Pfizer Vaccine on Reddit

Jinina Rei Garcia

Master's degree on Business Analytics, Hult International Business School

DAT-5317: Text Analytics and Natural Language Processing (NLP)

Prof. Thomas Kurnicki

December 5, 2021

Abstract

This study explored the Reddit posts and comments of Pfizer vaccinees and the side effects of Pfizer vaccine. Data is gathered from subreddit PfizerVaccine. Both posts and comments contain the following fields: title (relevant for posts), score (relevant for posts - based on impact, number of comments), id (unique id for posts/comments), URL (relevant for posts, URL of post thread), commns_num (relevant for post, number of comments to this post), created (date of creation), body (relevant for posts/comments, text of the post or comment), and timestamp. The data only contains the thread in 2021 and posts are not filtered. This study aims to determine the most common side effects that are experienced by vaccinees and the sentiments they have about the Pfizer vaccine. Upon doing text mining and analytics, results suggested that the most common side effects are mild sore throat, mild headache, exhaustion, body aches and fever. Results are discussed in terms of the four IDs with the highest scores and most number of comments and in terms of scores only. Findings also show the medicine that most vaccinees take after having side effects. This study also looked at the positive posts and comments of vaccinees. Overall, most vaccinees feel more safe, healthy, and positive in their outlook in life even after experiencing side effects.

Keywords: side effects, Pfizer vaccine, Tylenol, mild symptoms

Posts and comments on Pfizer Vaccine

This study focuses on finding the side effects caused by the vaccine through analyzing the posts and comments on Reddit. The task is to rank the side effects based on the greatest number of comments and scores. In addition, the purpose of the study is to determine the topics that most Reddit participants discuss regarding Pfizer vaccine. The research is also interested with the positive and negative sentiments about the Pfizer vaccine.

Materials and Method

This section tackles the methods of analysis using R. The analysis involves Text analytics and Natural Language Processing (NLP) solely. The frameworks used are Latent Dirichlet Algorithm, sentiment analysis, pair-wise correlations, N-grams and TF-IDF.

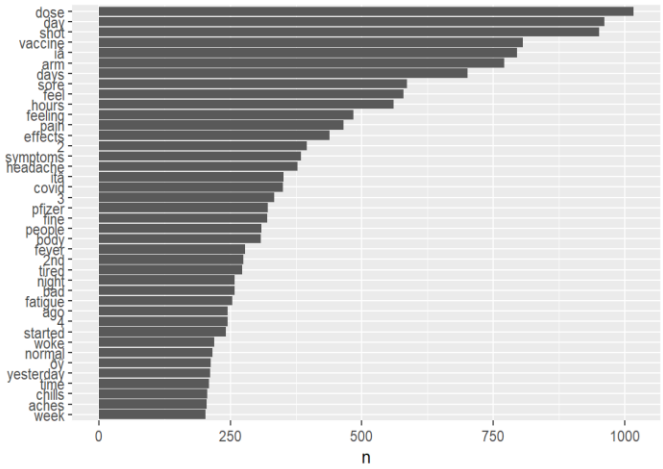
Results

Most frequent words in all the posts and comments

Upon tokenizing and removing the stop words of the body of posts and comments, the most frequent words found are dose, day, vaccine, shot, arm, days, sore, hours, feeling, pain etc. Figure 1 shows that most of the participants took the vaccine shot and experienced the side effects after hours or days of taking it.

Figure 1

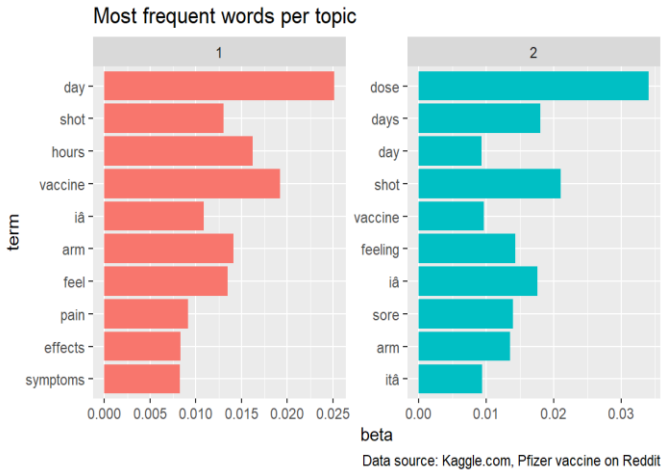
Token frequency histogram



Topics discussed in PfizerVaccine Subreddit

The study uses an LDA_VEM topic models to determine the two topics that are discussed in the Reddit thread. According to Figure 2, both topics have almost the same common words but considering only the top five words, the words that are different are hours and day. The data infers that the topics discussed revolves around the side effects of the vaccine and when they are experienced, in hours or days.

Figure 2



The Most Participant ID and their Score

Based on Figure 3, the IDs with the greatest number of comments and highest scores are lam09i, l8e8kj, l8skoz, and l7nc0f. These are the persons of interest in one section of this study. Focusing on lam09i whose post has the highest number of scores, he or she has experienced a little soreness in the shoulder the day after and exhaustion a week after. In relation to this, the researcher concludes that soreness in the shoulder and exhaustion are included in the most common side effects of Pfizer vaccine.

