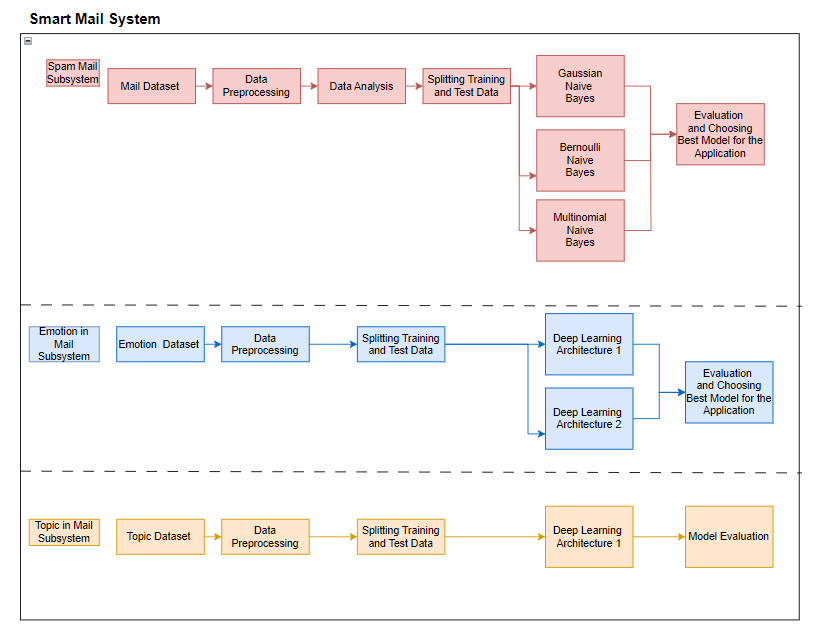
**Project team**

|  |  |  |  |
| --- | --- | --- | --- |
| Team leader | Team leader | 20P6483 | Mariam Essam |
| Team member | Team member | 20P4197 | Norhan waleed |
| Team member | Team member | 20P4322 | Rokaya Essam |
| Team member | Team member | 19P1472 | Mohamed elhajj |
| Team member | Team member | 19P2374 | Belal Mahmoud |
|  |  | | |

**Introduction and Background**   
Millions of emails are received daily. We want the user to know which mails to delete (spam mails), and which to keep (read/write mails). We also want the user to know which emotion is dominant in the mails text, and the topic of the mail.  
  
**Objectives**

* [Classify spam and ham mails with accuracy score greater than or equal 95%]
* [Classify emotion in mail with accuracy greater than or equal 80%]
* [Classify topics in mail with accuracy over 95%]

**System overview**



**Note: This section was added after the discussion**

**Methods**

[ List all the methods used to implement the system such as ML techniques, Searching algorithms or any AI algorithms]

**Spam and Ham Mail Classifier**

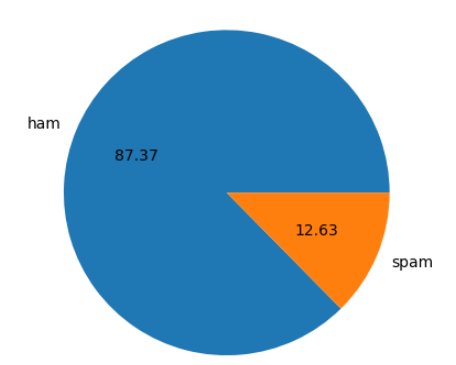
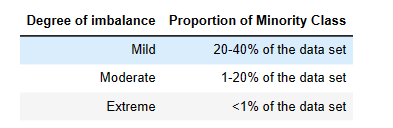
**Dataset used:** https://www.kaggle.com/datasets/mfaisalqureshi/spam-email

