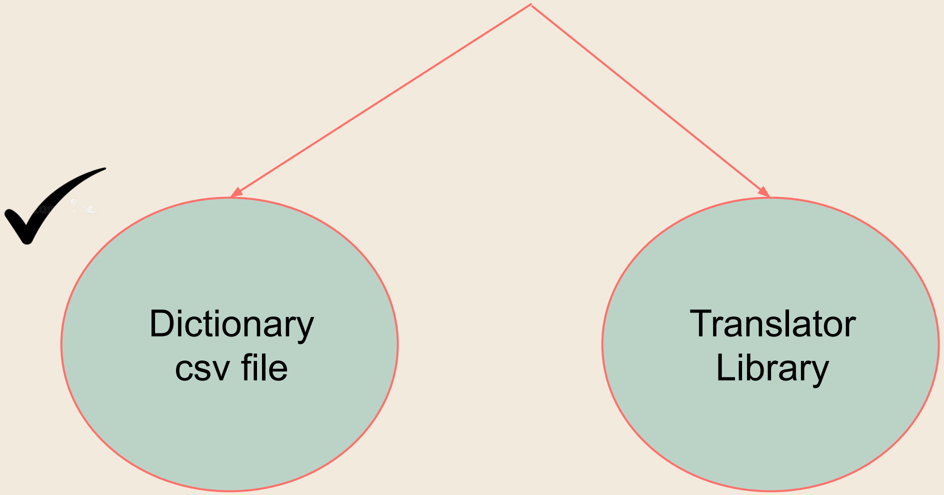
- average of 1704 answers per question
- average length of 36 words per response.



Id	EssaySet	Score1	Score2	EssayText
1	1	1	1	بعض المعلومات الإضافية التي سنحتاجها لتكرار التجربة هي مقدار الخل الذي يجب وضعه في كل حاوية متطابقة ، وكيف أو ما هي الأداة التي يجب استخدامها لقياس كتلة العينات الأربعة المختلفة وكمية الماء المقطر التي يجب استخدامها لشطف العينات الأربع بعد إخراجها من الخل.
2	1	1	1	بعد قراءة تاريخ انتهاء الصلاحية ، أدركت أن المعلومات الإضافية التي تحتاجها لتكرار انتهاء الصلاحية هي واحدة ، كمية الخل التي سكبتها النتائج دقيقة yarفي كل حاوية ، اثنتان ، قم بتسمية الحاويات قبل أن تبدأ في انتهاء الصلاحية وثلاثة ، اكتب خاتمة للتأكد من.
3	1	1	1	تحتاجه هو المزيد من التجارب ، وإعدادات تحكم ، وكمية محددة من الخل لتسكب في كل كوب / دورق. يمكنك أيضًا أخذ الكتلة وفحصها كل 30 دقيقة أو ساعة واحدة.

Code-Switch Dataset

From our dataset, a code switch dataset was generated, there were 2 approaches



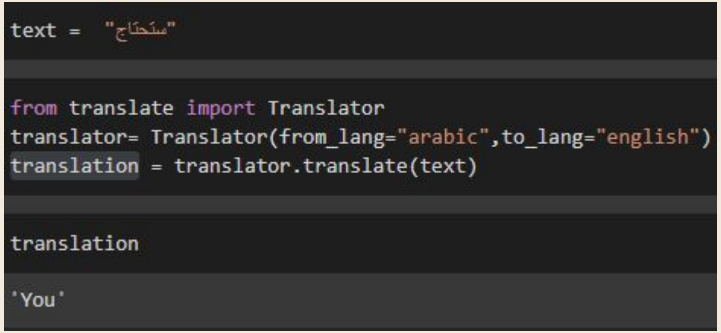
✓

Dictionary csv file

Translator Library

Dictionary Sample

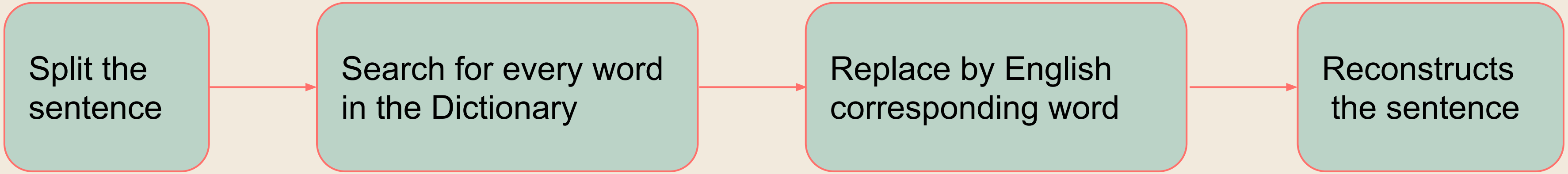
Arabic	English
سبب	reason
تكرار	repeating
التجربة	The experiment



```
text =  "ستحتاج"

from translate import Translator
translator= Translator(from_lang="arabic",to_lang="english")
translation = translator.translate(text)

translation

'You'
```

