# Exploratory data analysis - Classification of news headlines with impact on the probability of stock prices changes

Max Franke

March 31, 2020

Course: CIS-627 CAPSTONE

Term: Spring/T2

Last modified date: 03/31/2020

Data: news headlines and descriptions ~5k news articles pertaining to several publicly traded companies

## Approach of the analysis

This analysis aims to visualize the collected data to include subsequent results and relationships in the further process of the project. The following procedure is applied:

First, a data set is examined that evaluates the mood of headlines and their descriptions concerning different stocks in a labeled form.

Based on this, the most frequently used words from headline and description are linked to its frequency. The background to this approach is that the assumption is made that the defined words are representative of a word family that influences stocks so that the trend of these words can be examined more closely.

## Data presenation

In this section the data in its raw version is presented.

### Description of data

The dataset shows four columns. The first two columns are "text"-columns (title and description). The third column shows the stock symbol. The fourth column is the labeling of the first three columns. It is divided into three classes:

1 if the news is positive for the company and may encourage people to buy shares.

0 if the news is neural or not possible to identify it as positive or negative

-1 if the news is negative for the company, bad publicity, or would discourage people from owning shares.

Table 1: Dataset stocks with headlines and description

| title | description | symbol | sentiment |
|---|---|---|---|
| Apple updates iMac with Intel processors | (Reuters) - Apple Inc said it updated the iMac with fourth generation Intel Corp processors, better graphics, next generation Wi-Fi and faster flash storage options. | AAPL | 1 |
| Google pays $55 million tax in Britain on 2012 sales of $5 billion | LONDON (Reuters) - Google, which has been grilled twice in the past year by a UK parliamentary committee over its tax practices, had a UK tax bill of 35 million pounds ($55 million) in 2012, on sales | GOOGL | -1 |
| Microsoft plans to cut 1,000 jobs in Finland -newspaper | HELSINKI, July 16 (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources. | MSFT | 0 |
| Microsoft plans to cut 1,000 jobs in Finland: newspaper | HELSINKI (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources. | MSFT | 0 |
| Smartphone suit against Google plays into rivals' hands | SAN FRANCISCO (Reuters) - A U.S. consumer lawsuit accusing Google of monopolizing prime real estate on Android smartphones will help mobile rivals like Microsoft make their antitrust case with Europea | GOOGL | -1 |
| Apple should do more to tackle in-app purchases problem: EU | BRUSSELS (Reuters) - Apple has provided no concrete and immediate solutions to tackle the problem of adults and children racking up credit card bills by making "in-app" purchases on tablets and mobile | AAPL | -1 |

# Analyze the distribution

## Graph: Total numbers of symbols

The first graph shows the total numbers of symbols go give a visualization about the distribution of the stocks on a high level.

### Total Numbers of Symbols



Source: news headlines of ~5k news articles
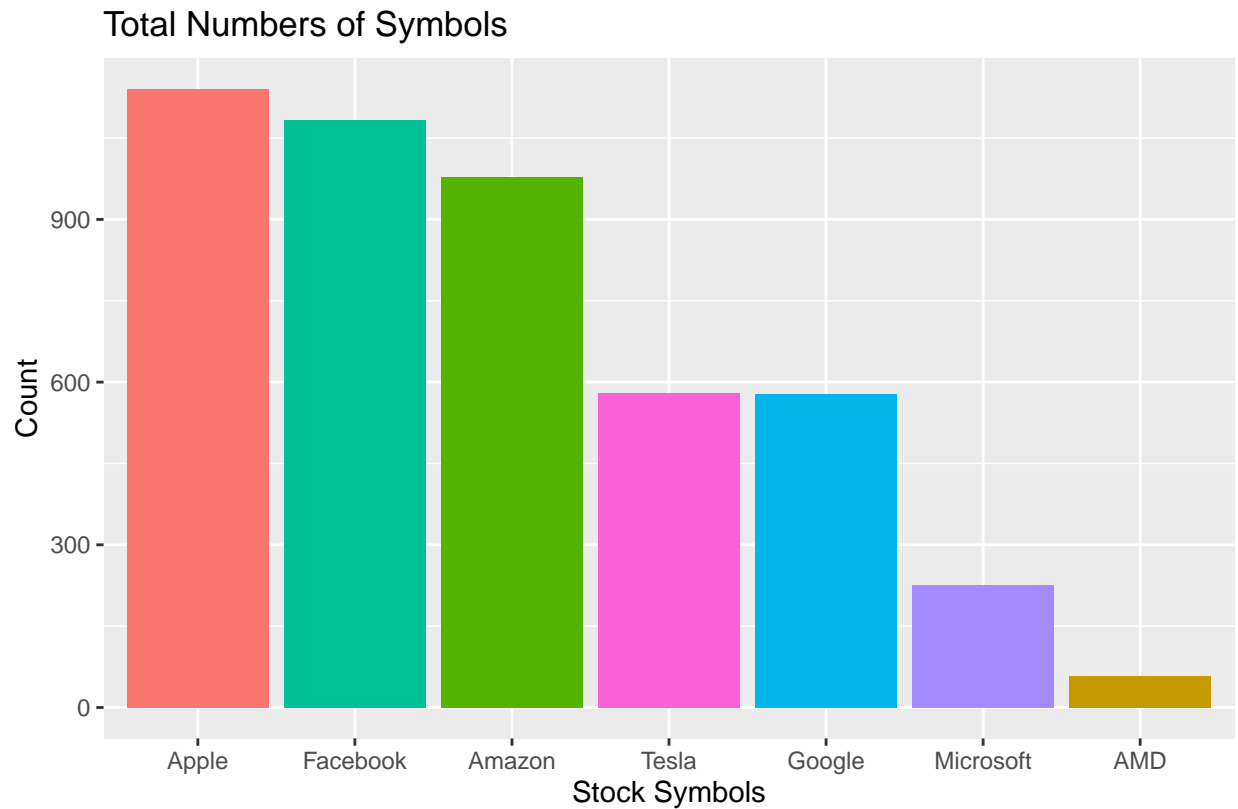
**Interpretation of the graph:**

