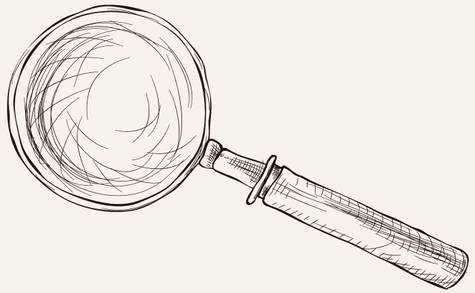
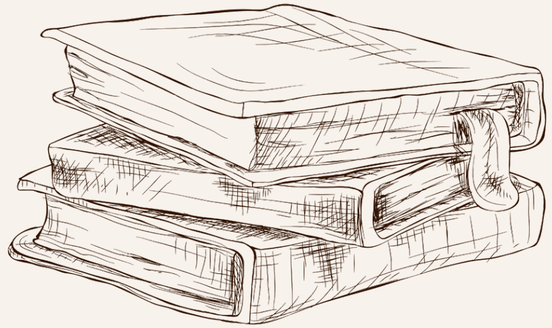
COMP 2501 Presentation



IDENTIFYING FAKE NEWS



Ng Kin Hei (3036067418)

WHAT IS FAKE NEWS?

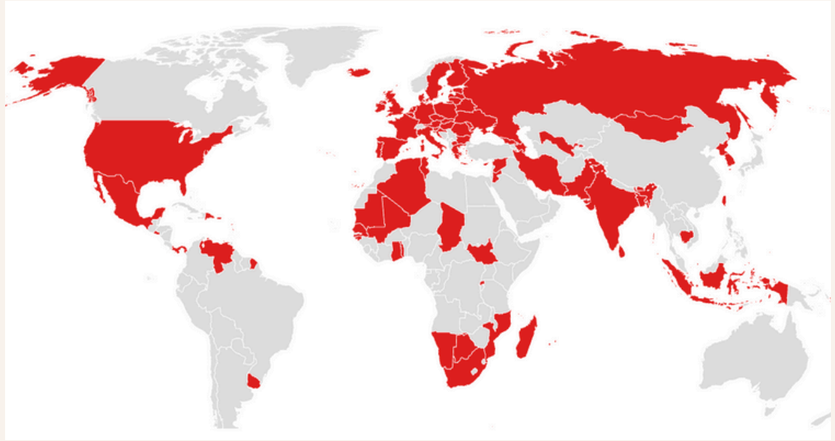
Misleading!!!

IMPORTANCE OF IDENTIFYING

We live in an age of information disorder

Instagram, Facebook, X, Tiktok...

The Ultimate Election Year



Project Objectives

Are there specific sentiment patterns associated with fake news articles that can be detected using sentiment analysis techniques?

Are there any specific patterns or linguistic cues that are more prevalent in fake news articles compared to genuine ones?

Dataset I use

Kaggle (WELFake_Dataset)

>72,000 news articles

Dataset contains four columns: Serial number (starting from 0); Title (about the text news heading); Text (about the news content); and Label (0 = real and 1 = fake).

Problems Considered

Language - (sentiment patterns can differ across cultural contexts)

Tools

RStudio (Version: 2023.12.0+369)

Library:

library(dplyr) [data manipulation]
library(ggplot2) [visualization]
library(textcat) [language detection]
library(wordcloud) [visualization]
library(tm) [text mining]
library(RColorBrewer) [color palettes]
library(wordcloud2) [visualization]
library(syuzhet) [sentiment analysis]
library(stringr) [string operations]

Data Cleaning

	X	title	text	label
1	0	LAW ENFORCEMENT ON HIGH ALERT Following Threat…	No comment is expected from Barack Obama Member…	1
2	1		Did they post their votes for Hillary already?	1
3	2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS M…	Now, most of the demonstrators gathered last night …	1
4	3	Bobby Jindal, raised Hindu, uses story of Christian co…	A dozen politically active pastors came here for a priv…	0

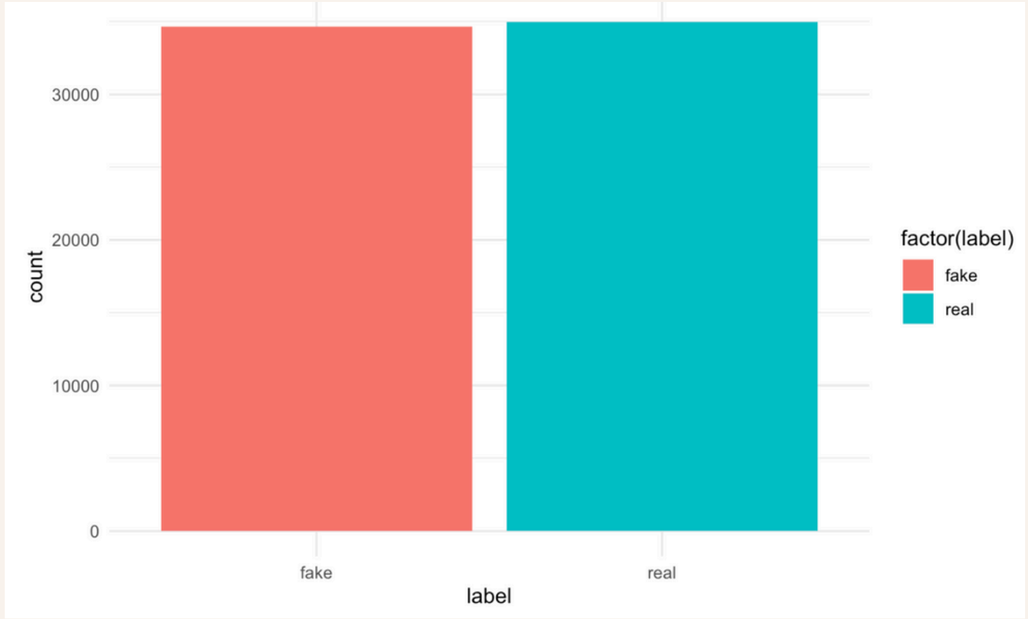
exclude rows with blank spaces in the "title" column

exclude rows with blank spaces in the "text" column

language detection for text and title -filter out all news that the text is not in english