Figure 3

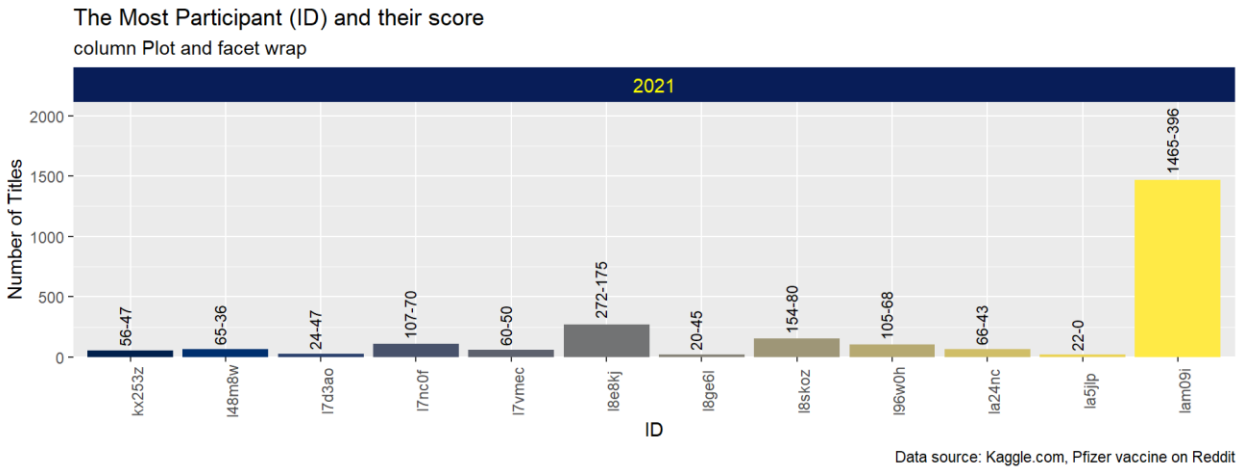


Table 1

Title	Text	ID	Comms_num	Score
Post Pfizer Vaccine Experience	Hey guys, I received the Pfizer Vaccine (only the first dose so far) and am looking for others to share experiences, mainly experience a week after the shot as well as related to physical activity and working out. I did feel a little soreness in the shoulder the day after but felt overall alright. However it's been over a week now and I feel generally more tired at all times, especially at the gym. I am a hobby bodybuilder and am In the sport for 5 years now. Ever since	lam09i	1465	396

	I received the vaccine, I feel like my workouts have been short, way more exhausting and weaker. Has anyone else experienced this?

Positive and negative posts and comments on Pfizer vaccine

The Afinn and Bing lexicons are used to do a sentiment analysis with the purpose of showing the positive and negative comments on Pfizer vaccine. Figure 4 shows that the most negative posts are harsh words which are not filtered but the most frequent negative words with scores of -2 and -1 relates to the negative side effects of the vaccine such as pain, tired, sore, dizzy, and numb. On the other hand, the most frequent positive words with scores 1 and 2 are feeling, immune, safe, fine, approved, hope. The data reveals that there are vaccinees that felt fine after the vaccine. It is also shown that it is important to the vaccinees that Pfizer vaccine is safe, and FDA approved. Pfizer vaccine makes them immune and brings hope to them. The same goes with the Bing lexicon which divides the sentiments into positive and negative. The most negative words are sore, pain, fever, headache and tired. It can be concluded that the symptoms dizziness, headache, and fever can be added to the side effects.

Figure 4

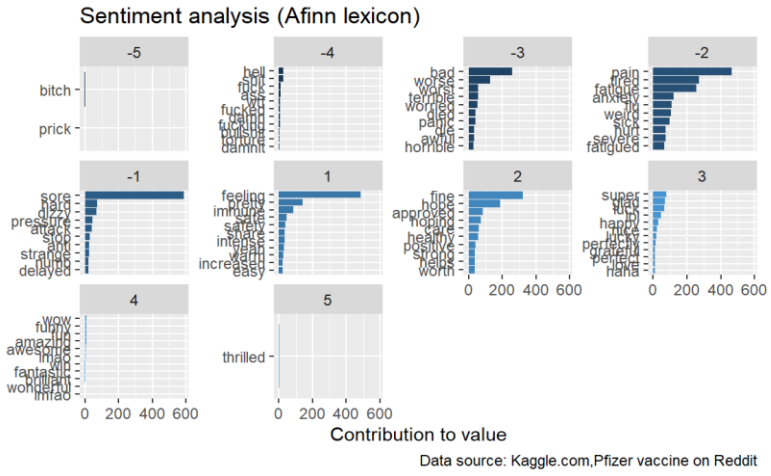
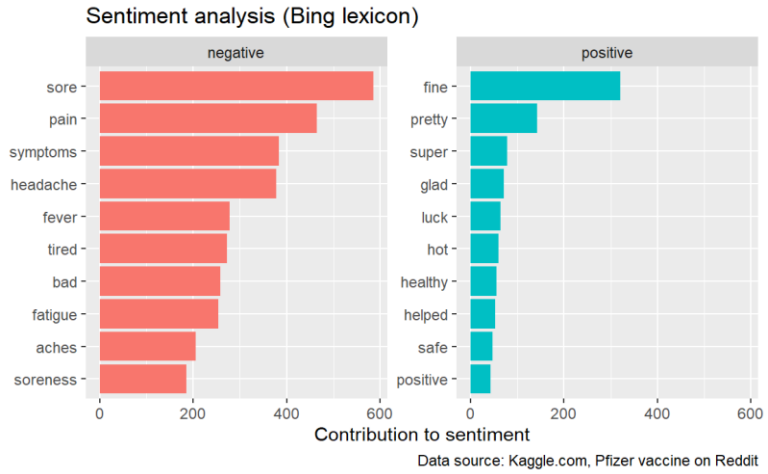


Figure 5



Other side effects of Pfizer Vaccine and their duration

The analysis is done with two groups: low score (score<3) and high score (score>3). Based on Figure 6, most of the correlated words of posts and comments with high score are brain fog, body aches, mild/slight and bad headache, heart rate, fell asleep, upset stomach, and sore throat. In terms of the duration, it can be inferred from the nodes formed that the side effects can start from 8-26 hours after getting the shot. It can last up to 2-3 days and worse 10 days. Meanwhile, 91% of the posts and comments have low scores so this is analyzed using trigrams. Some of the most correlated words are dizzy and weird bowel, elevated/increasing heart rate and fever/chills and body/muscle which confirms the other side effects found in Figure 6. In terms of duration, the biggest node includes the duration like 24 hours and 2-10 days. It can be concluded that the other side effects of the vaccine are brain fog, body aches, mild/slight and bad headache, increased heart rate, sleepiness, upset stomach, sore throat, and swollen lymph node in arm. The duration of these side effects can start within 8-24 hours and lasts to 2-10 days.