The library mistranslates some words.
So the Dictionary approach was chosen over it.

Code-Switch Dataset

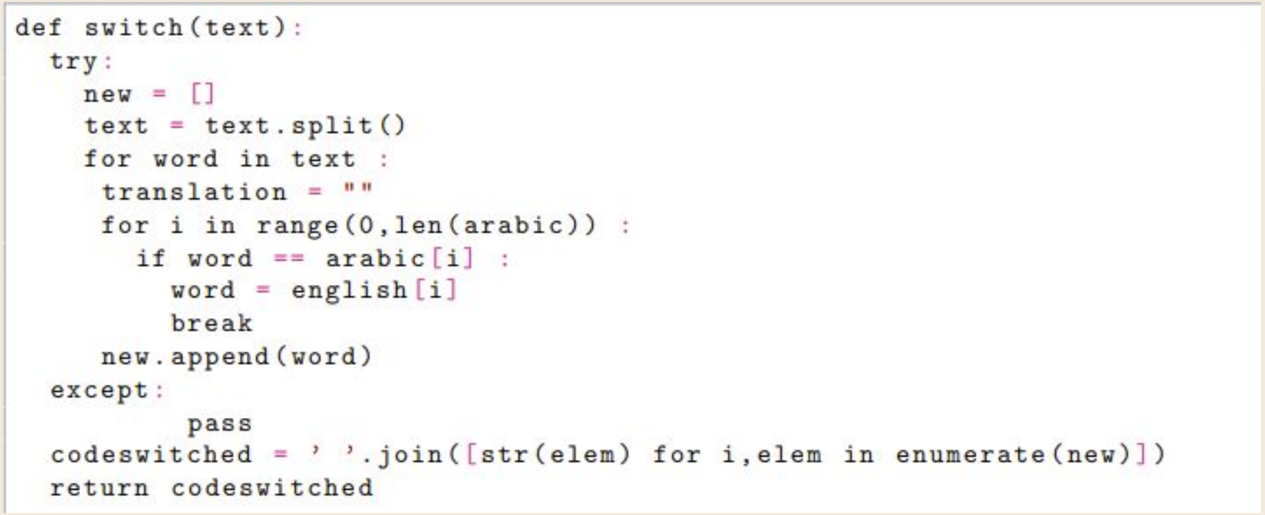


Split the sentence

Search for every word in the Dictionary

Replace by English corresponding word

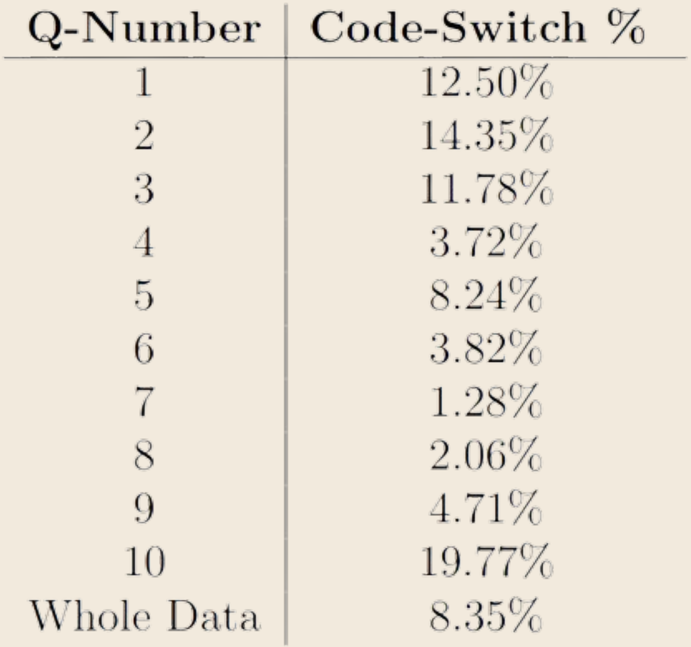
Reconstructs the sentence



```
def switch(text):
  try:
    new = []
    text = text.split()
    for word in text :
     translation = ""
     for i in range(0,len(arabic)) :
       if word == arabic[i] :
         word = english[i]
         break
     new.append(word)
  except:
          pass
  codeswitched = ’ ’.join([str(elem) for i,elem in enumerate(new)])
  return codeswitched
```

