Analysis in the classification of sentiments

Francisco de Asis dos Santos Silva, Josué Marinho Hinrichs and Leidy Milena Leal Abril

Universidade Federal do Ceará, Campus do Pici, 60451-970, Fortaleza, Ceará, Brazil

Abstract: In this report we show the results more relevant in the analysis of sentiments. We begin by doing the exploratory data analysis of the corresponding dataset and showing some results after the cleaning. We explore three models of machine learning to determine which one of them is the most adequate. We show some metrics to quantify the validity of a model.

Details about the complete computation are found in:

Link notebook: https://github.com/josuhinrichs/mandacaru-desafio2

Link API: https://github.com/josuhinrichs/mandacaru-api

Link ChatBot: https://github.com/LLEALABRIL/mandacaru-chatbot

Contents

We use the dataset posted in Kaggle which consists of news classified as positive, negative or neutral.

1 Exploratory Data Analysis

We begin exploring the percentage of elements of each sentiment as shown in Fig. 1. One can see that only the neutral sentiment comprises more than 50% of the dataset, while the remaining one is divided by positive and negative sentiments.

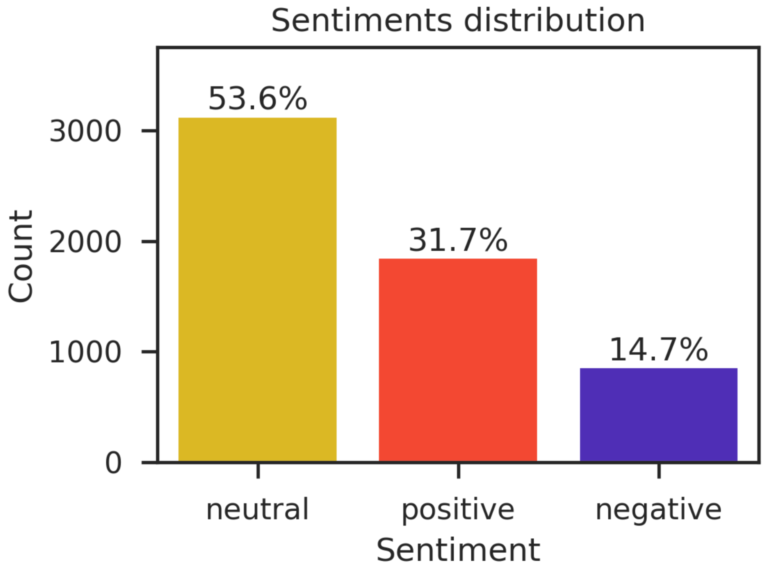


Figure 1: Distribution of sentiments of dataset.

Although there is a great variation in the percentage of sentiments, one can see that the text length average is similar among them as shown in Fig. 2.

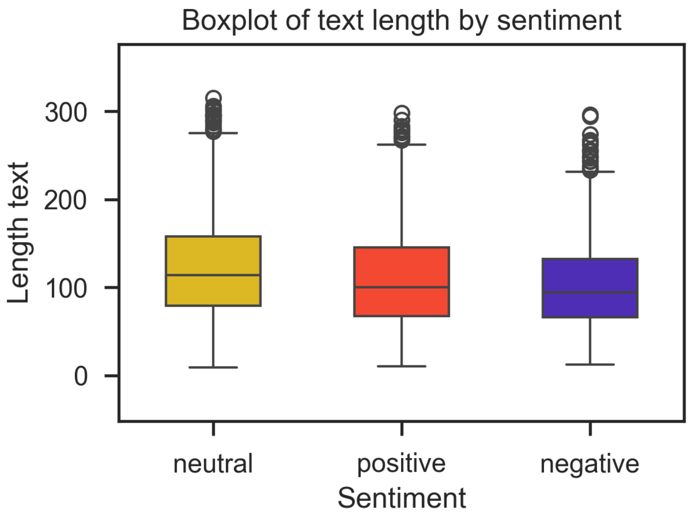


Figure 2: Box plot of the text length for each sentiment

Also, one can determine the words most common in the full dataset without considering the stopwords. We show the top 20 words most common in Fig. **??**.