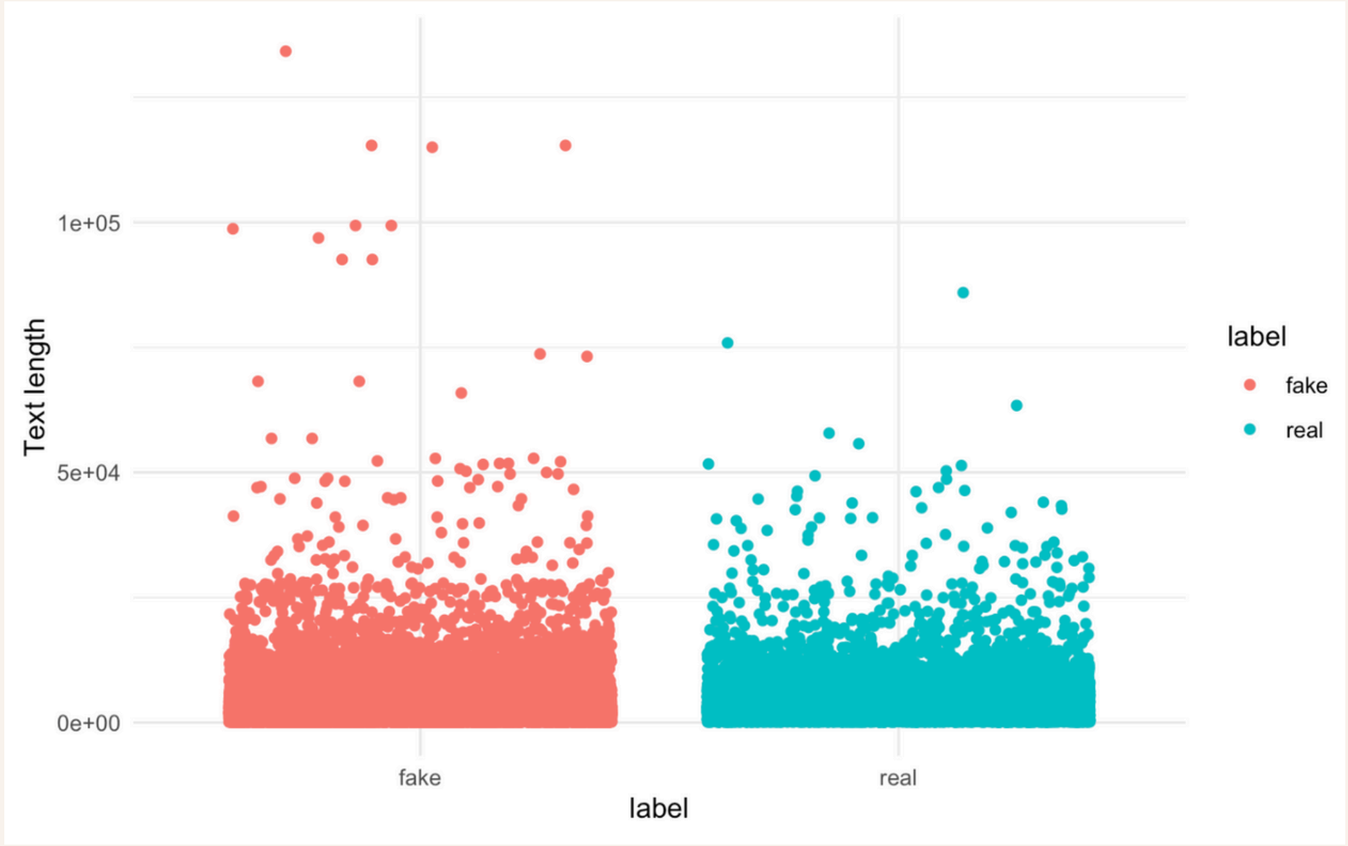
43	46	Mayday on the Carolina Queen - The New York Times	The seven fishermen aboard the Carolina Queen III, a …	real	scots	english
44	47	Illinois House passes $5 billion tax package, spendin…	CHICAGO (Reuters) - Illinois' Democratic-controlled …	real	english	english
45	48	AT&T, Time Warner and the Death of Privacy	AT&T, Time Warner and the Death of Privacy By Am…	fake	english	english
46	49	An Architect Who Built His Career on Resuscitating Ne…	Grand Central Terminal, the main building on Ellis Isl…	real	english	english
47	50	American dream, revisited	by Pepe Escobar for the Strategic Culture Foundation …	fake	english	english

label <chr>	count <int>
fake	34652
real	34951



Text Length

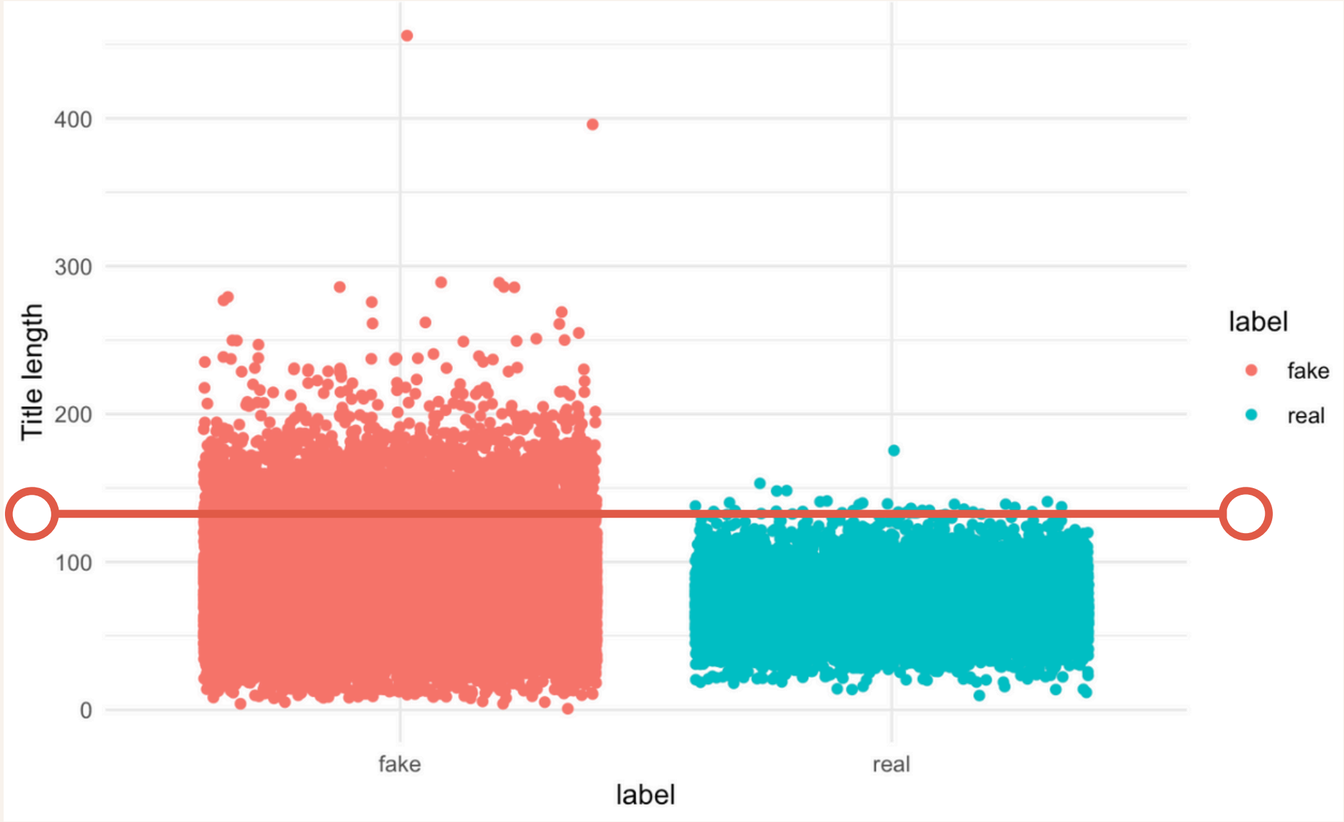
```
ggplot(filter_data, aes(x=label,
y=nchar(text), color=label)) +
    geom_jitter() +
    theme_minimal() +
    labs(y="Text length")
```



Not much different (How about title length?)

