Leibniz Institute
for the Social Sciences

Automatic Sampling and Analysis of YouTube Data

Sentiment Analysis of User Comments

Johannes Breuer, Annika Deubel, & M. Rohangis Mohseni

February 15th, 2023

gesis Leibniz Institute for the Social Sciences

Setup

Note: In previous versions of `R`, all strings were automatically translated to factors when reading data. This has been changed with version 4.0.0.

```
# Only necessary for R versions < 4.0.0
options(stringsAsFactors = FALSE)

# Load the data set
comments <- readRDS("./content/data/ParsedEmojiComments.rds")
```

Sentiment Analysis

- The basic task of sentiment analysis is to detect the positive and negative **valence** of a sentence or a collection of sentences
- Sentiment is often used in market research for product reviews
- For *YouTube* videos, we can look at the sentiment in the comments to quantify the valence of user reactions
- *Note*: There are also other methods for detecting:
    - emotional states
    - political stances
    - objectivity/subjectivity

gesis Leibniz Institute for the Social Sciences

Basic Idea of Sentiment Analysis

We compare each word in a sentence with a predefined dictionary

- Each word gets a score: For example a score between -1 (negative) and +1 (positive), with 0 being neutral
- We add up all the scores for a sentence to get an overall score for the sentence

The current health model focuses on disease care, favors the financial aspect, is fragmented, and each actor performs its functions without coordination with the others. In addition, in terms of health, the national Government does not convene, lead or coordinate public and private actors to guarantee rights. You are not taking care of your citizenship. Through a territorial model in health, we are going to guarantee a better quality in the provision of services and effective access for all inhabitants. The main debt of the system is to improve equity, which can be summed up in bringing true quality of services to the most unprotected sectors of society, especially the inhabitants who reside in the most distant territories of the country. Gradually we will seek to conform throughout the national territory what we call a territorial model in health. In each municipality and department of Colombia, work will be done for the adequate provision of services and the real articulation between the actors of the system, such as the EPS, IPS and district, municipal and departmental secretariats. To achieve this objective, we will focus on compliance with quality standards, technical, administrative and financial performance of the EPS. The insurance management will be accompanied by a strengthened stewardship at the national and territorial levels. In order to increase the efficiency of the system, we will promote management that takes advantage of associativity schemes between territorial entities, to account for ethnic and social similarities and facilitate integrated governance

Basic Sentiment Analysis

- Example from lexicon. Jockers is an example table with 10738 rows.

```
lexicon::hash_sentiment_jockers[sample(1:10738,10),]
```

```
##                   x     y
##  1:           hacks -0.60
##  2:    ruthlessness -1.00
##  3:         nutcase -1.00
##  4:       procedure  0.40
##  5:       jealously -1.00
##  6:         succes   0.60
##  7:        incensed -0.60
##  8:     overbalance -0.25
##  9:          heroes  0.80
## 10: altruistically  1.00
```

Basic Sentiment Analysis

- This simple approach is usually only a crude approximation
- It is limited for a multitude of reasons:
  - Negations ("This video is not bad, why would someone hate it?")
  - Adverbial modification ("I love it" vs. "I absolutely love it")
  - Context ("The horror movie was scary with many unsettling plot twists")
  - Domain specificity ("You should see their decadent dessert menu")
  - Slang ("Yeah, this is the shit!")
  - Sarcasm ("Superb reasoning, you must be really smart")
  - Abbreviations ("You sound like a real mofo")
  - Emoticons and emoji ("How nice of you... 😠")
  - ASCII Art ("(͡° ͜ʖ ͡°)")

Basic Sentiment Analysis

- These limitations can lead to inaccurate classifications, for example...
- Classified as negative (unpopular, hate)

```
comments$TextEmojiDeleted[7088]
```

```
## [1] "(unpopular opinion)\nWhy does everyone hate this movie"
```

- Classified as positive (genuinely, enjoy)

```
comments$TextEmojiDeleted[935]
```

```
## [1] "I don’t understand how people could genuinely enjoy this"
```

Sentiment Analysis of *YouTube* Comments

There are more sophisticated methods for sentiment analysis that yield better results. However, their complexity is beyond the scope of this workshop. We will do three things in this session and compare the respective results

1. Apply a basic sentiment analysis to our scraped *YouTube* comments
2. Use a slightly more elaborate out-of-the-box method for sentiment analysis
3. Extend the basic sentiment analysis to emoji

Sentiment Analysis of *YouTube* Comments

Word of Advice: Before using the more elaborate methods in your own research, make sure that you understand the underlying model, so you can make sense of your results. You should never blindly trust someone else's implementation without understanding it. Also: Always do sanity checks to see if you get any unexpected results.

Basic Comment Sentiments

First of all, we load the `syuzhet` package and try out the built-in basic sentiment tagger with an example sentence

```
# load data
comments <- readRDS("./content/data/ParsedEmojiComments.rds")

# load package
library(syuzhet)

# test simple tagger
get_sentiment("Superb reasoning, you must be really smart!")
```

```
## [1] 1.5
```

Basic Comment Sentiments

We can apply the basic sentiment tagger to the whole vector of comments in our data set. Remember that we need to use the text column without hyperlinks and emojis.

```
# create basic sentiment scores
BasicSentiment <- get_sentiment(comments$TextEmojiDeleted)

# summarize basic sentiment scores
summary(BasicSentiment)
```

```
##       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -34.45000  -0.50000   0.00000  -0.06446   0.25000  20.95000
```

Checking the documentation for the `get_sentiment()` function reveals that it can take different *methods* as arguments. These methods correspond to different dictionaries and might yield different results. The function also allows using a custom dictionary by providing a dataframe to the *lexicon* argument.

Basic Comment Sentiments

Let's compare the results of the different dictionaries (getting the sentiments for NRC can take a while)

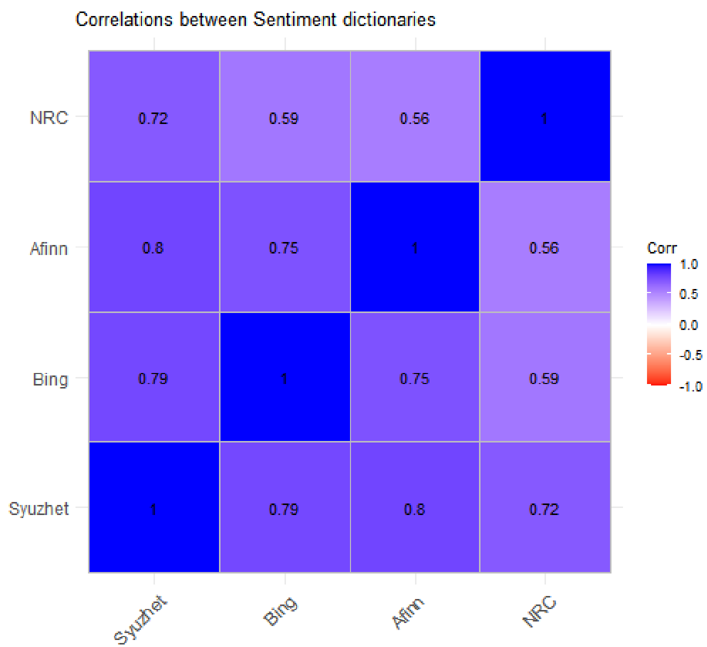
```
# compute sentiment scores with different dictionaries
BasicSentimentSyu <- get_sentiment(comments$TextEmojiDeleted,
                                   method = "syuzhet")
BasicSentimentBing <- get_sentiment(comments$TextEmojiDeleted,
                                    method = "bing")
BasicSentimentAfinn <- get_sentiment(comments$TextEmojiDeleted,
                                     method = "afinn")
BasicSentimentNRC <- get_sentiment(comments$TextEmojiDeleted,
                                   method = "nrc")
```