Figure 6

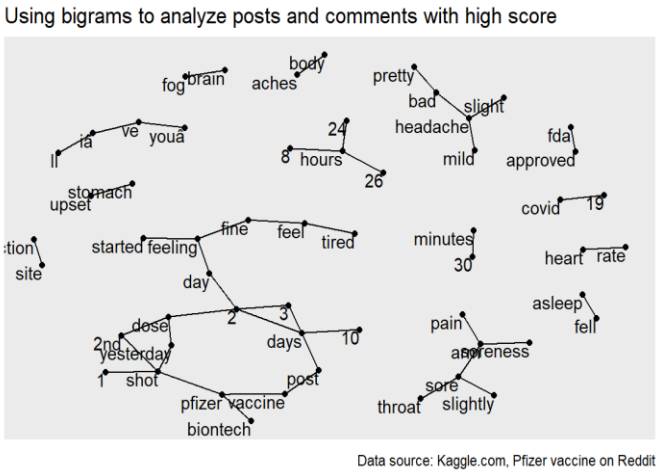
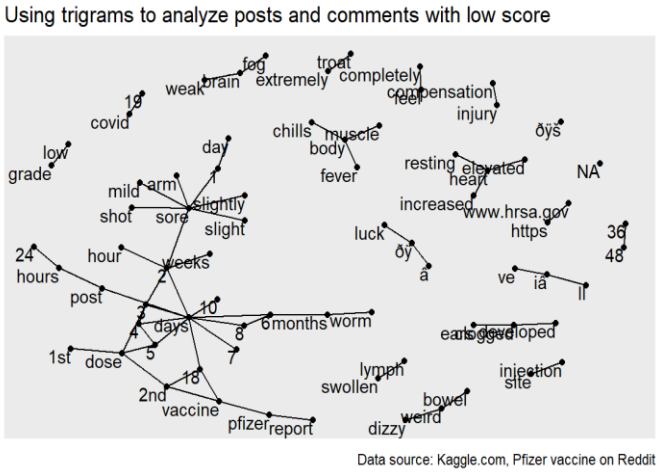


Figure 7

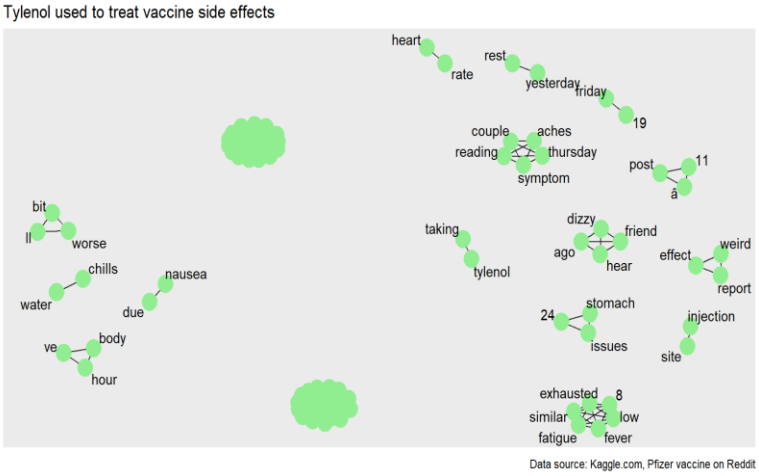


Tylenol used to treat side effects of vaccine

Figure 8 shows the pair-wise correlations between words. One of the most common word paired is Tylenol, an acetaminophen. It can be inferred that this is the medicine that most vaccinees take when experiencing the side effects. Miller, E. (2021) cites the side effects of Acetaminophen which are headaches, agitation, insomnia, constipation and vomiting. *Drug watch*. https://www.drugwatch.com/tylenol/. Based on Figure 8, most of the words paired with

Tylenol are dizzy, nausea, heart rate and stomach issues. It can be derived that the mentioned side effects can also be caused by the medicine and not the vaccine.

Figure 8



Most participant (ID) with highest scores as persons of interest

The four ID's with the highest scores are analyzed using TF-IDF analysis to check the less frequent words that can add more insights to the study. Based on Figure 11, the persons can be classified according to the side effects they have experienced. This can be used to determine the most common side effects experienced because these comments had the highest scores or votes. ID l7nc0f is a female who experienced increased heart rate. However, looking at other words like scared, the fright might have triggered her increased heart rate. ID l8sk0z experienced increased temperature. ID l8e8kj experienced headaches and achy joints. Lastly, ID lam09i, with the highest score, experienced exhaustion. It can be concluded that most vaccinees experienced exhaustion, headache, and fever.

