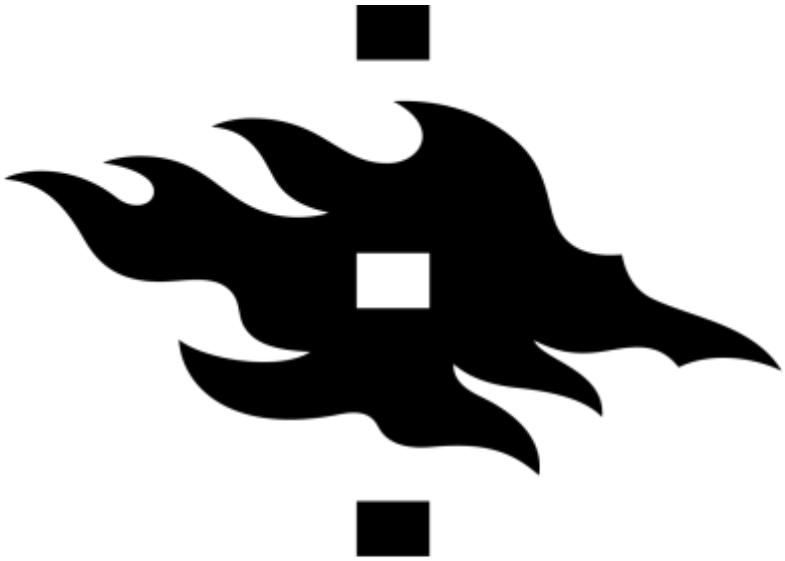


HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

English Studies Master's Program

Module: Computational Literacy

Learner: Jacqueline Guadalupe Armijos Monar

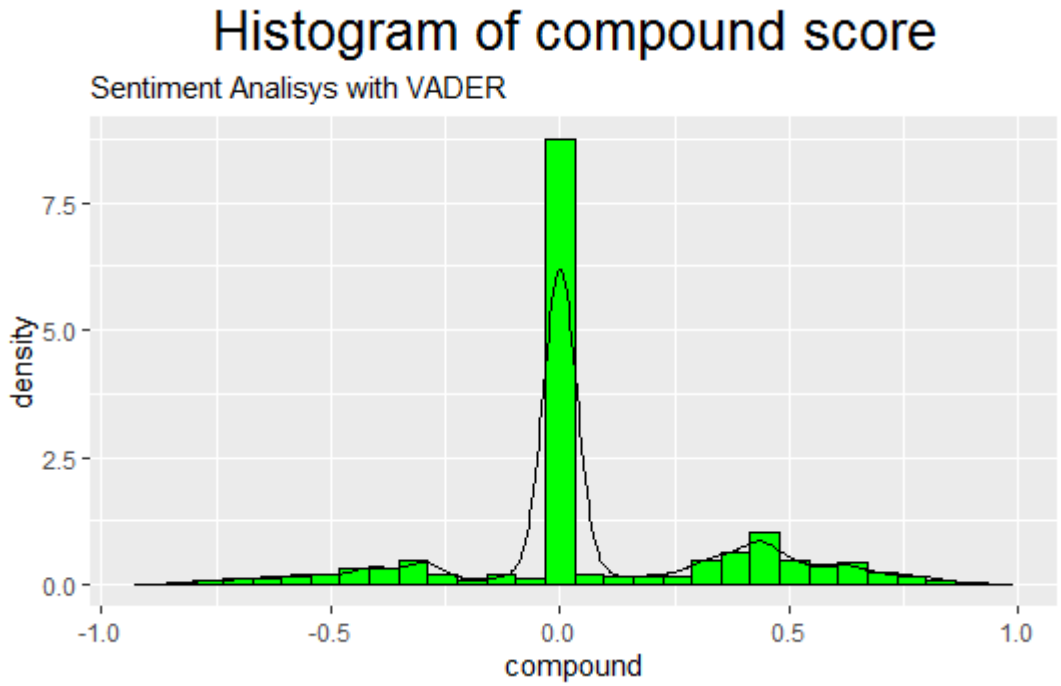
Period: II

Task: Extra assignment.