Basic Comment Sentiments

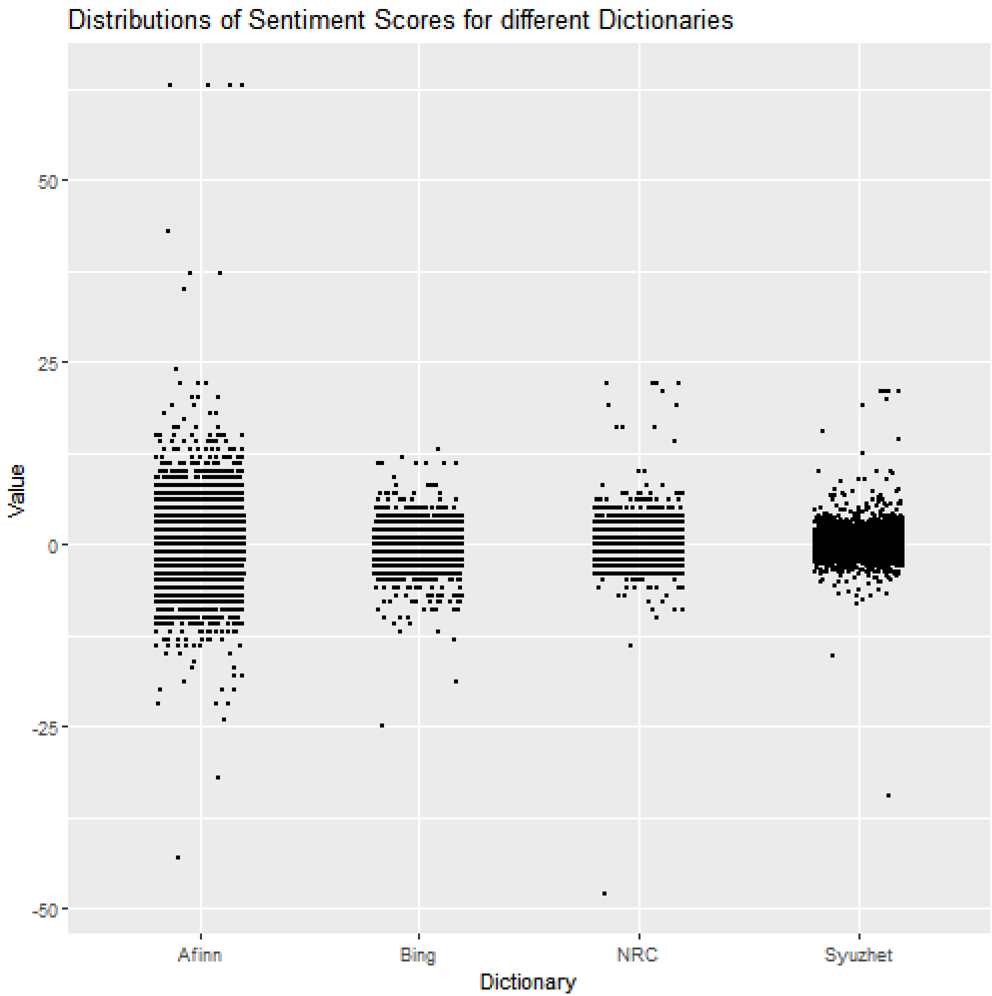
```
# combine them into a dataframe
Sentiments <- cbind.data.frame(BasicSentimentSyu,
                               BasicSentimentBing,
                               BasicSentimentAfinn,
                               BasicSentimentNRC,
                               1:dim(comments)[1])
# set column names
colnames(Sentiments) <- c("Syuzhet",
                          "Bing",
                          "Afinn",
                          "NRC",
                          "Comment")
```

Basic Comment Sentiments

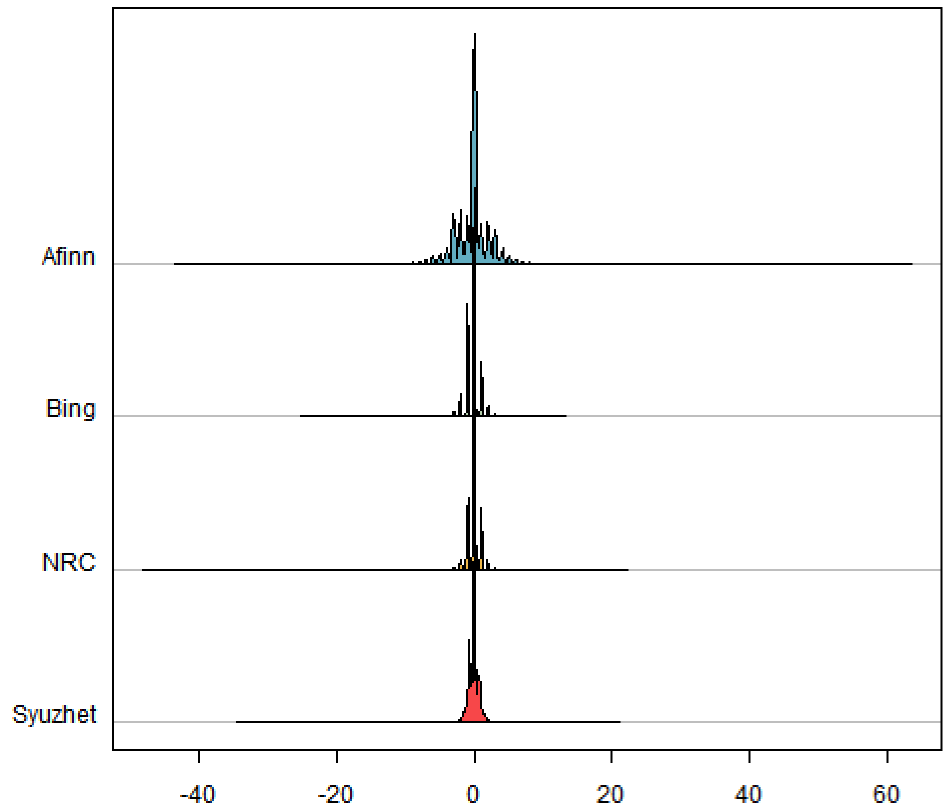
```r
# correlation matrix
library(ggcorrplot)
ggcorrplot(cor(Sentiments[,c(-5)]),
           color=c("red","white","blue"),
           title="Correlations between Sentiment dictionaries",
           lab=TRUE)
```



