Twitter Data Analysis: Text mining, Topic Modelling and Sentiment Analysis

Social, political and economic debate about Hong Kong

Marta Fattorel

1. Introduction

This year, the 23rd anniversary of Hong Kong's handover to China from British rule has been marked by the new *National Security Law*, which came into force on June 30.

Until 1997 Hong Kong was a British colony, then returned to China under the so called Basic National Law agreement and the "one country, two systems" principle. This arrangement allows the city to enjoy some autonomy and freedom from mainland China, at least until 2047. However, in recent times the Beijing Government has shown clear intentions to enhance its control over the city of Hong Kong (The New York Times, 2020; BBC, 2020).

Last June, the plan to allow extradition to mainland China, which, according to the opponents, would have exposed hongkongers to violent and unfair treatments and would have enhanced China's influence over Hong Kong, caused numerous pro-democracy protests carried out by Hong Kong citizens. Even though the extradition bill was suspended, the protests to claim more rights and democracy did not stop (BBC, 2020).

However, the recent National Security Law has changed the rules. Its 66 ambiguous articles impose harsh penalties to the communist party's dissidents and thus give the authorities extensive power to target activists. Furthermore, the law states that even foreigners who support Hong Kong's independence or criticize the Chinese government will be prosecuted upon entering China. Clearly, this measure will have a severe impact on freedom and security of hongkongers (The New York Times, 2020).

As a consequence, the approval of the National Security Law has triggered the reactions of many countries and companies worldwide. For instance, President Trump announced that he will end preferential treatments for Hong Kong in trade and travel, defining the new measure as a "tragedy". Canada and Australia have suspended their extradition treaties, while the UK has offered citizenship options to hongkongers. Furthermore, even some Tech giants have questioned the Chinese measure. Twitter, Google and Facebook stated that they would temporarily stop processing Hong Kong government requests for users' data. Tik Tok, instead decided to stop offering its social video app in Honk Kong (CNN, 2020; BBC, 2020).

Now, what if we want to know also people's thoughts and reactions to this political measure? Nowadays, thanks to the spread of the World Wide Web and social media networks, people have, on the one hand, the possibility to broaden their viewpoints and on the other hand, can share their

thoughts and polarized opinions. In addition, the development of Natural Language Processing and Text Mining techniques have allowed to retrieve, process and analyse user-generated content in an aggregate and systematic way (Cambria, Schuller, Xia, Havasi, 2013).

Taking this into account, the purpose of this paper is to explore and analyse the social and political debate about Hong Kong by applying techniques of Text Mining, Topic Modelling and Sentiment Analysis. In particular, I will retrieve data from Twitter since this micro-blogging platform is widely used by politicians and influential personalities as well as by ordinary people to debate about relevant political and social issues.

Thus, this work is going to be structured as follows: I will first briefly present some studies related to my research in terms of methodology and content and I will describe the two questions that guide my research design. Secondly, I will illustrate the methodology used to answer my questions and I will perform the chosen techniques in R on data retrieved from Twitter. Finally, I will interpret the obtained results and present some final considerations.

2. Related work and research questions

Recently, several researchers have combined techniques, especially Sentiment Analysis, in order to analyse people's opinions on Twitter with respect to different topics. For instance, a study carried out by three researchers from the University of Turin, explored the polarized political debate around the reform of marriage in France in 2012-2013, which addressed the topic of homosexual weddings (Lai, Virone, Bosco, Patti, 2013). Another interesting work investigated the public opinions and sentiments towards the Syrian refugees crisis through a comparative study of tweets in English and Turkish (Naza, Serkan, 2018). Furthermore, two researchers from the University of Munich, worked on the analysis of tweets' sentiments in the context of German federal elections to examine whether Twitter is used as a forum for political deliberation and whether online tweets mirror the offline political sentiment (Tumasjan, Sprenger, Sandner, Welpe, 2010). Even outside of academia, many companies use Sentiment Analysis to extract consumers' or clients' opinions about products and services from online reviews or again, they extract data from social media to examine the general consensus achieved by a political party or candidate to predict their electoral success. Finally, another study combined techniques of topic modelling (LDA) and sentiment analysis, respectively to identify the major topics on Twitter and to estimate their degree of polarity (Yoon, Kim, Kim, Song, 2016).

