

---

# Arabic Sentiment Analysis

Team ID : 68

Name	ID	Section	Department
Martina Al-Kes Angelos	20201700624	7	CS
Mohamed Hesham	20201701245	9	CS
Omar Yasser	20201701195	6	CS
Salma Ayman Mohamed	20201700345	4	CS
Salma Mahdy Ahmed	20201700353	4	CS
Yosef Mahmoud Madboly	20201701037	12	CS

# Introduction

In the era of digital communication, analyzing sentiments in **Arabic** text is crucial due to its status as **one of the most spoken languages**. This report examines the challenges of sentiment analysis in Arabic, addressing linguistic variations, cultural subtleties, and contextual obstacles. We delve into practical methodologies, tools, and key considerations for conducting robust sentiment analysis on Arabic text.

## Dataset Overview

### Training Dataset:

The training dataset consists of diverse Arabic text from reviews, spanning **32,036** rows. With two columns,

- **"review\_description"** captures opinions—these reviews may include English text or emojis.
- The **"rating"** column provides a numerical dimension: -1 for negative, 0 for neutral, and 1 for positive sentiments.

	review_description	rating
0	...شرکه زباله و سواقين بتبرشم و مفيش حتي رقم للشك	-1
1	...خدمة الدفع عن طريق الكي نت توقفت عندي اصبح فقط	1
2	...تطبيق غبي و جاري حذفه ، عاملين اكواد خصم و لما	-1
3	...فعلا تطبيق ممتاز بس لو فى امكانية يتيح لمستخدم	1
4	...سيء جدا ، اسعار رسوم التوصيل لا تمت للواقع ب ص	-1

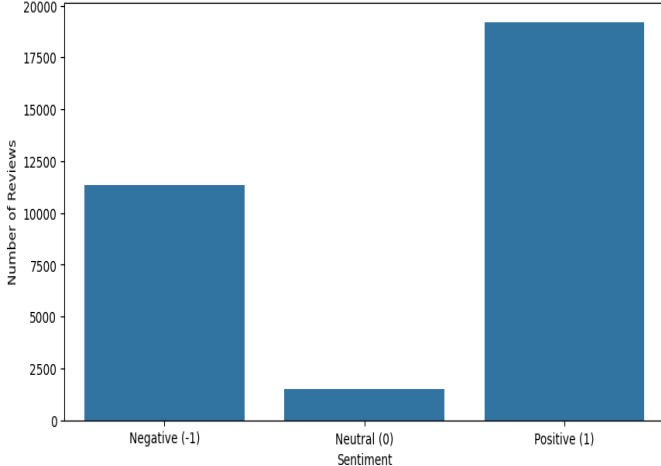
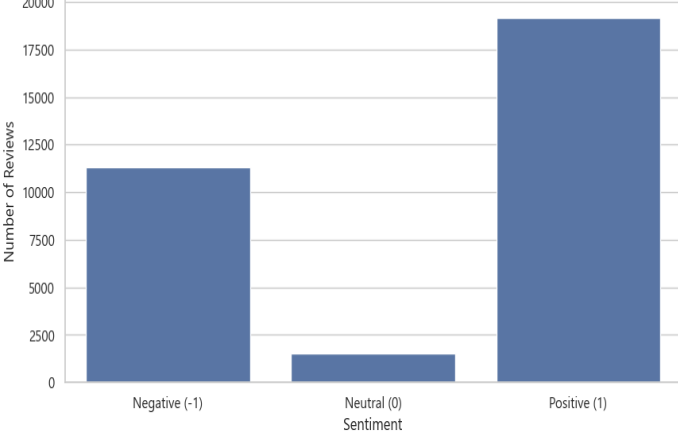
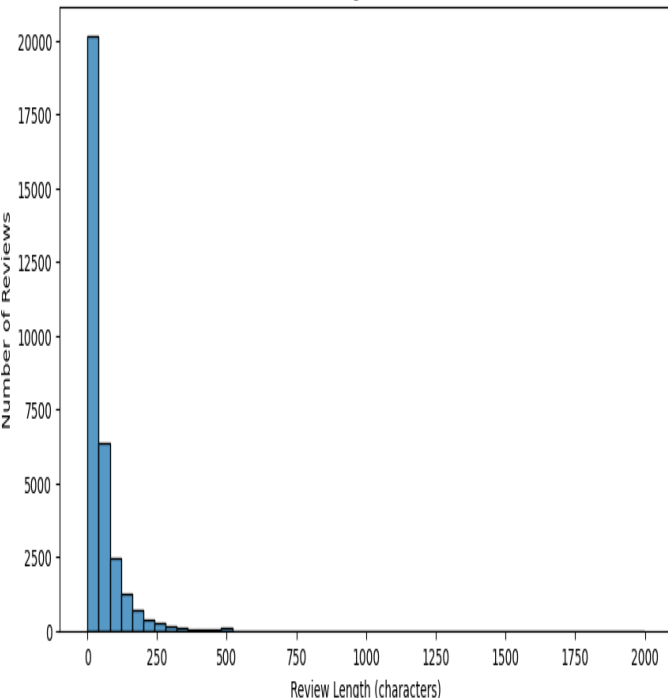
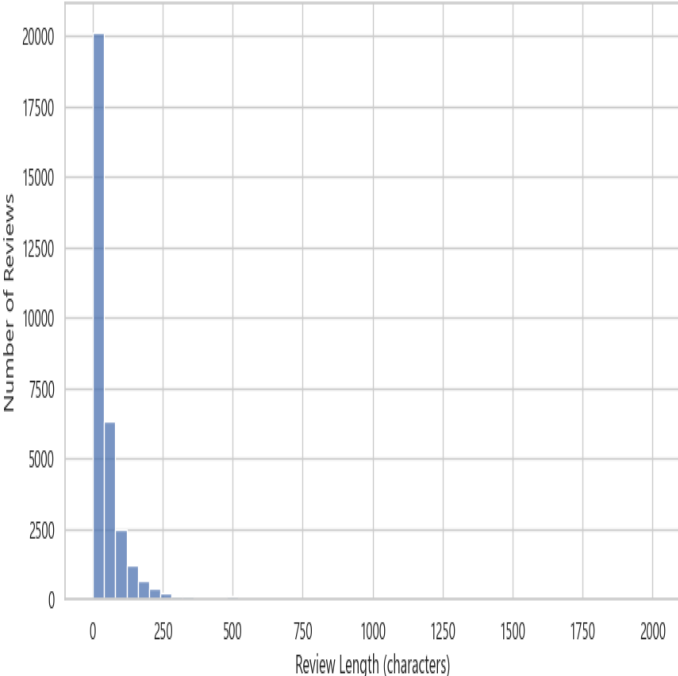
---