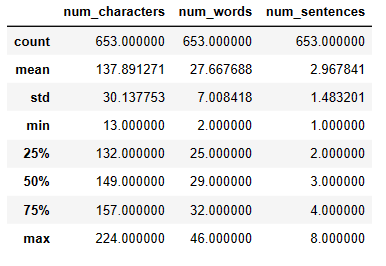
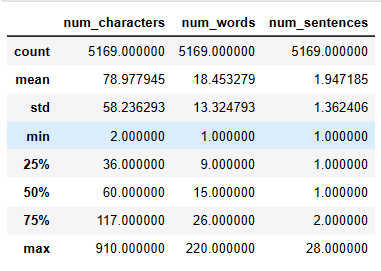
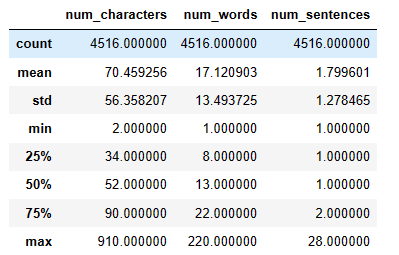
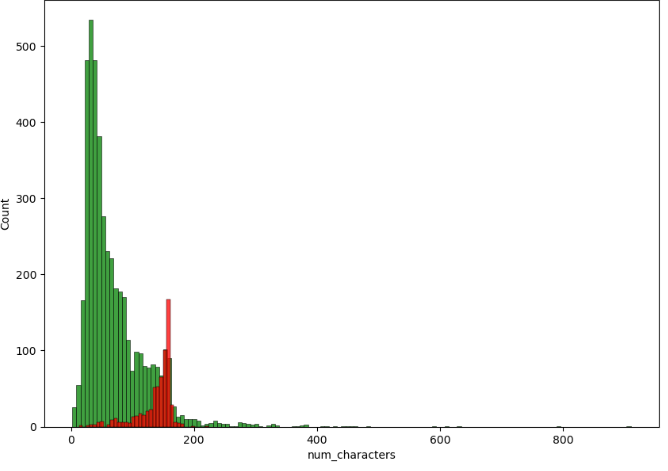
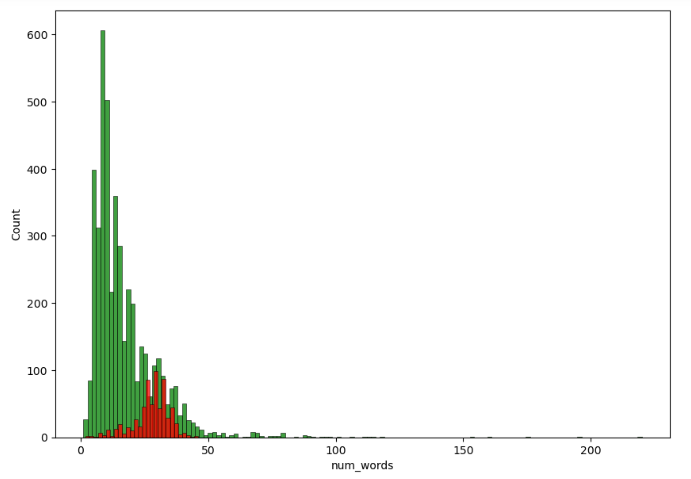
**1- Data Cleaning:**

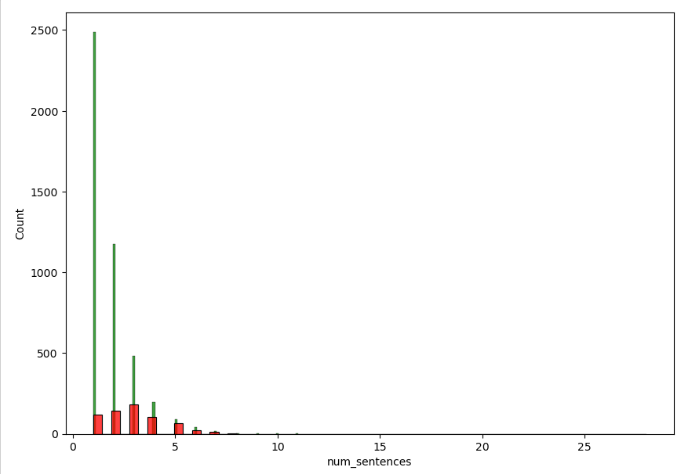
* Drop null columns in place
* Rename columns
* Encode Labels
* Handle missing values
* Remove duplicate values