Title Length

```
ggplot(filter_data, aes(x=label,
y=nchar(title), color=label)) +
    geom_jitter() +
    theme_minimal() +
    labs(y="Title length")
```



fake news headlines are longer than real news !!!

Most Used Word

Randomly select 30000 data for testing from real and fake news datasets

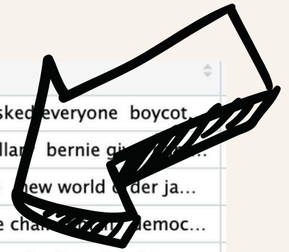
```
#dataset of real and fake news
real_news <- filter(filter_data, label == "real")
fake_news <- filter(filter_data, label == "fake")

real_news_testing <- real_news %>% sample_n(30000,
replace = FALSE)
fake_news_testing <- fake_news %>% sample_n(30000,
replace = FALSE)
```

Remove stop words

```
#Remove stop words
remove_stopwords <- function(text) {
  corpus <- Corpus(VectorSource(text))
  corpus <- tm_map(corpus, content_transformer(tolower))
  corpus <- tm_map(corpus, removePunctuation)
  corpus <- tm_map(corpus, removeNumbers)
  corpus <- tm_map(corpus, stripWhitespace)
  corpus <- tm_map(corpus, removeWords, c(stopwords("en"),
"", "'s"))

  # Get the cleaned text from the corpus
  cleaned_text <- sapply(corpus, function(x)
gsub("[|\\-\\--\\“\\”]|’S|’s", "", x))

  return(cleaned_text)
}

# Apply the remove_stopwords function to the text column
and create a new column
real_news_testing$cleaned_text <-
remove_stopwords(real_news_testing$text)
fake_news_testing$cleaned_text <-
remove_stopwords(fake_news_testing$text)
```

Create Wordcloud for real and fake news !!!

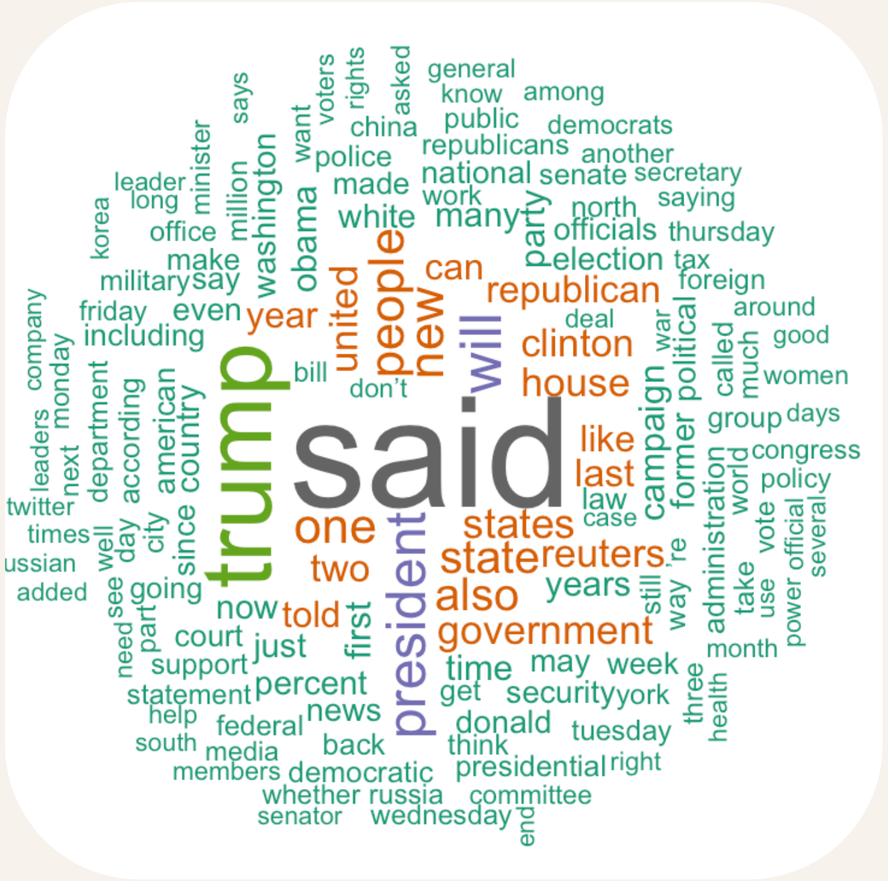
	X	title	text	label	title_language	text_language	cleaned_text
1	69390	BOX OFFICE BOMB: Seth Rogan Tweeted F*ck You To ...	Who s laughing now funny guy?We asked everyone t...	fake	scots	english	s laughing now funny guywe asked everyone boycot...
2	35747	HEY HILLARY...Who Are You Going To Blame For The ...	Does anyone believe for one second that Hillary or Be...	fake	english	english	anyone believe one second hillar bernie gi...
3	45263	James Ellroy's "American Tabloid" - Film, Literature a...	Corbett • 11/22/2016 This month on Film, Literatur...	fake	english	english	corbett • month film literature new world der ja...
4	47279	Comment on Democratic National Committee Shuts O...	According to The New York Times, the vice chairwom...	fake	english	english	according new york times vice cha democ...
5	25052	BUSTED! FBI Hid Clinton-Lynch Tarmac Meeting Docu...	The FBI is out of control. It is stunning that the FBI f...	fake	german	english	fbi control stunning fbi found clintonlynch tar...
6	6064	WATCH: Trump Calls CNN Reporter "Rude" For Questi...	WATCH: CNN Host Stuns Media, Says Clinton's Team ...	fake	english	english	watch cnn host stuns media says clinton team thinks ...
7	59610	WHOA! Did Donald Trump Just Imply Obama Is Worki...	And if Trump did indeed imply Obama was working o...	fake	german	english	trump indeed imply obama working behalf musli...
8	56556	Trump Explains Why He Fired James Comey, F*cks U...	After kicking the morning off with a tweet storm attac...	fake	german	english	kicking morning tweet storm attacking democrats ...
9	3631	"Donald Trump ""I'll overturn the shocking gay marria...	In a recent interview with Pat Robertsons television ne...	fake	scots	english	recent interview pat robertsons television network d...
10	60064	Donald Trump Retweets One Of His Fans, A 'WhiteGe...	Donald Trump woke up this morning, sleepily grabbe...	fake	scots	english	donald trump woke morning sleepily grabbed phon...
11	65328	Marco Rubio NAILS IT: Trump As President Would Ha...	Florida Senator Marco Rubio has been going after pre...	fake	catalan	english	florida senator marco rubio going presidential rival ...
12	67651	BREAKING: AP Finds Proof Melania Trump Illegally Wo...	In what may be the last bombshell to drop prior to ...	fake	german	english	may last bombshell drop prior election day asso...
13	48985	Supreme Court Decides to Weigh in On Transgender ...	By Adalia Woodbury on Sun, Oct 30th, 2016 at 11:31 ...	fake	english	english	adalia woodbury sun oct th pm friday supreme co...
14	19653	AG JEFF SESSIONS Warns Leakers...Taking Steps to Sto...	Attorney General Jeff Sessions announced today that t...	fake	english	english	attorney general jeff sessions announced today doj ...
15	22400	WHY UNEDUCATED SOMALI REFUGEES Who Don't Spe...	A few weeks ago, we reported on the first female Mus...	fake	middle_frisian	english	weeks ago reported first female muslim legislator ...
16	17901	TERROR AND DEATH THREATS CAUSE Carson And Tru...	Barack Obama already had secret service detail by thi...	fake	catalan	english	barack obama already secret service detail point colum
17	35750	Paul Ryan: The Republican Party Is More Important T...	On Sunday, House Speaker Paul Ryan stated that he fe...	fake	english	english	sunday house speaker paul ryan stated feels must ...
18	30726	WATCH: Hillary's Latest Anti-Trump Ad Absolutely B...	Hillary is starting to pull out all the stops in going aft...	fake	english	english	hillary starting pull stops going trump s bromanc...
19	25831	FBI SAYS DEM PRESIDENTIAL FRONTRUNNER HILLARY ...	The corrupt covering for the corrupt. Chicago politics...	fake	german	english	corrupt covering corrupt chicago politics best hilla...
20	48764	WHY IS THE MEDIA HIDING This Endorsement?...KKK K...	It s time for the media to start doing their job. Instea...	fake	middle_frisian	english	s time media start job instead making baseless cl...