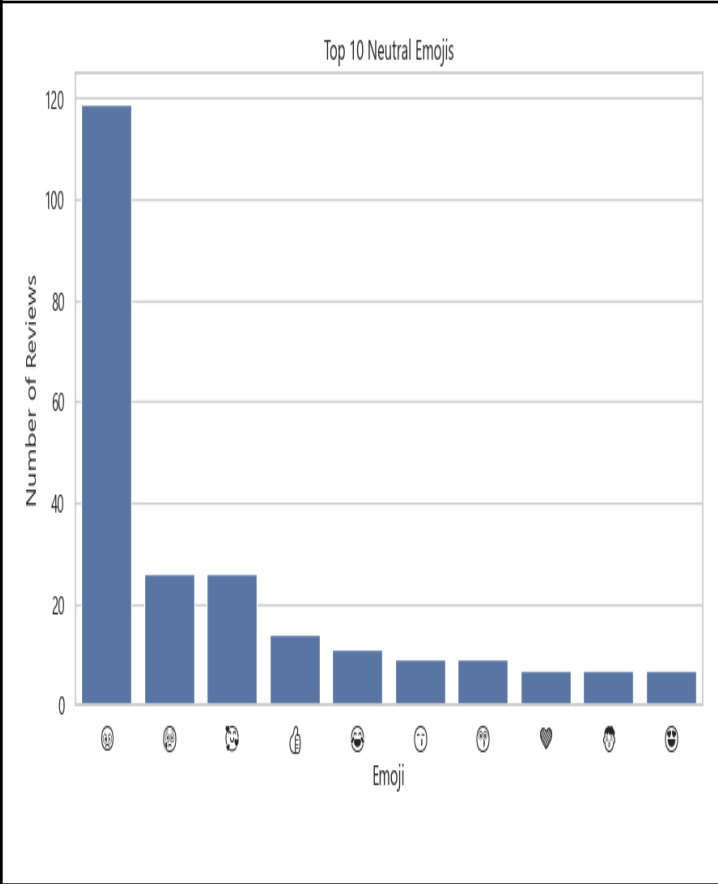
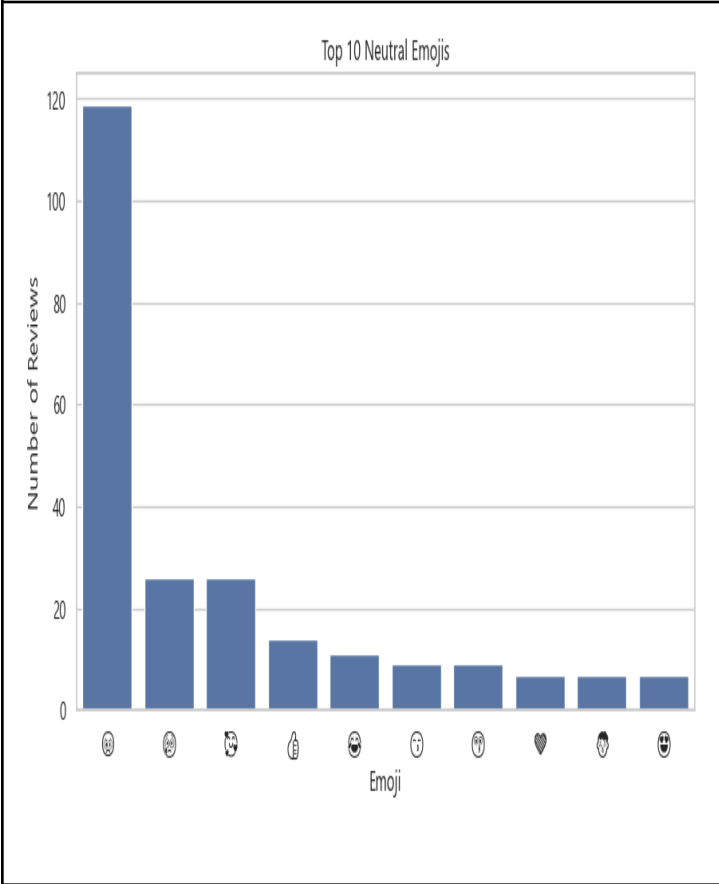
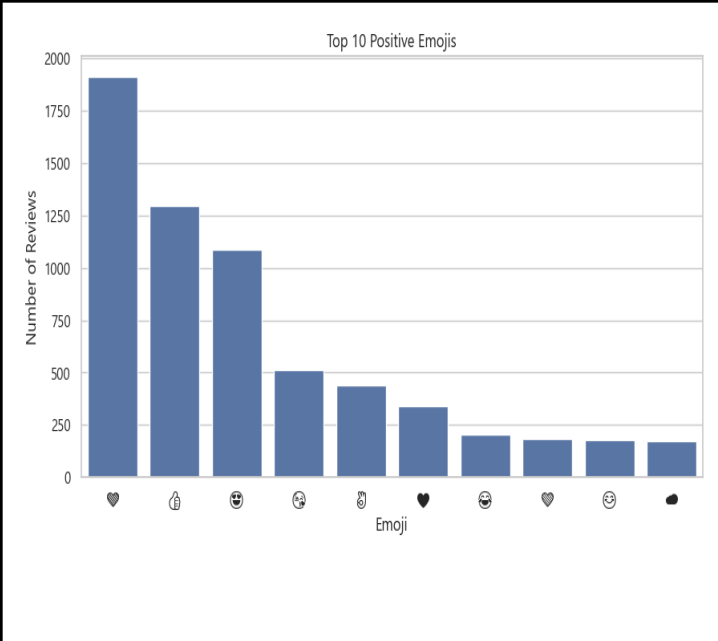
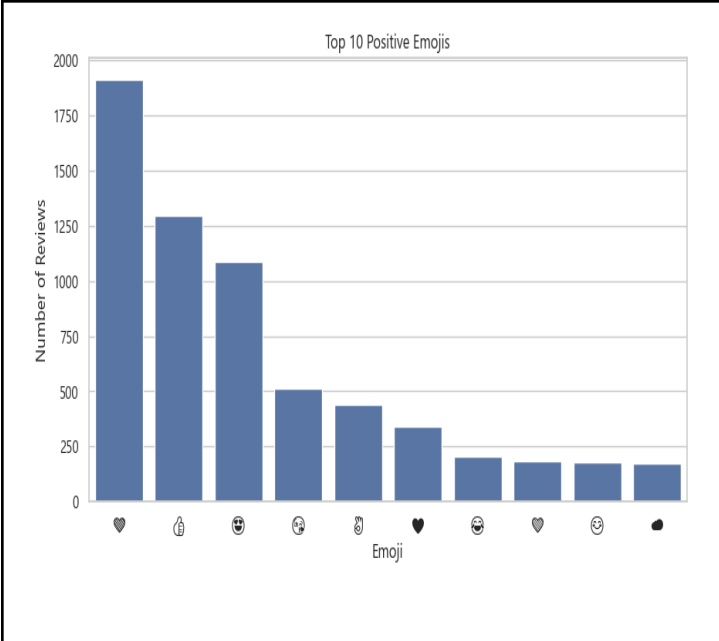
### Test Dataset:

The test dataset contains only "review\_description" and an "id" column.

ID		review_description
0	1	...اهنئكم على خدمه العملاء في المحادثه المباشره م
1	2	...ممتاز جدا ولكن اتمنى ان تكون هناك بعض المسابقا
2	3	كل محملته يقول تم ايقاف حطيت2 عشان تسوون الخطاء
3	4	شغل طيب
4	5	بعد ماجربت

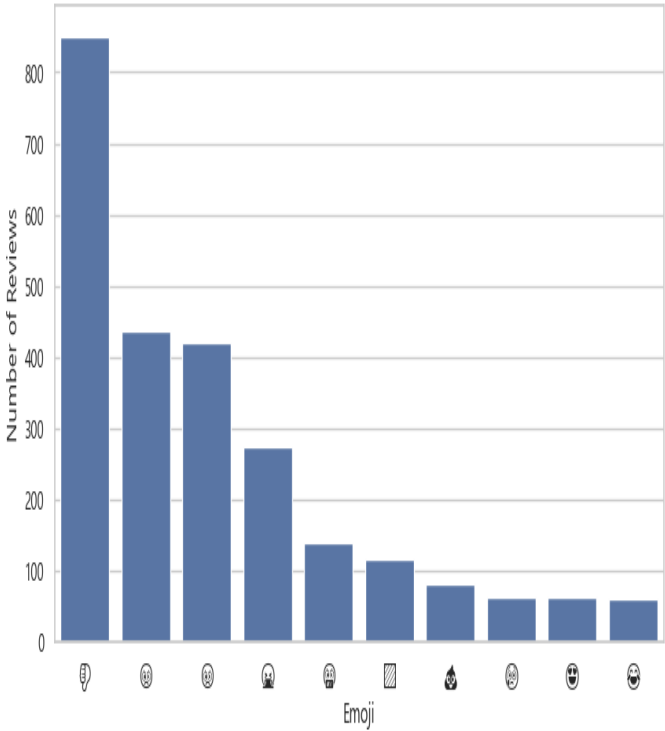
# Exploratory Data Analysis (EDA)