Apple, Facebook, and Amazon show the highest frequency. The sum of these three stocks is 46.9% of the total dataset.

In the next step, the relationship between the stocks per sentiment will be analyzed.

**Graph: Number of stock symbols per sentiment**

Number of Symbols per Sentiment
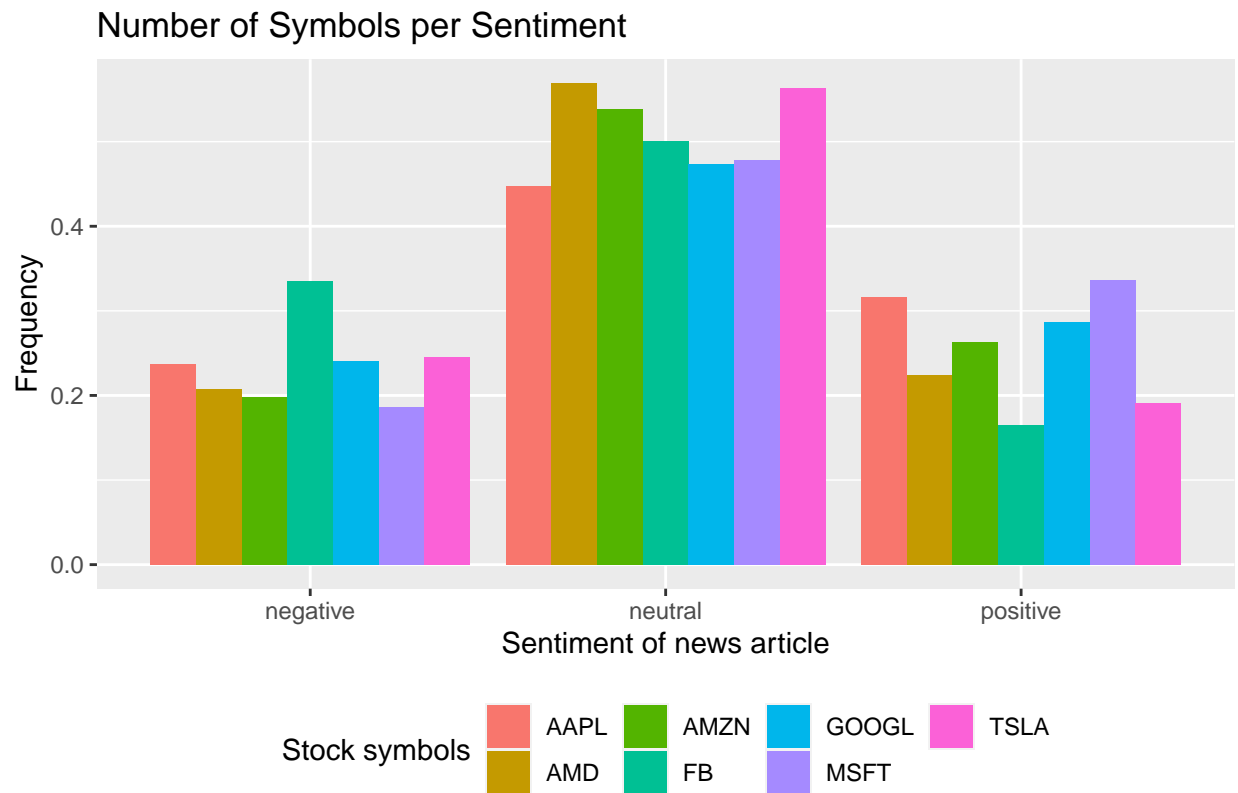


Source: news headlines of ~5k news articles

**Interpretation of the graph:**

The frequency of symbols per sentiment is shown above. It can be seen that facebook has a high number of bad publicity (33.6% of all facebook news are labeled as a negative sentiment).

In general, the frequency in the neutral area is the strongest (50% of all cases in the dataset) Apple and Microsoft show a high number with positive news.

# Word clouds

## Process of generating word clouds for the column titles

### Step 1: Extract and then load the titles from dataset

First, the column "title" is extracted and saved to a txt.file. Then, this txt. file is load into R. The following table shows the first 6 titles of this file:

Table 2: First 6 titles

| Number | Titles |
|---:|---|
| 1 | "Apple updates iMac with Intel processors" |
| 2 | "Google pays $55 million tax in Britain on 2012 sales of $5 billion" |
| 3 | "Microsoft plans to cut 1,000 jobs in Finland -newspaper" |
| 4 | "Microsoft plans to cut 1,000 jobs in Finland: newspaper" |
| 5 | "Smartphone suit against Google plays into rivals' hands" |
| 6 | "Apple should do more to tackle in-app purchases problem: EU" |

### Step 2: Required packages

For this analysis, following packages are required:

- tm (text mining)
- snowball (text stemming)
- wordcloud (generator for the visualization)
- RColorBrewer (palettes of colors)