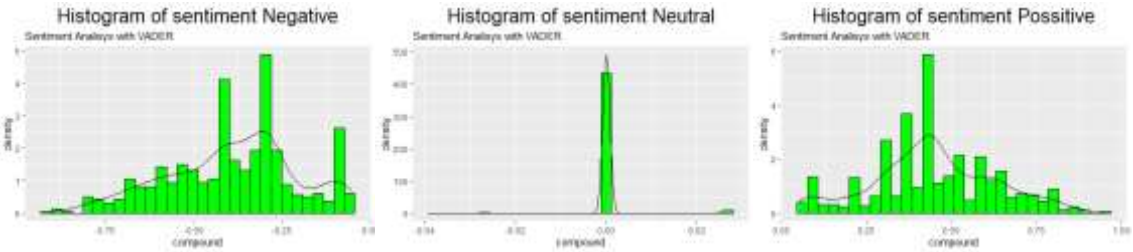
A brief explanation, I completed this extra assignment by selecting one-hundred tweets from the Kaggle corpus, which is a data repository. Essentially, after manually tagging the tweets as positive, negative, or neutral, I ran VADER on the same tweets to obtain their compound scores.

Analysis

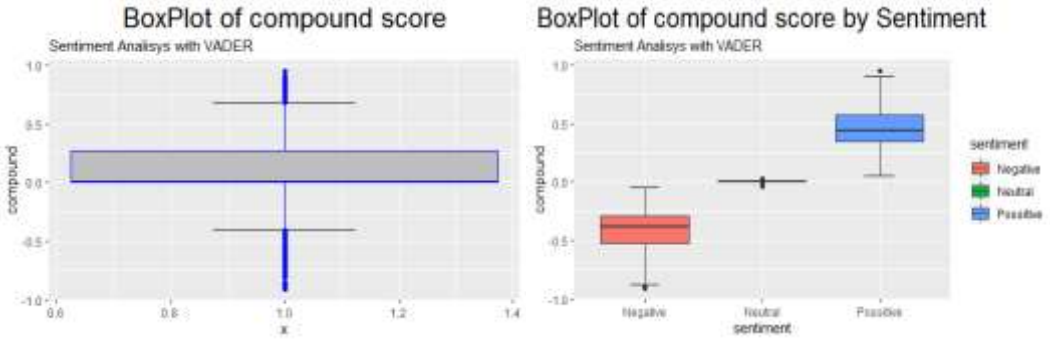
After employing the VADER dictionary for the Sentiment analysis, I obtained a compound score that reflected the sentiment of the analyzed phrase *vaccine.* Then, to ensure the reliability of the analysis, all data underwent a normality test. This step enabled me to generate a histogram with its density curve, which provided a visual representation of the sentiment distribution.



The chart depicts a density curve that does not follow a standard distribution. On the contrary, the compound histogram axis exhibits a greater concentration of data around zero, with minimal frequencies in the tails. Therefore, it suggests a relatively balanced distribution between positive and negative sentiment polarity, while neutral sentiment accounts for the most significant proportion.



When analyzing the sentiment score separately, it showed that the density curves did not resemble a standard curve. Consequently, it was necessary to evaluate the data through box plots so that I could identify data distribution.



After analyzing the global box plots, I observed an equal distribution between the first and the second quartiles. Additionally, the compound score by sentiment box plot shows most of the data accumulated to the left. Interestingly, the data distribution evidenced some atypical data; it means that some texts (tweets) had reached high scores when calculating their sentiment quantifications. On the other hand, I studied sentiment samples separately. They uncovered that the sentiments did not form a box similar to a normal distribution. Mainly, the negative sentiment data occupied the left side, while the positive sentiment data appeared on the opposite side. Based on the explained graphs, I elaborated a normality test for each one of the cases.