Before Preprocessing	After Preprocessing																																																
<div><p>Sentiment Distribution</p><table><tr><th>Sentiment</th><th>Number of Reviews</th></tr><tr><td>Negative (-1)</td><td>11500</td></tr><tr><td>Neutral (0)</td><td>1500</td></tr><tr><td>Positive (1)</td><td>19000</td></tr></table></div>	Sentiment	Number of Reviews	Negative (-1)	11500	Neutral (0)	1500	Positive (1)	19000	<div><p>Sentiment Distribution</p><table><tr><th>Sentiment</th><th>Number of Reviews</th></tr><tr><td>Negative (-1)</td><td>11500</td></tr><tr><td>Neutral (0)</td><td>1500</td></tr><tr><td>Positive (1)</td><td>19000</td></tr></table></div>	Sentiment	Number of Reviews	Negative (-1)	11500	Neutral (0)	1500	Positive (1)	19000																																
Sentiment	Number of Reviews																																																
Negative (-1)	11500																																																
Neutral (0)	1500																																																
Positive (1)	19000																																																
Sentiment	Number of Reviews																																																
Negative (-1)	11500																																																
Neutral (0)	1500																																																
Positive (1)	19000																																																
<div><p>Review Length Distribution</p><table><tr><th>Review Length (characters)</th><th>Number of Reviews</th></tr><tr><td>0</td><td>20000</td></tr><tr><td>50</td><td>6500</td></tr><tr><td>100</td><td>2500</td></tr><tr><td>150</td><td>1000</td></tr><tr><td>200</td><td>500</td></tr><tr><td>250</td><td>200</td></tr><tr><td>300</td><td>100</td></tr><tr><td>350</td><td>50</td></tr><tr><td>400</td><td>20</td></tr><tr><td>450</td><td>10</td></tr><tr><td>500</td><td>5</td></tr></table></div>	Review Length (characters)	Number of Reviews	0	20000	50	6500	100	2500	150	1000	200	500	250	200	300	100	350	50	400	20	450	10	500	5	<div><p>Review Length Distribution</p><table><tr><th>Review Length (characters)</th><th>Number of Reviews</th></tr><tr><td>0</td><td>20000</td></tr><tr><td>50</td><td>6500</td></tr><tr><td>100</td><td>2500</td></tr><tr><td>150</td><td>1000</td></tr><tr><td>200</td><td>500</td></tr><tr><td>250</td><td>200</td></tr><tr><td>300</td><td>100</td></tr><tr><td>350</td><td>50</td></tr><tr><td>400</td><td>20</td></tr><tr><td>450</td><td>10</td></tr><tr><td>500</td><td>5</td></tr></table></div>	Review Length (characters)	Number of Reviews	0	20000	50	6500	100	2500	150	1000	200	500	250	200	300	100	350	50	400	20	450	10	500	5
Review Length (characters)	Number of Reviews																																																
0	20000																																																
50	6500																																																
100	2500																																																
150	1000																																																
200	500																																																
250	200																																																
300	100																																																
350	50																																																
400	20																																																
450	10																																																
500	5																																																
Review Length (characters)	Number of Reviews																																																
0	20000																																																
50	6500																																																
100	2500																																																
150	1000																																																
200	500																																																
250	200																																																
300	100																																																
350	50																																																
400	20																																																
450	10																																																
500	5																																																

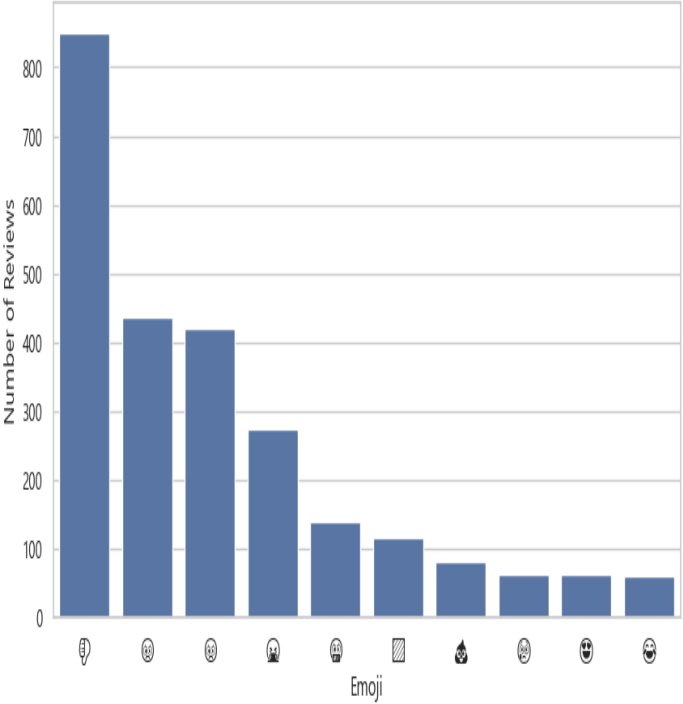




Top 10 Negative Emojis



Top 10 Negative Emojis



---

# Preprocessing:

In order to prepare our dataset for sentiment analysis, we implemented a comprehensive preprocessing pipeline using the following techniques:

- **Lowercasing:**
  - Convert all text to lowercase to ensure uniformity.
- **Duplicate Review Handling:**
  - Drop duplicate reviews based on
- **Text Cleaning:**
  - Remove specific punctuation marks, including ``÷×_-"..."'!|+|~{ } ', .?":/[,~] [%^&* () _<>:` and digits to focus on the linguistic content.
- **Diacritics Removal:**
  - Eliminate diacritics to standardize Arabic text.
- **Normalization:**
  - Normalize Arabic text to account for linguistic variations.
- **Repeating Character Removal:**
  - Remove consecutive repeating characters to enhance readability.
- **Long Word Removal:**
  - Eliminate excessively long words that might not contribute meaningfully.
- **Stopword Removal:**
  - Remove common stopwords to focus on meaningful content.
- **Emoji Handling:**
  - Replace emojis with the corresponding text to preserve emotional context.
- **Invalid Character Removal:**

- 
- Remove unhandled emojis and other invalid characters that might disrupt analysis.

- **Consecutive Space Reduction:**

- Collapse any consecutive spaces into a single space for consistency.

- **Lemmatization:**

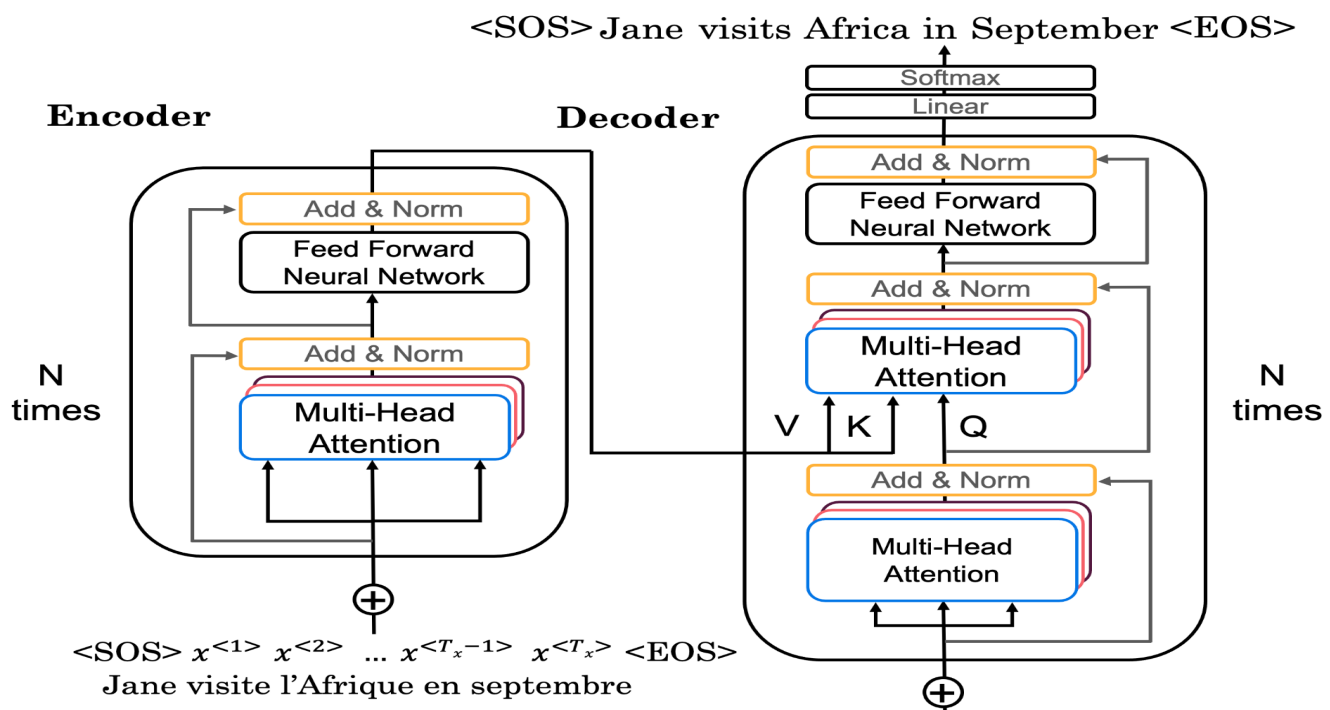
- Apply multilingual lemmatization to reduce words to their base form.