### Step 3: Text mining

1. Load the txt. file as a corpus

2. Inspect the corpus

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 10
##
##  [1] "Apple updates iMac with Intel processors"
##  [2] "Google pays $55 million tax in Britain on 2012 sales of $5 billion"
##  [3] "Microsoft plans to cut 1,000 jobs in Finland -newspaper"
##  [4] "Microsoft plans to cut 1,000 jobs in Finland: newspaper"
##  [5] "Smartphone suit against Google plays into rivals' hands"
##  [6] "Apple should do more to tackle in-app purchases problem: EU"
##  [7] "Federal appeals court set to hear Microsoft 'cloud' case"
##  [8] "Mystery of 'Gold Artifact' That Stumped Archaeologists Solved by FB User"
##  [9] "How Mark Zuckerberg could prevent gun violence"
## [10] "Banksy's Steve Jobs mural spotlights refugee crisis"
```

3. Text transformation

- Replace special characters from text

4. Cleaning the text

- Convert the text to lower case
- Remove punctuations
- Eliminate extra white spaces
- Remove english common stopwords
- Remove the name of the stocks

**Step 4: Building a term-document matrix**

A term-document matris is a table which contains the frequency of the words.

In the following table, the first 10 words of this term-document matrix are shown.

Table 3: Top 10 words in title by frequency

| word | freq |
| --- | --- |
| update | 720 |
| says | 372 |
| new | 315 |
| india | 175 |
| china | 172 |
| iphone | 152 |
| data | 139 |
| will | 138 |
| report | 137 |
| tech | 136 |

**Step 5: Generate the word cloud**

<div align="center">

### Wordcloud of news headlines (company names excluded)

</div>
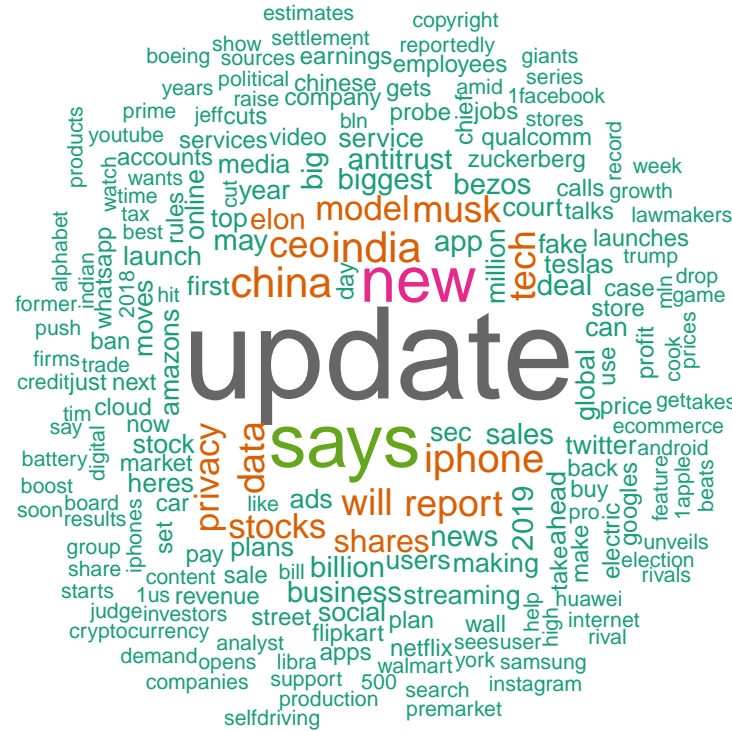


**Interpretation of the graph:**

Eyecatcher of this graph is the word "update", which will be analyzed in the following chapters. Also, "says" and "news" are highlighted.