Concerning my research topic, in order to shed light on the current debate about Hong Kong I will answer the following questions: 1) Concerning Hong Kong, what are people discussing? Are they talking about the National Security Law and its social, political and economic aspects? 2) Do people have polarized opinions about the topics?

3. Methodology

As mentioned before, in order to answer my research questions I will combine techniques of Text Mining, Topic Modelling and Sentiment Analysis on Twitter data. These analytic tools grew out of the need to gain insights from the increasing amount of online user-generated content deriving from

Internet forums, discussion groups, blogs and social media and they represent useful techniques to process textual data.

My analysis will follow this structure: data retrieval through Twitter Web API, text mining as a pre-processing step in order to clean and prepare the collected tweets, topic modelling aiming at exploring the most relevant sub-topics concerning the Hong Kong situation and finally sentiment analysis to examine the debate polarization. I will use the R programming language.

3.1 The dataset

My dataset consists of 7,000 tweets in English, which cover the time period from the 5th to the 12th of July and thus, refer to the week after the National Security Law approval. First of all, I retrieved 100,000 tweets that matched the expression 'Hong Kong' or the hashtags '#hongkongers' or '#HKPolice'. In this way I collected enough data to cover the whole week and then, due to the lack of powerful computational tools to process such a number of information, I selected a random subset of 7,000 tweets and I created a data frame keeping only two columns: text and date of tweet's creation.

```
# load up all the required libraries

library(rtweet)
library(ggplot2)
library(wordcloud)
library(tm)
library(topicmodels)
library(lubridate)
library(SentimentAnalysis)
library(quanteda)
library(ggpubr)
library(dplyr)
library(tidytext)

# Twitter authentication and data retrieval

consumer_key <- 'xxx'
consumer_secret <- 'xxx'
access_token <- 'xxx'
access_secret <- 'xxx'

token <- create_token(
  app = "xxx",
  consumer_key = consumer_key,
  consumer_secret = consumer_secret,
  access_token = access_token,
  access_secret = access_secret)

HK_tweets <- search_tweets(q = "Hong Kong OR #hongkongers OR #HKPolice",
                           n = 100000,
                           lang = "en",
                           include_rts = FALSE,
                           since = "2020-07-05",
                           until = "2020-07-12",
                           retryonratelimit = TRUE)
```

```r
HK_df <- data.frame(HK_tweets$text, HK_tweets$created_at)

samp <- sample(nrow(HK_df), 7000)
HK_df <- HK_df[samp,]

colnames(HK_df) <- c("Text", "Date")
row.names(HK_df) <- NULL
HK_df$Date <- as.Date(ymd_hms(HK_df$Date)) # remove time
write.csv(HK_df, "HongKong.csv", quote = TRUE)
```

3.2 Data analysis

Text Mining

Text Mining refers to the process of extracting patterns and knowledge from unstructured text documents. It is an interdisciplinary field of study that encompasses data mining, linguistics, computational statistics, and computer science techniques (Feinerer, Hornik, Meyer, 2008). It has several applications, but in this study, I will use it only in this initial stage as a data pre-processing technique.

```r
HK_df <- read.csv("HongKong.csv ", row.names = 1, stringsAsFactors = TRUE)

# remove duplicates
HK_df <- HK_df[unique(HK_df$Text),]
row.names(HK_df) <- NULL

# remove mentions
HK_df$Text <- gsub("@[[:alpha:]]*","", HK_df$Text)



# remove emoji in unicode
HK_df$Text <- gsub("[<].*[>]", "", HK_df$Text)

# remove tabs
HK_df$Text = gsub("[ |\t]{2,}", " ", HK_df$Text)

# leading blanks
HK_df$Text = gsub("^ ", "", HK_df$Text)

# lagging blankc
HK_df$Text = gsub(" $", "", HK_df$Text)

# general spaces
HK_df$Text = gsub(" +", " ", HK_df$Text)

# convert to basic ASCII text
HK_df$Text <- iconv(HK_df$Text, to = "ASCII", sub = " ")
```