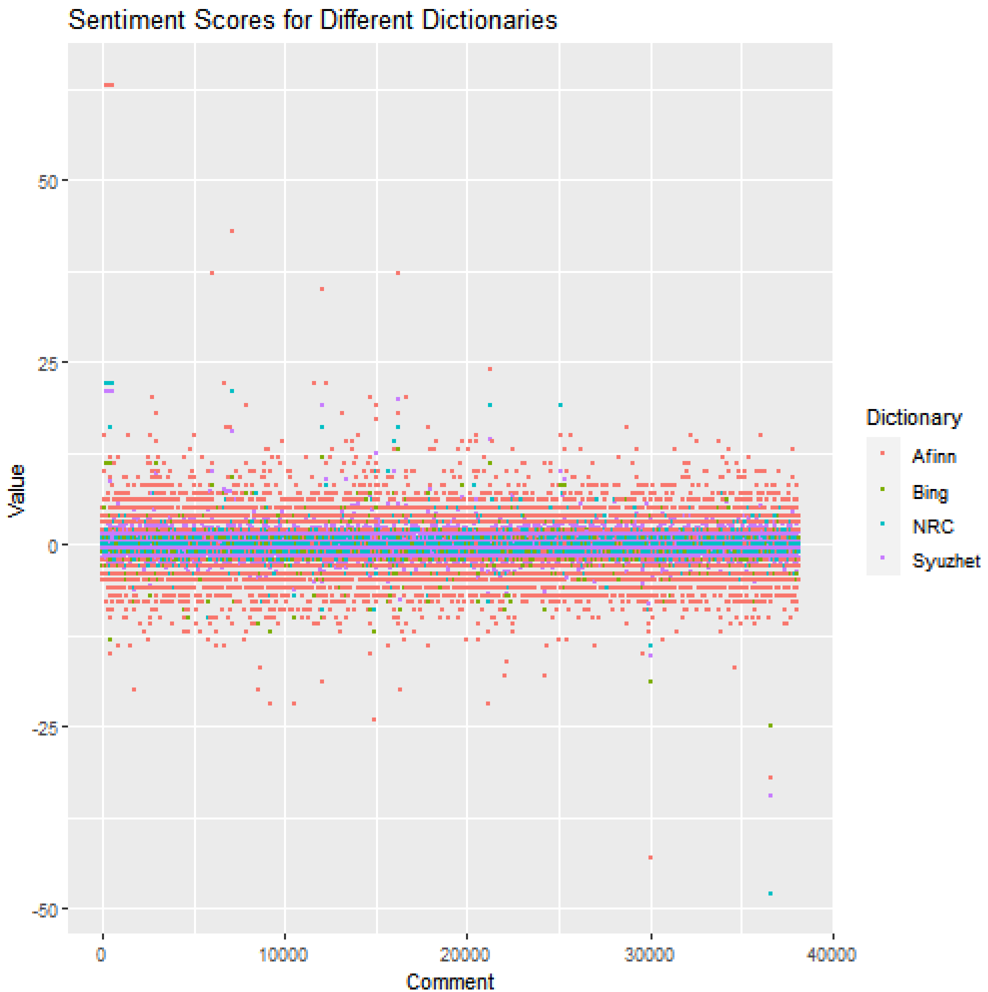
Basic Comment Sentiments: Jitter Plot



Basic Comment Sentiments: Ridgeline Plot



Basic Comment Sentiments: Scatter Plot



gesis Leibniz Institute for the Social Sciences

Basic Comment Sentiments

The choice of the dictionary can have an impact on your sentiment analysis. For this reason, it's crucial to select the dictionary with care and to be aware of how, by whom and for which purpose it was constructed. You can find more information on the specifics of the differnt dictionaries here.

It can also make sense to use more than one dictionary to check the robustness of the analysis

In this session, we will continue with the `syuzhet` dictionary

```
# add the syuzhet sentiment scores to our dataframe
comments$Sentiment <- Sentiments$Syuzhet
```

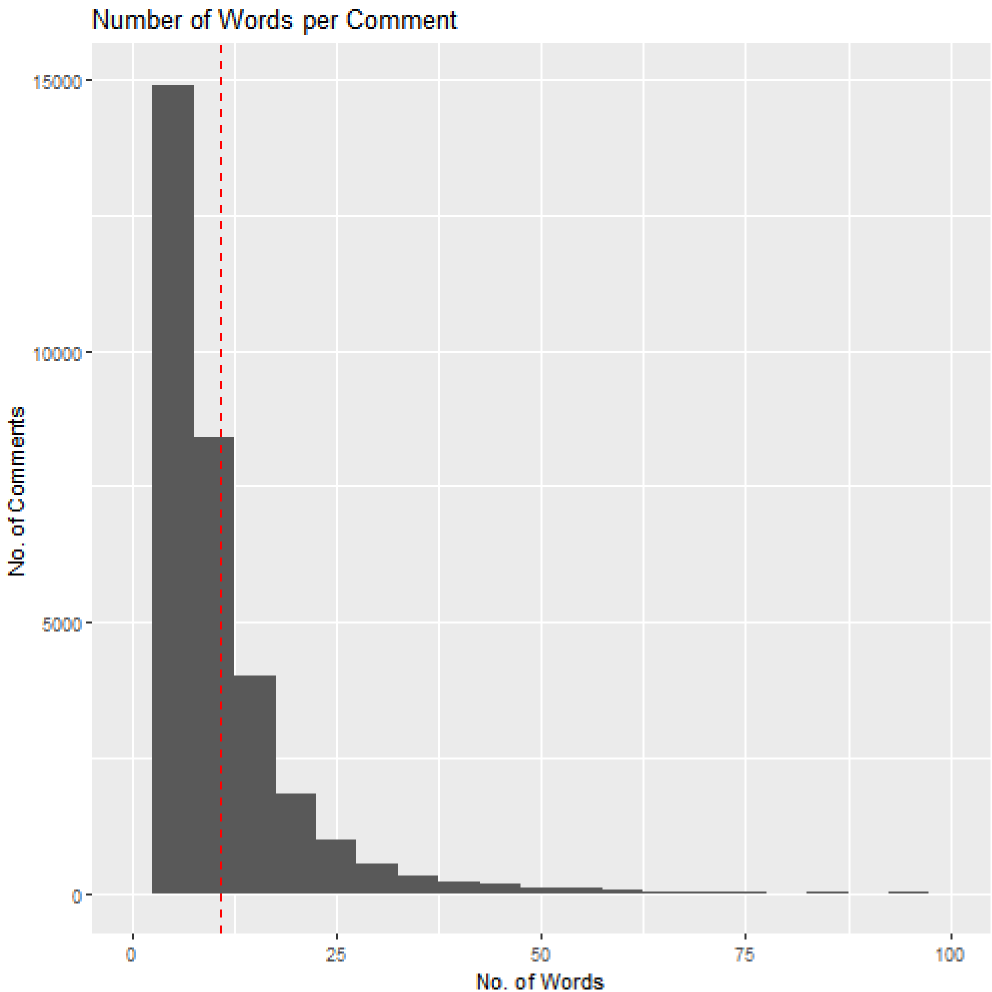
Basic Comment Sentiments

Another pitfall to be aware of is the length of the comments. Let's have a look at the distribution of words per comment.

```r
# compute number of words per comment
comments$Words <- sapply(comments$TextEmojiDeleted,
                         function(x){length(unlist(strsplit(x,
                                                            " ")))})
```

```r
# histogram
ggplot(comments, aes(x=Words)) +
  geom_histogram(binwidth = 5) +
  geom_vline(aes(xintercept=mean(Words)),
             color="red",
             linetype="dashed",
             size = 0.5) +
  ggtitle(label = "Number of Words per Comment") +
  xlab("No. of Words") +
  ylab("No. of Comments")+
  xlim(0, 100)
```