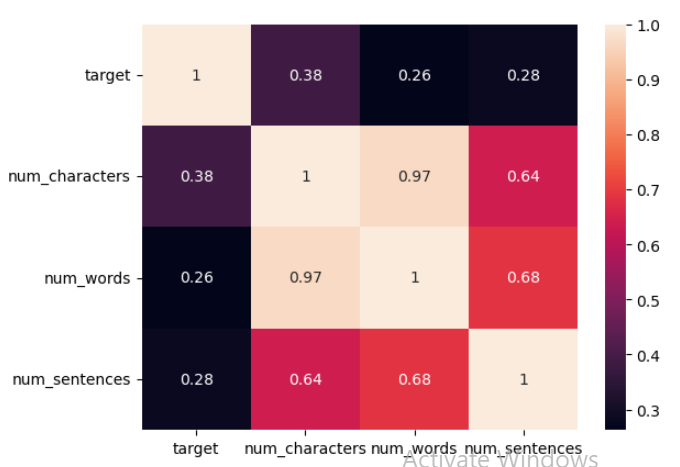
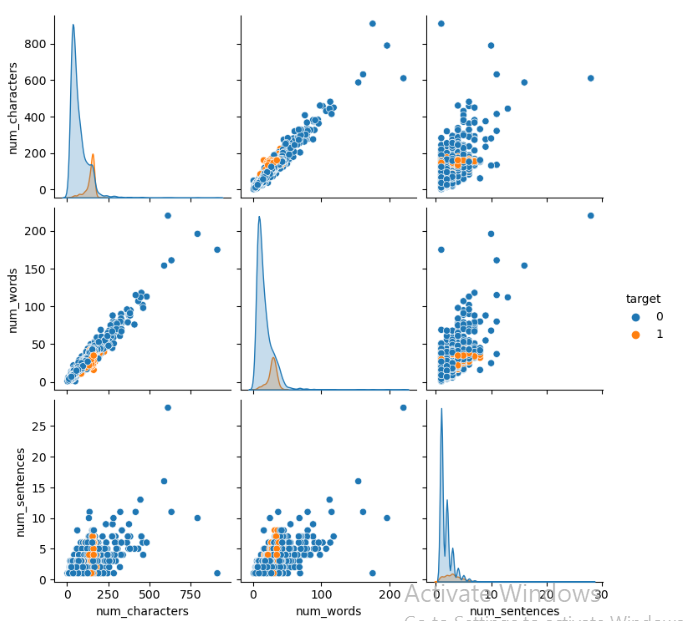
**2- EDA (Exploratory Data Analysis)**

* Get target value counts (Our data is imbalances)

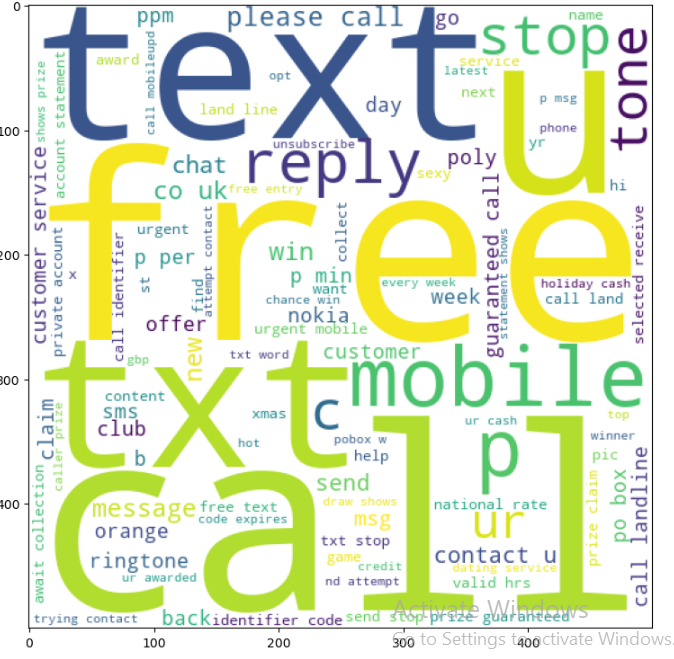


* Using NLTK(Natural Language Toolkit)
* Adding num\_characters column
* Adding num\_words column using nltk.word\_tokenize()
* Adding num\_sentencess column using nltk.sent\_tokenize()
* Describing statical details for all mails, hams only and spams only
* Plotting graphs and heatmap using seaborn (ham in green, spam in red)