Furthermore, words like model, tech and, data are highlighted which is interesting and can be interpreted as an indicator for the impact of the field "data science" in the stock market.

**Step 6: Frequent terms in the term-document matrix**

The following table shows the words which have the lowest frequence of 30. The table is ordered by column, which means that the first column presents the words 1-10.

<div align="center">

Table 4: Frequent words in title

</div>

| Words 1-10 | Words 11-20 | Words 21-30 | Words 31-40 | Words 41-50 | Words 51-60 |
|---|---|---|---|---|---|
| billion | set | support | report | talks | video |
| million | user | elon | shares | deal | trump |
| sales | zuckerberg | musk | twitter | internet | top |
| tax | back | calls | streaming | 1apple | wants |
| cut | take | street | data | bln | ads |

| Words 1-10 | Words 11-20 | Words 21-30 | Words 31-40 | Words 41-50 | Words 51-60 |
|---|---|---|---|---|---|
| jobs | iphone | tech | update | car | political |
| plans | moves | hit | fake | ceo | walmart |
| case | cook | store | stock | firms | board |
| cloud | tim | make | settlement | antitrust | 1us |
| court | says | plan | sources | record | lawmakers |

**Identify the frequency of top 10 words**

The following table shows the top 10 words by frequency in the title.

Table 5: Frequency of top 10 words in title

| word | freq |
|---|---|
| update | 720 |
| says | 372 |
| new | 315 |
| india | 175 |
| china | 172 |
| iphone | 152 |
| data | 139 |
| will | 138 |
| report | 137 |
| tech | 136 |

**Graph: Word frequency for the top 10 words in the title**

This graph visualize the top 10 words in the title

Frequency of the top 10 words in the title

company names excluded
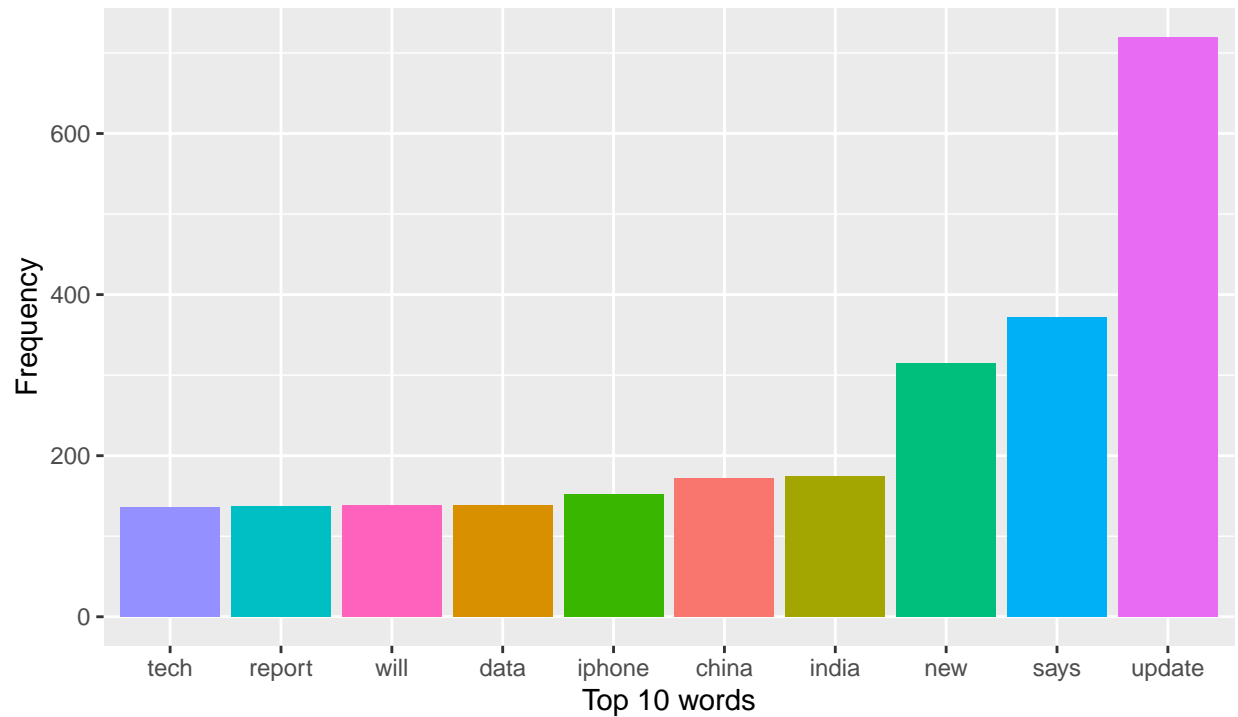


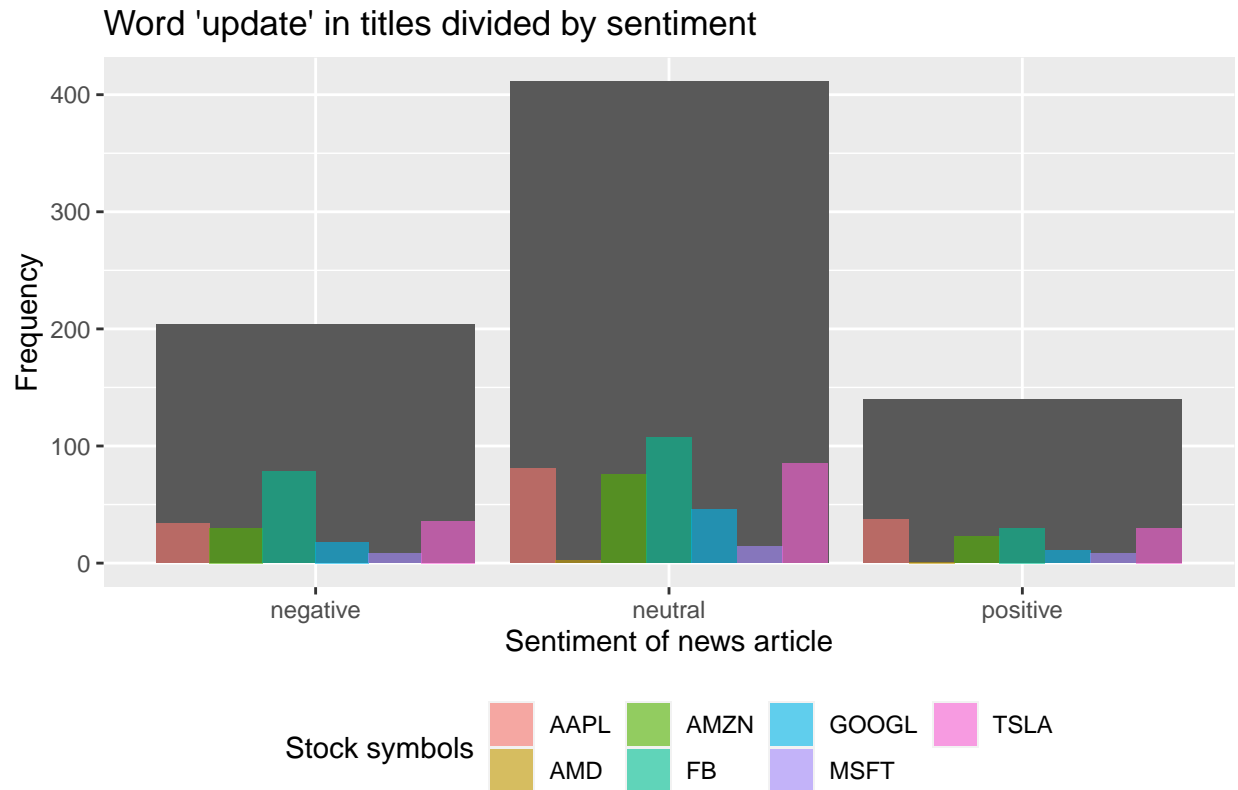Source: news headlines of ~5k news articles