For the normality test, I employed the following hypotheses:

H_0: All data stem from a normal distribution.

H_1: All data do not stem from a normal distribution.

Then, I used the Lilliefords normality test to make a final decision. It means to accept or reject the hypothesis according to the significance level, which is 5%. It is essential to mention that the Lilliefords is an improved test based on Kolmogorov Smirnov test. The result below came out from the global data.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  sentiment$compound
D = 0.29121, p-value < 2.2e-16
```

According to the p-value, it is close to zero and less than the level of significance. Hence, it can be concluded that the null hypothesis is rejected. In other words, the data is not distributed under normal law.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  filter(sentiment_text, sentiment == "Negative")$compound
D = 0.082241, p-value = 2.536e-09

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  filter(sentiment_text, sentiment == "Possitive")$compound
D = 0.092513, p-value < 2.2e-16

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  filter(sentiment_text, sentiment == "Neutral")$compound
D = 0.51834, p-value < 2.2e-16
```

When analyzing the data by divided sentiments, they showed a p-value less than the significance level, whereby I rejected the null hypothesis. In that case, I affirmed that the information about sentiments is not similar to a normal distribution.

In the endeavor of verifying the reliability of the analysis, I made use of the VADER dictionary. The employment of the dictionary allowed me to do an eye-hand study of 100 texts, where I verified the expressed sentiments in each published word (semantic resources). Thus, I obtained and verified the following crosstable.

	Negative	Neutral	Possitive
Negative	7	6	2
Neutral	3	49	4
Possitive	2	10	17

Consequently, I found 7 texts in the table above where the negative sentiment emerged. Then, 6 of them were identified as neutral, and 2 of them as positive. On the other hand, 3 texts were valued as neutral, turned as negative, and 4 as positive. On the contrary, 2 texts, which had been qualified as positive at the beginning, became negative manually; the other 10 texts turned neutral. According to this manual analysis, 73% of the calculated data is correct, indicating that the VADER dictionary is reliable. Its classification error was 27%.

Topic modeling

For this current and extra assignment, I used Topic modeling to evaluate users' opinions on Twitter. By employing topic modeling techniques, I generated visualizations that effectively depicted some phrases and words closely related to sentiment. To facilitate this analysis, all dataset was divided into three matrices with each matrix representing a distinct sentiment expressed in the text. Afterward, I processed the information by using two libraries (topic models) and (quanteda) along with the LDA function in R. Studio. The mentioned steps allowed me to categorize words into four groups, each group representing a unique topic. These identified topics showed to be valuable and potential topics related to the executed analysis at hand.

In the final step, I observed the most salient or frequent words in each topic appearing inside a cloud, (word cloud). Each word cloud represented the words that were linked or associated with each one of the sentiments or attitudes. After applying the sentiment analysis and the topic modeling, I recorded all the results and exported them to an Excel file. So, this process allowed me to demonstrate and visualize the classification of the texts based on the primary sentiments expressed in three groups, *positive, neutral,* and *negative*.

Negative

The classification of words depended on their frequency levels. The most frequent words produced greater or bigger word sizes, while the least frequent words occurred in a smaller size. In other words, the more frequently words appeared the larger the word size, while the less frequent words appeared, the smaller the word size resulted in. Therefore, this approach led to the identification of four distinct groups:

- **a)** The first group classified words related to vaccines as a *drug*. In this group of words, the information highlighted the financial aspect, including the necessity to pay for vaccine access. Specifically, India was noticeable. Furthermore, some concerns were raised regarding the potential risk of developing autism, as part of the side effects.
- **b)** The second group had a worldwide perspective and mainly discussed countries that had developed the Covid-19 vaccine. Thereby, criticism was against those powerful nations, who were involved in vaccine creation. The

criticism and people's discourses against the Covid-19 vaccine creation were revolving around political influence and political issues.

c) On the other hand, in the third group, the words showed that the use of Covid-19 vaccines had been used previously in clinical trials to counteract the effects of the disease.

d) Finally, the fourth group of words reflected some emotional concerns associated with the covid-19 vaccine. Notably, the words emphasized the negative polarity by showing delays in vaccination programs due to shortages of the available vaccines.

