Tweet Sentiment Analyzer Report

Anne-Jan Mein

April 21, 2025

1 Exploratory data analysis

When looking at the distribution of amount of tweets per sentiment in the dataset in Figure 1, it can be observed that the dataset is not balanced.

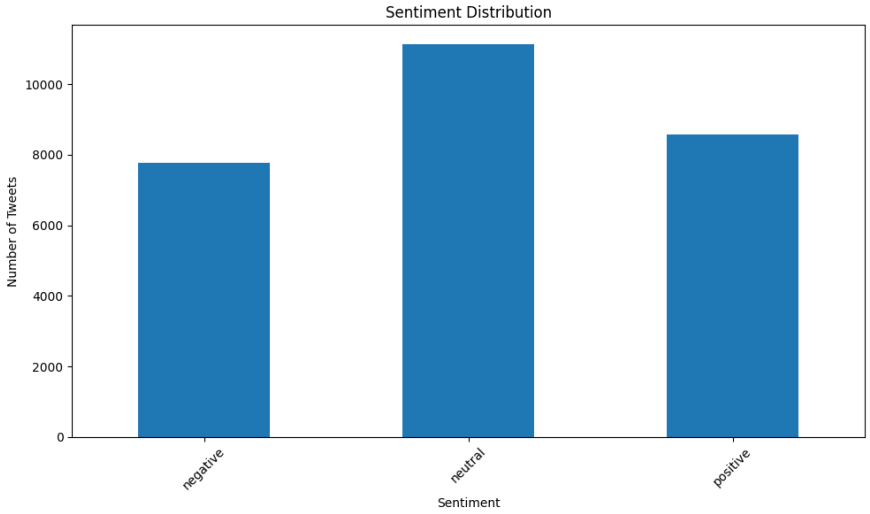
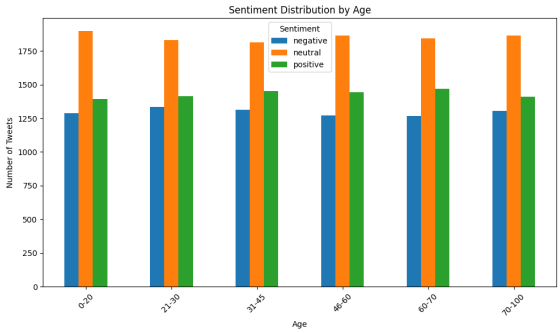


Figure 1: Sentiment distribution

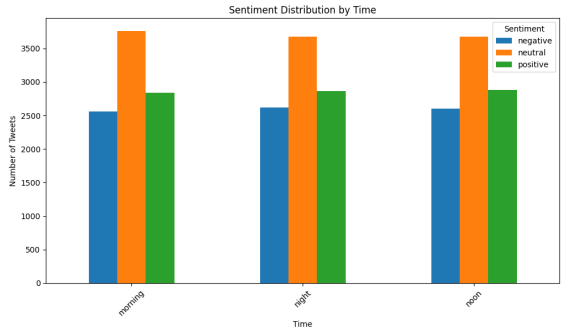
There are 7781 *Negative* tweets, 11117 *Neutral* tweets and 8582 *Positive* tweets. When further examining the dataset, it also contains the following information for each tweet: *textID, selected$_t$ext, Time of Tweet, Age of User, Country, Population -2020, Land Area (km^2) and Density (P/km^2)*. The columns containing *textID* and *selected$_t$exts* are dropped as they contain no information useful for training the sentiment classifier. While *selected$_t$exts* is useful for determining how a sentiment was given to a tweet, it cannot be used as there is no way of knowing how the author of the dataset pruned the original tweets to obtain the final text. It can, however, be used as an example for removing irrelevant parts of the tweet when building the pre-proccessing step.

Visualising both the *Age of User* and *Time of Tweet* in Figures 2a and 2b, it can be observed that these follow roughly the same sentiment distribution as the entire dataset, indicating that these are likely not contributing factors in determining the sentiment of a tweet, and are thus dropped from the dataset.