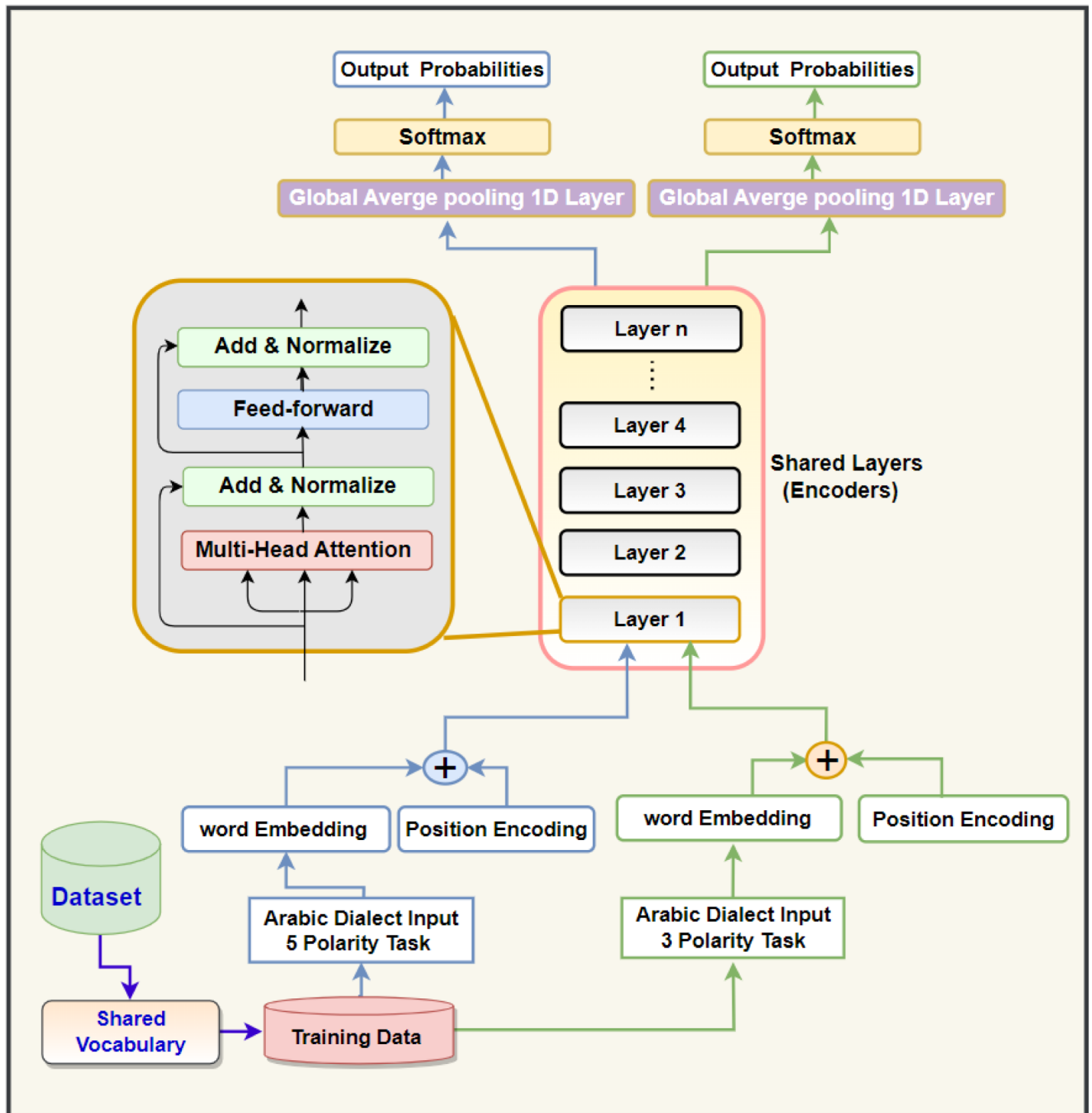


# Transformers

Transformers are renowned for their formidable capability to capture intricate patterns and long-range dependencies in sequential data. We implemented the model **from scratch** and achieved a Training Accuracy of **95.70%**, underscoring the effectiveness of Transformers in learning complex representations from data.

Adhering to the transformer architecture introduced in the seminal paper *"Attention is All You Need"* by Vaswani et al., we have witnessed a revolution in natural language processing (NLP) tasks. The Transformer model, comprising both the encoder and decoder, has extended its application beyond translation tasks to various sequence-to-sequence tasks, including text classification.