Code-Switch Dataset

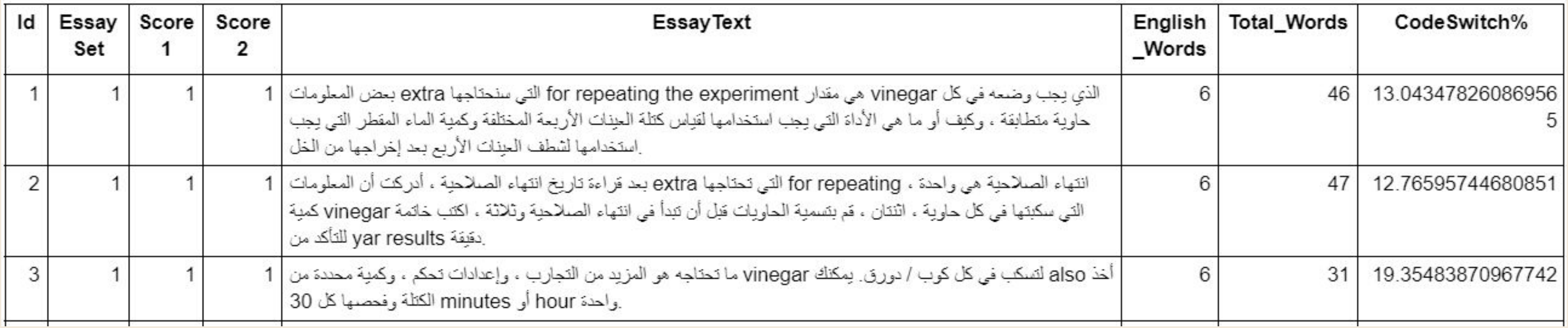
The percentage of the code-switch for every question and for the whole dataset was calculated.



Q-Number	Code-Switch %
1	12.50%
2	14.35%
3	11.78%
4	3.72%
5	8.24%
6	3.82%
7	1.28%
8	2.06%
9	4.71%
10	19.77%
Whole Data	8.35%

