I-590 Project Proposal

Sentiment Analysis on Tweets about WhatsApp and Telegram

Created by: Fauzan Isnaini

INDIANA UNIVERSITY BLOOMINGTON

SECTION 1

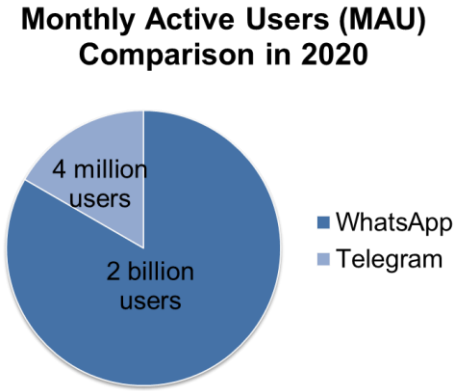
Overview

WhatsApp vs Telegram

1. WhatsApp has a clear advantage due to its head start. It boasts more than 2 billion MAU in 2020 , compared to Telegram's 4 million MAU (Bucher, 2020).

2. In January 2021, WhatsApp introduced a new privacy policy that would allow Facebook to aggregate all of its users' data across all of its services. The change of privacy policy in WhatsApp drove more than 100 million of its users to migrate to Telegram in January 2021 (Baloch, 2021).

3. Telegram gives some additional features to seize the market, such as a bigger file transfer limit, larger group size, integration with bots, multiple devices login, and flexibility in storing messages and files (Abbas, 2021).



Goals of the Study

1. Analyze the topics about WhatsApp and Telegram in social media and the real-world connection behind them.

2. Analyze the popularity of WhatsApp and Telegram in social media and how important are technological features behind their popularity.

3. Analyze the impact of the change in WhatsApp’s privacy policy to the popularity of these two applications.

Research Questions

1. What topics dominated the tweets about WhatsApp and Telegram?

- We used wordcloud and LDA to answer this.

2. How did Twitter reflect the sentiment for each instant messaging service?

- We used sentiment analysis to answer this.

3. Did the change in WhatsApp's privacy policy have an impact on WhatsApp and Telegram's sentiment polarity?

- We used T-test and Cohen's d effect size to answer this.

A priori Hypotheses

1. Telegram overall has a higher level of sentiment polarity than WhatsApp in January 2021.
2. Telegram has a higher level of sentiment polarity than WhatsApp in terms of privacy.
3. Telegram has a higher level of sentiment polarity than WhatsApp in terms of features.
4. WhatsApp has a higher level of sentiment polarity in December 2020 than January 2021.
5. Telegram has a higher level of sentiment polarity in January 2021 than December 2020.

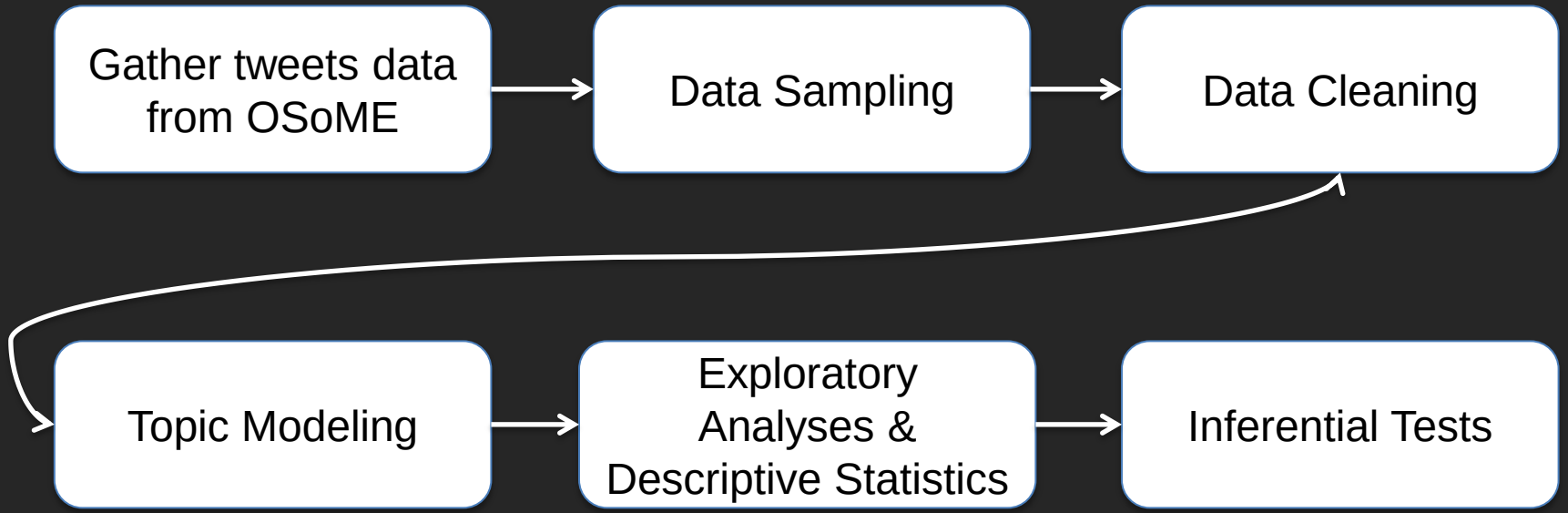
Null Hypotheses

1. Telegram overall has the same level of sentiment polarity as WhatsApp in January 2021.
2. Telegram has the same level of sentiment polarity as WhatsApp in terms of privacy.
3. Telegram has the same level of sentiment polarity as WhatsApp in terms of features.
4. WhatsApp has the same level of sentiment polarity in December 2020 and January 2021.
5. Telegram has the same level of sentiment polarity in December 2020 and January 2021.

