Natural Language Processing Project

Team 7

Team Members:

Name	Sec	Bn
Hussein Mustafa Hussein	1	27
Mahmoud El-Sayed Mahmoud	2	17
Mariem Ashraf Malak	2	20
Mina Emad Fakhry	2	30

A. Pipeline

- Pre-processing
- Feature Extraction
- Model Training
- Testing and Evaluating

B. Preprocessing

Used regex to:

1. Remove unwanted tokens from the corpus such as: non words, Arabic stop-words and diacritics, links, mentions, tags, English characters, and numbers.
2. Have consistent "ي,ه,ا" throughout the corpus.

This leaves us a clean corpus containing only Arabic words.

C. Feature Extraction

Out of the corpus we extracted:

1. Bag of words
2. Tf-Idf Transformer
3. Count vectorizer
4. word embeddings and wordvec

D. Model Training

- Classical machine learning classifiers:
  - SVM
  - Multinomial Naive bayes
  - Logistic Regression

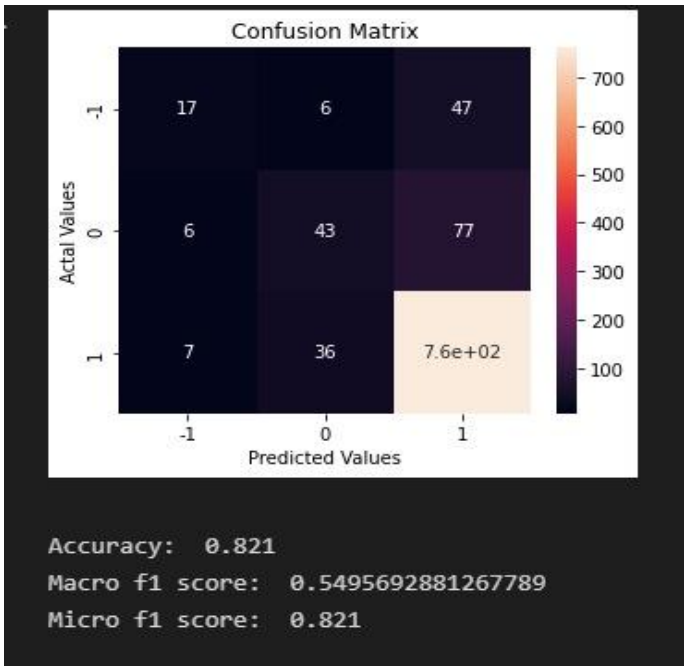
- Sequence classifiers:
  - LSTM
  - RNN

E. Evaluation

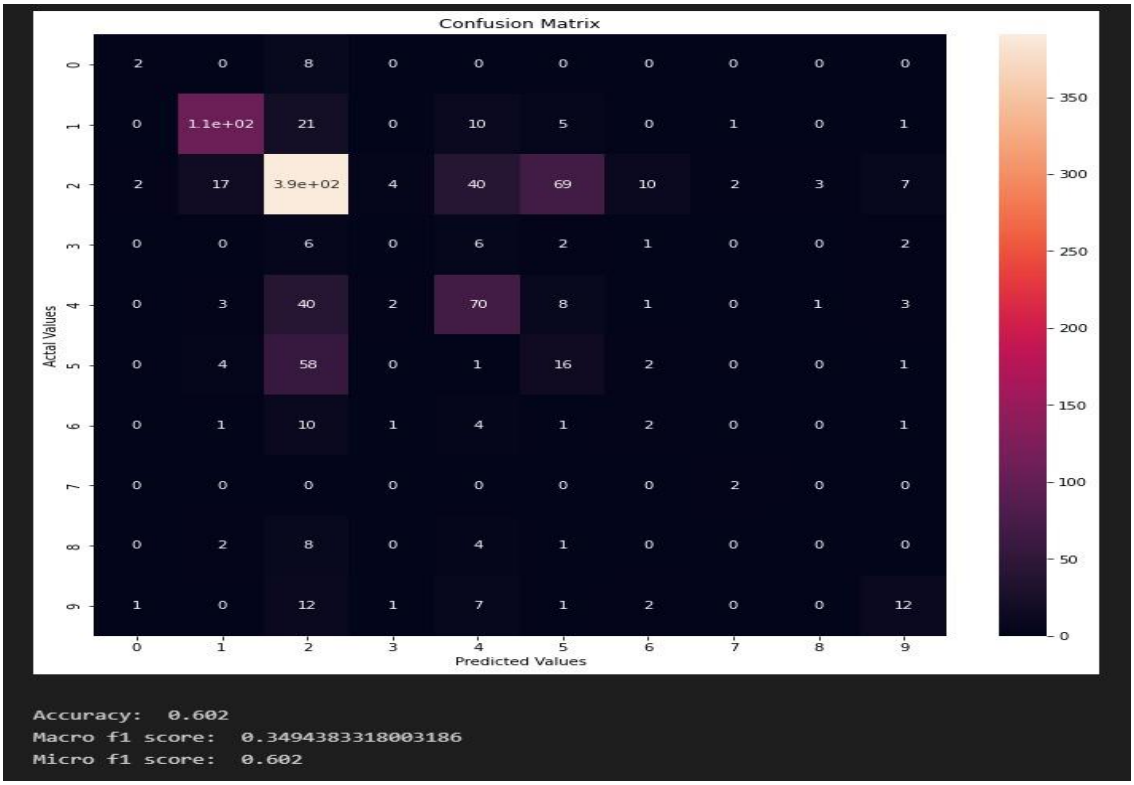
We used these metrics to evaluate each model:

(Accuracy, macro and micro f1-score)

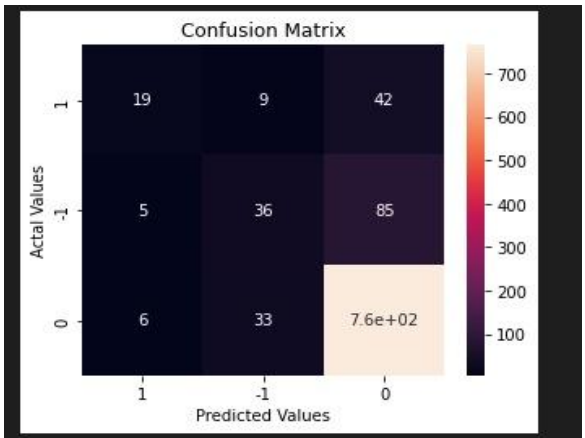
1.a. SVM (stance)



1.b. SVM (category)

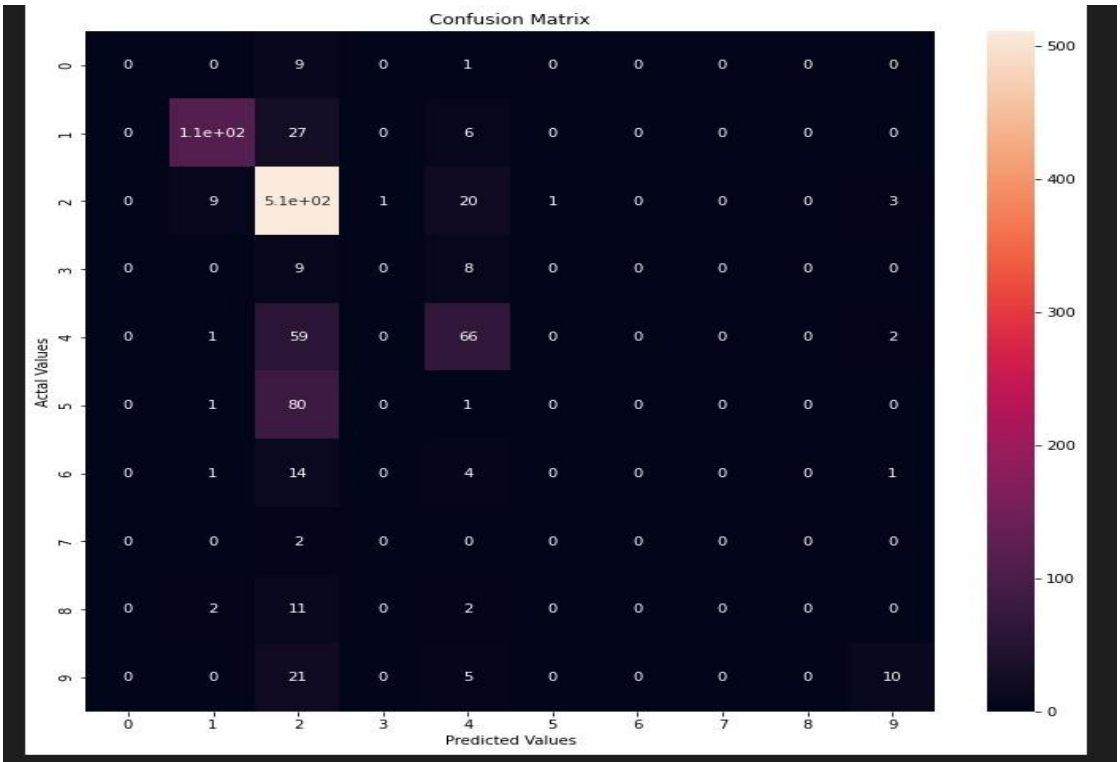


2.a. Naive Bayes (stance)



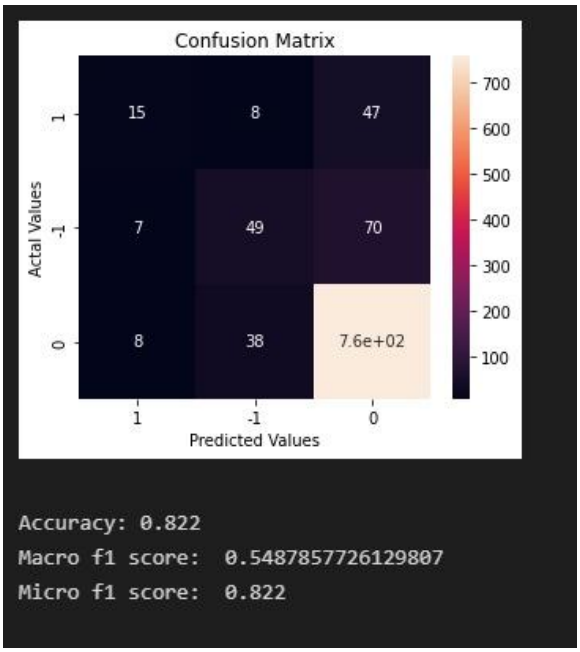

```