I am now going to transform the tweets' text into a corpus object and use the *tm* library to perform some more data cleaning on it.

```
text_corpus <- Corpus(VectorSource(HK_df$Text))

# convert to lower case
text_corpus <- tm_map(text_corpus, content_transformer(tolower))

# words to remove
text_corpus <- tm_map(text_corpus, removeWords,
c("will", "just", "can", "amp", "via", "also", "yet", "per", "hong", "kong",
"s", "t", "get", "hk", "because", "hongkongers", "hkpolice", "hongkong"))

# stem words
text_corpus <- tm_map(text_corpus, stemDocument)

# remove stopwords
text_corpus <- tm_map(text_corpus, removeWords,
                      stopwords("english"))

# remove punctuation
text_corpus <- tm_map(text_corpus, removePunctuation)

# remove URLs
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
text_corpus <- tm_map(text_corpus, content_transformer(removeURL))

# remove extra whitespace
text_corpus <- tm_map(text_corpus, stripWhitespace)

# clean corpus back to the HK_df
text_df <- data.frame(text_clean = get("content", text_corpus),
                      stringsAsFactors = FALSE)
HK_df$Text <- text_df$text_clean
```

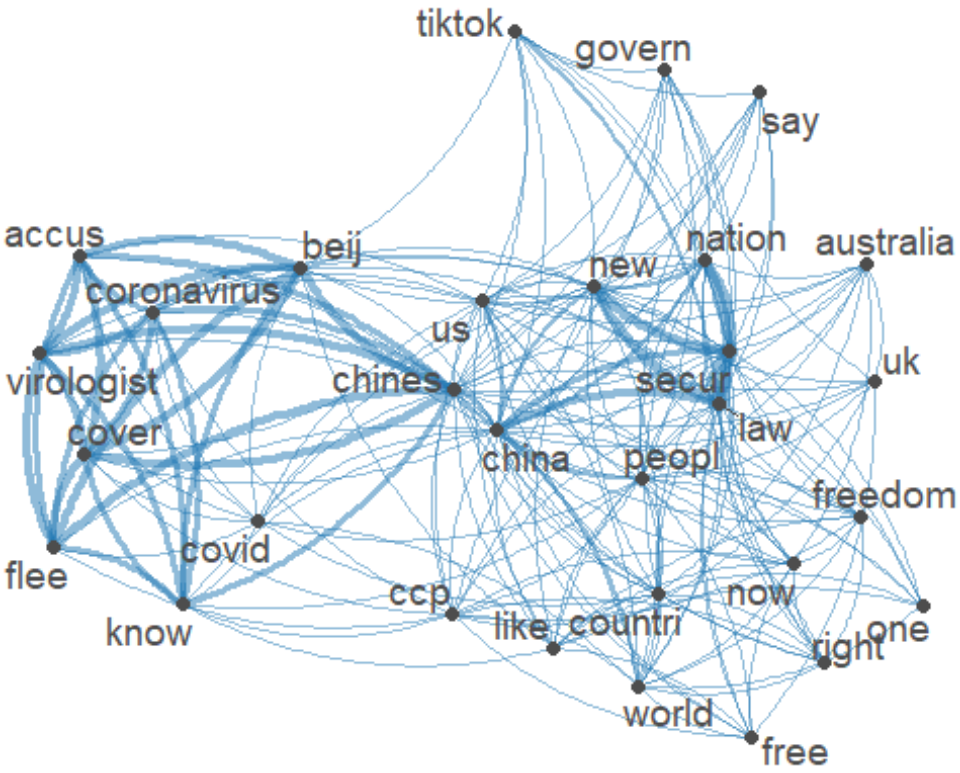
I create here a document term matrix, which describes the frequency of terms that occur in my coll ection of tweets.

```
dtm <- TermDocumentMatrix(text_corpus)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix), decreasing = TRUE)
df <- data.frame(word = names(words), freq = words)

set.seed(1234)

wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words = 200, rando
m.order = FALSE, rot.per = 0.35, colors = brewer.pal(12, "Paired"))
```

The word cloud gives a first exploratory visualization of the tweets' content. As expected, among the most frequent words, it is possible to recognize terms related to the National Security Law such as 'newsecur', 'law', 'nation', 'china', 'ccp', 'protest'. However, along with these, the connection of some relevant words with the coronavirus topic is visible.

