A picture containing text, soup, dish, clipart

Description automatically generated

**Cairo University Department of Computer**

**Faculty of Engineering Engineering**

**Natural Language Processing**

Project

**Arabic Text Diacritizer**

|  |  |  |
| --- | --- | --- |
| Name | Student Code | Email |
| Bemoi Erian Ayad | 9202391 | bemoi.tawadros00@eng-st.cu.edu.eg |
| Peter Atef | 9202395 | peter.zaki00@eng-st.cu.edu.eg |
| Doaa Ashraf | 9202519 | doaa.ahmed01@eng-st.cu.edu.eg |
| Mark Yasser | 9203106 | mark.ibrahim00@eng-st.cu.edu.eg |

**Project Pipeline**

From draw.io [[Link](https://app.diagrams.net/#G1fdamfAf2iVt9KHhXpsHx8fIPzKBRVPT3)]

**Detailed Description**

1. **Data preprocessing**

Preprocessing

1. **Feature extraction**

Feature

1. **Model training**

training

**Evaluation**

evaluation

**Models Used**

Models

**Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Strategy** | **Word Embedding** | **Character Embedding** | **Classifier** | **Accuracy** |
| Split on lines | - | One Hot Encoding | LSTM | 51.2% |
| Split on lines | - | One Hot Encoding include space | LSTM | 61% |
| Split on lines | FastText | One Hot Encoding include space | LSTM | 61% |
| Split on lines | Word2Vec | One Hot Encoding | LSTM | 53% |
| Split on lines | FastText | One Hot Encoding | RNN | 54.7% |
| Split on Sentence | FastText | One Hot Encoding | LSTM | 63.3% |
| Split on lines | FastText | One Hot Encoding | LSTM | 64.1% |
| Split on lines | FastText | FastText | LSTM | 51.6% |