Basic Comment Sentiments



Basic Comment Sentiments

Because longer comments also contain more words, they have a higher likelihood of getting more extreme sentiment scores. Lets' look at one of the most negatively scored and one of the most positively scored comments.

```
# Very positively scored comment
comments$TextEmojiDeleted[7135]

# Very negatively scored comment
comments$TextEmojiDeleted[29916]
```

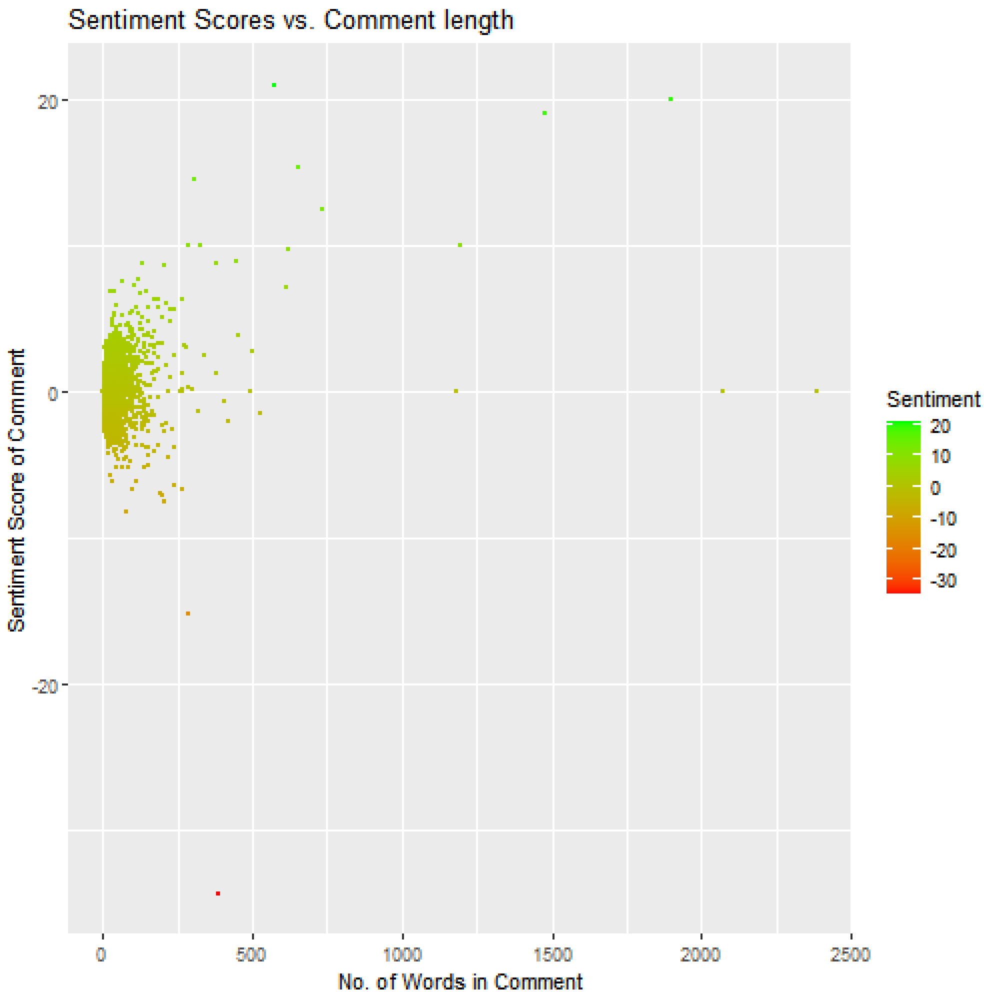
Basic Comment Sentiments

Positively scored comment: "The emoji movie is a beautifully executed, well layered experience that has been the victim of a barrage of such biased reviews such as “The worst movie of 2017” , “ a very formulaic story” and “9% on Rotten Tomatoes.” So, in defense of this movie, i will give you, the reader, why The Emoji Movie is in fact the best movie of in 2017. Remember, though, that this essay has some of my opinions, and takes on certain parts of the film. The Story. The story is a beautifully written, homage to many fairy tales. If you have not watched this movie yet, here's the rundown of the story. Gene is a meh emoji in a world of other emoji’s on a teenager’s smartphone. Gene, not wanting to be like the other emoji’s, goes out to find himself. Along the way, [...]"

Basic Comment Sentiments

Negatively scored comment: "How far we've fallen from the light of Yeshua? Only for our arrogance were we to be consumed by the Leviathon and trapped within the belly of the beast as it slowly bleeds out as our wicked and unwashed cries of agony fall on deaf ears, cursed to toil within and gaze back on our decaying with envious eyes as death hadn't taken us first. Heh, I dunno... Emojis are pretty much a line of communication for minorities drug deals/hookups. [...]"

Basic Comment Sentiments



Basic Comment Sentiments

To control for the effect of comment length, we can divide the sentiment score by the number of words in the comment to get a new indicator: *SentimentPerWord*

```r
# normalize for number of words
comments$SentimentPerWord <- comments$Sentiment / comments$Words

# most positive comment
head(comments$TextEmojiDeleted[comments$Sentiment ==
                                 max(comments$SentimentPerWord,
                                     na.rm = T)],1)
```

When i was like 8 or 7 i liked this movie

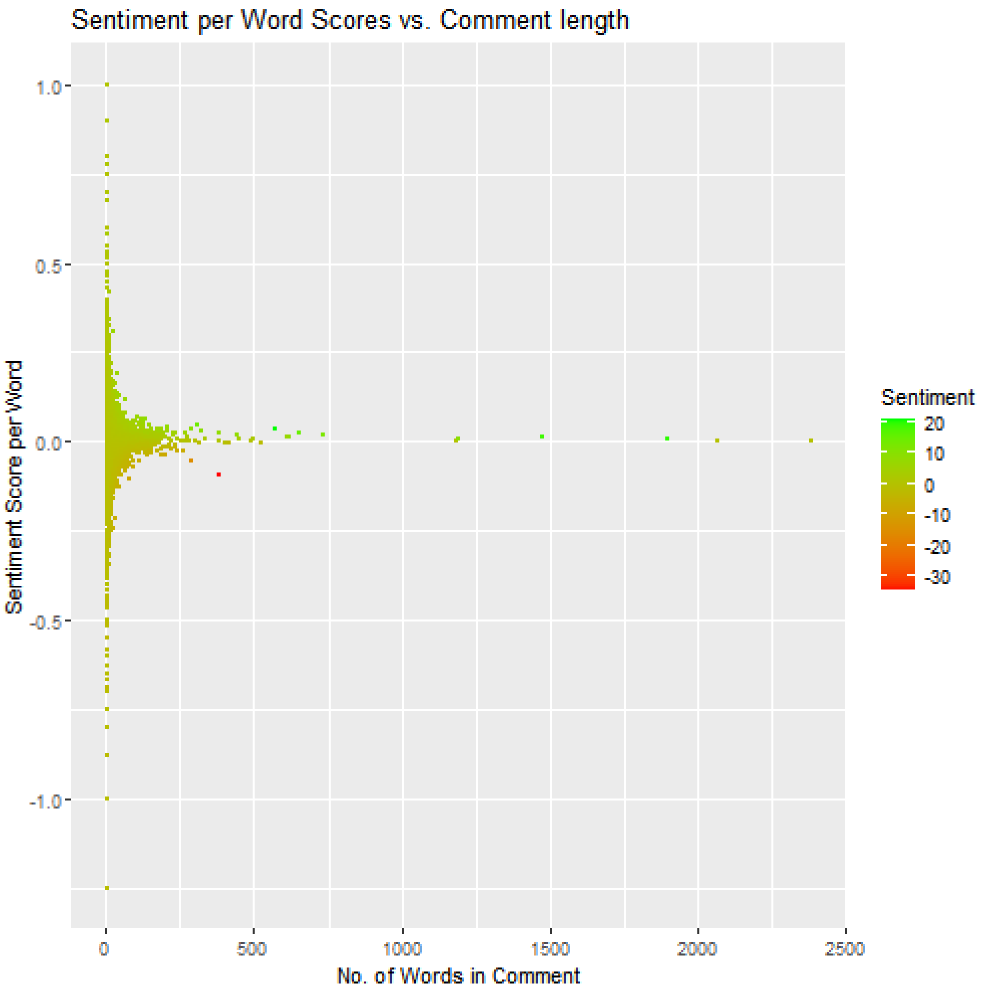
```r
# most negative comment
head(comments$TextEmojiDeleted[comments$Sentiment ==
                                 min(comments$SentimentPerWord,
                                      na.rm = T)],1)
```

