Sentiment Analysis Notebook

library(dplyr) # for data wrangling & manipulation

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidytext) # for unnest\_tokens  
library(stringr) # to manage text  
library(ggplot2) # For data visualizations & graphs  
  
#read the data file  
articles <- read.csv("articles.csv")  
#clean the data - remove unwanted variables fields  
df\_art <- subset(articles, select = -c(X))  
  
#dropping entries with missing values  
df\_art<- na.omit(df\_art)  
  
#add date column  
df<- df\_art%>%  
 mutate(published\_date = format (as.Date(publication\_date, format= "%d-%b-%y"), "%B-%Y"))  
#load stop words data  
#data("stop\_words")

## Article Titles

The analysis shows that among the most popular words used in article titles are: COVID-19, SARS, 2020, and United States. The top 10 commonly used words and the number of times they appeared on the article titles are shown below.

#art\_title  
title <- data\_frame(text = df$art\_title) %>%  
 unnest\_tokens(word, text) %>% # split words  
 anti\_join(stop\_words) %>% # take out "a", "an", "the", etc.  
 count(word, sort = TRUE) # count occurrences

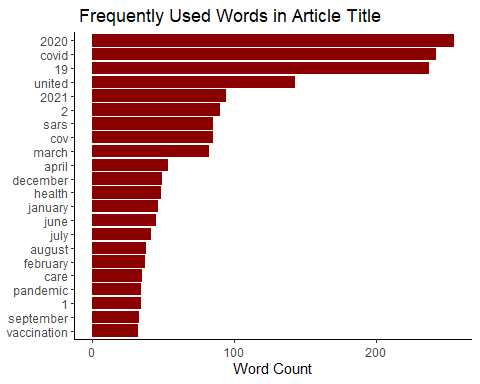
## Warning: `data\_frame()` was deprecated in tibble 1.1.0.  
## Please use `tibble()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

## Joining, by = "word"

head(title, 10)

## # A tibble: 10 × 2  
## word n  
## <chr> <int>  
## 1 2020 255  
## 2 covid 242  
## 3 19 237  
## 4 united 143  
## 5 2021 94  
## 6 2 90  
## 7 cov 85  
## 8 sars 85  
## 9 march 82  
## 10 april 53

title %>% filter(n > 30)%>%  
 mutate(word = reorder(word, n)) %>%  
 ggplot(aes(word, n)) +  
 geom\_col(fill = "darkred") + theme\_classic() +  
 xlab(NULL) +  
 ylab("Word Count") +  
 coord\_flip() +  
 ggtitle(" Frequently Used Words in Article Title")



## What is Known

The variable represent further information about the titles. Analysis of the variable shows that similar words are used like in the headlines. COVID, SARS, health, and transmission are some of th most commonly used terms. A list of top 10 words is shown below accompanied with a visualization of the commonly used words with more than 30 entries.

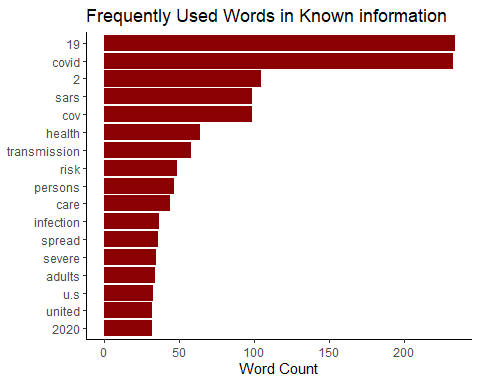
#what\_is\_known  
known <- data\_frame(text = df$what\_is\_known) %>%   
 unnest\_tokens(word, text) %>% # split words  
 anti\_join(stop\_words) %>% # take out "a", "an", "the", etc.  
 count(word, sort = TRUE) # count occurrences

## Joining, by = "word"

head(known,10)

## # A tibble: 10 × 2  
## word n  
## <chr> <int>  
## 1 19 234  
## 2 covid 233  
## 3 2 105  
## 4 cov 99  
## 5 sars 99  
## 6 health 64  
## 7 transmission 58  
## 8 risk 49  
## 9 persons 47  
## 10 care 44

known%>% filter(n > 30)%>%  
 mutate(word = reorder(word, n)) %>%  
 ggplot(aes(word, n)) +  
 geom\_col(fill = "darkred") + theme\_classic() +  
 xlab(NULL) +  
 ylab("Word Count") +  
 coord\_flip() +  
 ggtitle("Frequently Used Words in Known information")



## What is Added

Additional information included in the articles was analyzed to determine how it impacted the audience. Sentiment analysis shows that most of the terms used affected the readers negatively. Below is a list of words and how they could have affected the audience. The analysis shows more negative sentiments than positive.