**Interpretation of the graph:**

The word "update" is by a wide range the most frequent word in the title.

Consequently, a further analysis concerning update and sentiment would be interesting.

**Graph: The word "update" in titles, split by sentiment**

This graph selected the cases in the dataset, which includes in the column "title" the word "update" in any form. The big grey bar charts represents the overall frequency of the word "update", split by the three classes in the sentiment. The colorful bar charts shows the distribution per each stock, split by the three classes in the sentiment.

## Word 'update' in titles divided by sentiment



Source: news headlines of ~5k news articles

**Interpretation of the graph:**

The word "update" shows concerning the different values of sentiment kind of a normal distribution, which means that the most titles including the word "update" have a neutral sentiment. But there is a small trend that the word update leads to negative sentiment in the news title.

Split by stocks, it is shown that facebook shows a high negative frequency, which was already shown in the Graph: Number of stock symbols per sentiment.

## Process of generating word clouds for the column description

### Step 1: Extract and then load the descriptions from dataset

First, the column "description" are extracted and saved to a txt.file. Then, this txt. file is load into R.

Table 6: First 6 descriptions

| Number | descriptions |
|---:|---|
| 1 | "(Reuters) - Apple Inc said it updated the iMac with fourth generation Intel Corp processors, better graphics, next generation Wi-Fi and faster flash storage options." |
| 2 | "LONDON (Reuters) - Google, which has been grilled twice in the past year by a UK parliamentary committee over its tax practices, had a UK tax bill of 35 million pounds ($55 million) in 2012, on sales" |