2017 was fucking wild

Basic Comment Sentiments

```
# plot SentimentPerWord vs. Number of Words
ggplot(comments, aes(x = Words, y = SentimentPerWord, col = Sentiment)) +
    geom_point(size = 0.5) +
    ggtitle("Sentiment per Word Scores vs. Comment length") +
    xlab("No. of Words in Comment") +
    ylab("Sentiment Score per Word") +
    scale_color_gradient(low = "red", high = "green")
```

Basic Comment Sentiments



More Elaborate Method(s)

Although no sentiment detection method is perfect, some are more sophisticated than others. Two of those options are

- the `sentimentR` package
- the *Stanford coreNLP* utilities set

`SentimentR` attempts to take **valence shifters** into account:

- negators ("not")
- amplifiers / intensifiers ("very"),
- de-amplifiers / downtoners ("hardly", "somewhat"),
- adversative conjunctions ("but", "however")

More Elaborate Method(s)

Negators appear ~20% of the time a polarized word appears in a sentence. Conversely, adversative conjunctions appear with polarized words ~10% of the time. Hence, not accounting for **valence shifters** could significantly impact the modeling of the text sentiment.

More Elaborate Method(s)

Stanford coreNLP utilities set:

- build in Java
- very performant
- good documentation
- but tricky to get to work in `R`

We will be using `sentimentR` for this session as it represents a good trade-off between usability, speed, and performance

SentimentR

First, we need to install and load the package

```
if ("sentimentr" %in% installed.packages() == FALSE) {
  install.packages("sentimentr")
}

library(sentimentr)
```

Then we can compute sentiment scores

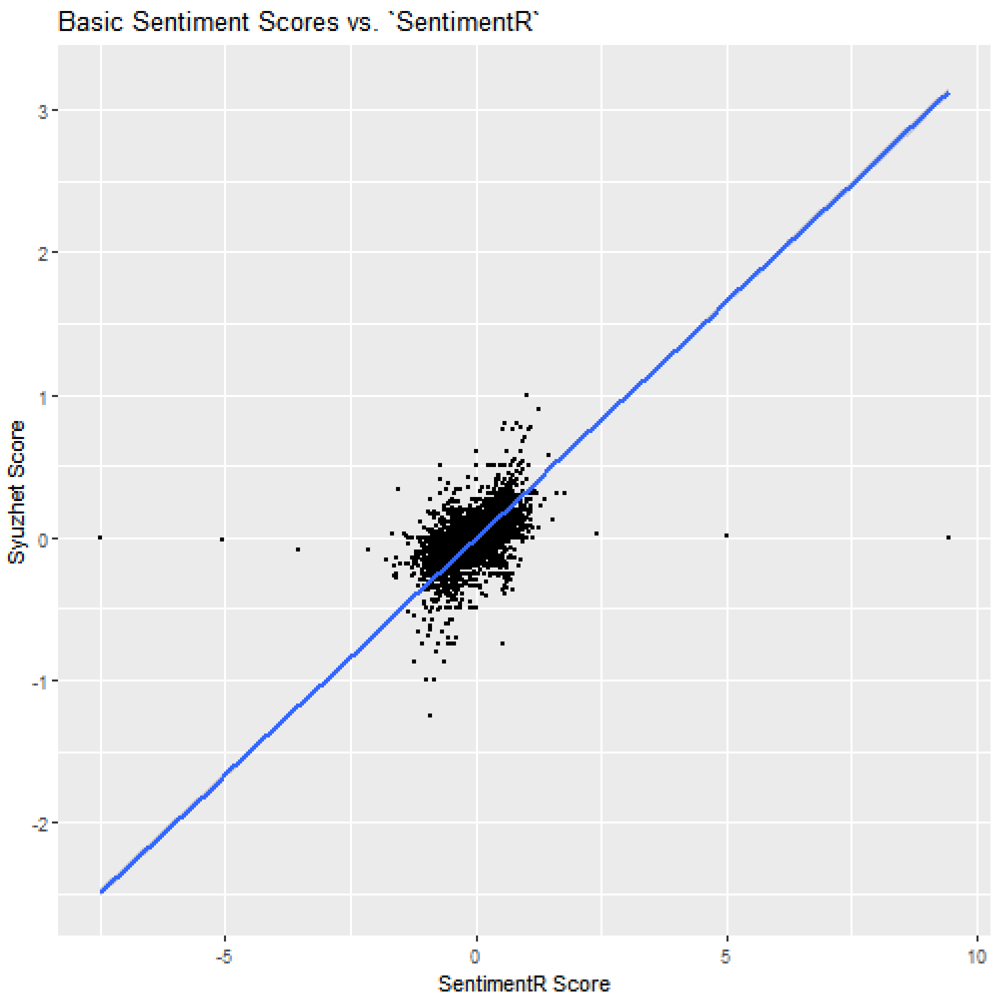
```
# compute sentiment scores
Sentences <- get_sentences(comments$TextEmojiDeleted)
SentDF <- sentiment_by(Sentences)
comments <- cbind.data.frame(comments,SentDF[,c(2,3,4)])
colnames(comments)[c(15,16,17)] <- c("word_count",
                                      "sd",
                                      "ave_sentiment")
```

SentimentR

Let's check if the sentiment scoring for sentimentR correlates with the simpler approach

```
# plot SentimentPerWord vs. SentimentR
ggplot(comments, aes(x=ave_sentiment, y=SentimentPerWord)) +
    geom_point(size =0.5) +
    ggtitle("Basic Sentiment Scores vs. `SentimentR`") +
    xlab("SentimentR Score") +
    ylab("Syuzhet Score") +
    geom_smooth(method=lm, se = TRUE)
```