**3- Data Preprocessing**

* Lower case
* Tokenization
* Removal of special characters
* Removal of stop words and punctuation
* Stemming (process of reducing a word to its stem that affixes to suffixes and prefixes or the roots)
* Generating word cloud for spam and ham mails
* Putting spam and words in a list (spam\_corpus, ham\_corpus)
* Applying collection counter on the corpuses and showing most common 30 words
* Using a bar plot

**4- Model Building and Evaluation**

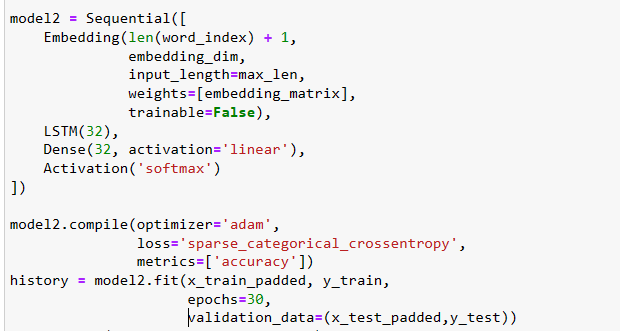
* Feature extraction using TFidVectorizer
* Train Test Split
* Model used is Naïve Bayes
* 3 different types of algorithms (Bernoulli, Multinomial, Gaussian)
* Fitting the data
* Printing accuracy score, confusion matrix and precision score
* Saving TFidVectorizer and Multinomial Naïve Bayes

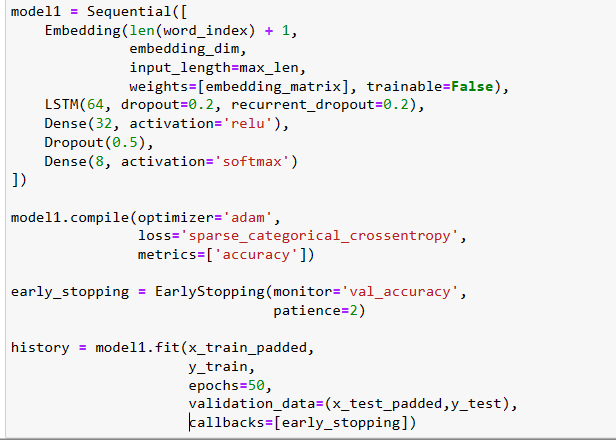
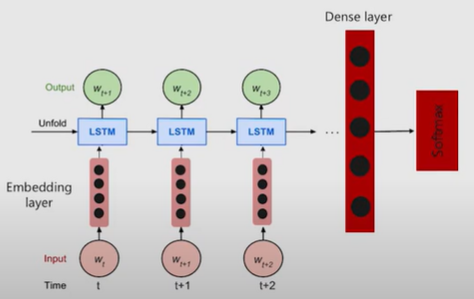
**Emotion in Mail Classifier**

# Dataset used: <https://www.kaggle.com/datasets/faisalsanto007/isear-dataset> (ISEAR Dataset)

**1- Data Preprocessing**

* Lower case
* Removal of stop words (using NLTK)
* Lemmatization (using NLTK WordNwtLemmatizer)
* Train Test Split
* using keras Tokenizer
* Fitting tokenizer on training text
* Training text to training sequence
* Finding max length of training sequence
* Padding training data with max length (using keras pad\_sequence)
* Test text to test sequence
* Padding test data with max length
* Label Encoding(1-> anger, 2-> fear, 3->joy, 4->Sadness)

**2- Model Building and Evaluation**

* Building embedding matrix (using Global Vectors using glove-wiki-gigaword-300)
* 2 Deep learning architectures were used



* Both architectures have embedding layer, LSTM and dense layer
* 3 different types of algorithms (Bernoulli, Multinomial, Gaussian)
* Fitting the data
* Printing accuracy score, precision score,f1 score, recall score and confusion matrix
* Saving the first model and the tokenizer