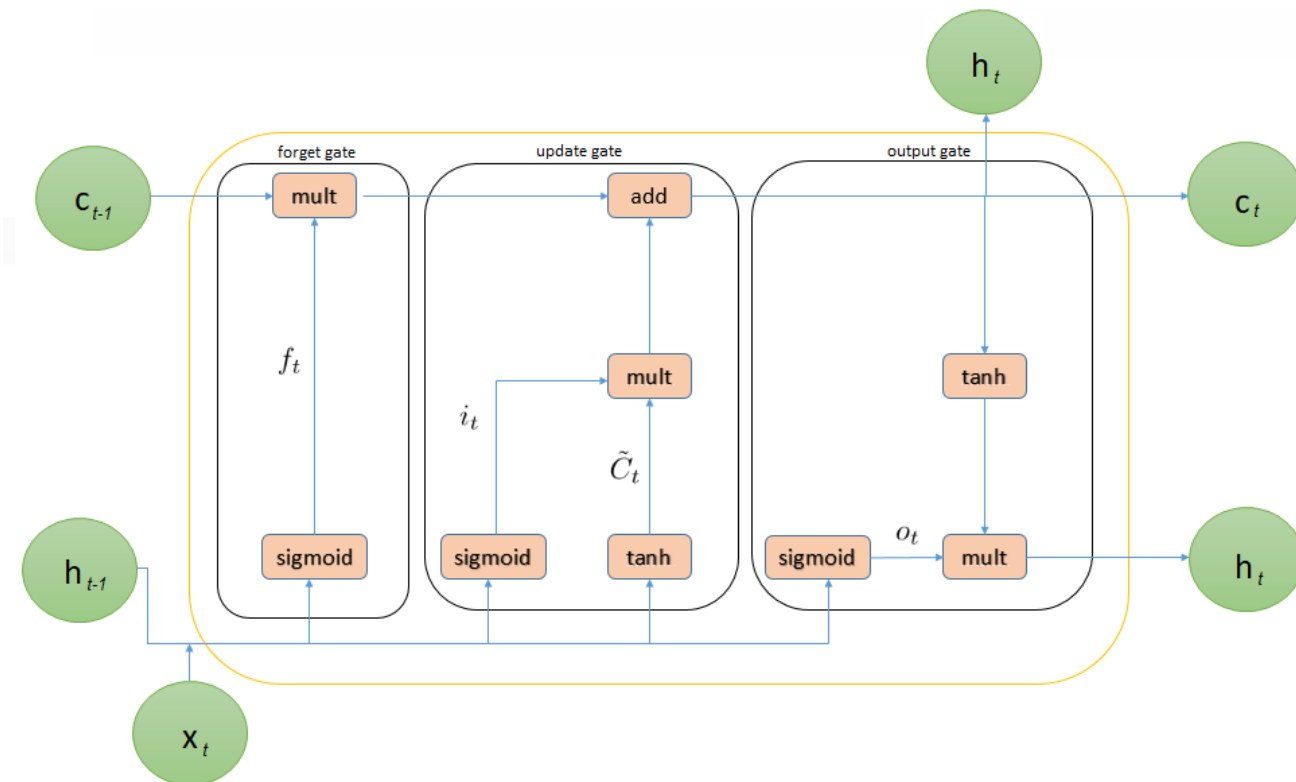




Transformers Architecture

# LSTM

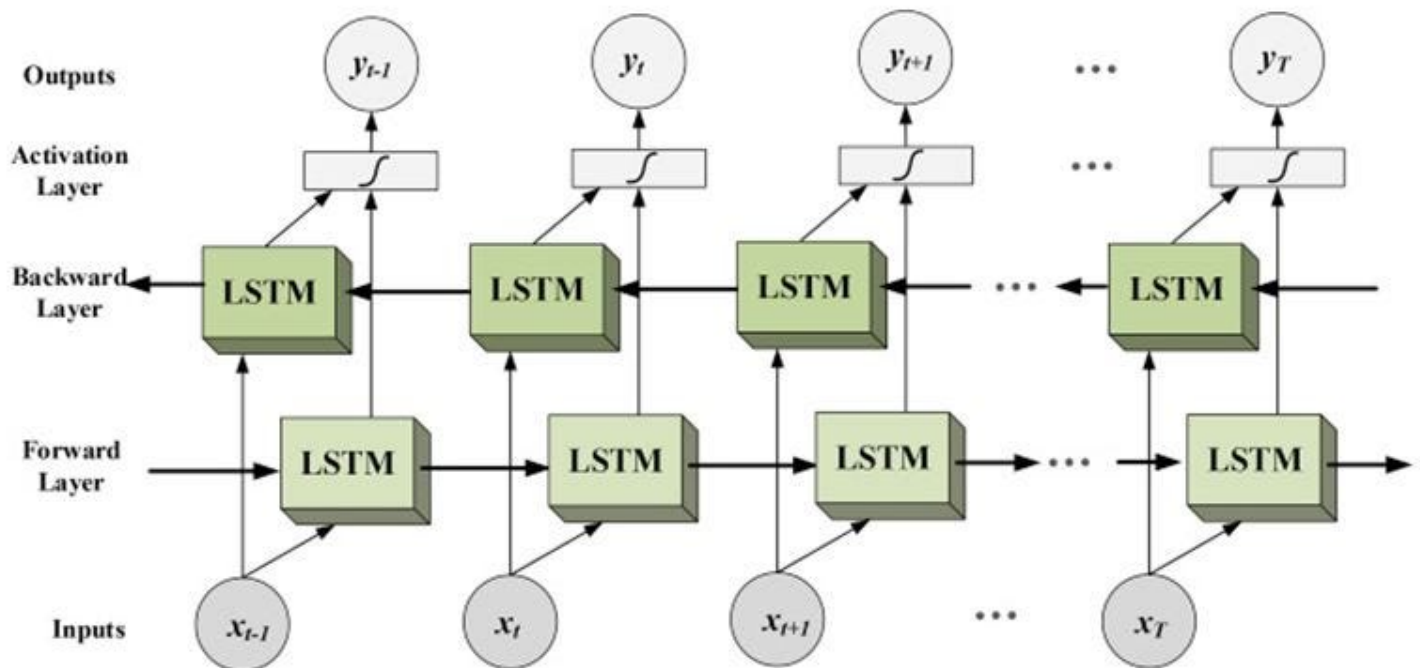
LSTM is a powerful sequence processing technique renowned for its capture of contextual dependencies. The model was configured with **Adam optimization** and underwent training for **five epochs** with a selected **batch size of 32**, and a judicious **validation split of 20%** was implemented to monitor generalization performance. The model showcased a commendable **training accuracy of 92.84%**.