It can be said that the categorization of the words into four different groups provided insights into different situations or aspects related to the discourse about Covid-19 vaccines.

Positive

a)

b)

c)

d)

The word cloud charts showed salient positive words extracted from tweets. They were divided into four distinct topics or groups. This analysis identified the most dominant words according to their internal frequency scale.

a) The first cloud of words highlighted the importance of the carried work to combat the Covid-19 virus. The realized work was portrayed as good and necessary to face and battle the pandemic.

b) The second cloud of words indicated a positive perception of the Covid-19 tests and their efficacy. Therefore, the tests showed an agreement with the Covid-19 vaccination efforts.

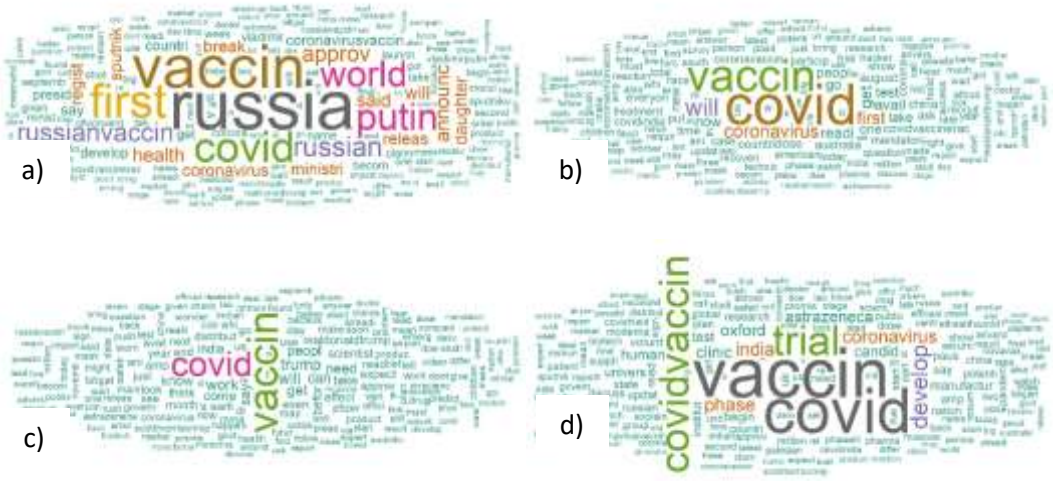
c) The third-word cloud depicted a positive appreciation for Russia. This country appeared as a pioneering country in developing a vaccine and giving hope to the world.

Finally, **d)** In this group of words, it could be seen expressions related to various vaccination phases, covering from its initial stages to the current ones. These words demonstrated a significant interest in obtaining the vaccines. Additionally, these words highlighted a widespread sentiment to get protected against infection.

To sum up, the word cloud analysis demonstrated the apparent sentiments and dominant topics in the examined tweets. In addition, they captured positive attitudes

towards the Covid-19 calamity battle, and the vaccination efforts to proceed with the campaign.

Neutral



In the frame of neutral sentiments, the word clouds did not reflect either a pertain or a specific theme. In general, the word clouds presented information as follows:
a) The words did not delimit a specific topic; however, a high proportion of words mentioned Russia, Covid, World, and Putin. Those words did not communicate positive or negative sentiments specifically regard to the keyword *vaccine*, or the Covid-19 disease itself, as it occurs in items: **b)**, **c)**, and **d)**. As a result, the observed words did not determine a coherent relationship or a specific topic of discussion within this present study.