Accuracy:  0.82
Macro f1 score:  0.5450212726600073
Micro f1 score:  0.82
```

2.b. Naive bayes (category)

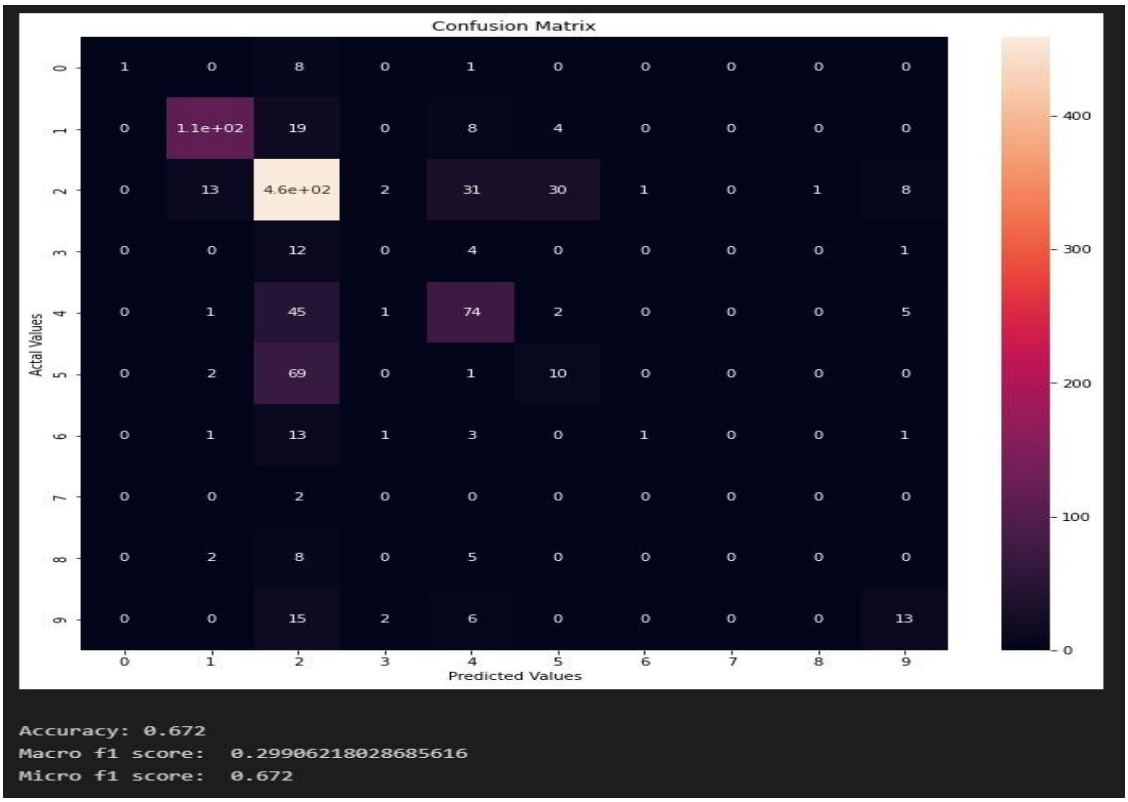



```
Accuracy:  0.699
Macro f1 score:  0.25523797534912585
Micro f1 score:  0.699
```