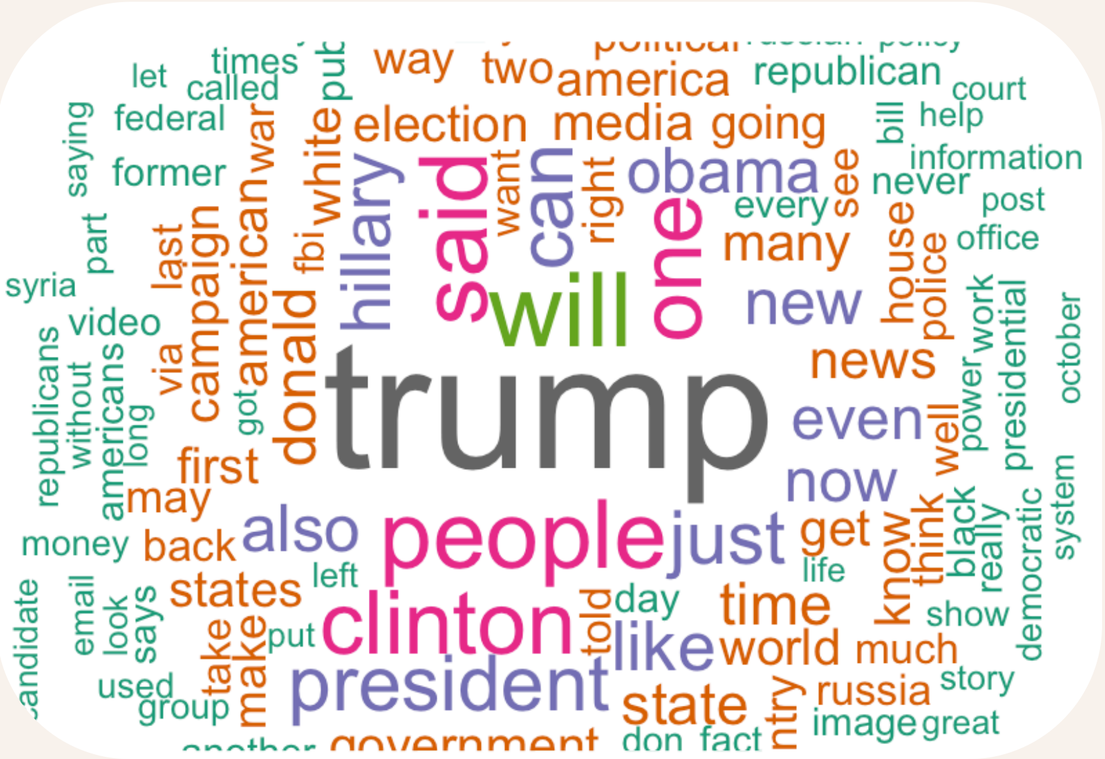
```
#Wordcloud of fake news
wordcloud(fake_news_testing$cleaned_text, min.freq = 1,
          max.words=150, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

WordCloud

Real



Fake

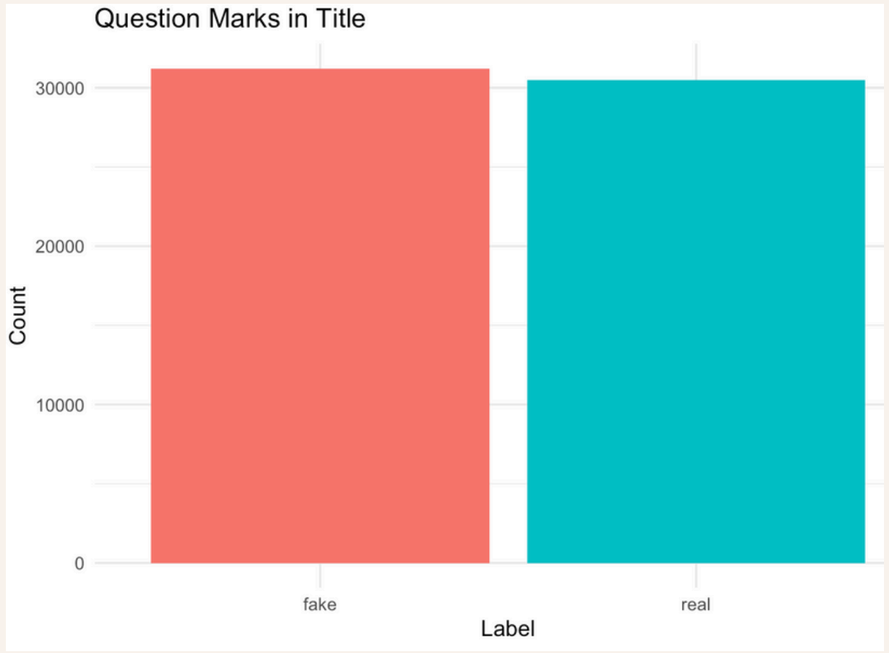


It seems that Fake news text often contains repeated vocabulary

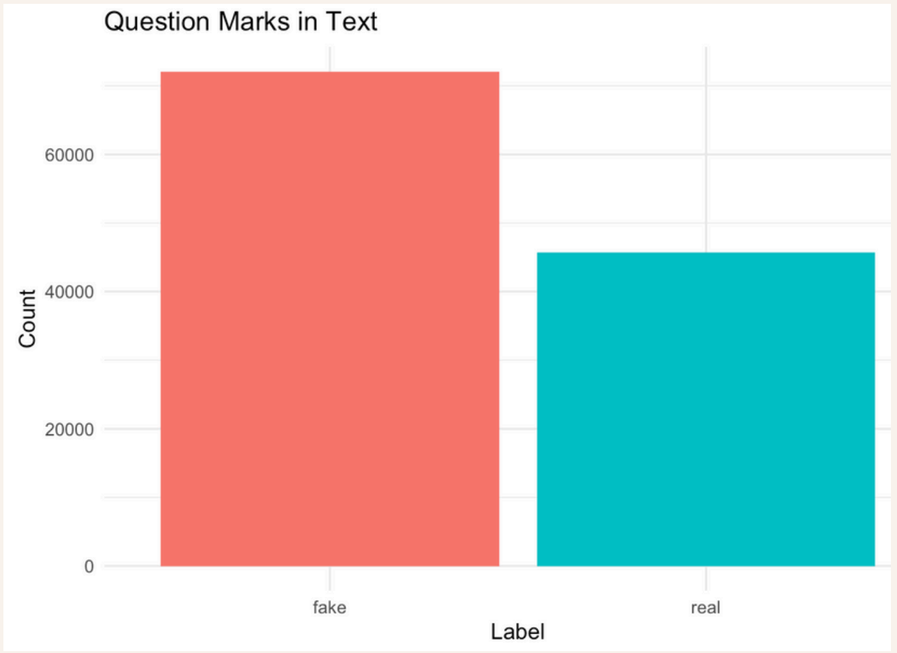
Amount of Question Mark