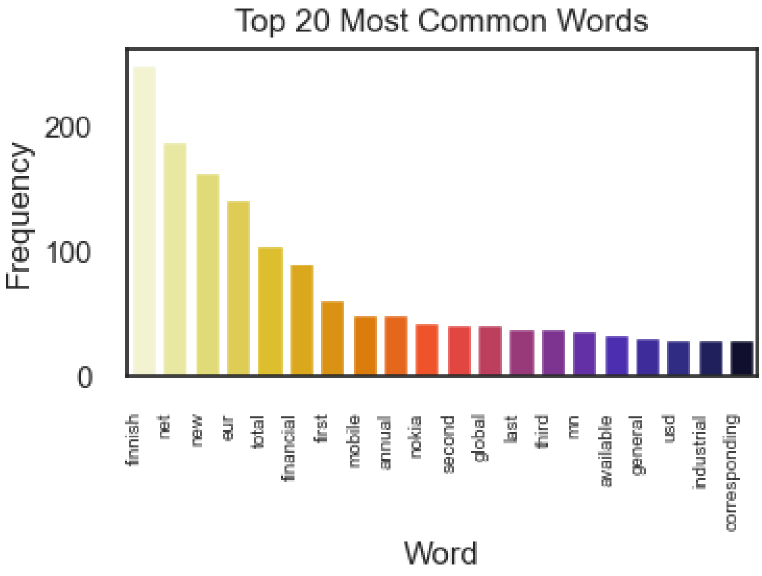


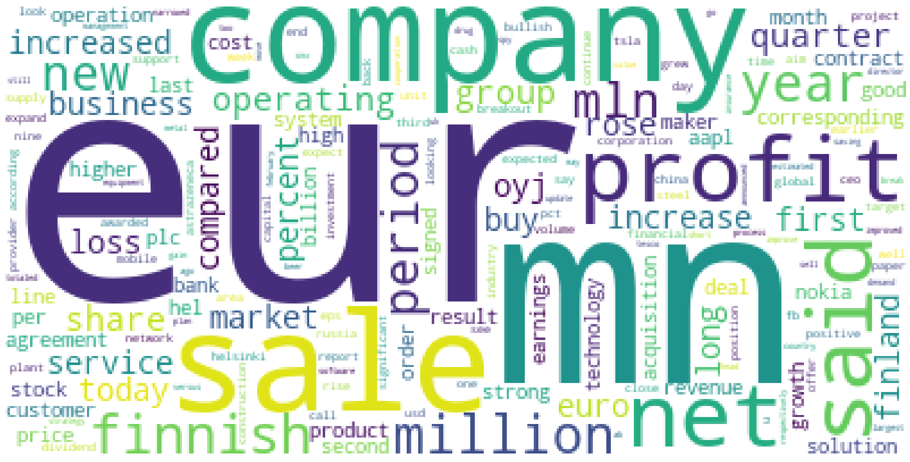
Figure 3: Words most common of the full text

In the EDA we also find rows repeated which are removed to avoid overfit in the model of machine learning.

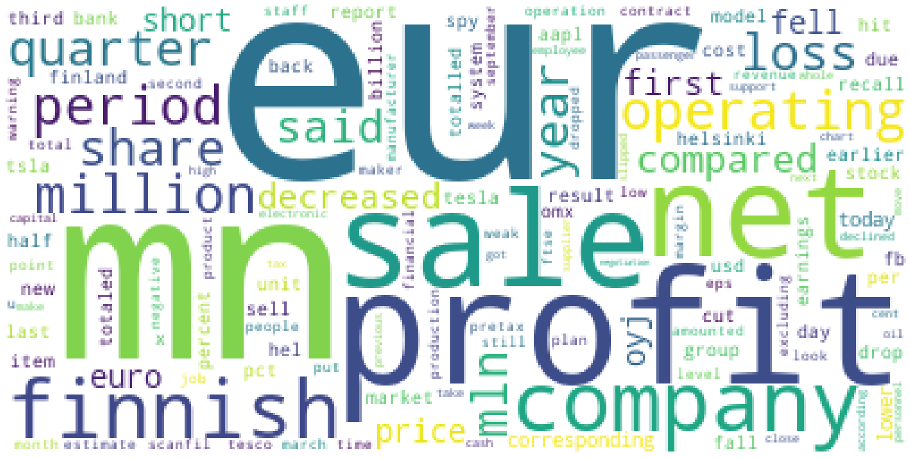
2 Cleaning the data

After tokenizing, removing the stopwords and lematize, we determine the words more common by each sentiment. We plot wordclouds to observe the words most common.