| 3 | "HELSINKI, July 16 (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources." |
| 4 | "HELSINKI (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources." |
| 5 | "SAN FRANCISCO (Reuters) - A U.S. consumer lawsuit accusing Google of monopolizing prime real estate on Android smartphones will help mobile rivals like Microsoft make their antitrust case with Europea" |
| 6 | "BRUSSELS (Reuters) - Apple has provided no concrete and immediate solutions to tackle the problem of adults and children racking up credit card bills by making \"in-app\" purchases on tablets and mobile" |

### Step 2: Required packages

For this analysis, the same packages as shown in the chapter for the titles are required.

### Step 3: Text mining

1. Load the txt. file as a corpus

2. Inspect the corpus

3. Text transformation

- Replace special characters from text

4. Cleaning the text

- Convert the text to lower case
- Remove punctuations
- Eliminate extra white spaces
- Remove english common stopwords
- Remove the name of the stocks

**Step 4: Building a term-document matrix**

A term-document matris is a table which contains the frequency of the words.

In the following table, the first 10 words of this term-document matrix are shown.

Table 7: Top 10 words in description by frequency

| word | freq |
|---|---|
| said | 1066 |
| will | 627 |
| new | 611 |
| company | 445 |
| thursday | 390 |
| tuesday | 381 |
| wednesday | 355 |
| monday | 321 |
| incs | 318 |
| chief | 299 |

**Step 5: Generate the word cloud**

Wordcloud of news description (company names excluded)
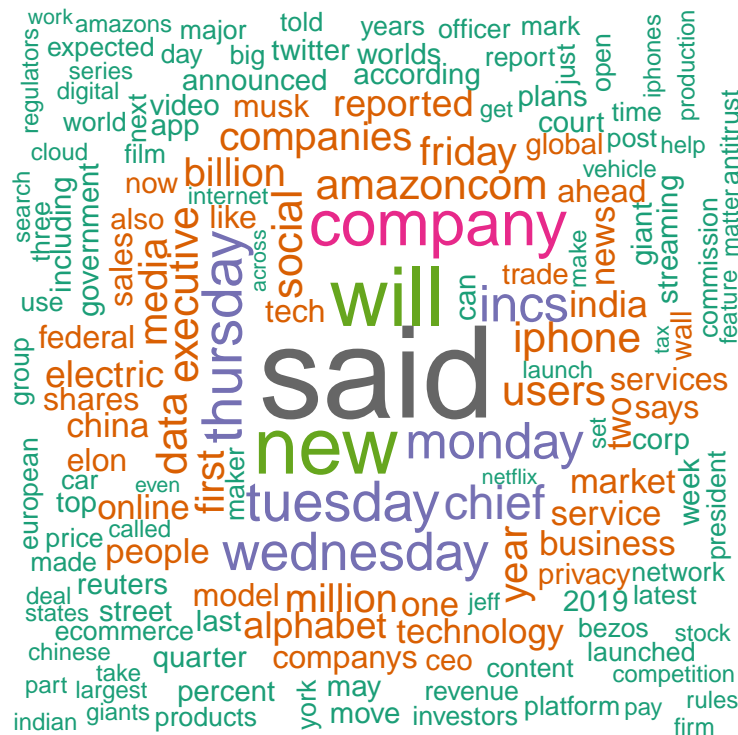
**Interpretation of the graph:**

Eyecatcher of this graph is the word "said", which will be analyzed in the following chapters. Also, "will" and "new" are highlighted.