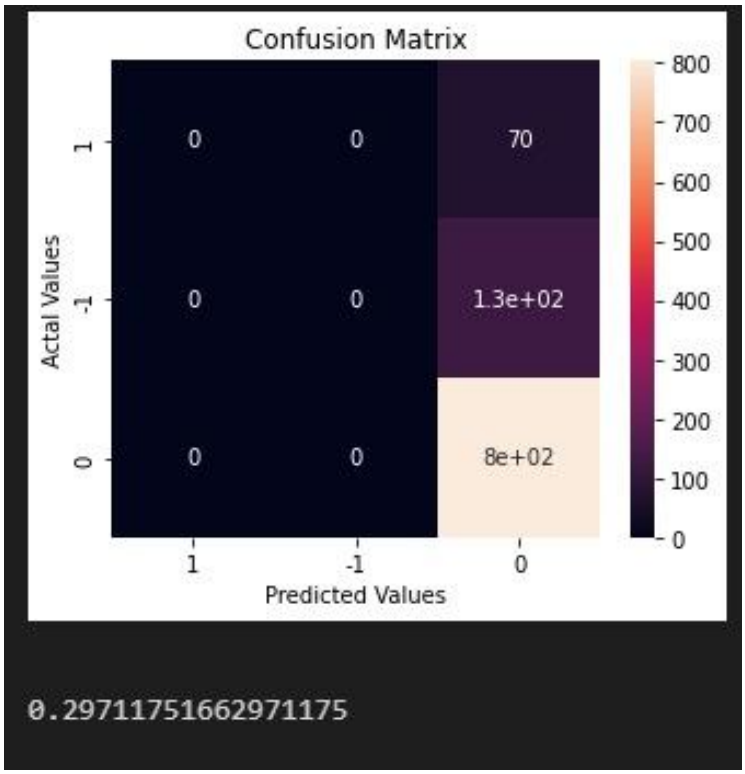
3.a. Logistic Regression (stance)



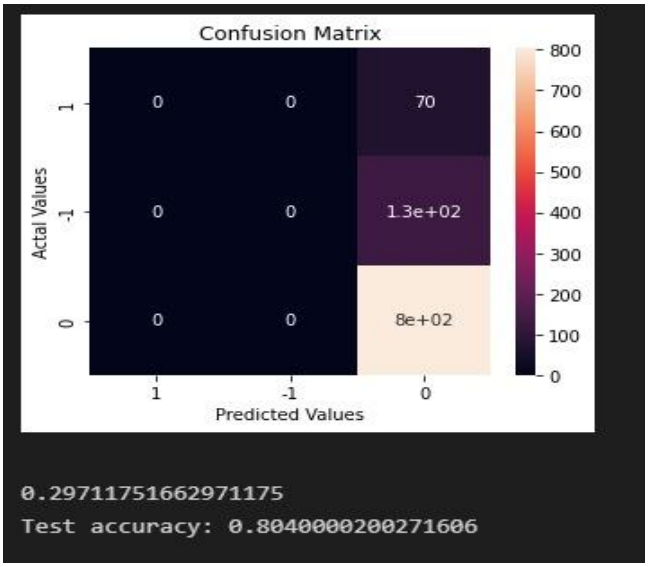
3.b. Logistic Regression (category)