label <chr>	question_marks_in_text <int>	question_marks_in_title <int>
real	45681	30513
fake	72107	31229

In Title



In text

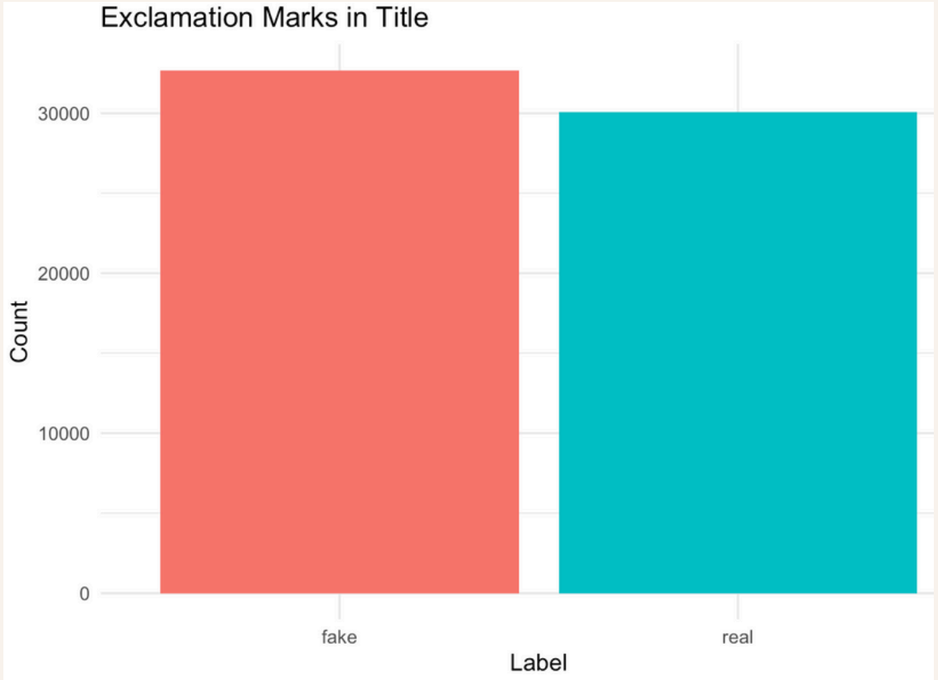


How about Exclamation Mark???

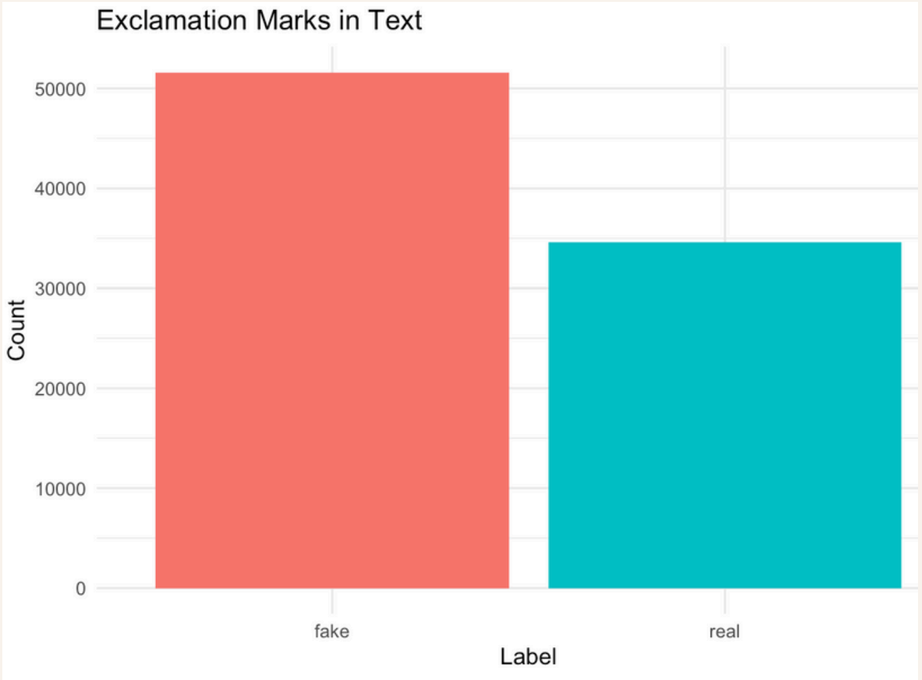
Amount of Exclamation Mark

label <chr>	exclamations_in_text <int>	exclamations_in_title <int>
real	34605	30062
fake	51615	32698

In Title