Next, *population -2020, Land Area (km^2) and Density (P/km^2)* are dropped in favor of working with the *Country* information instead, as this can easily be encoded to 52 labels if desired. The *Country* of which the user of the tweet lives in does not always follow the same distribution as the entire dataset as can be seen in Figure 3. However, in order to keep the complexity of the models low, this information was also dropped from the dataset in order to only use the tweet for training the classifiers.



(a) Sentiment by age



(b) Sentiment by time

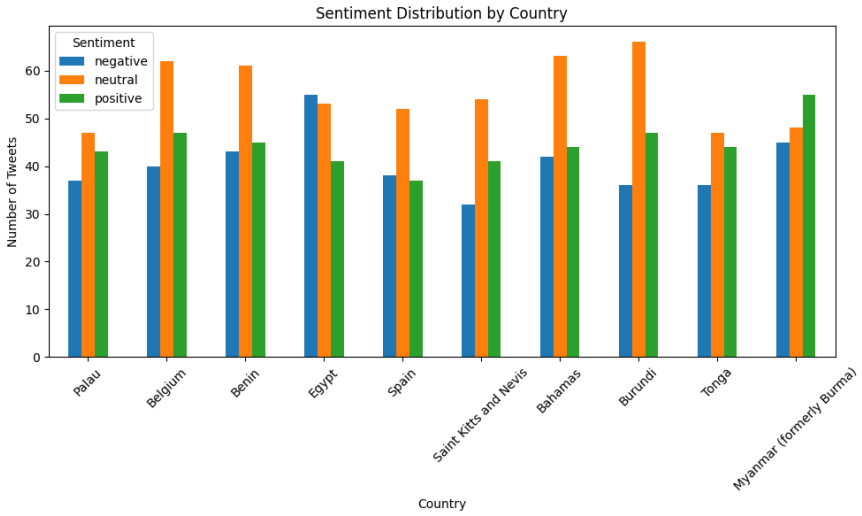


Figure 3: Sentiment by country (10 random country's were sampled for legibility)

2 Data Pre-Processing

The text of the tweets was pruned in order to remove parts that are not helpful for determining its sentiment. The following steps were taken:

- Conversion to lowercase
- Removal of URLS
- Postive and negative emoticons replaced with the words *pos* and *neg* respectively
- Removal of digits
- Removal of punctuation with the exception of the asterisk symbol
- Removal of excess whitespace
- Removal of stopwords
- Lemmatization of words

3 Evaluation

Two models were trained and evaluated, a *Logistic Regression*-based classifier combined with a *Tfidf-Vectorizer* and a *LSTM*-based classifier.

3.1 Logistic Regression

The classifier was trained with a *5-fold cross-validation* with a 80%/20% train/test split. The mean accuracy of the model after training was 69%. Next the classifier was evaluated on the validation dataset, tabelized in Tabel 1

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.66	0.66	0.66	1001
Neutral	0.64	0.67	0.65	1419
Positive	0.78	0.72	0.75	1103
Accuracy			0.68	3525
Macro Avg	0.69	0.68	0.69	3525
Weighted Avg	0.69	0.68	0.69	3525

Table 1: Classification report for the Logistic Regression model on the evaluation set

These scores show that the classifier struggles most with the classification of *neutral* and *negative* tweets.

3.2 LSTM

The *LSTM*-based classifier was trained for 10 epochs with a vocabulary of the 2500 most common words. The classifier was also trained with a *5-fold cross-validation* with a 80%/20% train/test split. For each fold, the model was trained for 10 epochs, as longer training led to a stagnating or decreasing accuracy. The results of the *LSTM* were very similar to the previous classifier, also obtaining a mean accuracy of 69% after training. The results on the validation set can be seen in Table 2 and show that, while it achieves a slightly higher performance, it also struggles most with the classification of *neutral* and *negative* tweets.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.72	0.63	0.67	1001
Neutral	0.63	0.72	0.67	1419
Positive	0.77	0.72	0.74	1103
Accuracy			0.69	3525
Macro Avg	0.71	0.69	0.7	3525
Weighted Avg	0.7	0.69	0.7	3525

Table 2: Classification report for the LSTM on the evaluation set