Let's see if a network of words co-occurrences gives additional information.

```
# Generate Tokens
toks <- tokens(HK_df$Text)

set.seed(30)

fcmat <- fcm(toks, context = "document", tri = FALSE)
feat <- names(topfeatures(fcmat, 30))
fcm_select(fcmat, pattern = feat) %>%
textplot_network(min_freq = 0.5)
```



The plot displays strong ties, on the one hand among words strictly related to the National Security Law, on the other hand among coronavirus connected terms, meaning that often the words within these two 'clusters' occur together. These two sub-topics appear to be connected to each other by the words 'china' and 'chines'. I will further explore the different aspects of the debate in the next sections.

Topic Modelling

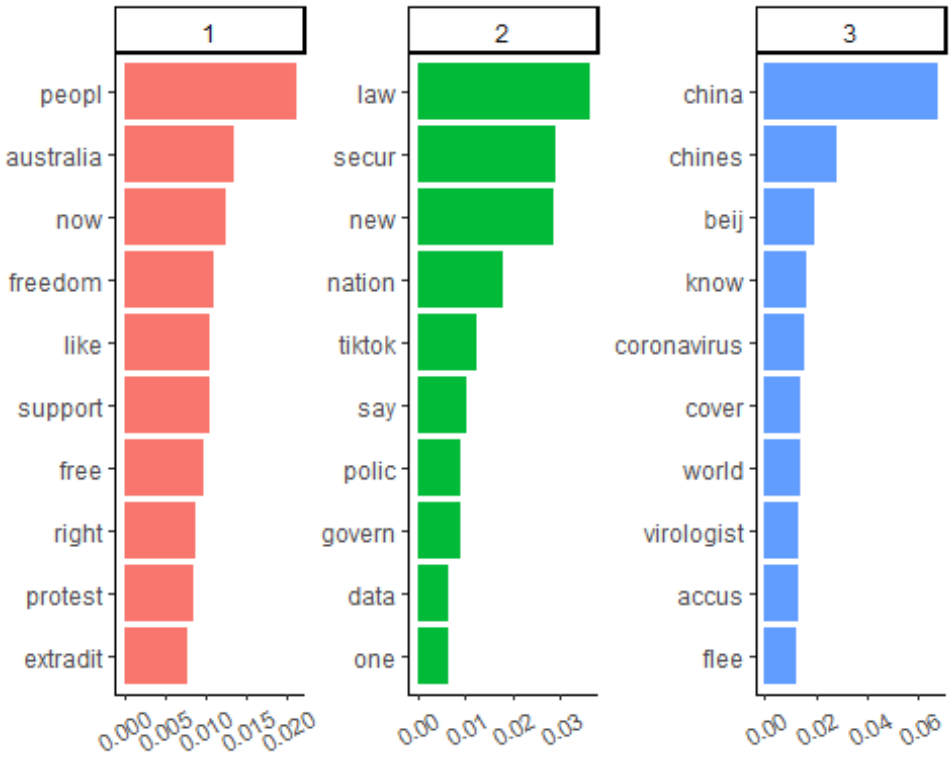
In Machine Learning and NLP, topic models are generative models which provide a probabilistic framework for the term frequency occurrences in documents in a given corpus. The fitted models can then be used to estimate the similarity between documents (Bettina, Kurt, 2011).

To answer the first research question whose purpose was to investigate the aspects of the current debate around Hong Kong, I am going to use the library *topicmodels* in order to perform a Latent Dirichlet Allocation analysis aiming at detecting sub-topics from the general debate. This time I will consider only half of the tweets, insofar the process is computationally expensive.

```
set.seed(100)
samp <- sample(nrow(HK_df), 3500)

corpus_sub <- Corpus(VectorSource(HK_df$Text[samp]))
dtm_sub = DocumentTermMatrix(corpus_sub)
doc.length = apply(dtm_sub, 1, sum) # find the sum of words in each document
dtm_sub = dtm_sub[doc.length > 0,]  # remove all docs without words
```

