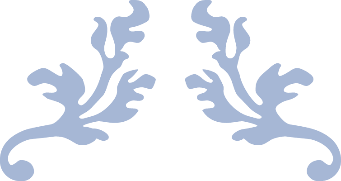
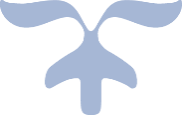


Cairo University Computer Engineering Department

Faculty of Engineering Fourth year

**natural language processing**

**Project Document**

**Team #21 Members**

|  |  |  |
| --- | --- | --- |
| **Name** | **Section** | **B.N.** |
| Mostafa Wael Kamal | 2 | 28 |
| Mostafa Mohamed Elgendy | 2 | 26 |
| Micheal Aziz | 2 |  |
| Robert Mounir | 1 |  |

**December 2022**

**Project Pipeline:**

**Diagram

Description automatically generated**

**Project Phases:**

**Data preprocessing:**

This module performs the following to every tweet in each of the training and the dev datasets:

* Remove twitter handlers
* Remove digits
* Remove all the special characters, punctuations, and emojis
* Remove all English alphabet
* Substitute multiple spaces with a single space
* Remove all empty spaces
* Remove all stop words (downloaded from nltk)

It also does the following for the training dataset only after performing the previous steps:

* Remove all empty tweets
* Remove duplicate tweets
* Remove the tweets with less than 10 characters
* Remove the tweets with less than 4 characters

**Feature Extractor:**

**Data Balancing:**

Chart

Description automatically generatedChart, pie chart

Description automatically generated