SentimentR



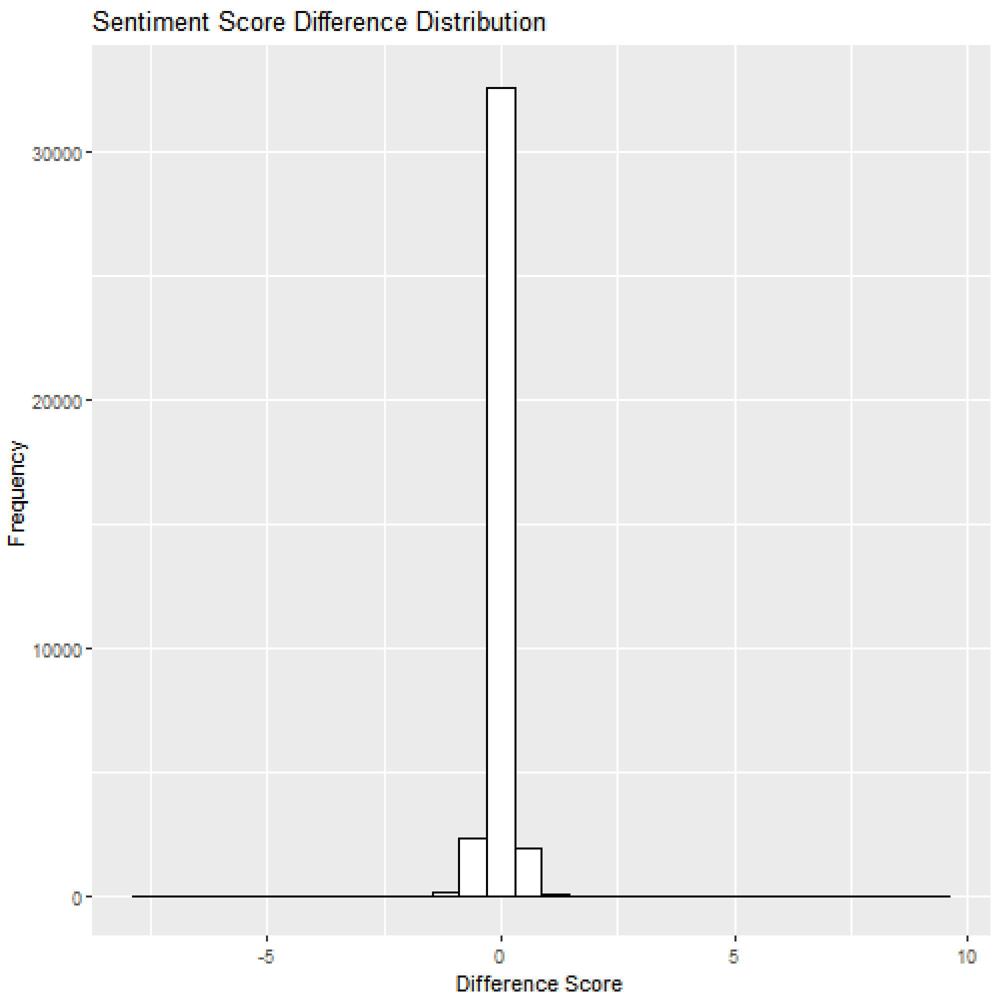
Basic Sentiment Scores vs. `SentimentR`
Syuzhet Score
SentimentR Score
3
2
1
0
-1
-2
-5
0
5
10

SentimentR

We can also look at the difference scores for the two methods

```
#compute difference score
comments$SentiDiff <- comments$ave_sentiment-
                      comments$SentimentPerWord

ggplot(comments, aes(x=SentiDiff)) +
  geom_histogram(color="black", fill="white")+
  labs(title="Sentiment Score Difference Distribution",
       x="Difference Score",
       y = "Frequency")
```

SentimentR


Sentiment Score Difference Distribution
30000
20000
10000
0
Frequency
-5
0
5
10
Difference Score

SentimentR

Let's check for which comments we get large differences between the two methods. *Note*: We use an absolute difference score here and order from largest to smallest difference.

```
# illustrate scoring differences
Diff_comments <- comments[order(-abs(comments$SentiDiff)),
                          ][c(2,12:18)]
Diff_comments[c(1,2,3,4), c(8, Syuzhet = 2, SentimentR = 7, 1)]
```

```
##        SentiDiff Sentiment ave_sentiment
## 5759    9.436887      0.00      9.436887
## 14599 -7.492500     -0.75     -7.500000
## 14613 -5.034112     -0.75     -5.048001
## 3130    5.007107      1.00      5.019011
##
## 5759  This screams money money money money money money money money money money money money
## 14599
## 14613
## 3130
```

SentimentR

Compared to the basic approach, `sentimentR` is:

- better at dealing with negations
- better at detecting fixed expressions
- better at detecting adverbs
- better at detecting slang and abbreviations
- relatively easy to implement
- quite fast

gesis Leibniz Institute for the Social Sciences

Sentiments for Emojis

Emojis are often used to confer emotions (hence the name), so they might be a valuable addition to assess the sentiment of a comment. This is less straight-forward than assessing sentiments based on word dictionaries for multiple reasons:

- Emojis can have multiple meanings: 🙏
- They are highly context-dependent: 🍆
- They are culture-dependent: 🍑
- They are person-dependent: 🤣 😂

Sentiments for Emojis

In addition, emojis are rendered differently on different platforms, meaning that they can potentially elicit different emotions

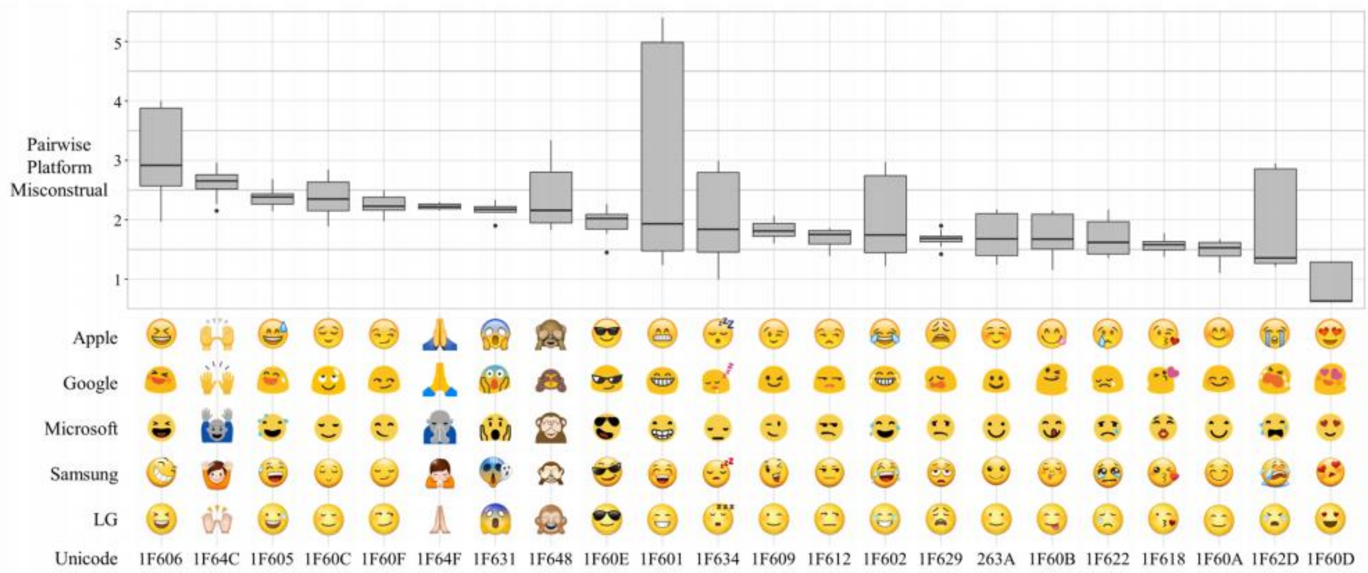


Figure 1. Across-platform sentiment misconstrual scores grouped by Unicode. Each boxplot shows the range of sentiment misconstrual scores across the five platforms. They are ordered by decreasing median platform-pair sentiment misconstrual, from left to right.