I tried to run the algorithm with different k options, but in terms of words coherence, 3 topics resulted to be the best choice.

```r
# Number of topics
k <- 3

# Run LDA using Gibbs sampling
ldaOut3 <-LDA(dtm_sub, k, method = "Gibbs", control =
                list(nstart = 5, seed = list(2003, 5, 63, 100001, 765),
                     best = TRUE, burnin = 4000, iter = 2000, thin = 500))


topics <- as.matrix(topics(ldaOut3))
terms <- as.matrix(terms(ldaOut3, 10))
topics_prob <- as.matrix(ldaOut3@gamma)

theme_set(theme_classic())
topics_beta <- tidy(ldaOut3, matrix = "beta")

top_terms_b <- topics_beta %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms_b %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  labs(x = NULL, y = NULL) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  theme(axis.text.x = element_text(angle = 30, vjust = 0.5, size = 8)) +
  coord_flip()
```

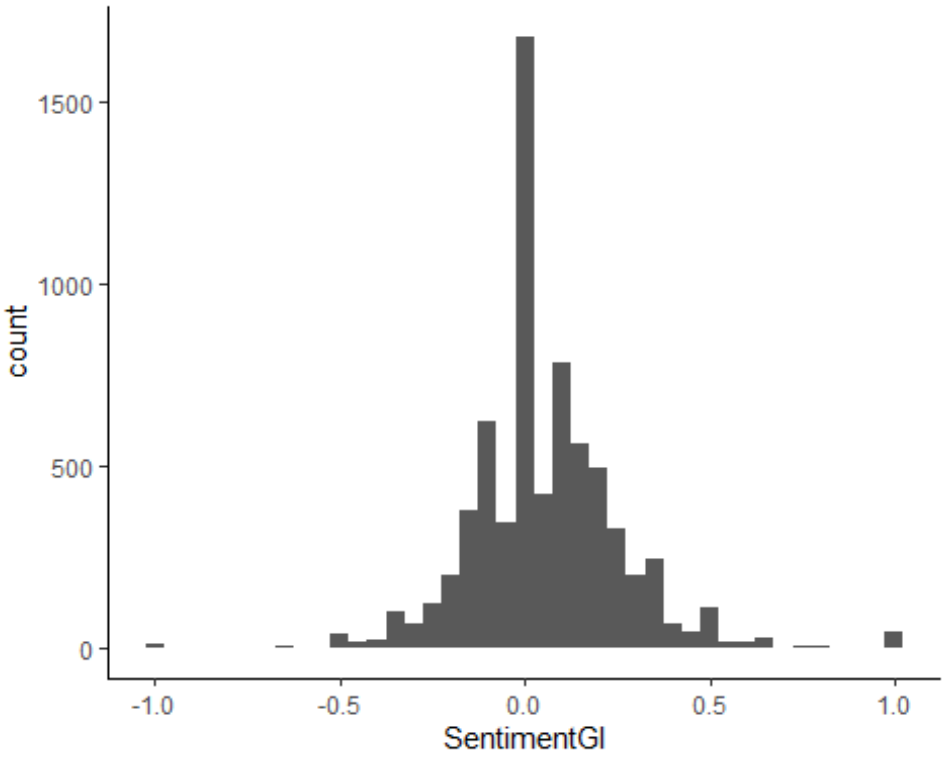


The plot shows the 10 most common words within each topic considering 'beta' which refers to the per-topic-per-word probability. Looking at the words it is possible to assign some labels to identify the sub-topics obtained. In particular, the first one seems to be related to the Hong Kong pro-democracy protests as underlined by words like 'people', 'freedom', 'support', 'right', 'protest'. The second sub-topic appears to be linked to the National Security Law. In fact, beyond the words that define the law, we see 'politic', 'govern', 'data' and 'tiktok' that are part of the same debate. Finally, the third sub-topic is quite interesting since before seeing the data I did not think about the possibility to retrieve tweets related to coronavirus. However, the outcome makes sense. In fact, on July 10, an hongkonger virologist named Li-Meng Yan, told in an interview with Fox News, that the government knew about the coronavirus before it claimed it did. It also ignored the research she was doing about the novel virus. Li-Meng Yan is now in hiding in the US because she fears her life is in danger after this statement (Fox News, 2020).