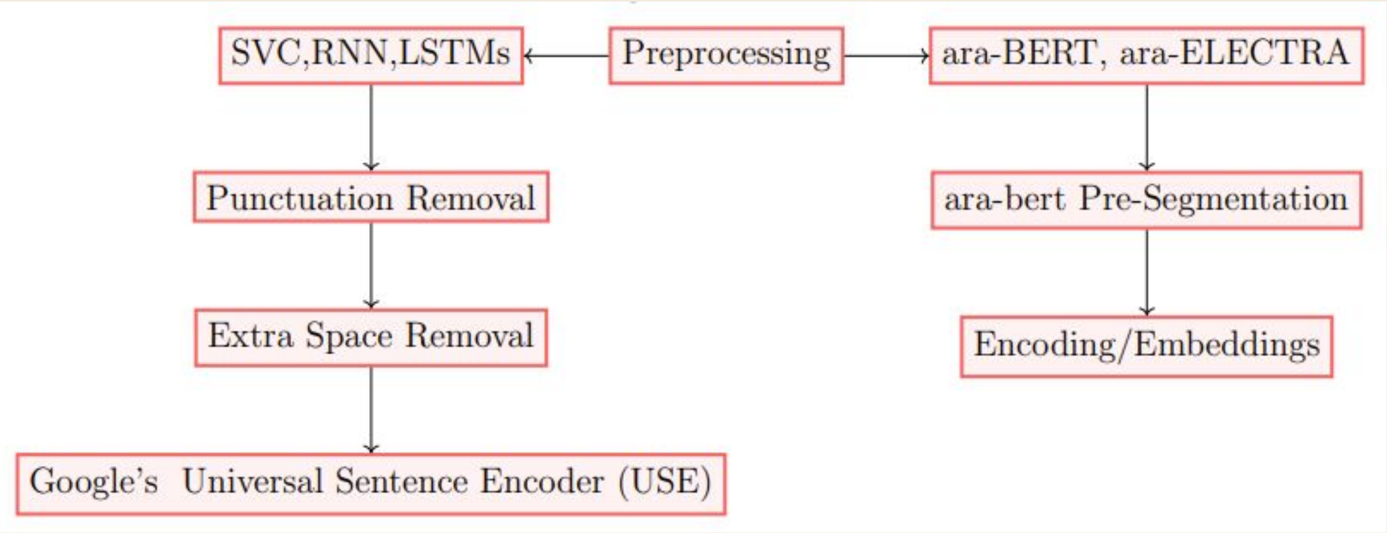
Code-Switch Dataset

The table shows the final representation of the data before the preprocessing phase.



Id	Essay Set	Score 1	Score 2	EssayText	English _Words	Total_Words	CodeSwitch%
1	1	1	1	الذي يجب وضعه في كل vinegar هي مقدار for repeating the experiment التي سنحتاجها extra بعض المعلومات حاوية متطابقة ، وكيف أو ما هي الأداة التي يجب استخدامها لقياس كتلة العينات الأربعة المختلفة وكمية الماء المقطر التي يجب استخدامها لشطف العينات الأربع بعد إخراجها من الخل.	6	46	13.043478260869565
2	1	1	1	انتهاء الصلاحية هي واحدة ، for repeating التي تحتاجها extra بعد قراءة تاريخ انتهاء الصلاحية ، أدركت أن المعلومات التي سكبتها في كل حاوية ، اثنتان ، قم بتسمية الحاويات قبل أن تبدأ في انتهاء الصلاحية وثلاثة ، اكتب خاتمة vinegar كمية دقيقة. yar results للتأكد من	6	47	12.76595744680851
3	1	1	1	أخذ also لتسكب في كل كوب / دورق. يمكنك vinegar ما تحتاجه هو المزيد من التجارب ، وإعدادات تحكم ، وكمية محددة من واحدة. hour أو minutes الكتلة وفحصها كل 30	6	31	19.35483870967742

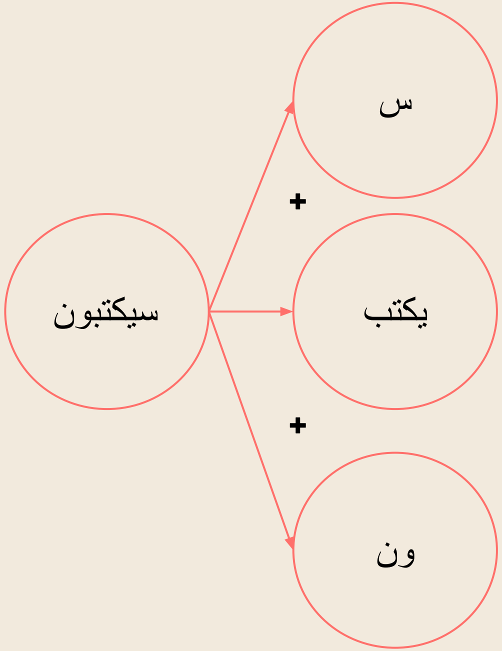
Data Preprocessing



- USE : it encodes the sentence into high-dimensional vectors to be in an understandable format to the model, and each sentence has a vector of size 512.

- arabert/araelectra: The sentence was encoded with a maximum length of 450. Each sentence is encoded into arrays of Input IDs and Attention Mask.

AraBERT Pre-segmentation



This helps when we have the phrase: “يكتب” and “سيكتب”

BERT word segmentation in the word vocabulary it will take three places:

1. 2.

سيكتب = س + ##يكتب

3.

يكتب = يكتب

using this pre-segmentation we will have 2 tokens instead of 3 tokens.

Data Preprocessing

The data for the two processes were divided into 80-20%

80% were used for the training set
20% were used for the testing set.

The training set was divided into 90-10 training and validation sets, 90% were used for the training set and 10% for the validation

Models Architecture

According to the work represented by Omar Nael in [1], [2]. These are the models that performs the best with our dataset

01 Lazy Classifier + SVC + USE

02 RNN + USE

03 LSTM + USE

04 Bidirectional LSTM + USE

05 BERT + AraBERT

06 ELECTRA + AraELECTRA

Lazy Classifier + SVC + USE

Lazy classifier was used across all the popular classifiers.