Source: Miller et al., 2016

Sentiments for Emojis

Emojis are also notoriously difficult to deal with from a technical perspective due to the infamous character encoding hell

- Emojis can come in one of multiple completely different encodings
- Your operating system has a default encoding that is used when opening/writing files in a text editor
- Your `R` installation has a default encoding that gets used when opening/writing files

Sentiments for Emojis

If either of those mismatch at any point, you can accidentally overwrite the original encoding in a non-recoverable way. For us, this happened quite often with UTF-8 encoded files on Windows (the default encoding there is Latin-1252).

Hex Dec Oct

Native	Symbola [1]	Code	UTF-8	UTF-16 LE	Surrogates	Name
😁	😁	1F601	F0 9F 98 81	3D D8 01 DE	D83D DE01	GRINNING FACE WITH SMILING EYES

Leibniz Institute for the Social Sciences

Sentiments for Emojis

Luckily, we already saved our emojis in a textual description format and can simply treat them as a character strings for our sentiment analysis. We can therefore proceed in 3 steps:

1. Create a suitable sentiment dictionary for textual descriptions of emojis
2. Compute sentiment scores for comments only based on emojis
3. Compare the emojis sentiment scores with the text-based sentiments

Emoji Sentiment Dictionary

We will use the emoji sentiment dictionary from the `lexicon` package. It only contains the 734 most frequent emojis, but since the distribution of emojis follows Zipf's Law, it should cover most of the used emojis.

```
# emoji sentiments
EmojiSentiments <- lexicon::emojis_sentiment
EmojiSentiments[1:5,c(1,2,4)]
```

```
##                 byte                            name sentiment
## 1 <f0><9f><98><80>                  grinning face 0.5717540
## 2 <f0><9f><98><81>  beaming face with smiling eyes 0.4499772
## 3 <f0><9f><98><82>            face with tears of joy 0.2209684
## 4 <f0><9f><98><83>       grinning face with big eyes 0.5580431
## 5 <f0><9f><98><84> grinning face with smiling eyes 0.4220315
```

Leibniz Institute for the Social Sciences

Emoji Sentiment Dictionary

By comparison, our data looks like this:

```
# example from our data
comments$TextEmojiReplaced[5999]
```

```
## [1] "Amazing movieEMOJI_GrinningFace "
```

Emoji Sentiment Dictionary

We bring the textual description in the dictionary in line with the formatting in our data so we can replace one with the other using standard text manipulation techniques

```
# change formatting in the emoji dictionary
EmojiNames <- paste0("emoji_",gsub(" ","",EmojiSentiments$name))
EmojiSentiment <- cbind.data.frame(EmojiNames,
                                   EmojiSentiments$sentiment,
                                   EmojiSentiments$polarity)
names(EmojiSentiment) <- c("word","sentiment","valence")
EmojiSentiment[1:5,]
```

```
##                                 word sentiment  valence
## 1                  emoji_grinningface 0.5717540 positive
## 2   emoji_beamingfacewithsmilingeyes 0.4499772 positive
## 3           emoji_facewithtearsofjoy 0.2209684 positive
## 4      emoji_grinningfacewithbigeyes 0.5580431 positive
## 5 emoji_grinningfacewithsmilingeyes 0.4220315 positive
```

Emoji Sentiment Dictionary

```r
# tokenize the emoji-only column in our formatted dataframe
library(quanteda)
EmojiToks <- tokens(tolower(as.character(unlist(comments$Emoji))))
comments$Text[11557]
```

```
## [1] "😀+🎥=💩"
```

```r
EmojiToks[11557]
```

```
## Tokens consisting of 1 document.
## text11557 :
## [1] "emoji_grinningface" "emoji_moviecamera"  "emoji_pileofpoo"
```

Computing Sentiment Scores

We can now replace the emojis that appear in the dictionary with the corresponding sentiment scores

```r
# create the dictionary object
EmojiSentDict <- as.dictionary(EmojiSentiment[,1:2])

# replace emoji with sentiment scores
EmojiToksSent <- tokens_lookup(x = EmojiToks,
                               dictionary = EmojiSentDict)
comments$Text[11557]
```

```
## [1] "😀+🎥=💩"
```

```r
EmojiToksSent[11557]
```

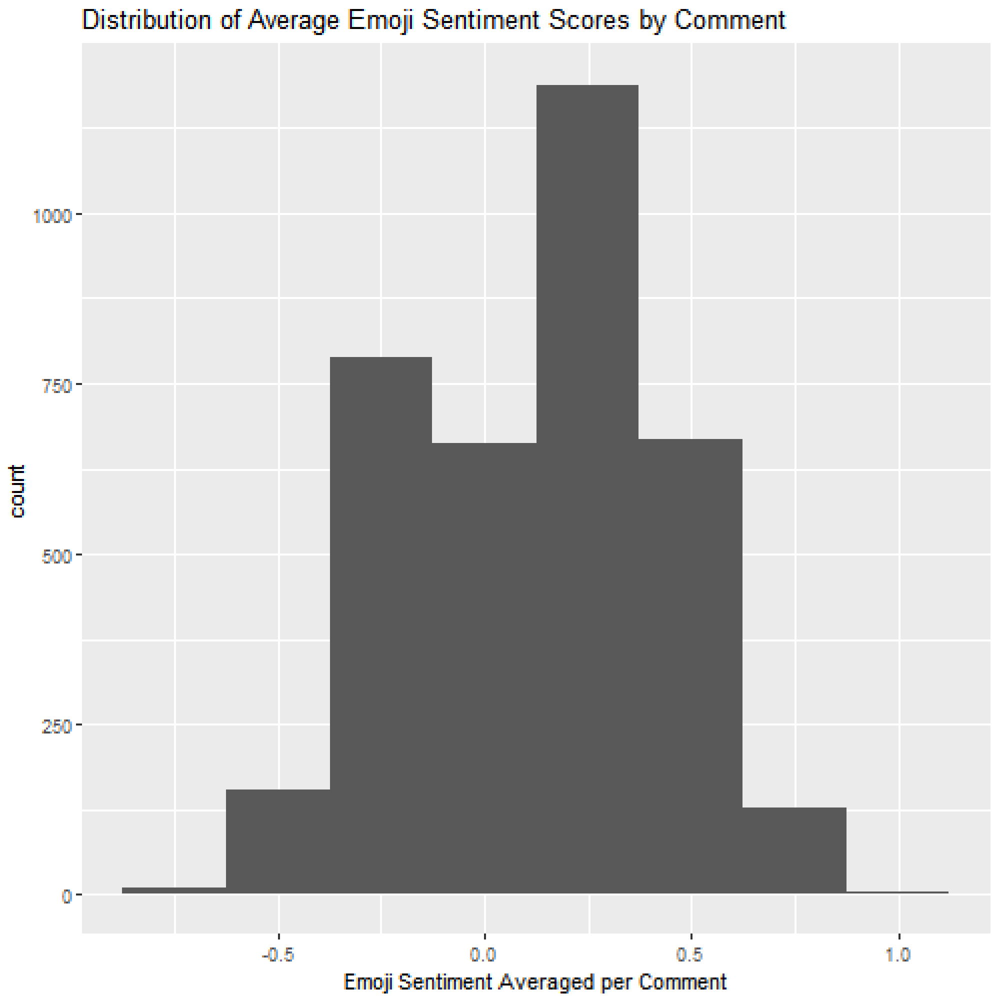
```
## Tokens consisting of 1 document.
## text11557 :
## [1] "0.571753986332574"  "0.303030303030303"  "-0.117903930131004"
```

