# Team: CS\_28

|  |  |  |
| --- | --- | --- |
| Sec | Name | ID |
| 7 | 2021170642 | يوسف محمد سيد محمد |
| 5 | 2021170480 | محمد عوض طلعت محمد |
| 7 | 2021170654 | يوسف وجيه وديع ناشد |
| 7 | 2021170649 | يوسف مصطفى محمد يوسف البنى |
| 5 | 2021170483 | محمد مبروك فكرى عبدالفتاح |
| 7 | 2021170638 | يوسف عصام عزت اخنوخ |

# Preprocessing

**Column ‘Discussion’**

* Drop NULL values
* Remove non-alphanumeric characters
* Remove URLs
* Convert to lowercase
* Split to tokens
* Remove punctuations
* Remove stop words
* We tried 5 ways for cleaning the tokens
  1. Steaming
  2. Lemmatization
  3. Lemmatization with tagging
  4. Lemmatization then Steaming
  5. Lemmatization with tagging then Steaming

**Column ‘Category’**

* Map to numeric value
* Politics : 0
* Sports : 1
* Media : 2
* Market & Economy : 3
* STEM : 4

# Feature Extraction

**FFNN (Feedforward Neural Network)**

* Intialize TfidfVectorizer
* Fit TfidfVectorizer on the preprocessed data
* Transforming the data into a sparse matrix
  + Each row represents a document.
  + Each column corresponds to a term in the vocabulary.
  + The values represent the TF-IDF score of each term in each document.
* We used this matrix to fit the model

**RNN (Recurrent Neural Network)**

* Intialize Tokenizer with num\_words = **20000.**
* Fit the Tokenizer on the preprocessed data.
* Convert the text data into sequences of integers, where each integer represents a specific word in the vocabulary.
* Standardizing sequence lengths to **100**, where longer sequences are trimmed and shorter sequences are padded with zeroes.

**Transformers**

* Intialize TokenizerandFit on the preprocessed data.
* Convert the text data into sequences of integers, where each integer represents a specific word in the vocabulary.
* Standardizing sequence lengths, where longer sequences are trimmed and shorter sequences are padded with zeroes.
* Load GloVe Embedding.

# Trials

**FFNN (Feedforward Neural Network)**

|  |  |  |
| --- | --- | --- |
| **Preprocessing** | **epochs** | **Train accuracy** |
| Steaming | 1 | 0.91 |
| Lemmatization | 2 | 0.58 |
| 3 | 0.87 |
| 10 | 0.92 |
| Lemmatization and Steaming | 2 | 0.92 |
| 3 | 0.93 |

**RNN (Recurrent Neural Network)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Preprocessing** | **num\_words** | **Seq\_len** | **epochs** | **RNN layer** | **Train accuracy** |
| Lemmatization | 10000 | 27 | 5 | Simple RNN | 0.83 |
| 20000 | 100 | 10 | GRU | 0.81 |
| 11 | 0.82 |
| 12 | 0.83 |
| 15 | 0.84 |
| 18 | 0.85 |

**Transformers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Preprocessing** | **num\_words** | **Seq\_len** | **Embedding**  **diminsion** | **epochs** | **Train accuracy** |
| Lemmatization | 20000 | 100 | 100 | 9 | 0.69 |
| 15 | 0.72 |
|  |  |  |  |  |
|  |  |  |  |  |