**Topic in Mail Classifier**

# Dataset used: [Dataset](https://www.kaggle.com/datasets/faisalsanto007/isear-dataset) of topics from CNN news database

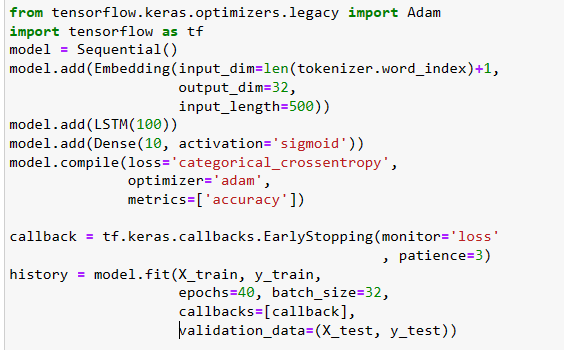
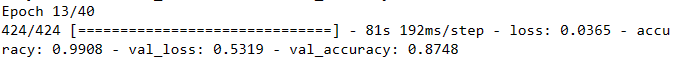
**1- Data Cleaning:**

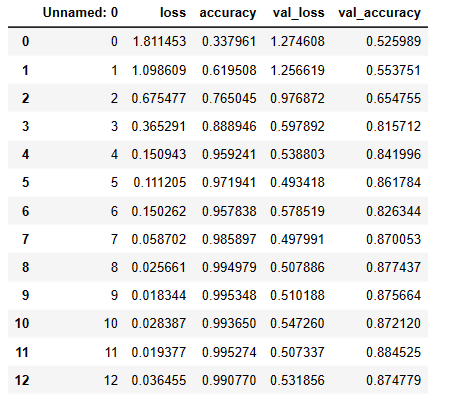
* Reading data from files
* Lower case
* Removal of html tags, emails, URLs and numbers
* using NLTK word\_tokenize
* Removal of stop words
* Joining words

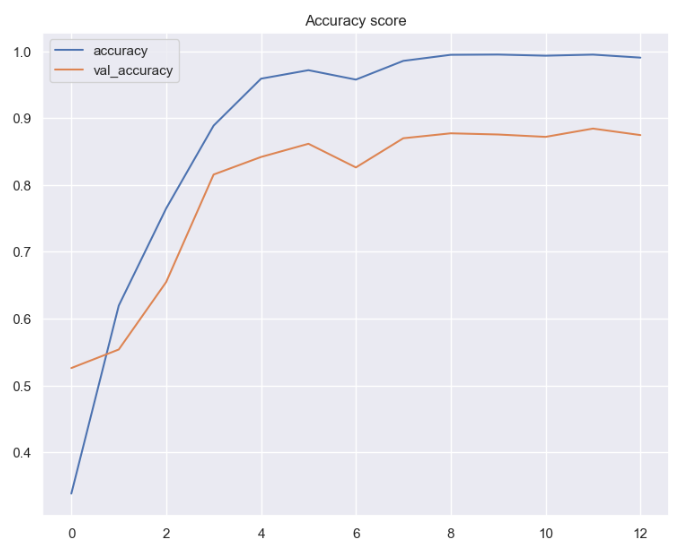
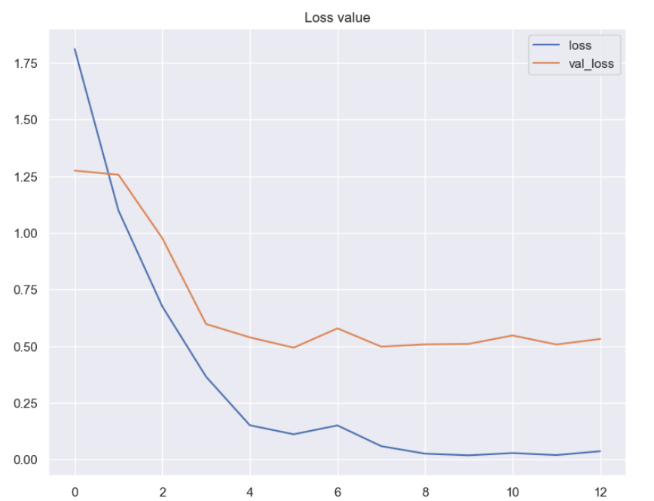
**2- Data Preprocessing**