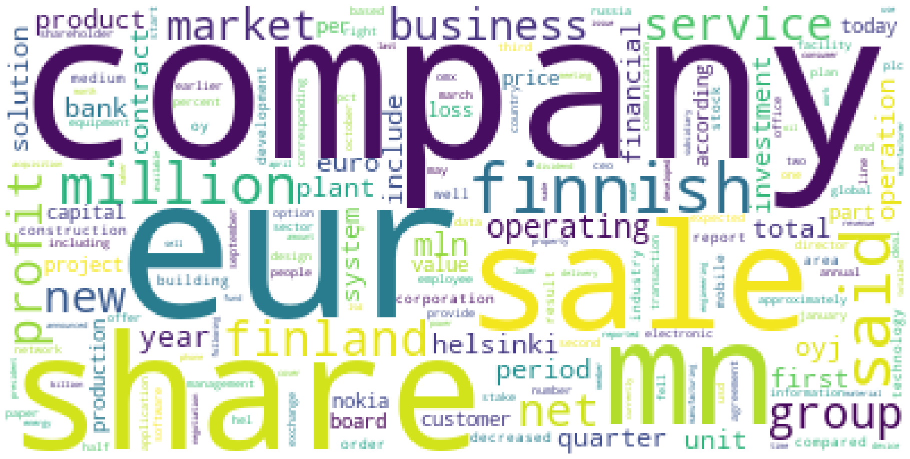
- Positive sentiment:



- Negative sentiment:



- Neutral sentiment:



3 Models of machine learning

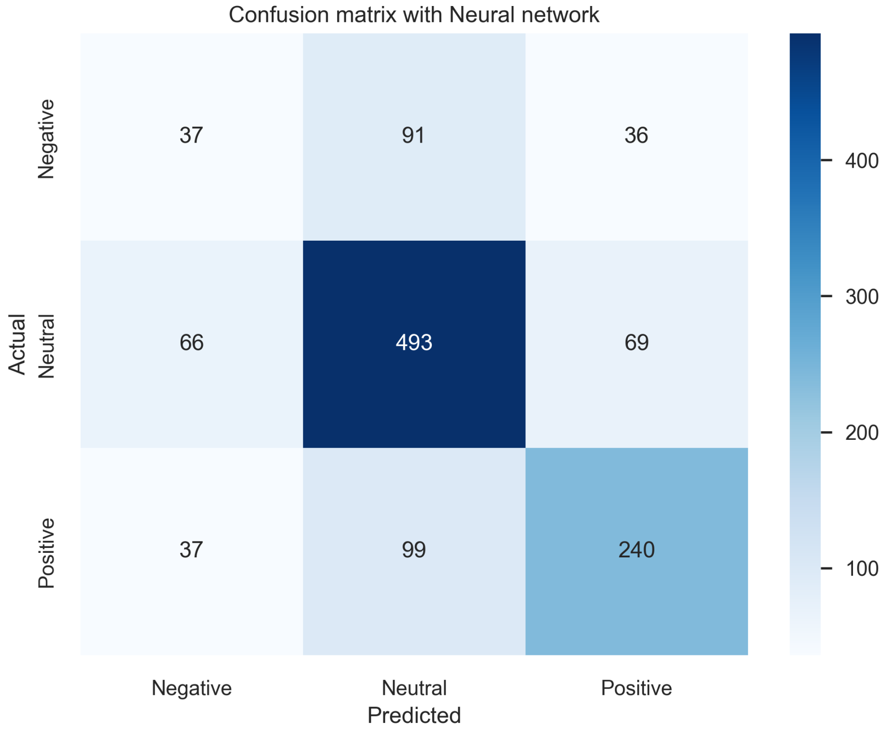
We explore three methods of machine learning: Neural networks (NN), Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM).

We can examine some metrics to determine the best model for classifying sentiments:

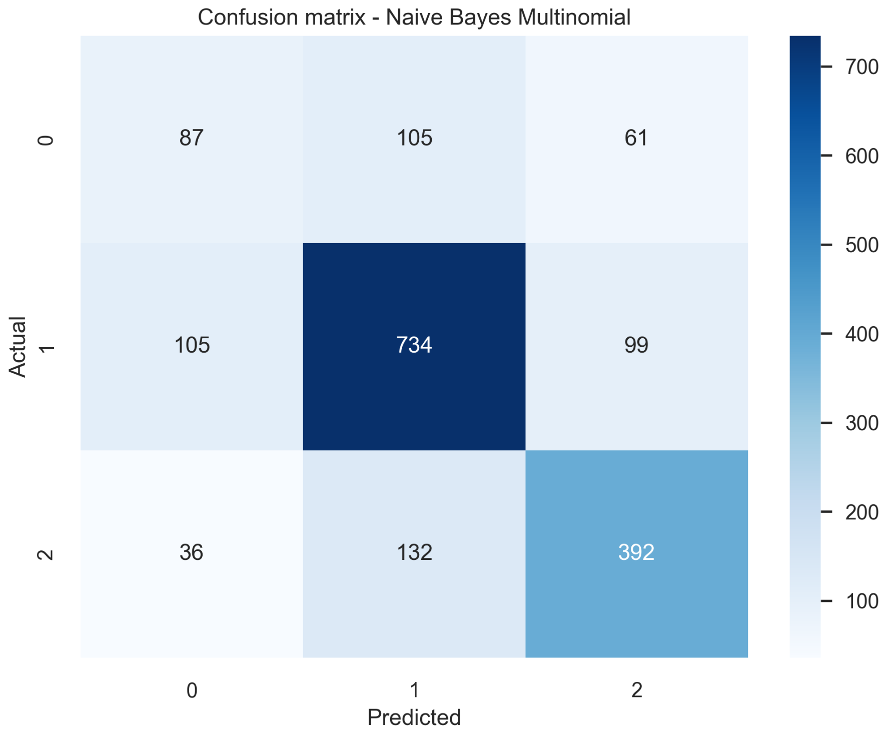
	Neural networks	Multinomial Naive Bayes	Support Vector Machine
Accuracy	0.659	0.692	0.693
Precision*	0.56	0.62	0.64
Recall*	0.55	0.61	0.67
F1-score*	0.55	0.61	0.65
Cross-validation*	-	0.68	0.69

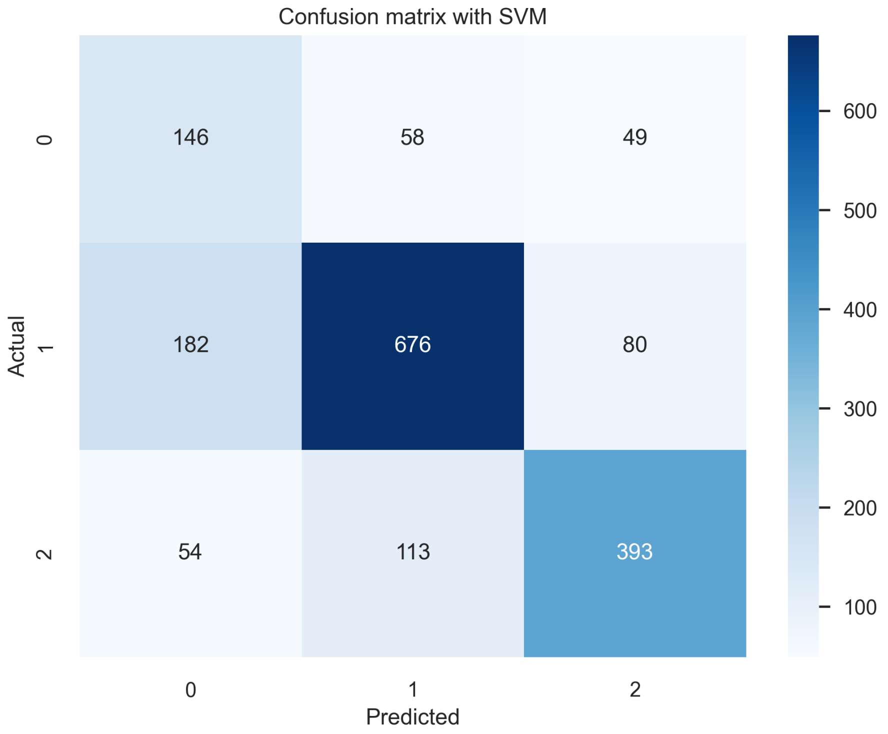
Table 1: Metrics to determine the best model: The symbol $*$ indicates the average over the three sentiments.

The confusion matrix can be calculated for the three methods:



One can observe that the best result in the metrics corresponds to the SVM model.


Confusion matrix - Naive Bayes Multinomial
Actual
0
1
2
87
105
61
105
734
99
36
132
392
0
1
2
Predicted
700
600
500
400
300
200
100


Confusion matrix with SVM
Actual
0
1
2
146
58
49
182
676
80
54
113
393
0
1
2
Predicted
600
500
400
300
200
100