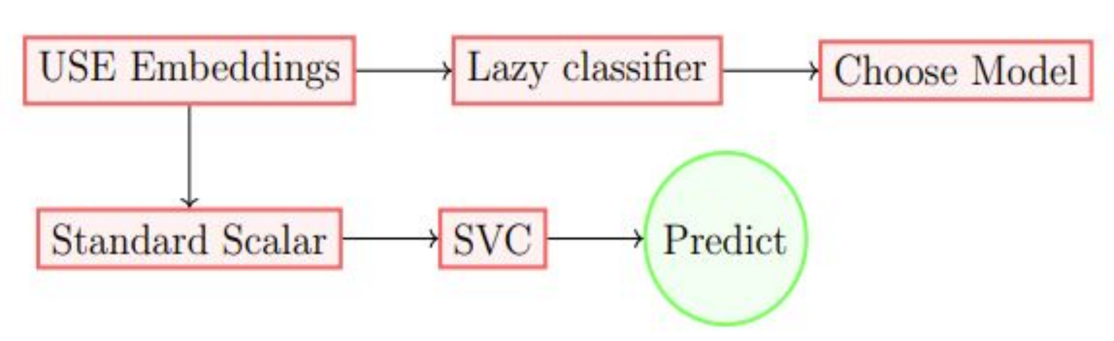
- Choose the model with the highest accuracy.

SVC is the one with highest accuracy.

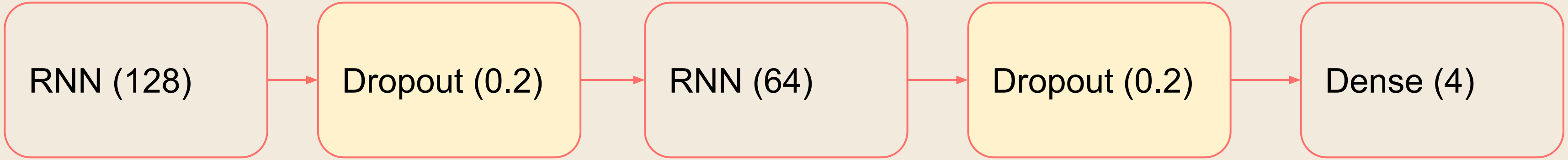
- USE produce 3-dimensions embeddings, so we transformed it to 2-dimensions to fit in with Machine Learning models.

- The StandardScaler transformer:

  standardizes the data by subtracting the mean and dividing it by the standard deviation of each feature.



RNN + USE



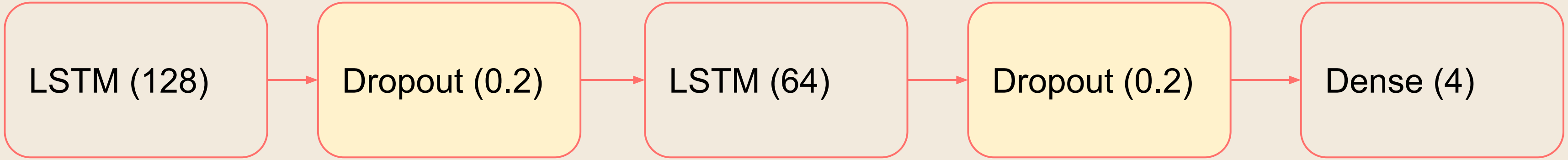
RNN (128) → Dropout (0.2) → RNN (64) → Dropout (0.2) → Dense (4)

- Sigmoid function were used for RNN layers.
- A dropout layer of 0.2 was added
- Dense layer with 4 output units that correspond to the 4 classes of the dataset (Softmax function)

The dropout layer prevent the dataset from overfitting too early

Adam and sparse categorical cross-entropy were used. The batch size used was 32 and learning rate of 1e-4

LSTM + USE
Bidirectional LSTM + USE



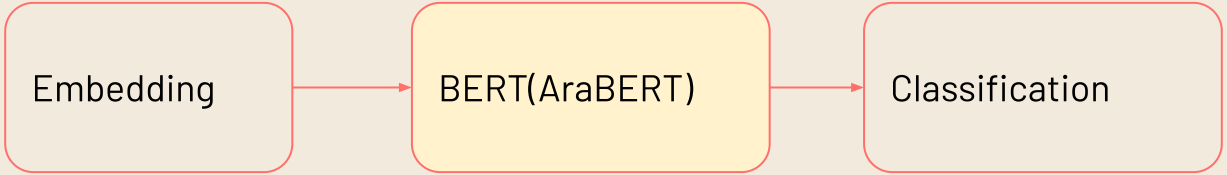
LSTM (128) → Dropout (0.2) → LSTM (64) → Dropout (0.2) → Dense (4)