Figure 9

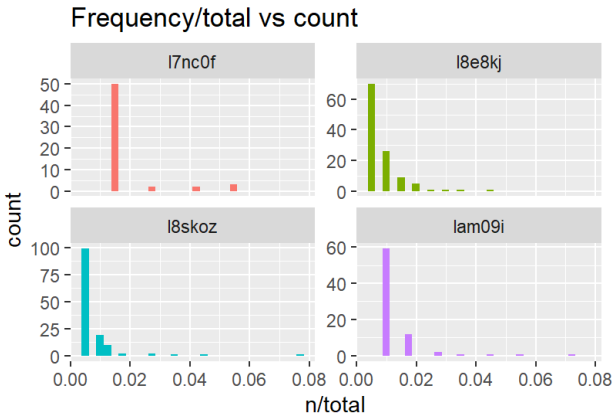


Figure 10

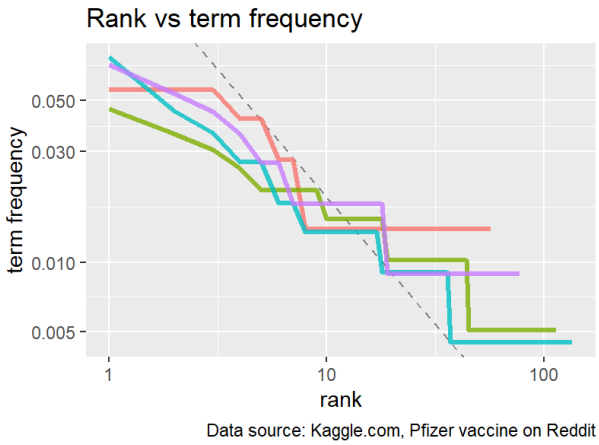
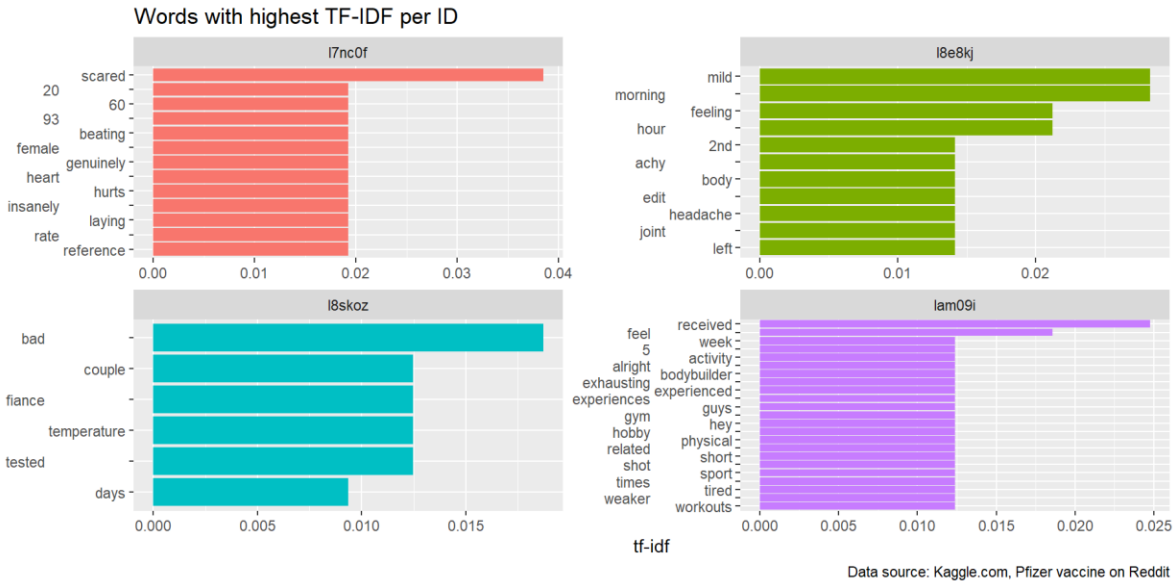


Figure 11



References

Preda, G. (2021). *Pfizer Vaccine on Reddit*. https://www.kaggle.com/gpreda/pfizer-vaccine-on-reddit

Miller, E. (2021). *Tylenol*. Drugwatch. https://www.drugwatch.com/tylenol/

Appendix

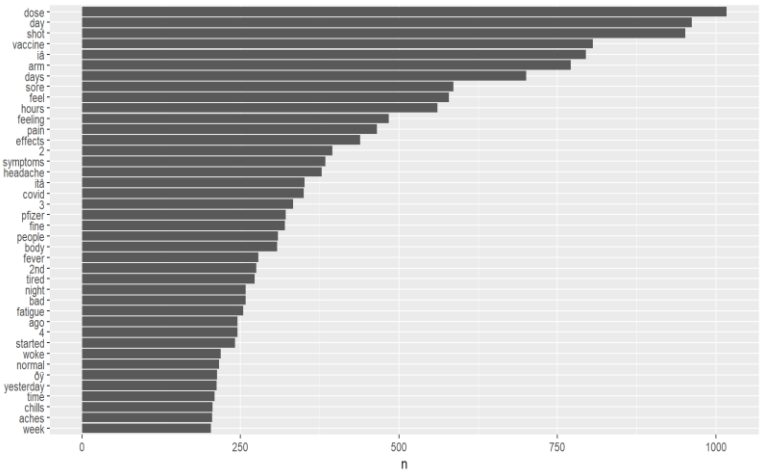
```
################################
### Individual Business Insight Report
### MSBA HULT 2021-2022
### Created by: Jinina Rei Garcia
### Date: 12.04.2021
### Version 1.0
################################

####calling all libraries

library (mongolite)
library(dplyr)
library(tidytext)
library(ggplot2)
library(scales)
library(janeaustenr)
library(stringr)
library(tidyr)
library(widyr)
library(textdata)
library(janitor)
library(lubridate)
library(viridis)
library(formattable)
library(igraph)
library(ggraph)
library(ggplot2)
library(topicmodels)

####importing the pfizer dataset
reddit_pfizer_vac <- read.csv("C:/Users/Jinina Rei
Garcia/OneDrive/Documents/HULT/STUDIES/Text Analytics/A3 business insight
report/reddit_pfizer_vaccine.csv")
colnames(reddit_pfizer_vac)[7] <- "text"

####cleaning the data (dates)
pfizer_vac <- reddit_pfizer_vac %>%
  mutate(across(where(is.character), tolower)) %>%
  mutate(date = as_date(timestamp)) %>%
  clean_names()

glimpse(pfizer_vac)

pfizer_vac <- pfizer_vac %>% select(text, comms_num, date, score, id)
```

```
pfizer_vac_tidy <- pfizer_vac %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)

print(pfizer_vac_tidy)
```

Console:

```
             word    n
1            dose 1017
2             day  962
3            shot  952
4         vaccine  806
5             iâ  795
6             arm  771
7            days  701
8            sore  586
9            feel  579
10          hours  561
11        feeling  484
12           pain  465
13        effects  439
14              2  395
```

```
####plotting the token frequencies
freq_hist <- pfizer_vac_tidy %>%
  filter(n>200) %>%
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  xlab(NULL)+
  coord_flip()

print(freq_hist)
```




```
##################################################
###### Latent Dirichlet algorithm ##############
##################################################
pfizer_vac_text <- reddit_pfizer_vac %>% select(title, text)
```

```
pfizer_dtm <- pfizer_vac_text %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(title, word) %>%
  cast_dtm(title, word, n)
```

Console:

```
<<DocumentTermMatrix (documents: 14, terms: 6654)>>
Non-/sparse entries: 7345/85811
Sparsity           : 92%
Maximal term length: 31
Weighting          : term frequency (tf)
```

```
ap_lda <- LDA(pfizer_dtm, k=2, control=list(seed=123))
ap_lda
```

Console:

```
A LDA_VEM topic model with 2 topics.
```

```
#now we are looking for the per topic per word probabilities aka. beta
#beta - what is the probability that "this term" will be generated by "this topic"
ap_topics <- tidy(ap_lda, matrix="beta")
ap_topics
```