LSTM Architecture

## BiLSTM (Bidirectional LSTM)

BiLSTM is a powerful sequence processing technique renowned for its capture of contextual dependencies in both forward and backward directions. The model was configured with **Adam optimization** and underwent training for **tem epochs** with a selected **batch size of 32**, and a judicious **validation split of 20%** was implemented to monitor generalization performance. The model showcased a commendable **training accuracy of 93.88%**.



*Bidirectional LSTM Architecture*

---

# Pre-trained Arabic BERT Integration for Sentiment Analysis

In the process of integrating a pre-trained model for sentiment analysis, we opted for the '**aubmindlab/bert-large-arabertv02-twitter**' model.

Initialization involved using the **BertTokenizer** and **BertForSequenceClassification** classes. Further customization was achieved with the **ArabertPreprocessor** for specific Arabic text preprocessing. To facilitate model training, a custom dataset class, **SimpleDataset**, was defined to handle input data, attention masks, and labels efficiently. Subsequently, necessary adjustments, such as normalizing labels, were applied. The dataset was then prepared using the defined **SimpleDataset** class and loaded into a **DataLoader** for batch processing.

The model fine-tuning process was conducted using the **Trainer** class from the Transformers library, employing specific training arguments. Following successful fine-tuning, the model was applied to predict sentiment labels for a test dataset, generating predictions.

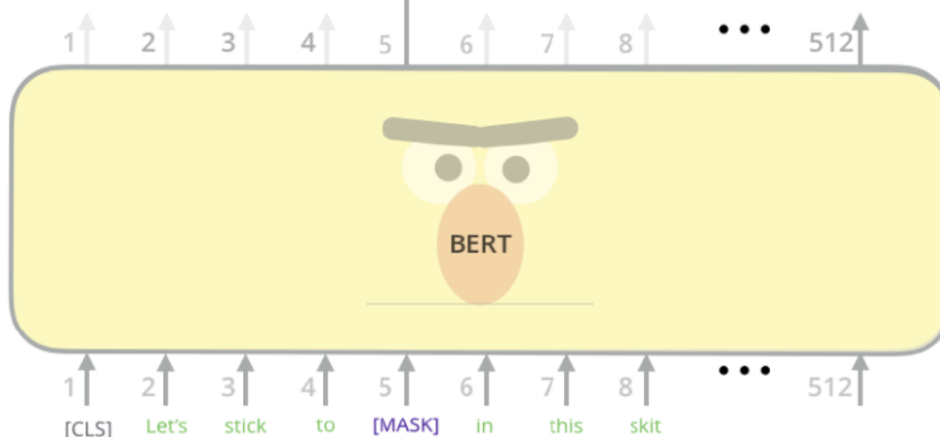
The model showcased a commendable **test accuracy** of **87.50%**.

Use the output of the masked word's position to predict the masked word

Possible classes:  
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzzzyva

FFNN + Softmax



Randomly mask  
15% of tokens

Input

[CLS] Let's stick to improvisation in this skit

---

# Results and Performance Metrics

	BiLSTM	LSTM	Transformers	Arabic BERT
Training Accuracy	93.88%	92.84%	95.70%	
Test Accuracy		80.06%	80.65%	87.50%

## Different Trials

For Arabic sentiment analysis, Softmax optimally predicts class probabilities in multi-label scenarios, while ReLU in hidden layers enhances the model's capacity to grasp intricate sentiment nuances efficiently. These choices form the core of our approach. Explore our hyperparameters below, refined through iterative trials for optimal module performance.

---

# 1- LSTM

LSTM UNITS	DROP OUT RATE	Dense Layers	Epochs	Training Accuracy
100	0.2	1	5	92.55%
50	0.3	2	8	93.61%
150	0.4	2	6	92.67%

# 2- Bidirectional - LSTM

LSTM UNITS L1	LSTM UNITS L2	Drop Out Rate	Output Dense Layer	Hidden Dense Layer	Epochs	Training Accuracy
64	32	0.2	32	3	10	93.88%
100	32	0.2	2	3	5	91.10%
50	64	0.3	16	3	8	93.66%

# 3- Transformers

MAX Sequence Length	DROP OUT RATE	Dense Layers	Epochs	Training Accuracy
100	1e-6	1	5	92.55%
50	0.3	2	8	93.61%
150	0.4	2	6	92.67%