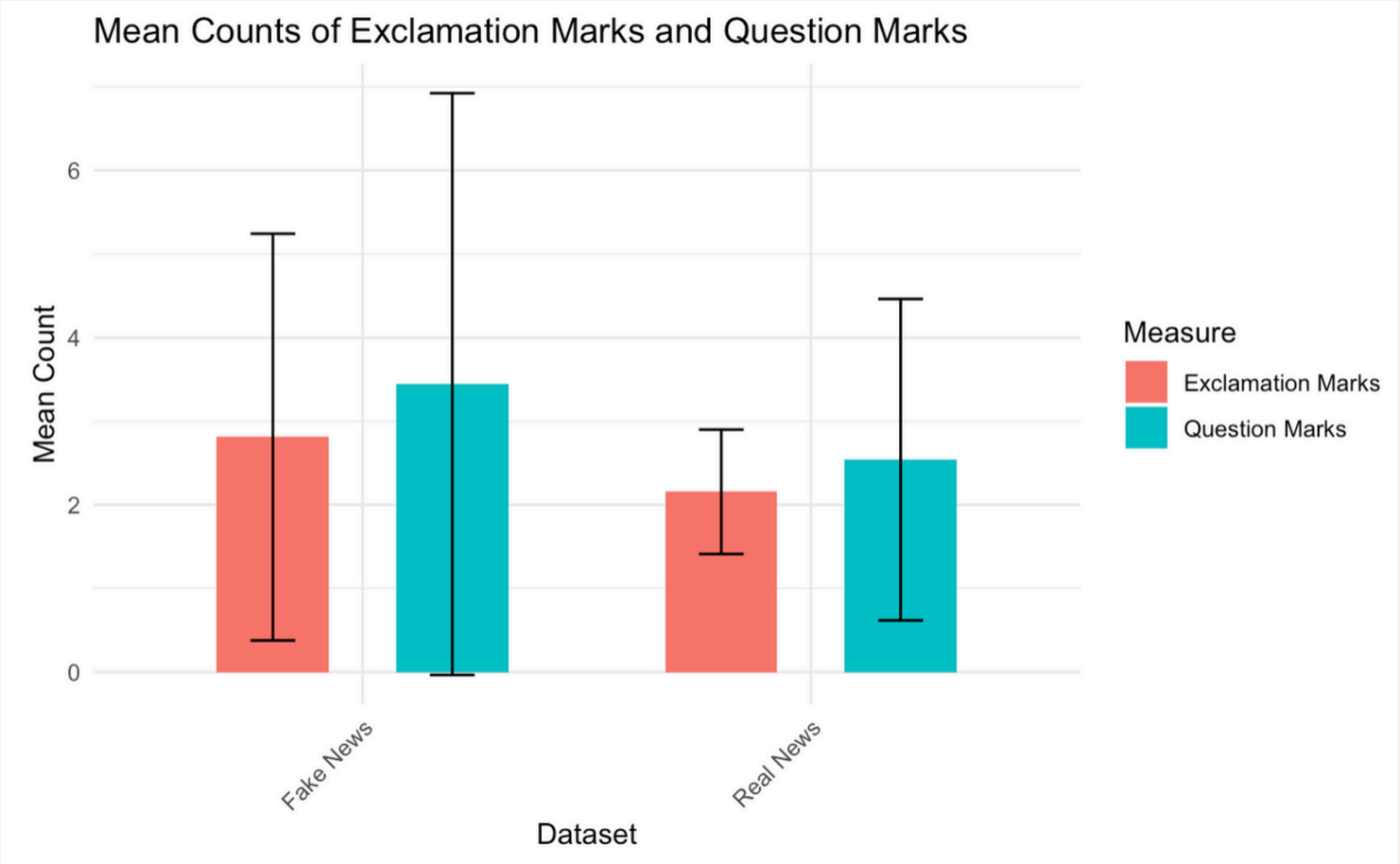
In text



Comparison

Fake news tends to have more ? and ! than real news

The difference is mainly in text but not title.



Capitalized words in title

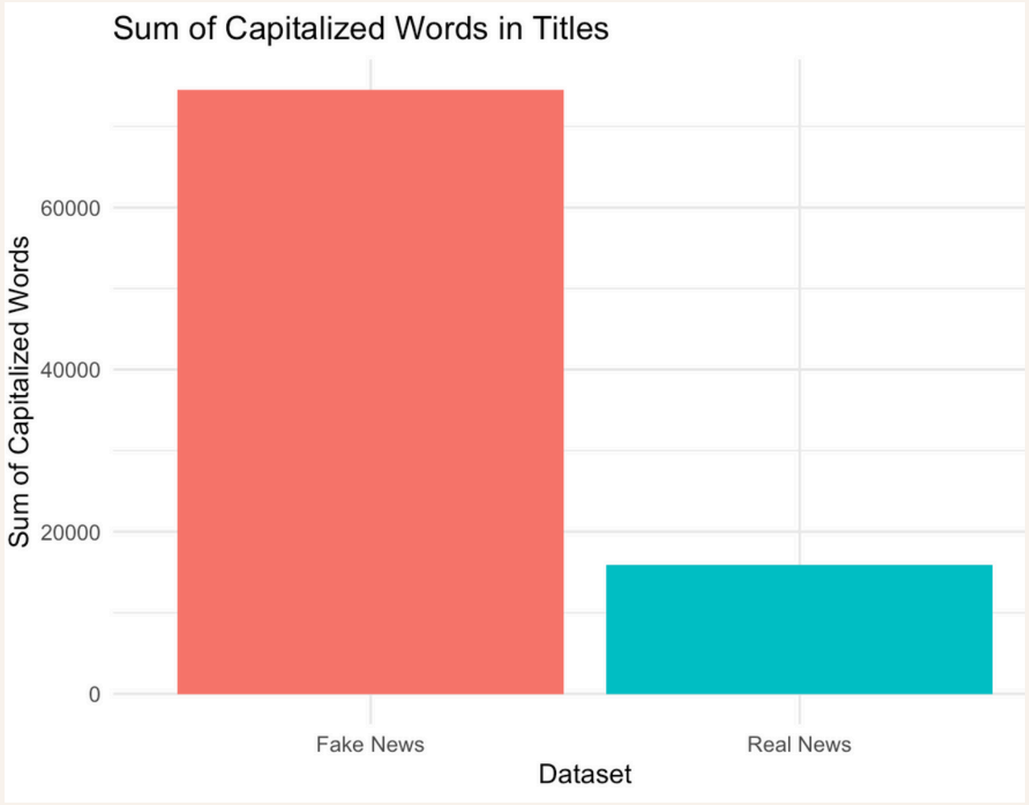
Count capitalized words in the title and add it to the new column

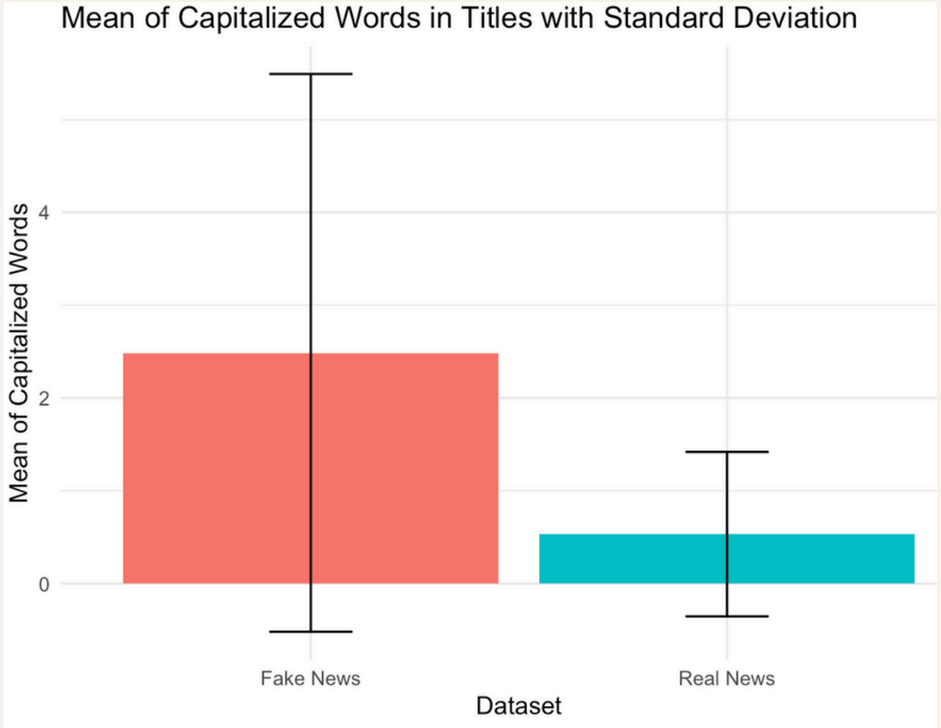
capitalized_words_count
4
2
0
0
2
2
2
2