**Step 6: Frequent terms in the term-document matrix**

The following table shows the words which have the lowest frequence of 30. The table is ordered by column, which means that the first column presents the words 1-10.

Table 8: Frequent words in description

| Words 1-10 | Words 11-20 | Words 21-30 | Words 31-40 | Words 41-50 | Words 51-60 |
|---|---|---|---|---|---|
| better | tax | antitrust | smartphones | united | use |
| corp | year | case | will | week | appeared |
| next | cut | consumer | card | york | ceo |
| reuters | jobs | help | credit | according | late |
| said | mobile | lawsuit | making | user | latest |
| bill | phone | like | court | big | one |
| million | sources | make | customers | business | allow |
| past | unit | prime | federal | data | comments |
| practices | wednesday | rivals | new | founder | group |
| sales | android | san | states | says | made |

**Identify the frequency of top 10 words**

The following table shows the top 10 words by frequency in the description.

Table 9: Frequency of top 10 words in description

| word | freq |
|---|---|
| said | 1066 |
| will | 627 |
| new | 611 |
| company | 445 |
| thursday | 390 |
| tuesday | 381 |
| wednesday | 355 |
| monday | 321 |
| incs | 318 |
| chief | 299 |

**Graph: Word frequency for the top 10 words in the description**

This graph visualize the top 10 words in the descriptions

## Frequency of the top 10 words in the description
company names excluded
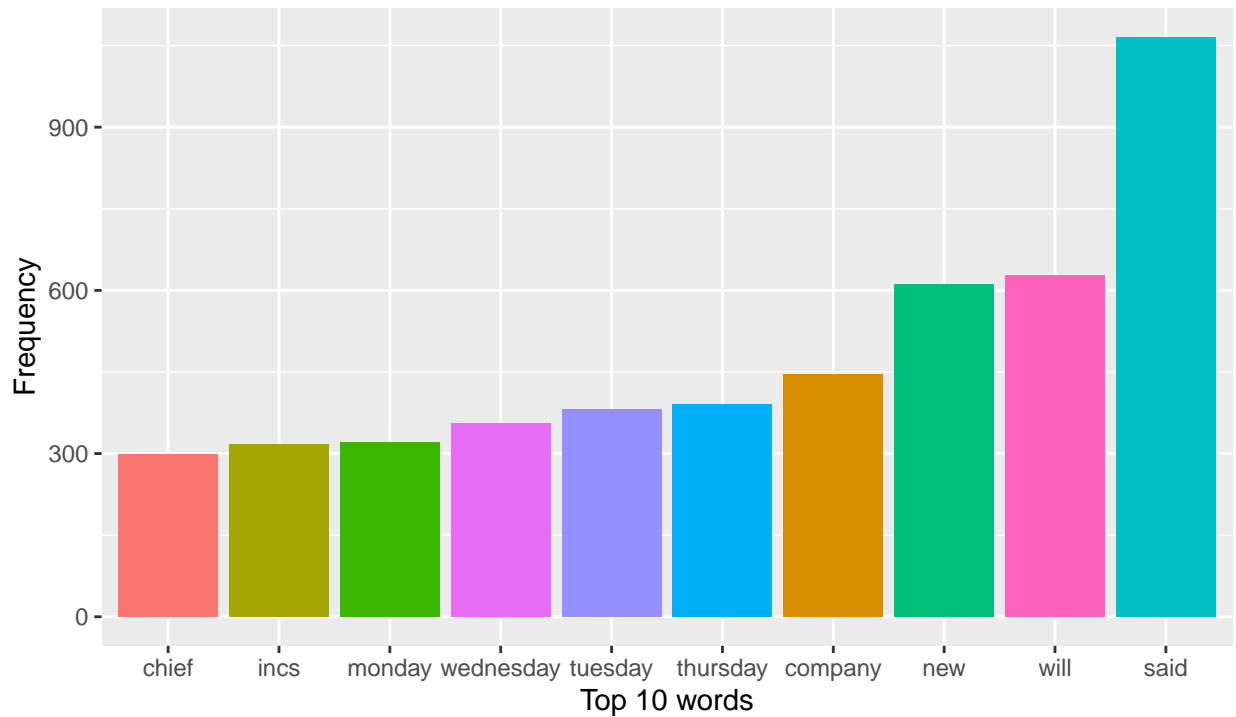


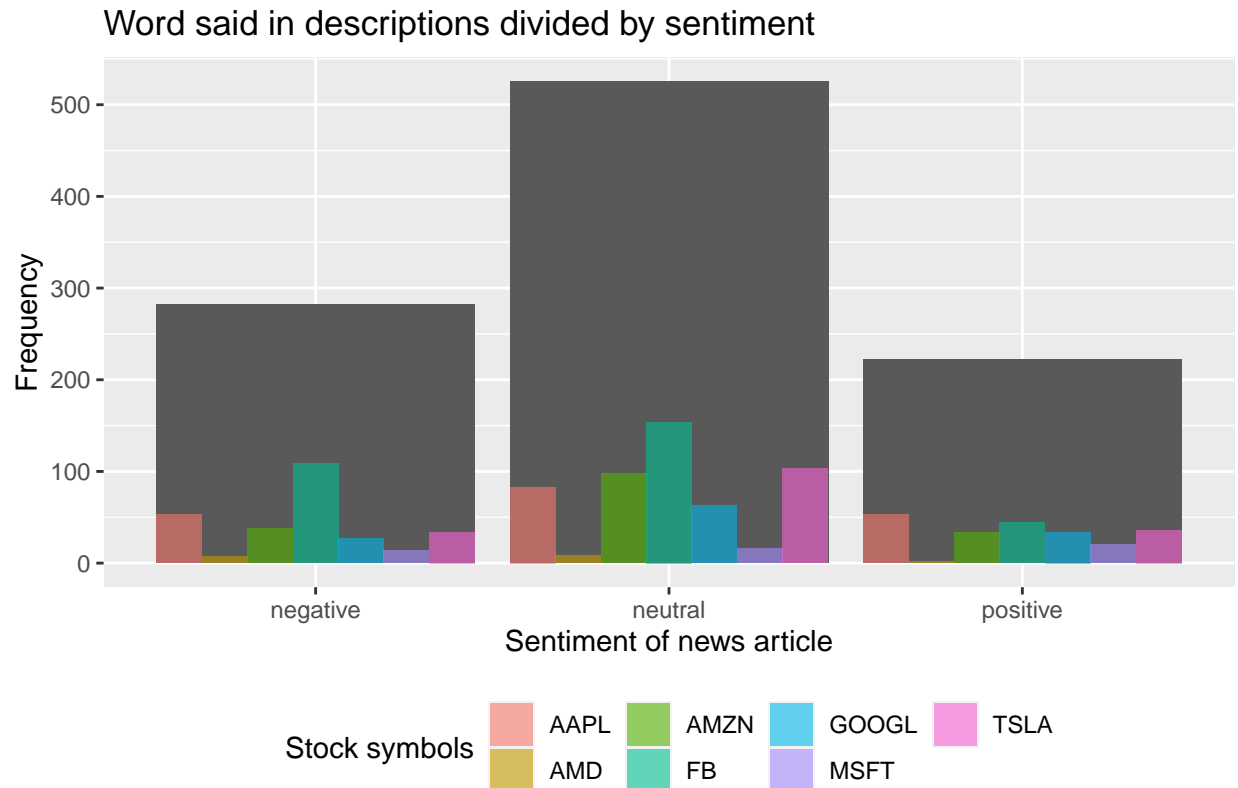Source: news headlines of ~5k news articles

**Interpretation of the graph:**

The word "said" is by a wide range the most frequent word in the description. So, somebody said something which has an impact on the stock