Sentiment Analysis

To answer the second research question, I will perform Sentiment Analysis on my data. In particular, Sentiment Analysis or Opinion Mining, is defined as the computational study of opinions, sentiments and emotions expressed in text (Cambria, Schuller, Xia, Havasi, 2013). I will perform here just the basic task of opinion mining i.e. polarity classification to detect the polarization of the debate about Hong Kong. I am going to use the *analyzeSentiment()* function from the *SentimentAnalysis* library and select only the sentiment score according to the dictionary GI since I think that this general purpose dictionary is more suitable for my analysis compared to the others, which are focused on finance. For each tweet, the algorithm computes a sentiment score ('SentimentGI') by averaging over all the scores (1, 0 or -1) given to each single counted word within the tweet.

```
HK_sentiment <- analyzeSentiment(HK_df$Text)

# keep only 'SentimentGI' and 'WorCount' columns
HK_sentiment <- dplyr::select(HK_sentiment,
                              SentimentGI,
                              WordCount)

# add to the HK_df dataset
HK_df <- cbind.data.frame(HK_df, HK_sentiment)
HK_df <- HK_df[!is.na(HK_df$SentimentGI),] # remove NA

theme_set(theme_classic())

ggplot(HK_df[, 3:4], aes(SentimentGI)) +
  geom_histogram(binwidth = .05)
```



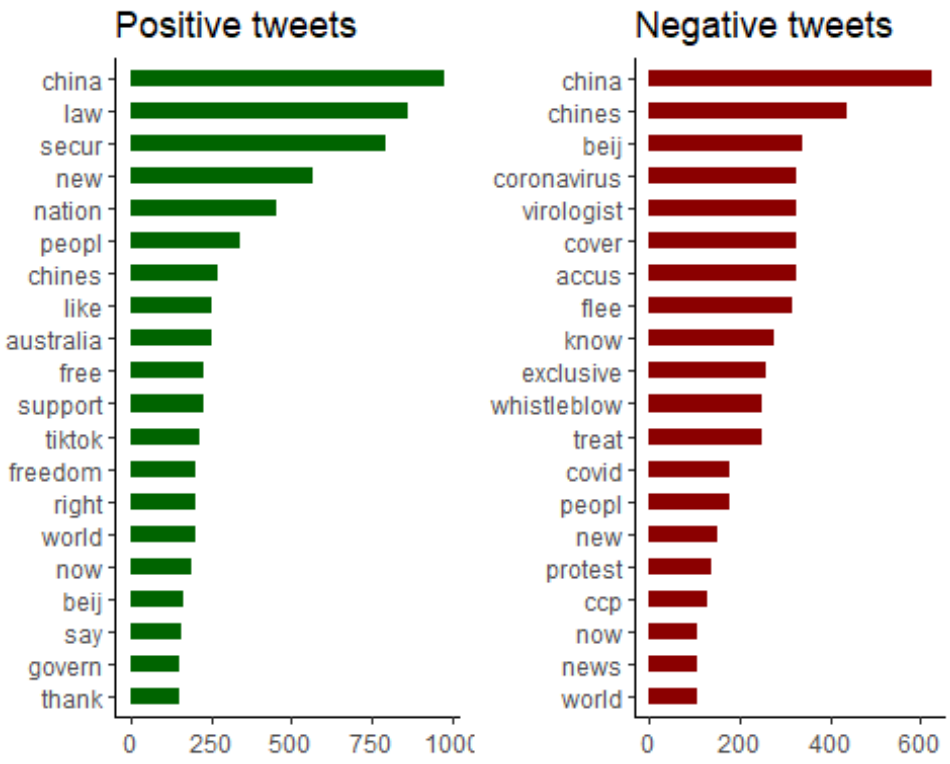
The histogram shows the distribution of the tweets' sentiment scores. It is possible to notice that there seems to be more positive tweets than negative, but the majority of them are close to 0, meaning that most of the tweets are neutral. This may be due to the fact that often, in each tweet some words are labelled as positive, while others as negative and then the average results to be close to 0. Furthermore, the function considers only single words, which means that expressions as 'not good' are labelled with 1 i.e. completely positive. Thus, we can state that 'SentimentGI' is an approximation of the whole tweet's sentiment, but many errors may have occurred.