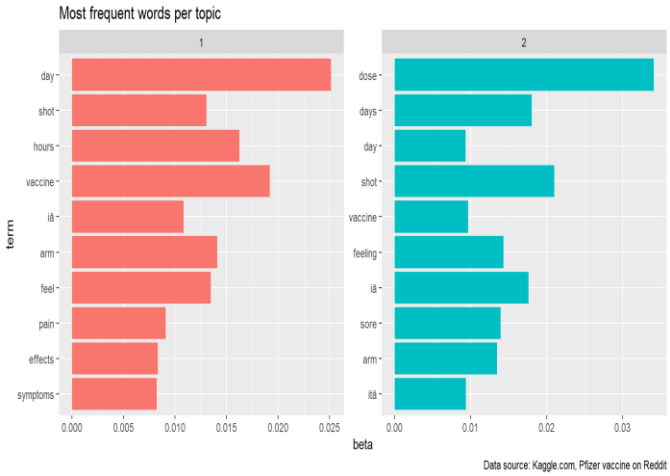
```
top_terms <- ap_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>% #selecting top 10 for beta
  ungroup() %>%
  arrange(topic, -beta)
top_terms
```

Console:

```
# A tibble: 20 x 3
   topic term        beta
   <int> <chr>      <dbl>
 1     1 day      0.0251
 2     1 vaccine  0.0192
 3     1 hours    0.0162
 4     1 arm      0.0141
 5     1 feel     0.0135
 6     1 shot     0.0131
 7     1 iâ       0.0108
 8     1 pain     0.00912
 9     1 effects  0.00833
10     1 symptoms 0.00824
```

```
#lets plot the term frequencies by topic
top_terms %>%
  mutate(term=reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend=FALSE) +
```

```
  facet_wrap(~topic, scales = "free") +
  labs(
    title = "Most frequent words per topic",
    caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit",
  )+
  coord_flip()
```




```
#lets calculate the relative difference between the betas for words in topic 1
#and words in topic 2
beta_spread <- ap_topics %>%
  mutate(topic=paste0("topic", topic)) %>%
  spread(topic, beta) %>% #1st column for topic 1 and 2nd for topic 2
  filter(topic1>0.01 | topic2 >0.01) %>%
  mutate(log_rate = log2(topic2/topic1))

beta_spread
```

Console:

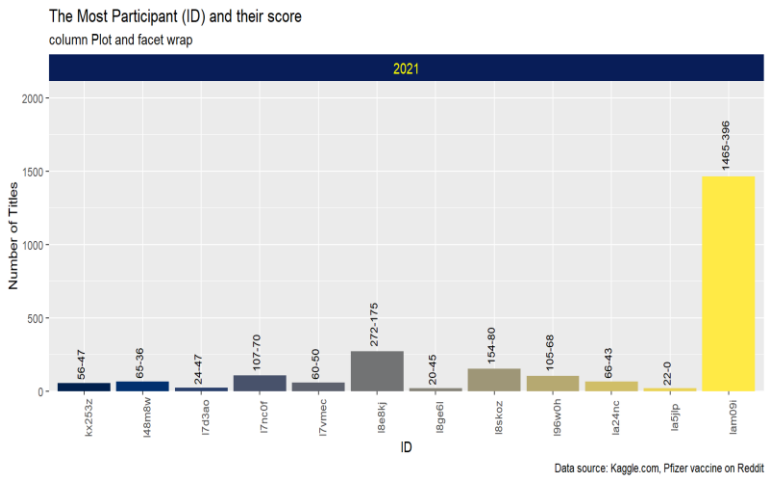
```
# A tibble: 11 x 4
   term     topic1  topic2 log_rate
   <chr>     <dbl>   <dbl>    <dbl>
 1 arm     0.0141  0.0135   -0.0638
 2 day     0.0251  0.00933  -1.43
 3 days    0.00703 0.0180    1.36
 4 dose    0.00226 0.0341    3.92
 5 feel    0.0135  0.00726  -0.891
 6 feeling 0.00297 0.0143    2.27
 7 hours   0.0162  0.00387  -2.07
 8 iâ      0.0108  0.0176    0.700
 9 shot    0.0131  0.0210    0.685
10 sore    0.00703 0.0139    0.985
11 vaccine 0.0192  0.00965  -0.995
```

```
##check for ID duplication
pfizer_vac %>% select(id) %>%
  anyDuplicated()
```

Console:

```
[1] 0
```

```
##most participant ID and score
pfizer_vac %>% filter(comms_num > 10) %>%
  mutate(year = year(date),
         comms_numscore = paste(comms_num, score, sep = "-")) %>%
  count(comms_num, id, year, comms_numscore) %>%
  ggplot(aes(
    x = id,
    y = comms_num,
    label = comms_numscore,
    fill = id
  )) +
  geom_col(show.legend = FALSE) +
  geom_text(vjust = 0, hjust = -0.1, size = 3, angle = 90) +
  facet_wrap(vars(year), ncol = 2, scales = "free_x") +
  scale_fill_viridis(discrete = TRUE, option = "E") +
  scale_y_continuous(expand = expansion(add = c(0, 650))) +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(strip.background = element_rect(fill = "#081d58")) +
  theme(strip.text = element_text(colour = 'yellow', size = 11)) +
  labs(
    title = "The Most Participant (ID) and their score",
    subtitle = "column Plot and facet wrap",
    caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit",
    x = "ID",
    y = "Number of Titles"
  )
```




```
###post of the highest score participant
ID_lam09i <- reddit_pfizer_vac %>% filter(id == "lam09i") %>% select(title, text, id,
comms_num, score)

formattable(ID_lam09i)
```