- Sigmoid function were used for LSTM layers.
- A dropout layer of 0.2 was added
- Dense layer with 4 output units that correspond to the 4 classes of the dataset (Softmax function)

The dropout layer prevent the dataset from overfitting too early

Adam and sparse categorical cross-entropy were used. The batch size used was 32 and learning rate of 1e-4

AraBERT

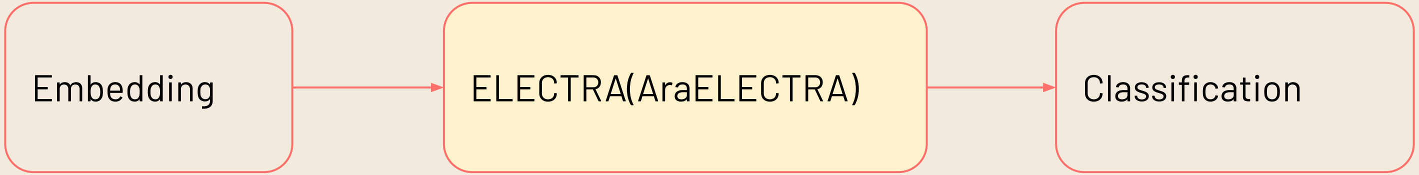


From hugging face library, **BertForSequenceClassification** was used, the pre-trained ELECTRA model with a single linear classification layer on top.

- Passing the model “araBERT-base-v2”.
- Number of labels to 4 (0,1,2,3).
- Adam was used as an optimizer with a learning rate of 1e-4.
- Number of epochs was set to 4 and the batch size was set to 8

Adam used to adjust the learning rate for each parameter individually, rather than using a fixed learning rate for all parameters. The learning rate is adjusted based on the past gradients and the current gradient.

AraELECTRA



Embedding

ELECTRA(AraELECTRA)

Classification

From hugging face library, **ElectraForSequenceClassification** was used, the pre-trained ELECTRA model with a single linear classification layer on top.

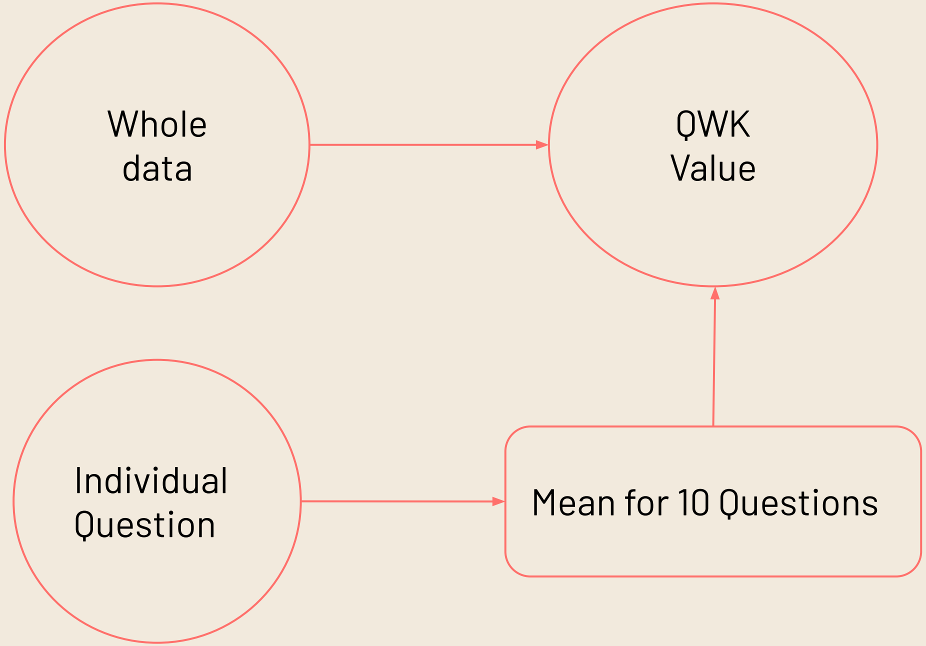
- Passing the model “araelectra-base”.
- Number of labels to 4 (0,1,2,3).
- Adam was used as an optimizer with a learning rate of 1e-4.
- Number of epochs was set to 4 and the batch size was set to 8

Adam used to adjust the learning rate for each parameter individually, rather than using a fixed learning rate for all parameters. The learning rate is adjusted based on the past gradients and the current gradient.