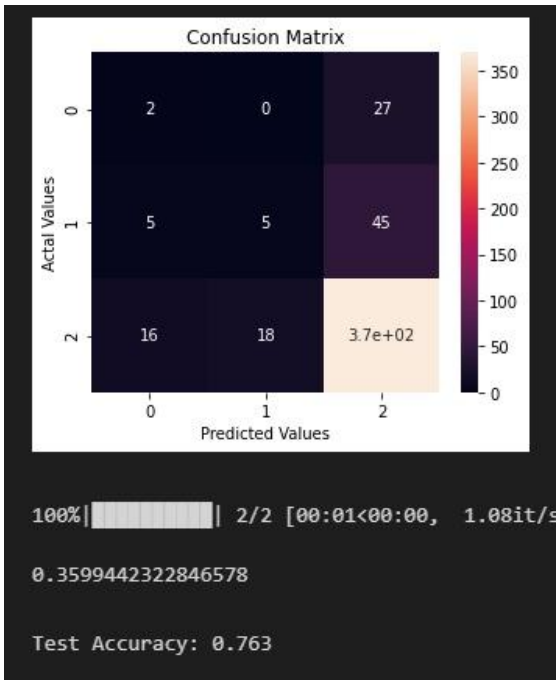
4. SVM with input word2vec “tockenizer”



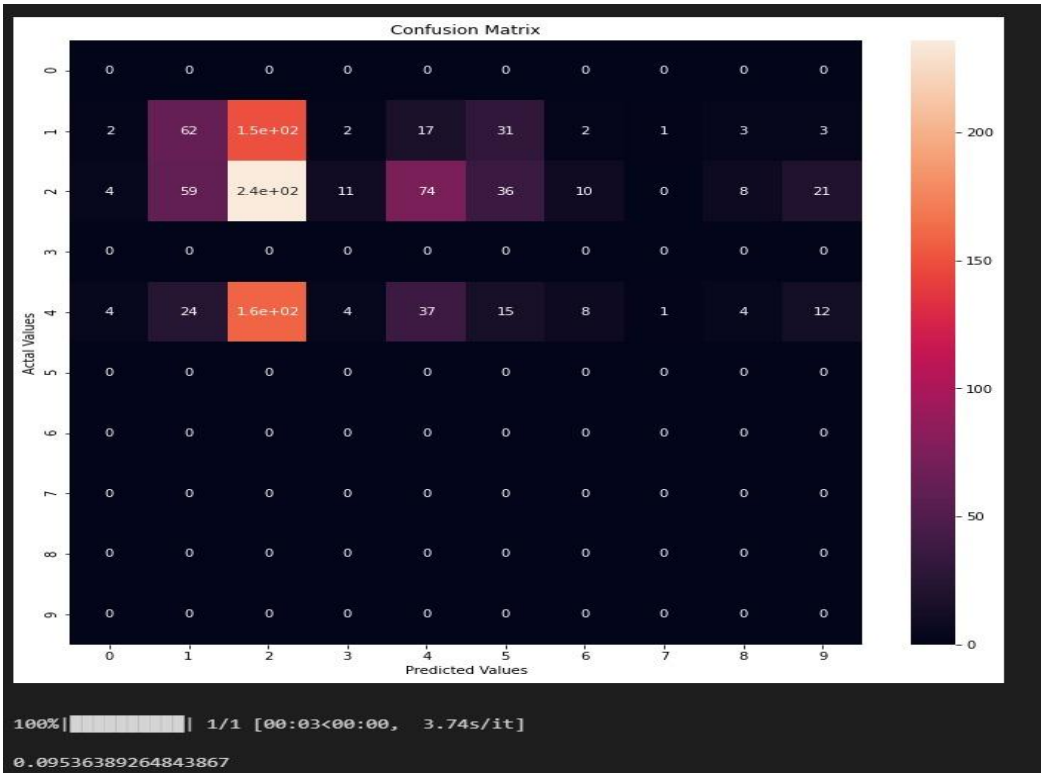
5. RNN with input word2vec “tockenizer”



6.a. LSTM (stance)



6.b. LSTM (category)



F. Submitted model

We chose the Logistic Regression model as it had the highest accuracy beside the macro f1 score, the SVM model is very close concerning the f1 score and may be better but significantly worse concerning the accuracy.