```
# Count capitalized words in the title

uppers <- "\\b[:upper:]+\\b"

fake_news_testing$capitalized_words_count
<- sapply(fake_news_testing$title,
function(title) {
  str_count(title, uppers)
})

real_news_testing$capitalized_words_count
<- sapply(real_news_testing$title,
function(title) {
  str_count(title, uppers)
})
```

Dataset <chr>	Sum_Capitalized_Words <int>
Real News	15958
Fake News	74566

Dataset <chr>	Mean_Capitalized_Words <dbl>	SD_Capitalized_Words <dbl>
Real News	0.5319333	0.8866077
Fake News	2.4855333	3.0050888





Fake news headlines often appear in capitalized words !!!

Sentiment Analysis

Use Bing lexicon (from Bing Liu and collaborators)

categorizes words in a binary fashion into positive and negative categories

Use GetSentiment(cleaned_text) to find out the net sentiment of each news

Sentiment = Positive - negative

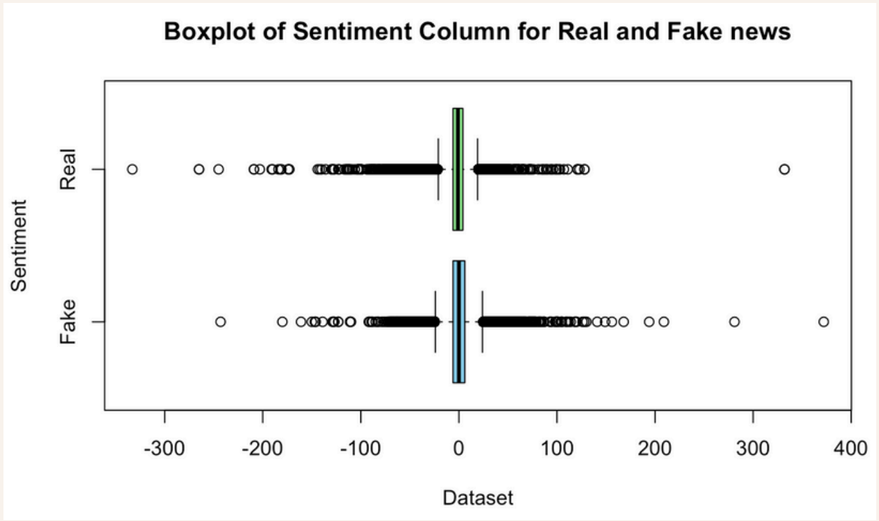
```r
# Define the function to perform sentiment analysis
GetSentiment <- function(cleaned_text){
    # Tokenize the text column
    tokens <- data_frame(text = cleaned_text) %>% unnest_tokens(word, text)
    # Get the sentiment from the text
    sentiment_table <- tokens %>%
      inner_join(get_sentiments("bing")) %>%
      count(sentiment) %>%
      spread(sentiment, n, fill = 0)

    if (sum(sentiment_table$negative) == 0) {
      return(0)}

    if (!"negative" %in% colnames(sentiment_table)) {
      # If it doesn't exist, add it and assign 0 to all its values
      sentiment_table <- sentiment_table %>%
      mutate(negative = 0)
    }
    if (!"positive" %in% colnames(sentiment_table)) {
      # If it doesn't exist, add it and assign 0 to all its values
      sentiment_table <- sentiment_table %>%
      mutate(positive = 0)}
    sentiment_table <- sentiment_table %>%  mutate(sentiment = positive - negative)
    rm(tokens)
    # Return the modified data fram
    return(sentiment_table$sentiment)
}
fake_news_testing$sentiment <- sapply(fake_news_testing$cleaned_text, GetSentiment)
real_news_testing$sentiment <- sapply(real_news_testing$cleaned_text, GetSentiment)
```

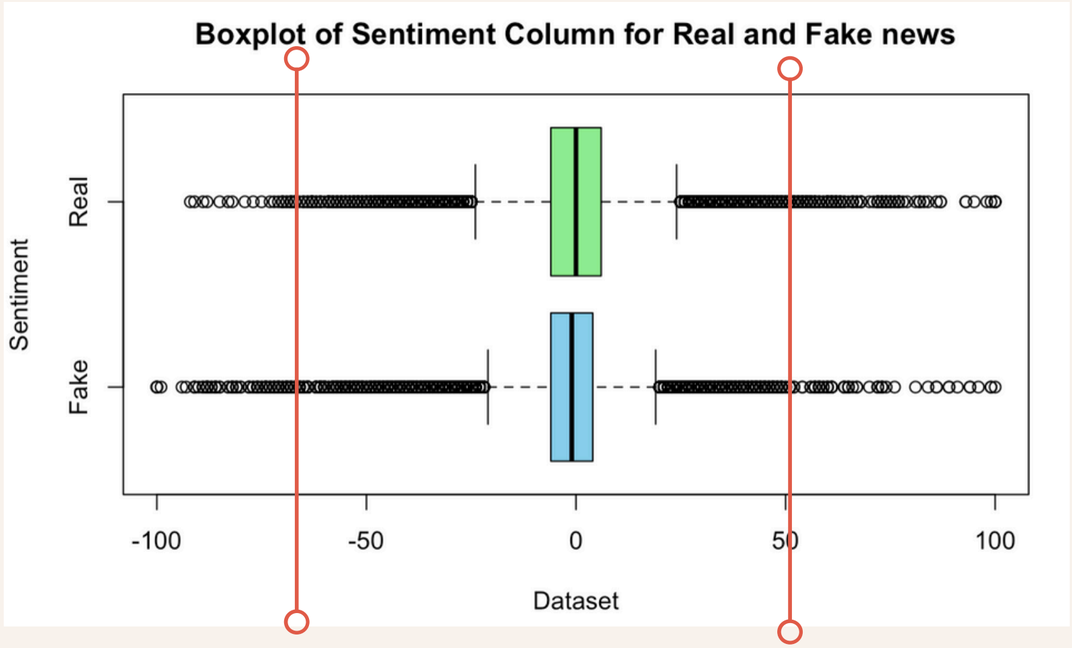
negative	positive	sentiment
117	240	123

Sentiment Analysis

Overall sentiment of Fake news is more negative than real news.