Computing Sentiment Scores

```
# only keep the assigned sentiment scores for the emoji vector
AllEmojiSentiments <- tokens_select(EmojiToksSent,EmojiSentiment$sentiment,
                                    "keep")
AllEmojiSentiments <- as.list(AllEmojiSentiments)

# define function to average emoji sentiment scores  per comment
MeanEmojiSentiments <- function(x){

  x <- mean(as.numeric(as.character(x)))
  return(x)

}

# apply the function to every comment that contains emojis
MeanEmojiSentiment <- lapply(AllEmojiSentiments,MeanEmojiSentiments)
MeanEmojiSentiment[MeanEmojiSentiment == 0] <- NA
MeanEmojiSentiment <- unlist(MeanEmojiSentiment)
MeanEmojiSentiment[11557]
```

```
## text11557
## 0.2522935
```

Emoji Sentiment Scores



gesis Leibniz Institute for the Social Sciences

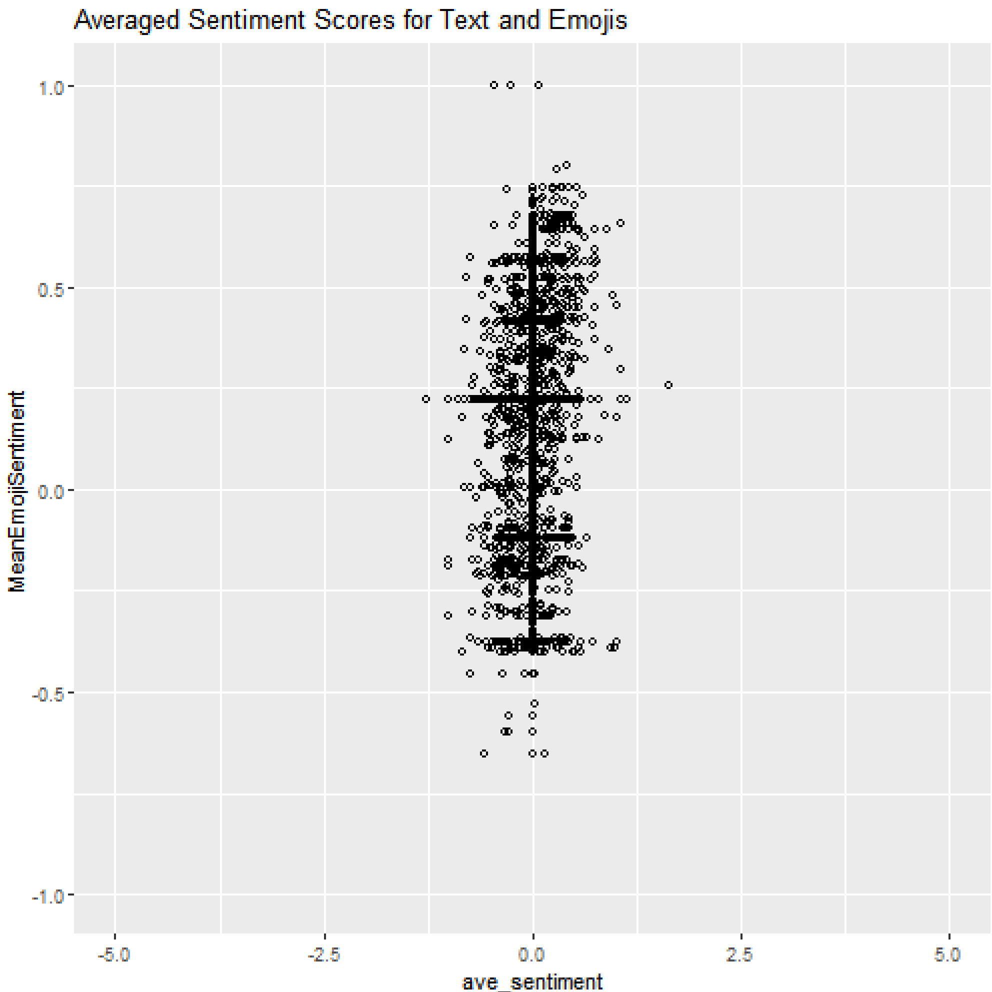
Emoji Sentiment vs. Word Sentiment

```
comments <- cbind.data.frame(comments,MeanEmojiSentiment)

# correlation between average emoji sentiment score
#   and average text sentiment score
cor(comments$ave_sentiment,
    comments$MeanEmojiSentiment,
    use="complete.obs")
```

```
## [1] 0.1347397
```

```
# plot the relationship
ggplot(comments, aes(x = ave_sentiment,
                     y = MeanEmojiSentiment))+
  geom_point(shape = 1) +
  labs(title = "Averaged Sentiment Scores for Text and Emojis") +
  scale_x_continuous(limits = c(-5,5)) +
  scale_y_continuous(limits = c(-1,1))
```

Emoji Sentiment vs. Word Sentiment



Emoji Sentiment vs. Word Sentiment

As we can see, there seems to be no meaningful relationship between the sentiment scores of the text and the sentiment of the used emojis. This can have multiple reasons:

- Comments that score very high (positive) on emoji sentiment typically contain very little text
- Comments that score very low (negative) on emoji sentiment typically contain very little text
- Dictionary-based bag-of-words/-emojis sentiment analysis is not perfect - there is a lot of room for error in both metrics
- Most comment texts are scored as neutral
- Emojis are very much context-dependent, but we only consider a single sentiment score for each emoji
- We only have sentiment scores for the most common emoji
- If comments contain an emoji, it's usually exactly one emoji

Takeaway

From the examples in this data, we have seen multiple things:

- Sentiment detection is not perfect
- Bag-of-words approaches are often too simplistic
- Even more sophisticated methods can often misclassify comments
- The choice of dictionary and sentiment detection method is highly important and can change the results substantially
- Emojis seem to be even more challenging for classifying sentiments

Exercise time 🏋️💪🏃🚴

Solutions