* using keras Tokenizer
* Fitting tokenizer on all cleaned text
* Padding with max length of 500 (using keras pad\_sequence)
* Using One Hot Encoder and fitting data

**3- Model Building and Evaluation**

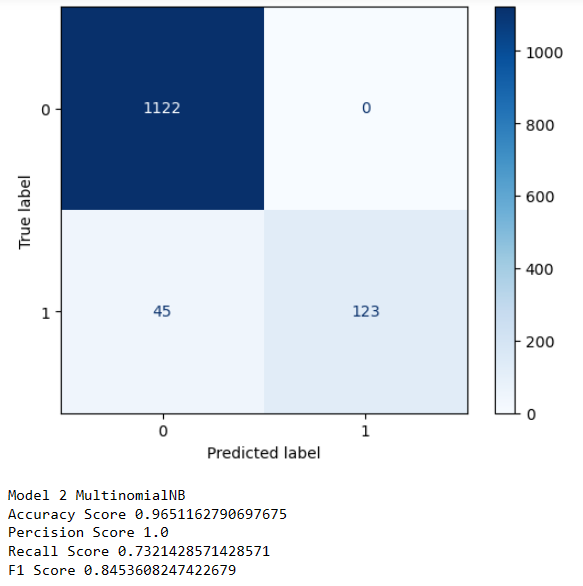
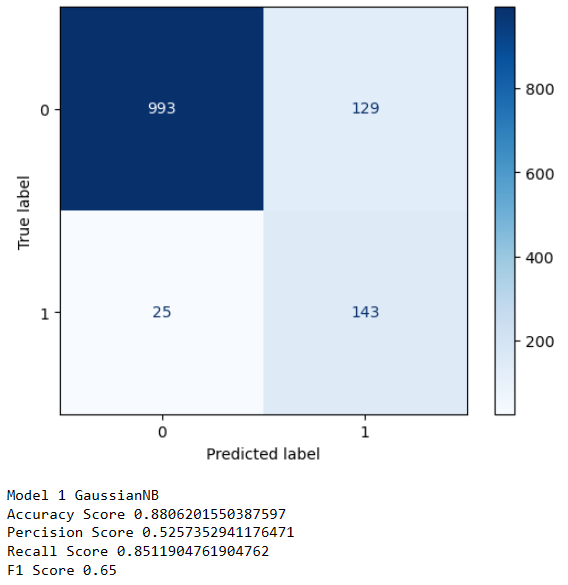
* Train Test Split
* 1 Deep learning architecture was used
* Plotting accuracy versus val\_accuraccy and loss versus val\_loss
* Saving the model

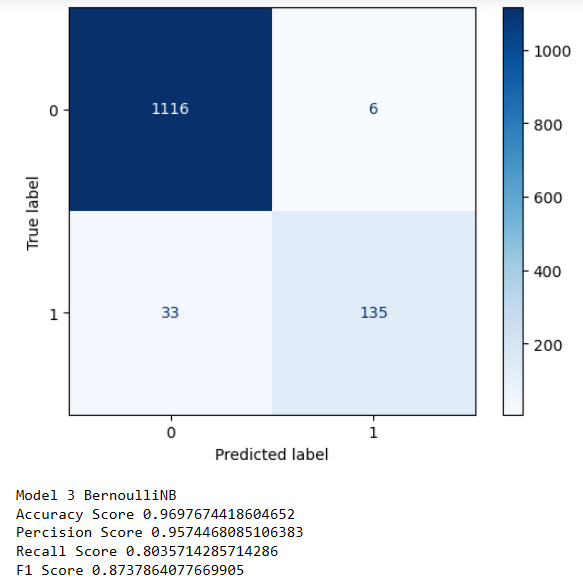




**Results and Evaluation**[Describe how results are evaluated at the end of the project. Formulate clear indicators for objectives and result. Additionally, mention the evaluation criteria used and figure out samples of the output.]

