

A picture containing text

Description automatically generated

[**Fake News Detection**]

Icon

Description automatically generated with medium confidence

***Scientific Computing Department***

***Faculty Of Computer and Information Science***

***Ain Shams University***

***Team* (69)**

***NLP Project***

***Under Supervision:***

***Dr Sally Saad***

*From* ***OT3A*** *Team*

|  |  |  |
| --- | --- | --- |
| Name | ID | Section |
| Osama Anter Mohamed Afify | 20191700091 | 1 |
| Tarek Ashraf Mahmoud Hussein | 20191700322 | 2 |
| Ahmed Mohamed Ibrahim Mohamed | 20191700059 | 1 |
| Adham Mohamed Tawfik Mohamed | 20191700086 | 1 |
| Ahmed Mohamed Ali Abdelrahman | 20191700068 | 1 |

**Idea of Project:**

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims and may end up being virtualized by algorithms, and users may end up in a filter bubble.

**Dataset:**

We have a new CSV file contains 7796 rows and 4 columns:

1. The first column identifies the news.
2. title: represent title of news.
3. text: have news data.
4. label: representing new Belong to fake or real Class.

**We Used Columns: Text & Label Only**.

**1-Preprocessing:**

We Use This Techniques to Clean Data and Preprocessing Text:

* Remove Rows Have Nulls Cells
* Removing Duplicate Rows
* text cleaning
  + Expand Contractions.
  + Lower Case.
  + Remove punctuation.
  + Remove words and digits containing digits.
  + Remove Stopwords.
  + Rephrase text 🡪 URL.
  + Stemming
  + Lemmatization.
  + Remove Extra Spaces.
* Label Encoder.

**2- Techniques and Feature Extraction:**

1. we will use **‘TF-IDF Vectorizer’** in our “news” data.
2. We will initialize the classifier, transform, and fit the model and calculate the performance of the model using the appropriate performance matrix/matrices to see how well our model performs.

**TF-IDF Vectorizer:**

* **TF (Term Frequency):**  In the document, words are present so many times that is called term frequency. If you get the largest values, it means that word is present so many times with respect to other words. When you get word is parts of a speech word that means the document is a very nice match.



* **IDF (Inverse Document Frequency):** in a single document, words are present so many times, but also available so many times in another document also which is not relevant. IDF is a proportion of how critical a term is in the whole corpus.

collection of word Documents will convert into the matrix which contains TF-IDF features using TF-IDF Vectorizer.



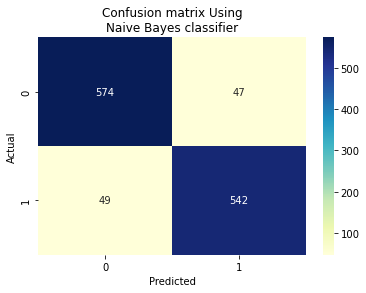
**Count Vectorizer:**

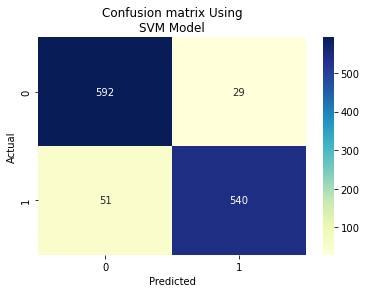
* Count vectorizer is a method to convert text to numerical data. It Convert a collection of text documents to a matrix of token counts.

**3- Classification Models:**

* Multinomial Naive Bayes classifier
* Support Vector Machine Classifier
* Logistic Regression Classifier

**4- Results visualization:**

1. **Multinomial Naive Bayes classifier:**
   * Accuracy: 92.07920792079209 %.
   * 
2. **Support Vector Machine Classifier:**

* Accuracy: 92.07920792079209 %
* :

1. **Logistic Regression Classifier:**

* Accuracy: 91.996699669967 %
* A screenshot of a computer

  Description automatically generated with medium confidence

**Real News Words:**

A close-up of words

Description automatically generated with low confidence

**Feck News Words:**