```
[1] "Number of Neg>Pos (real):  13768"
[1] "Number of Neg>Pos (fake):  15067"
Summary for Real News Sentiment:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-243.0000   -6.0000    0.0000    0.1726    6.0000  372.0000

Summary for Fake News Sentiment:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-333.000   -6.000   -1.000   -1.693    4.000  332.000
```





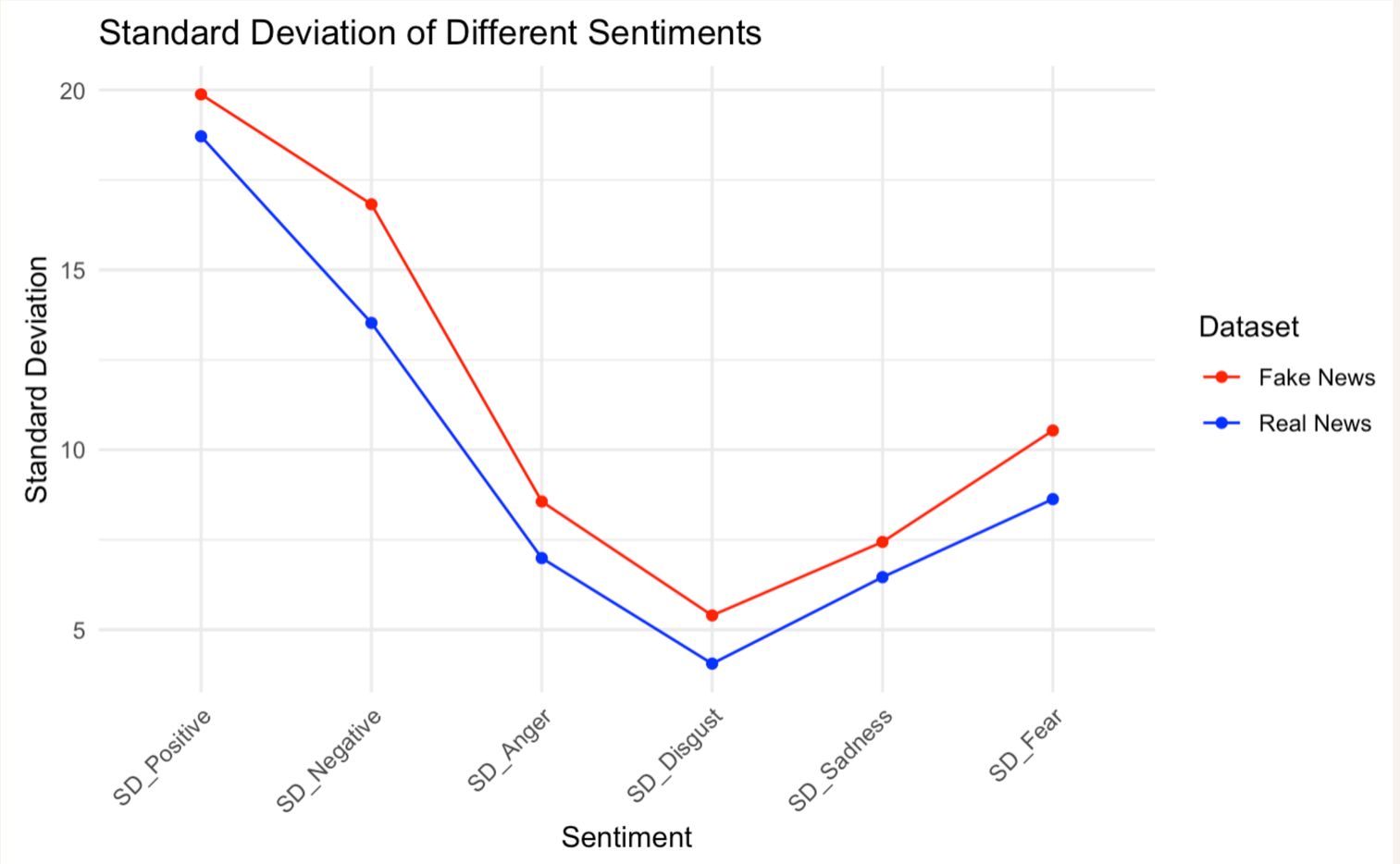
Sentiment Analysis

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive	sentiment
11	10	5	11	6	6	4	11	23	15	-12
4	11	4	7	9	4	8	13	12	27	22
1	2	1	2	0	1	0	2	5	2	-5
7	8	4	8	7	8	2	14	19	19	-4
7	16	3	9	7	4	4	22	11	31	6
3	3	2	3	1	2	1	2	2	3	2
0	0	0	0	0	0	1	0	0	1	0
10	6	6	11	4	8	5	10	15	18	-1
1	4	0	2	1	0	2	5	0	6	9
8	9	9	8	8	4	8	7	17	15	-2
13	8	6	13	5	10	4	14	23	19	-5
8	6	5	9	3	4	3	10	11	14	2
4	10	4	6	7	4	4	15	8	20	12
5	1	1	8	1	2	0	4	7	7	-9
14	24	7	18	10	18	7	26	27	42	-9
8	8	5	11	5	7	5	23	18	26	14
5	7	3	6	3	3	2	13	5	10	11

create sentiment table for text column

Summarize standard deviation of positive, negative, anger, disgust, sadness, and fear for real and fake datasets

Sentiment Analysis



Fake news tends to carry exaggerated sentiments to attract the reader

Conclusion

Through investigating textual-based features of news articles.

There's indeed specific sentiment patterns or linguistic cues that are more prevalent in fake news articles compared to genuine ones.

- Longer title length
- More ? ! in text
- Capitalised Words in Title
- Exaggerated sentiment
- Overall more repeated words

Thank you!

Misleading!!!