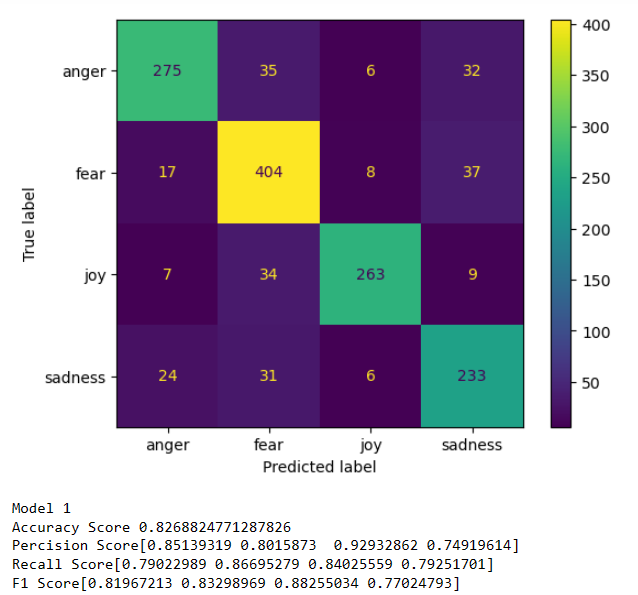
**Spam and Ham Mail Classifier**



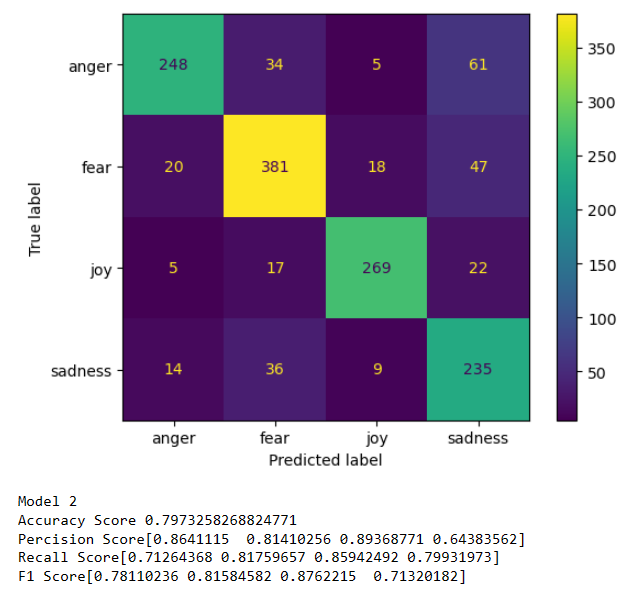




* [Classify spam and ham mails with accuracy score greater than or equal 95%] Objective is reached
* **Multinomial Naïve Bayes model was chosen** to be used in prediction application due to its high accuracy score and precision score