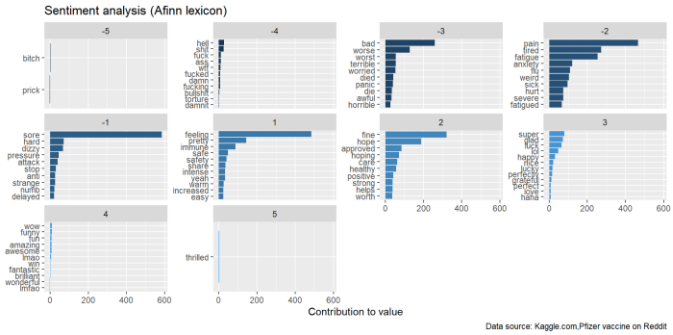
title	text	id	comms_num	score
Post Pfizer Vaccine Experience	Hey guys, I received the Pfizer Vaccine (only the first dose so far) and am looking for others to share experiences, mainly experience a week after the shot as well as related to physical activity and working out. I did feel a little soreness in the shoulder the day after but felt overall alright. However it's been over a week now and I feel generally more tired at all times, especially at the gym. I am a hobby bodybuilder and am In the sport for 5 years now. Ever since I received the vaccine, I feel like my workouts have been short, way more exhausting and weaker. Has anyone else experienced this?	lam09i	1465	396

```
################################################
###########Doing sentiment analysis############
####afinn lexicon
afinn_counts <- pfizer_vac %>%
  select(text) %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("afinn")) %>%
  count(word, value, sort=T) %>%
  ungroup()

afinn_counts
```

Console:

```
         word value   n
1        sore    -1 586
2     feeling     1 484
3        pain    -2 465
4        fine     2 320
5       tired    -2 272
6         bad    -3 258
7     fatigue    -2 254
8        hope     2 188
9      pretty     1 143
10      worse    -3 127
11    anxiety    -2 121
12        flu    -2 109
13      weird    -2 104
14       sick    -2  96
15     immune     1  87
16   approved     2  84
17      super     3  79
```

```
afinn_counts %>%
  group_by(value) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=value)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~value, scales = "free_y")+
  labs(title = "Sentiment analysis (Afinn lexicon)",
       caption = "Data source: Kaggle.com,Pfizer vaccine on Reddit",
       y="Contribution to value", x=NULL)+
  coord_flip()
```

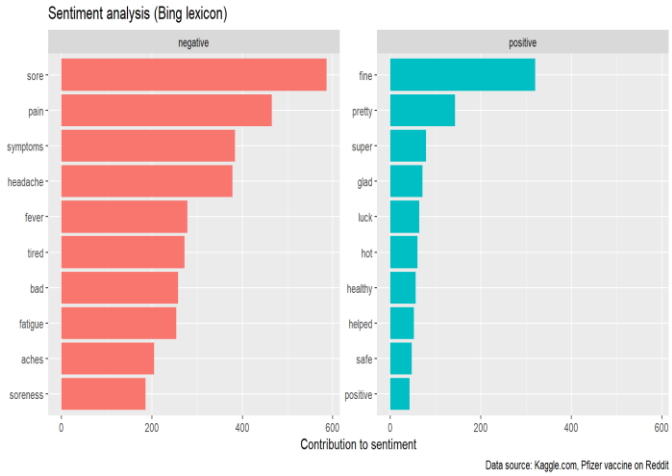



```
####bing lexicon
bing_counts <- pfizer_vac %>%
  select(text) %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_counts
```

Console:

```
          word sentiment   n
1         sore  negative 586
2         pain  negative 465
3     symptoms  negative 384
4     headache  negative 378
5         fine  positive 320
6        fever  negative 278
7        tired  negative 272
8          bad  negative 258
9      fatigue  negative 254
10       aches  negative 205
11   soreness  negative 186
12     pretty  positive 143
13      worse  negative 127
14    anxiety  negative 121
15      weird  negative 104
16  headaches  negative  97
17       sick  negative  96
```

```
bing_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(title = "Sentiment analysis (Bing lexicon)",
       caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit",
       y="Contribution to sentiment", x=NULL)+
  coord_flip()
```




```
################################################
###########Analyzing using ngrams############
#### creating a tidy format for score<3
low_score <- pfizer_vac %>%
  filter(score<3)

#### creating a tidy format for score>3
high_score <- pfizer_vac %>%
  filter(score>3)

###########using bigrams to analyze high score###########
pfizer_bigrams <- high_score %>%
  unnest_tokens(bigram, text, token = "ngrams", n=2)

pfizer_bigrams
```

Console:

```
    comms_num       date score      id           bigram
1           2 2021-01-28    17  l6vz2x      received my
2           2 2021-01-28    17  l6vz2x         my first
3           2 2021-01-28    17  l6vz2x       first dose
4           2 2021-01-28    17  l6vz2x        dose last
5           2 2021-01-28    17  l6vz2x       last night
6           2 2021-01-28    17  l6vz2x         night at
7           2 2021-01-28    17  l6vz2x            at 530
8           2 2021-01-28    17  l6vz2x         530 very
9           2 2021-01-28    17  l6vz2x       very quick
10          2 2021-01-28    17  l6vz2x        quick and
```

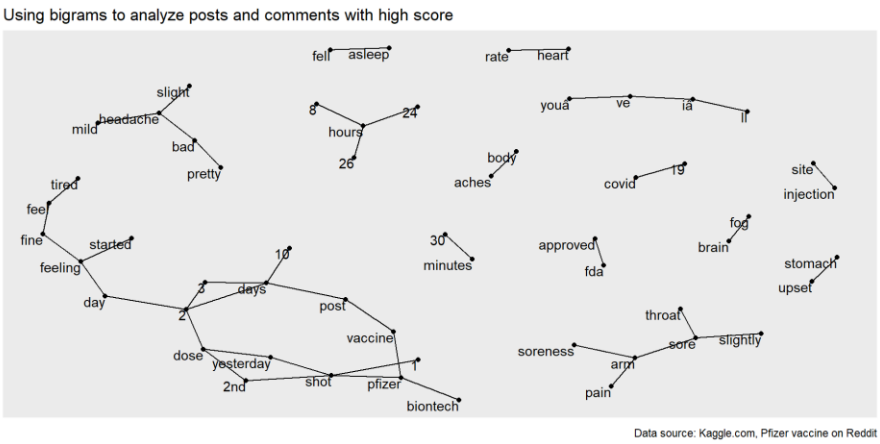
```
pfizer_bigrams  %>%
  count(bigram, sort = TRUE)

bigrams_separated <- pfizer_bigrams  %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
```

```
  filter(!word2 %in% stop_words$word)

#creating the new bigram, "no-stop-words":
bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)
#want to see the new bigrams
bigram_counts
```