#what\_is\_added  
added <- data\_frame(text = df$what\_is\_added) %>%   
 unnest\_tokens(word, text) %>% # split words  
 anti\_join(stop\_words) %>% # take out "a", "an", "the", etc.  
 count(word, sort = TRUE) # count occurrences

## Joining, by = "word"

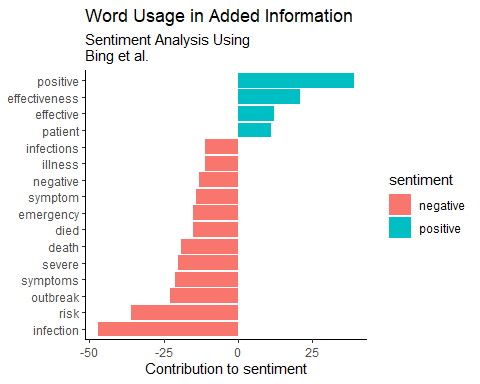
added\_bing <- added %>%  
 inner\_join(get\_sentiments("bing"))

## Joining, by = "word"

head(added\_bing, 10)

## # A tibble: 10 × 3  
## word n sentiment  
## <chr> <int> <chr>   
## 1 infection 47 negative   
## 2 positive 39 positive   
## 3 risk 36 negative   
## 4 outbreak 23 negative   
## 5 effectiveness 21 positive   
## 6 symptoms 21 negative   
## 7 severe 20 negative   
## 8 death 19 negative   
## 9 died 15 negative   
## 10 emergency 15 negative

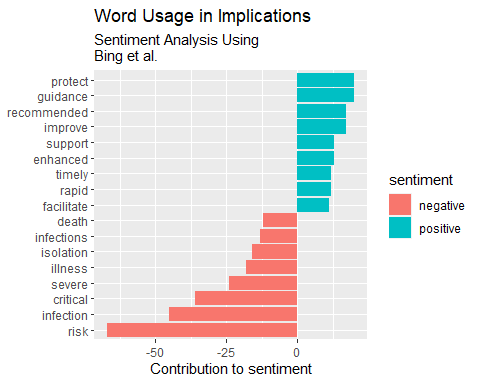
added\_bing %>%  
 filter(n > 10) %>%  
 mutate(n = ifelse(sentiment == "negative", -n, n)) %>%  
 mutate(word = reorder(word, n)) %>%  
 ggplot(aes(word, n, fill=sentiment)) +  
 theme\_classic() +  
 geom\_col() +  
 xlab(NULL) +  
 coord\_flip() +  
 ylab("Contribution to sentiment") +  
 ggtitle("Word Usage in Added Information", subtitle = "Sentiment Analysis Using  
Bing et al.")



## Implications

Similarly, sentiments analysis is used to further determine effects of the information on the audience and healthcare sector. The analysis shows more negative effects than positive. The visualization shows the sentiments from the implications variable.

## # A tibble: 10 × 3  
## word n sentiment  
## <chr> <int> <chr>   
## 1 risk 67 negative   
## 2 infection 45 negative   
## 3 critical 36 negative   
## 4 severe 24 negative   
## 5 guidance 20 positive   
## 6 protect 20 positive   
## 7 illness 18 negative   
## 8 improve 17 positive   
## 9 recommended 17 positive   
## 10 isolation 16 negative



## Publication Dates

We analyzed the publications dates by determining the number of publication done in each month. April 2021 had the most publications followed by October 2020. The publication date data is visualized below for clear understanding of the differences in number of articles published between 2020 and 2021.

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

dates <- df %>%  
 group\_by(published\_date) %>%  
 count()%>%  
 arrange(desc(n))  
  
  
  
head(dates, 10)

## # A tibble: 10 × 2  
## # Groups: published\_date [10]  
## published\_date n  
## <chr> <int>  
## 1 April-2021 24  
## 2 October-2020 22  
## 3 January-2021 21  
## 4 November-2020 20  
## 5 September-2021 20  
## 6 July-2020 18  
## 7 March-2021 18  
## 8 September-2020 18  
## 9 December-2020 15  
## 10 February-2021 15

dates %>%  
 ggplot(aes(published\_date, n, fill = n)) +  
 theme\_classic()+  
 geom\_col() +  
 xlab("Date of publication") +  
 coord\_flip() +  
 ylab("Number of publications") +  
 ggtitle("Dates of Publication and Number of Publications")