04

Results & Analysis

Evaluating



The Quadratic Weighted Kappa (QWK) was used as an evaluation metric across all models,

it measures the agreement between 2 raters so it is more suitable for the task.

SVC Results

QNum	CodeSwitch%	SVC
1	12.5	0.73
2	14.4	0.55
3	11.8	0.58
4	3.7	0.62
5	8.2	0.53
6	3.8	0.77
7	1.3	0.40
8	2.1	0.49
9	4.7	0.71
10	18.8	0.70
Mean	-	0.61
Whole	8.4	0.71

From the table : training the whole dataset to the mean results, we can observe that,

Training SVC model on the whole dataset outperforms training on each question individually

QNum	CodeSwitch%	RNN	LSTM	Bidirectional	araBERT	araELECTRA
1	12.5	0.74	0.74	0.73	0.82	0.79
2	14.4	0.56	0.56	0.56	0.72	0.73
3	11.8	0.63	0.65	0.64	0.66	0.64
4	3.7	0.67	0.67	0.67	0.64	0.65
5	8.2	0.67	0.70	0.69	0.75	0.77
6	3.8	0.81	0.82	0.82	0.83	0.81
7	1.3	0.43	0.43	0.42	0.68	0.71
8	2.1	0.53	0.55	0.51	0.62	0.58
9	4.7	0.73	0.74	0.73	0.81	0.78
10	18.8	0.70	0.70	0.69	0.71	0.79
Mean	-	0.65	0.66	0.65	0.72	0.71
Whole	8.4	0.72	0.72	0.72	0.74	0.74

Deep Learning Models Analysis

- No significant difference between RNN, LSTM, and Bi-directional LSTM on either training on individual questions or the entire dataset.

- RNN requires less time to train than LSTM and Bidirectional LSTM, using RNN.

- Questions 2, 7, and 8 gave bad results in LSTM, RNN, and Bi-directional LSTM,

However, they gave better results when we used ara-BERT and ara-ELECTRA.

QNum	RNN	LSTM	Bidirectional
2	0.56	0.56	0.56
7	0.43	0.43	0.42
8	0.53	0.55	0.51
Mean	0.65	0.66	0.65
Whole	0.72	0.72	0.72

QNum	araBERT	araELECTRA
2	0.72	0.73
7	0.68	0.71
8	0.62	0.58

Deep Learning Models Analysis

- When we compared the results without preprocessing for araElectra, we found out that it performs better without preprocessing as shown in the table.

QNum	With	Without
1	0.72	0.79
8	0.56	0.58
9	0.72	0.78
10	0.69	0.79

Table contains: araELECTRA with/ without preprocesing

- No significant difference between ara-BERT and ara-ELECTRA.

QNum	RNN	LSTM	Bidirectional	araBERT	araELECTRA
Mean	0.65	0.66	0.65	0.72	0.71
Whole	0.72	0.72	0.72	0.74	0.74

Results Overview

- Comparing training SVC on the whole dataset with RNN,LSTM and Bidirectional LSTM, will conclude that there is no significant difference between the results.

  However, SVC is a machine learning model, which means it is much faster than all other models.

QNum	RNN	LSTM	Bidirectional	SVC
Whole	0.72	0.72	0.72	0.71

- The **best** results were achieved by using araELECTRA and araBERT either by training each question individually or by training the whole dataset.
  - Both require a lot more time to train than the other models, which is the trade-off.

05

Conclusion

Conclusion

- Automatic Grading of Short Answers models were built using SVC, RNN, LSTM, Bidirectional LSTM, AraBERT, and AraELECTRA.
- Code Switch dataset was built.
- Google's USE was used as an embedding layer for SVC,RNN,LSTM, and Bidirectional LSTM.
- Best results were achieved using araBERT and araELECTRA scoring 0.74 on the whole dataset.
- For future work, I recommend gathering code-switch data from native Arab people

Scan

References

Colab Notebooks

Thank you

Any Questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**