Console:

```
            word1          word2  n
1            sore            arm 11
2       injection           site 10
3             iâ             ve  9
4             iâ             ll  8
5          pfizer        vaccine  8
6            sore         throat  8
7             2nd           dose  6
8             2nd           shot  6
9             arm           pain  5
10            arm       soreness  5
11          brain            fog  5
12            fda       approved  5
13           feel          tired  5
14           mild       headache  5
15           post        vaccine  5
```

```
####visualizing bigraph
bigram_graph <- bigram_counts %>%
  filter(n>2) %>%
  graph_from_data_frame()

##arrows indicate the direction
ggraph(bigram_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1)+
  labs(title = "Using bigrams to analyze posts and comments with high score",
       caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit")
```

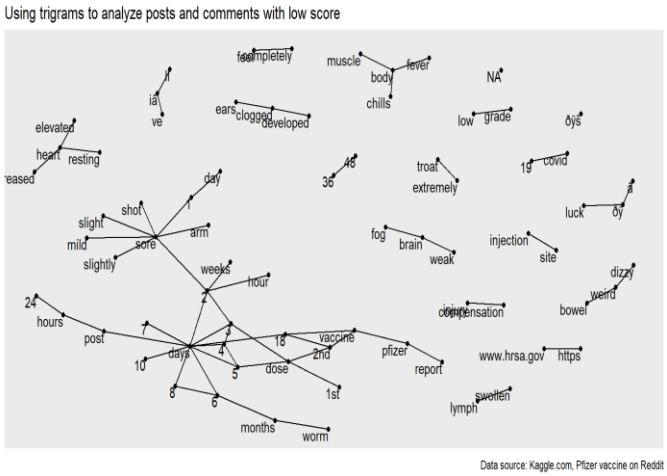


```
###########using trigrams to analyze low score###########
pfizer_trigrams <- low_score %>%
  unnest_tokens(trigram, text, token = "ngrams", n=3)

pfizer_trigrams
```

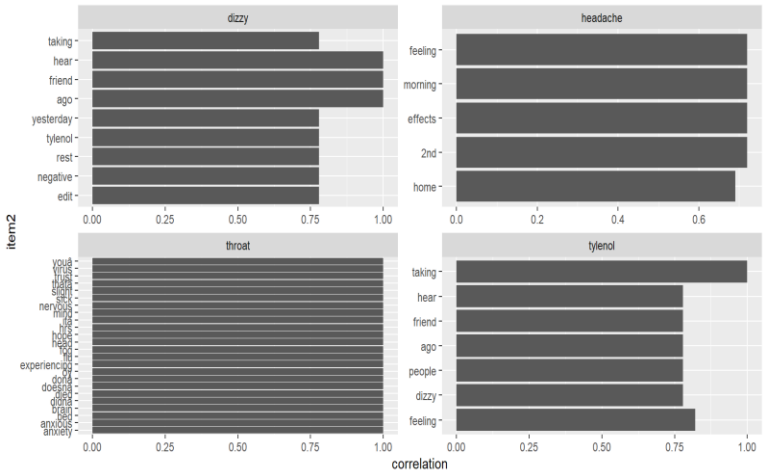
Console:

```
                         trigram
1                           <NA>
2                     there is no
3                       is no fda
4                  no fda approved
5             fda approved vaccine
6             approved vaccine for
7                vaccine for covid
8           for covid authorization
9           covid authorization is
10            authorization is not
11                     is not fda
12                not fda approval
```

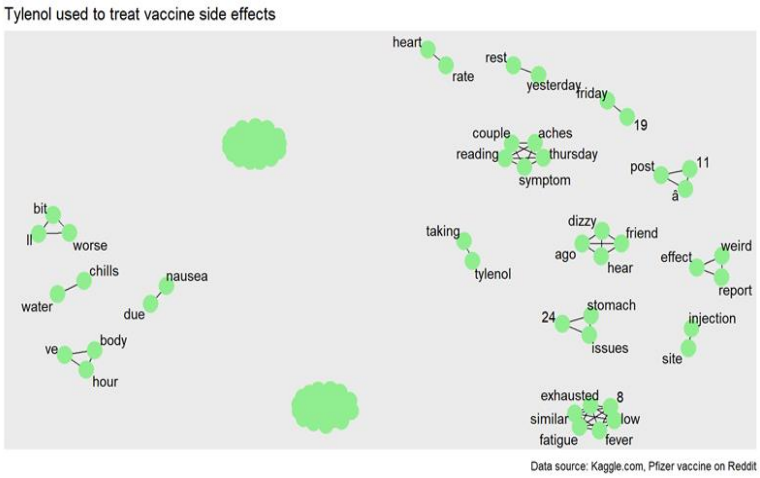
```
pfizer_trigrams  %>%
  count(trigram, sort = TRUE)

trigrams_separated <- pfizer_trigrams  %>%
  separate(trigram, c("word1", "word2", "word3"), sep = " ")

trigrams_filtered <- trigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word)

#creating the new trigram, "no-stop-words":
trigram_counts <- trigrams_filtered %>%
  count(word1, word2, word3, sort = TRUE)

#want to see the new trigrams
trigram_counts

####visualizing bigraph
trigram_graph <- trigram_counts %>%
  filter(n>3) %>%
  graph_from_data_frame()

ggraph(trigram_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1)+
  labs(title = "Using trigrams to analyze posts and comments with low score",
       caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit")
```




```
################################################
###Pairwise correlations between words #########
################################################
pfizer_vac_hscore <- reddit_pfizer_vac %>%
  filter(score>3) %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words)

word_cors <- pfizer_vac_hscore %>%
  group_by(word) %>%
  filter(n() >= 5) %>%
  pairwise_cor(word, title, sort=TRUE)

word_cors %>%
  filter(item1=="tylenol")

word_cors %>%
  filter(item1 %in% c("tylenol", "throat","dizzy","headache")) %>%
  group_by(item1) %>%
  top_n(5) %>%
  ungroup() %>%
  mutate(item2 = reorder(item2, correlation)) %>%
  ggplot(aes(item2, correlation)) + ##plotting a ggplot for item2 and the correlation
  geom_bar(stat = "identity")+
  facet_wrap(~item1, scales = "free")+
  coord_flip()
```




```
word_cors %>%
  filter(correlation >0.95) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr")+
  geom_edge_link(aes(edge_alpha = correlation), show.legend=F)+
  geom_node_point(color = "lightgreen", size=6)+
  geom_node_text(aes(label=name), repel=T)+
  labs(title = "Tylenol used to treat vaccine side effects",
       caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit")
  theme_void()
```