Consequently, a further analysis concerning the word "said" and sentiment would be interesting.

**Graph: The word "said" in descriptions, split by sentiment**

This graph selected the cases in the dataset, which includes in the column "description" the word "said" in any form. The big grey bar charts represents the overall frequency of the word "said", split by the three classes in the sentiment. The colorful bar charts shows the distribution per each stock, split by the three classes in the sentiment.

## Word said in descriptions divided by sentiment



Source: news headlines of ~5k news articles

**Interpretation of the graph:**

The word "said" shows concerning the different values of sentiment kind of a normal distribution, which means that the most descriptions including the word "said" have a neutral sentiment. But there is a small trend that the word "said" leads to negative sentiment in the news description.

Split by stocks, it is shown that facebook shows a high negative frequency, which was already shown before.

# Quick Summary

- In this data set, which contains news, the three companies (Apple, Facebook, Amazon() were most talked about
- The news was most often neutral in terms of sentiment (50% of all cases in the dataset)
- Facebook shows a high frequency in the negative sentiment
- The analysis of the titles shows that the word "update" is often used, although the distributions between the sentiments are also similarly normally distributed
- A word family (e.g. "model", "tech" and "data") which can be connected to the field of 'Data Science' are highlighted in teh word cloud
- The analysis of the descriptions shows that the word "said" is often used. This can mean that the description mentions that someone has said something and that this statement has an impact on the stock market.