**Emotion in Mail Classifier**

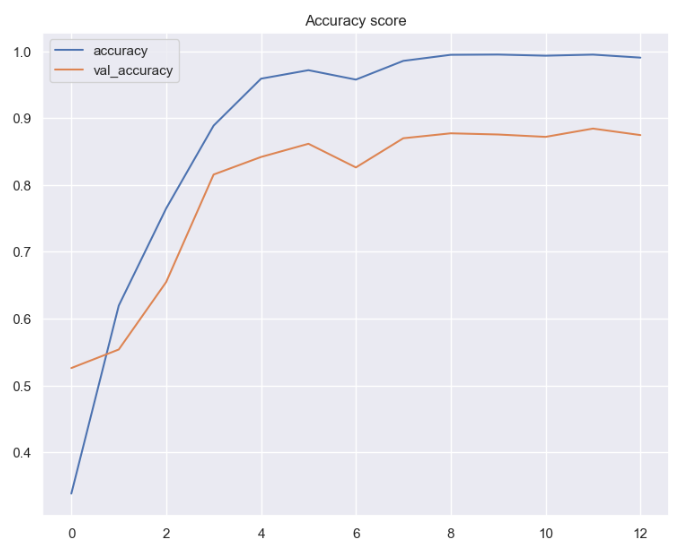


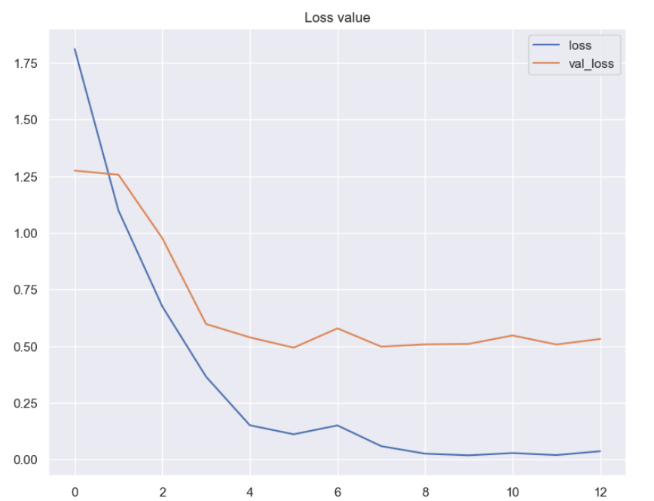


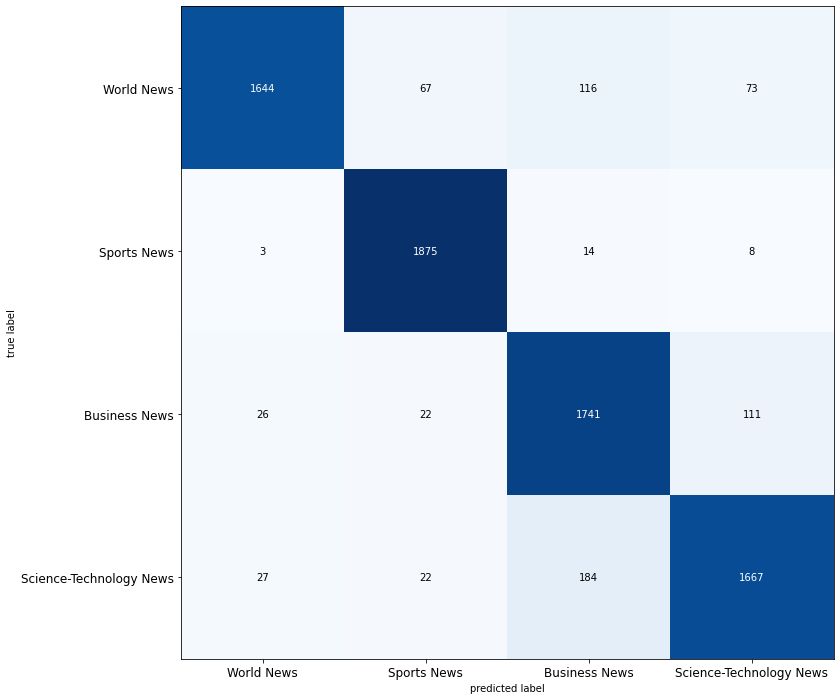


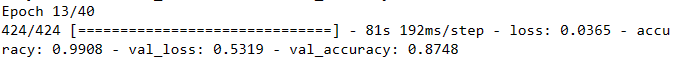


* [Classify emotion in mail with accuracy greater than or equal 80%] Objective is reached
* **First model was chosen** to be used in prediction application

**Topic in Mail Classifier**

****

****



Accuracy Score= 0.9114473684

Precision Score [0.96705 0.94411 0.84720 0.89672]

Recall Score [0.86526 0.948684 0.91632 0.87739]

F1 Score [0.91338 0.946391 0.880405 0.88695]



* [Classify topics in mail with accuracy over 95%] Objective is reached

**Grading Criteria**

You should follow the minimum requirements of the project and the grading is based on the discussion as a group as well as each team member in the group. The grading distribution can be follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Out of** | **Grade** | **Notes** |
| Does the report follow the required outline? | **3** |  |  |
| Application/System is running & all required modules are implemented | **3** |  |  |
| Is the dataset presented as a benchmark or collected? | **3** |  |  |
| Does the model/system clearly specify the loss/metric used to evaluate the quality of results? | **3** |  |  |
| Is the final model used to solve the problem obtained by selecting one of at least 2 candidate models? | **3** |  |  |
| Quality of discussion/conclusions. | **5** |  |  |

**Approval Signatures**

|  |  |  |
| --- | --- | --- |
| Supervisor |  | Teaching Assistance |
| Associate Professor / Walaa H. Elashmawi |  | Eng. Ziad Elgayar |

**Grade:**

20