NLP Project Report: COVID-19 Tweets Sentiment Analysis

By: Team 5

Taher Mohamed - Andrew Boshra

Omar Mohamed Ahmed - Peter Michael Azmy

Project Pipeline

1. Data preprocessing

The preprocessing phase is composed of several steps:

1. The data is read from files into lists of “text”, “category”, and “stance”. The preprocessing is done on the “text” only
2. Using **ArabicLightStemmer** from **tashaphyne.stemming** package, we have a **normalizer** and a **stemmer** functions
3. We first pass all text through the normalizer to convert characters with many forms to a single form and remove strange symbols
4. We pass all the text through the stemmer to remove suffixes and prefixes from the word and reduce its form to the original form
5. For BERT the preprocessing was done using **ArabertPreprocessor** tokenized using **AutoTokenizer**

2. Feature extraction

The feature extraction phase is composed of many steps:

1. We build the vocabulary by adding all the words in all text to a single vocabulary set
2. Removal of stop words, using the **nltk** stopwords set we removed all stop words from the text
3. We tried extracting several features as
    a. **Bag of words**
    Using **CountVectorizer** from **sklearn.feature_extraction.text**, we trained the countVector on the corpus and used the BOG vector as a feature for training the model
    b. **TF-IDF**
    Using **TfidfVectorizer** from **sklearn.feature_extraction.text**, we trained the TfidfVectorizer on the corpus and used the TF-IDF vector as a feature for training the model
    c. **Frequency Vector**

Counting the number of positive, negative, and neutral words in each sentence, we build a vector of length 3 which is used in training the model

d. **Word Embeddings**
We used a pre-trained Word2Vec word embeddings for the LSTM model

3. SMOTE

We noticed that the dataset was biased so we tried to solve this problem using SMOTE ,but it did not help for most cases.

Model training

We tried training many models some of which didn't get a good accuracy of F1 score as, But other models achieved higher accuracy and F1 score as

- Naive Bayes

	precision	recall	f1-score	support
-1	0.56	0.43	0.48	70
0	0.43	0.43	0.43	126
1	0.89	0.91	0.90	804
accuracy			0.81	1000
macro avg	0.62	0.59	0.60	1000
weighted avg	0.81	0.81	0.81	1000

	precision	recall	f1-score	support
advice	1.00	0.10	0.18	10
celebrity	0.80	0.86	0.83	145
info_news	0.73	0.78	0.76	545
others	0.00	0.00	0.00	17
personal	0.50	0.62	0.56	128
plan	0.22	0.17	0.19	82
requests	0.20	0.10	0.13	20
restrictions	0.00	0.00	0.00	2
rumors	0.00	0.00	0.00	15
unrelated	0.57	0.33	0.42	36
accuracy			0.66	1000
macro avg	0.40	0.30	0.31	1000
weighted avg	0.63	0.66	0.64	1000

- LSTM

	precision	recall	f1-score	support
-1	0.27	0.40	0.32	70
0	0.25	0.35	0.29	126
1	0.91	0.82	0.86	804
accuracy			0.73	1000
macro avg	0.48	0.52	0.49	1000
weighted avg	0.78	0.73	0.75	1000

accuracy			0.21	1000
macro avg	0.16	0.33	0.17	1000
weighted avg	0.16	0.21	0.16	1000

- BERT

	precision	recall	f1-score	support
NEGATIVE	0.60	0.54	0.57	70
NEUTRAL	0.55	0.48	0.52	126
POSITIVE	0.91	0.93	0.92	804
accuracy			0.85	1000
macro avg	0.69	0.65	0.67	1000
weighted avg	0.84	0.85	0.84	1000

	precision	recall	f1-score	support
advice	0.40	0.20	0.27	10
celebrity	0.86	0.88	0.87	145
info_news	0.74	0.80	0.77	545
others	0.07	0.06	0.06	17
personal	0.54	0.56	0.55	128
plan	0.35	0.27	0.31	82
requests	0.23	0.15	0.18	20
restrictions	1.00	0.50	0.67	2
rumors	0.38	0.20	0.26	15
unrelated	0.54	0.42	0.47	36
accuracy			0.68	1000
macro avg	0.51	0.40	0.44	1000
weighted avg	0.66	0.68	0.67	1000

The model used for the competition is the BERT transformer because it has highest F1 score (0.67)

Our Trials

- **Naive bayes (alpha = .31) + SMOTE using TF/IDF**

	precision	recall	f1-score	support
-1	0.41	0.56	0.47	70
0	0.40	0.51	0.45	126
1	0.92	0.86	0.89	804
accuracy			0.79	1000
macro avg	0.58	0.64	0.60	1000
weighted avg	0.82	0.79	0.80	1000

- **Naive bayes + SMOTE using BOG**

	precision	recall	f1-score	support
-1	0.41	0.36	0.39	80
0	0.47	0.42	0.44	140
1	0.88	0.91	0.89	780
accuracy			0.80	1000
macro avg	0.59	0.56	0.57	1000
weighted avg	0.79	0.80	0.79	1000

- **SVM + SMOTE using TF/IDF**

	precision	recall	f1-score	support
-1	0.41	0.36	0.39	80
0	0.47	0.42	0.44	140
1	0.88	0.91	0.89	780
accuracy			0.80	1000
macro avg	0.59	0.56	0.57	1000
weighted avg	0.79	0.80	0.79	1000

	precision	recall	f1-score	support
advice	0.50	0.20	0.29	10
celebrity	0.84	0.86	0.85	145
info_news	0.74	0.72	0.73	545
others	0.00	0.00	0.00	17
personal	0.56	0.66	0.60	128
plan	0.19	0.24	0.21	82
requests	0.19	0.15	0.17	20
restrictions	1.00	0.50	0.67	2
rumors	0.17	0.07	0.10	15
unrelated	0.44	0.44	0.44	36
accuracy			0.64	1000
macro avg	0.46	0.38	0.41	1000
weighted avg	0.64	0.64	0.64	1000

- **Naive Bayes using BOG**

```
max f1= 0.5927599454092305 @ alpha= 0.16600000000000012
              precision    recall  f1-score   support

          -1       0.47      0.47      0.47        70
           0       0.43      0.40      0.41       135
           1       0.89      0.90      0.89       795

    accuracy                           0.80      1000
   macro avg       0.60      0.59      0.59      1000
weighted avg       0.80      0.80      0.80      1000

max f1= 0.36997395662325017 @ alpha= 0.014000000000000000
              precision    recall  f1-score   support

      advice       0.30      0.38      0.33         8
   celebrity       0.82      0.84      0.83       142
   info_news       0.72      0.74      0.73       531
      others       0.06      0.07      0.06        14
    personal       0.57      0.52      0.54       140
        plan       0.29      0.24      0.26       102
    requests       0.15      0.14      0.14        22
restrictions       0.50      0.25      0.33         4
      rumors       0.07      0.09      0.08        11
   unrelated       0.33      0.46      0.39        26
...
    accuracy                           0.63      1000
   macro avg       0.38      0.37      0.37      1000
weighted avg       0.63      0.63      0.63      1000
```