At this point, I want to investigate which are the most frequent words within the positive and the negative tweets and try to find an association with the sub-topics previously built with LDA. In order to do so, I create two different document term matrices, one to compute the frequency of words within the tweets labelled as positive and the other for the negative ones (neutral tweets are excluded).

```
text_corpus_pos <- Corpus(VectorSource(HK_df[HK_df$SentimentGI > 0, "Text"]))
dtm.pos <- TermDocumentMatrix(text_corpus_pos)
mat.pos <- as.matrix(dtm.pos)
words.pos <- sort(rowSums(mat.pos), decreasing = TRUE)

text_corpus_neg <- Corpus(VectorSource(HK_df[HK_df$SentimentGI < 0, "Text"]))
dtm.neg <- TermDocumentMatrix(text_corpus_neg)
mat.neg <- as.matrix(dtm.neg)
words.neg <- sort(rowSums(mat.neg), decreasing = TRUE)

# create the data frames
df_sent_pos <- data.frame(word = names(words.pos), freq = words.pos)
df_sent_neg <- data.frame(word = names(words.neg), freq = words.neg)
```

```
theme_set(theme_classic())
df_sent_pos$word <- factor(df_sent_pos$word, levels = df_sent_pos$word)
g1 <- ggplot(df_sent_pos[1:20,], aes(reorder(word, freq), y=freq)) +
```

```
  geom_bar(stat ='identity', width =.5, fill = "darkgreen") +
  labs(x = NULL, y = NULL, title = "Positive tweets") + coord_flip()

df_sent_neg$word <- factor(df_sent_neg$word, levels = df_sent_neg$word)
g2 <- ggplot(df_sent_neg[1:20,], aes(reorder(word, freq), y=freq)) +
  geom_bar(stat ='identity', width = .5, fill = "darkred") +
  labs(x = NULL, y = NULL, title = "Negative tweets") + coord_flip()

ggarrange(g1, g2, ncol = 2, nrow = 1)
```



According to the bar plots, it appears that in general positive tweets concern the Hong Kong debate about the National Security Law, but words related to the protests are present as well, while the negative tweets relate mostly to the coronavirus sub-topic, even if words as ‘protest’ and ‘ccp’ are present as well.

3.3 Results

At this point it is possible to discuss the results of the analysis and answer the initial research questions.

1) Concerning Hong Kong, what are people discussing? Are they talking about the National Security Law and its social, political and economic aspects?

The first two exploratory visualizations (word cloud and network of words co-occurrences) gave an intuitive idea of the general content that characterizes the retrieved tweets. Surprisingly, beyond the National Security Law and the protests topics, a debate linked to coronavirus appeared. By applying

Topic Modelling with LDA, 3 distinct sub-topics connected to Hong Kong emerged. 1) Debate about the Hong Kong protests, as suggested by the most frequent words 'people', 'freedom', 'support', 'right', 'protest'. We can define it as the social aspect of the discussion. 2) Debate about the National Security Law approval underlined by 'law', 'secur', 'nation', 'politic', 'govern', 'data', 'tiktok'. Here we can recognize references to both the political and economic aspects. 3) Debate about coronavirus, related to the virologist who accused the government to have hidden, in its initial stage, the known information about the novel coronavirus ('coronavirus', 'know', 'virologist', 'accus').

2) Do people have polarized opinions about the topics?

In order to estimate the tweets' polarization I performed Sentiment Analysis. The majority of the tweets resulted to be neutral, with more positive than negative ones. I will discuss the drawbacks of this algorithm later on, but in general it showed that the overall debate is quite neutral. Furthermore, by displaying the most frequent words in tweets labelled as positive and as negative, it is possible to say that positive ones relate to the debate about the National Security Law and the protests, while the negative tweets are mostly linked to the coronavirus sub-topic.