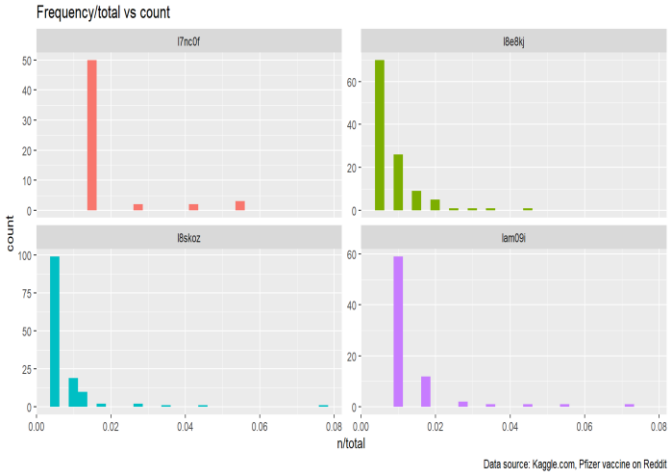

```
#########################################################
####### TF-IDF framework in Pfizer #######
#########################################################
#we're grouping by the ID this time
pfizer_token <- reddit_pfizer_vac %>%
  unnest_tokens(word, text) %>%
  count(id, word, sort=TRUE) %>%
  ungroup()
```

```
total_words <- pfizer_token %>%
  group_by(id) %>% ##doing for IDF
  summarize(total=sum(n)) ##sum of frequencies per ID

pfizer_words <- left_join(pfizer_token, total_words)%>%
  filter(id %in% c("lam09i", "l8e8kj", "l8skoz", "l7nc0f"))

print(pfizer_words)
```

Console:

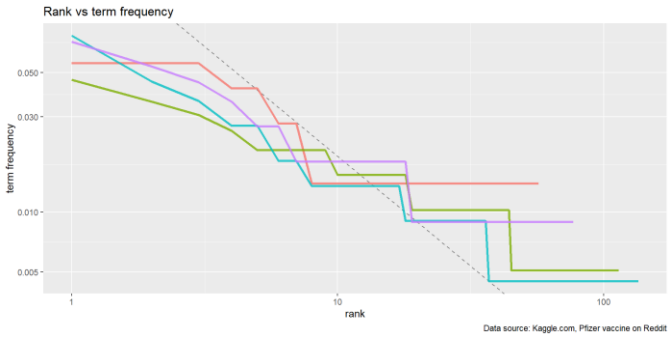
```
> print(pfizer_words)
        id        word  n total
1   l8skoz           i 17   222
2   l8skoz         and 10   222
3   l8e8kj         and  9   196
4   l8skoz         the  8   222
5   lam09i         the  8   112
6   l8e8kj         the  7   196
7   l8e8kj           i  6   196
8   l8skoz          my  6   222
9   l8skoz         was  6   222
10  lam09i           i  6   112
```

```
ggplot(pfizer_words, aes(n/total, fill = id))+
  geom_histogram(show.legend=FALSE)+
  labs(title = "Frequency/total vs count",
       caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit")+
  facet_wrap(~id, ncol=2, scales="free_y")
```




```
########################################
########## ZIPF's law ##################
########################################
freq_by_rank <- pfizer_words %>%
  group_by(id) %>%
  mutate(rank = row_number(),
         `term frequency` = n/total)
freq_by_rank
```

```
#let's plot ZIPF's Law
freq_by_rank %>%
  ggplot(aes(rank, `term frequency`, color=id))+
  #let's add a tangent line , the first derivative, and see what the slop is
  geom_abline(intercept=-0.62, slope= -1.1, color='gray50', linetype=2)+
  geom_line(size= 1.1, alpha = 0.8, show.legend = FALSE)+
  labs(title = "Rank vs term frequency ",
       caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit")+
  scale_x_log10()+
  scale_y_log10()
```




```
##upper- noise,low rank, high frequency, lower-highrank, low frequency

################## TF_IDF
id_words <- pfizer_words %>%
  bind_tf_idf(word, id, n)

id_words # we get all the zeroes because we are looking at stop words ... too common
```

Console:

```
        id        word  n total          tf       idf
1   l8skoz           i 17   222 0.076576577 0.0000000
2   l8skoz         and 10   222 0.045045045 0.0000000
3   l8e8kj         and  9   196 0.045918367 0.0000000
4   l8skoz         the  8   222 0.036036036 0.0000000
5   lam09i         the  8   112 0.071428571 0.0000000
6   l8e8kj         the  7   196 0.035714286 0.0000000
7   l8e8kj           i  6   196 0.030612245 0.0000000
8   l8skoz          my  6   222 0.027027027 0.0000000
9   l8skoz         was  6   222 0.027027027 0.2876821
10  lam09i           i  6   112 0.053571429 0.0000000
11  l8e8kj          my  5   196 0.025510204 0.0000000
12  lam09i         and  5   112 0.044642857 0.0000000
13  l7nc0f         and  4    72 0.055555556 0.0000000
```

```
uniqueness <- id_words %>%
  arrange(desc(tf_idf))

## looking at the graphical approach:
id_words %>%
  arrange(desc(tf_idf)) %>%
  anti_join(stop_words) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
```

```
group_by(id) %>%
top_n(6) %>%
ungroup %>%
ggplot(aes(word, tf_idf, fill=id))+
geom_col(show.legend=FALSE)+
labs(x=NULL, y="tf-idf")+
facet_wrap(~id, ncol=2, scales="free")+
scale_x_discrete(guide = guide_axis(n.dodge=2))+
labs(title = "Words with highest TF-IDF per ID ",
     caption = "Data source: Kaggle.com, Pfizer vaccine on Reddit")+
coord_flip()
```