4. Final considerations and conclusion

To recap, the main purpose of this research was to shed light on the current debate about Hong Kong on Twitter. Since this micro-blogging platform is widely used by people to share their knowledge and opinions about the surrounding world's events, I decided to analyse Twitter data. In particular, I applied techniques of Text Mining, Topic Modelling and Sentiment Analysis. My main findings are the following. The general debate about Hong Kong events is divided into: debate about the famous pro-democracy protests in Hong Kong (social aspect), debate about the National Security Law approval (political and economic aspects) and coronavirus discussion. In addition, tweets with a positive connotation seem to be connected with the first two sub-topics, instead the ones labelled as negative are more likely to refer to the coronavirus sub-topic. However, I want to underline that the fact that tweets about the National Security Law and the protests are labelled as positive does not mean that people support them. This also because if you are in favour of the protests, for sure you do not have a positive opinion about the repressive law. Further analyses are needed to clarify these aspects.

The main advantages of this work regard the possibility, given by the performed algorithms, to easily and rapidly analyse a quite large amount of tweets, which would not have been possible to process by hand. Furthermore, these analytic tools allow to extend data mining and in general quantitative and structured data analysis techniques, to process unstructured text in a systematic way. Moreover, it is interesting that during the research some unexpected aspects, such as the coronavirus sub-topic, may emerge, giving the researcher the possibility to gain new knowledge and insights into the topic of interest. With respect to the disadvantages, I previously underlined some limitations concerning the algorithm to implement Sentiment Analysis. They mainly refer to the fact that each word is considered in isolation, while we know that the context is really important. This leads to misclassify negations such as 'not good', which is given a positive score. In addition, the algorithm faces problems which the whole NLP field is struggling with: coreference resolution, anaphora resolution, named-entity recognition and word-sense disambiguation (Cambria, Schuller, Xia, Havasi, 2013). Moreover, since we are dealing with tweets, another challenge is represented by slang words. An improvement, to reduce the randomness of the Topic Modelling results could be to perform hyperparameter tuning.

In conclusion, future developments could consider data which cover a broader range of time and explore the evolution of people's sentiments over time. Furthermore, it would be interesting to differentiate between verified and not verified accounts to find out whether the general debate carried out by more influential people differs from the rest of the population.

References

BBC News (2020), *Hong Kong security law: What is it and is it worrying?* https://www.bbc.com/news/world-asia-china-52765838

BBC News (2020), *The Hong Kong protests explained in 100 and 500 words*: https://www.bbc.com/news/world-asia-china-49317695

BBC News (2020), *National security law: Australia suspends Hong Kong extradition treaty*: https://www.bbc.com/news/world-australia-53344013

Bettina G., Kurt H. (2011), *topicmodels: An R Package for Fitting Topic Models*, Journal of Statistical Software, 40 (13), p. 1

Cambria E, Schuller B, Xia Y, Havasi C. (2013), *New avenues in opinion mining and sentiment analysis*, IEEE Intelligent Systems, Vol. 28, pp. 15–21

CNN Business (2020), *Following controversial national security law, TikTok is leaving Hong Kong*: https://edition.cnn.com/2020/07/07/tech/tiktok-leaving-hong-kong-intl-hnk/index.html

Feinerer I., Hornik K., Meyer D. (2008), *Text Mining Infrastructure in R*, Journal of Statistical Software, 25(5), pp. 1–54

Fox News (2020), *EXCLUSIVE: Chinese virologist accuses Beijing of coronavirus cover-up, flees Hong Kong: 'I know how they treat whistleblowers'*: https://www.foxnews.com/world/chinese-virologist-coronavirus-cover-up-flee-hong-kong-whistleblower

Lai M., Virone D., Bosco C., Patti V. (2015), *Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization*, In Proc. of 2015 IEEE International Conference on Data Science and Advanced Analytics, pp. 1–9, Paris, France

Nazan Ö ztü rk, Serkan A. (2018), *Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis*, Telematics and Informatics, Vol. 35, 1, pp. 136-147

The New York Times (2020), *Harsh Penalties, Vaguely Defined Crimes: Hong Kong's Security Law Explained*: https://www.nytimes.com/2020/06/30/world/asia/hong-kong-security-law-explain.html

Tumasjan A., Sprenger T. O., Sandner P. G., Welpe I. M. (2010), *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*, In Proc. 4th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)

Yoon H. G., Kim H., Kim C. O., Song M. (2016), *Opinion polarity detection in Twitter data combining shrinkage regression and topic modelling*, J. Informetrics, Vol. 10, pp. 634-644