SECTION 2

Research Method

Workflow


Gather tweets data from OSoME
Data Sampling
Data Cleaning
Topic Modeling
Exploratory Analyses & Descriptive Statistics
Inferential Tests

Data Source

1. We collected Tweets data from December 2020 and January 2021. This time frame will allow us to analyze the effect of WhatsApp's change of privacy policy

2. Tweets data were collected from IU's OSoMe platform.

3. We used power analysis to determine the required sample size. With an effect size of 0.2, alpha of 0.05, power of 0.80, and an allocation ratio of 1, the required sample size per group are 394 tweets.

The Queries that were Used

1. “#whatsapp” → This query was used to get the overall sentiment polarity of WhatsApp and compare the sentiment between December 2020 and January 2021
2. “#telegram”. This query was used to get the overall sentiment polarity of Telegram and compare the sentiment between December 2020 and January 2021.
3. “#whatsapp”,“#privacy”. This query will be used to get the sentiment polarity of WhatsApp in terms of privacy in January 2021.
4. “#telegram”,“#privacy”. This query will be used to get the sentiment polarity of Telegram in terms of privacy in January 2021.
5. “#whatsapp”,“#feature”. This query will be used to get the sentiment polarity of WhatsApp’s features in January 2021.
6. “#telegram”,“#feature”. This query will be used to get the sentiment polarity of Telegram’s features in January 2021.

Programming Tools

1. Although the tweets datasets have json object “language”, most of the values are “NA”. So we used automatic language detection in R with textcat package.

```
> my_data %>% filter(is.na(lang)) %>% dim_desc()
[1] "[72,811 x 24]"
> my_data %>% filter(lang=="en") %>% dim_desc()
[1] "[361 x 24]"
```

1. We used R for language detection, data sampling, generating wordcloud, exploratory analyses, descriptive statistics, and inferential tests.

2. We used Python for data cleaning and LDA topic modeling

Exploring the Topics

1. One of our research questions was “What topics dominated the tweets about WhatsApp and Telegram?”

2. We used wordcloud to get the big idea of the types of words that occur frequently in tweets about WhatsApp and Telegram. One thing should be remembered: wordcloud uses a random generator to create the visualization, so we need a more advanced method to validate the takeaways.

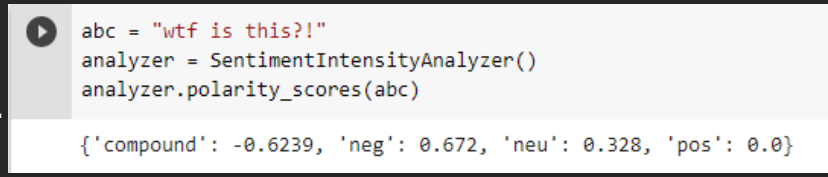
3. Latent Dirichlet Allocation (LDA) is more advanced than wordcloud. It does not only count the number of occurrence of each word. Instead, it analyzed the types of words that frequently occured together and categorize them as unique topics. This way we can explore the topics that were discussed in social media about WhatsApp and Telegram and understand the connection between their sentiment polarities and the real-world situation.

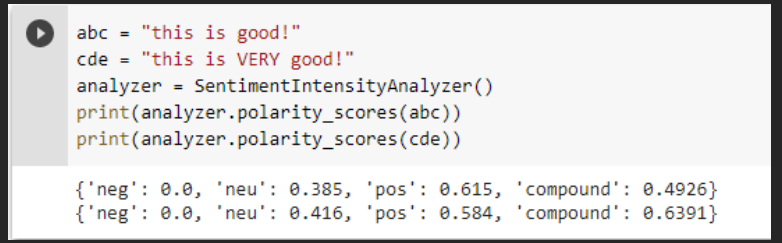
Sentiment Analysis

1. We used VADER for our sentiment analysis, which is specifically attuned for social media expressions.

2. It uses lexicon and rule-based sentiment analysis, which means it refers to a specifically constructed dictionary to calculate the sentiment polarity score. The dictionary also contains slang expressions, which are common in the social media.

3. It also incorporated several grammatical rules. For example: “This computer is good” will have a lower sentiment polarity score than “This computer is very good.”





VADER Score Calculation

1. VADER will first scan the text and match the words with the dictionary .
2. It then modifies the intensity and polarity according to the heuristic rules
3. The polarity scores for each word then are summed, and normalized between -1 for the most negative sentiment, and 1 for the most positive sentiment using function $\frac{x}{\sqrt{x^2+\alpha}}$; where x is the pre-normalized polarity score, and α is set to 15.

SECTION 3

Topic Modeling

Collected Data

Our queries originally used words-based approach (e.g. whatsapp instead of #whatsapp), but it returned some product advertisement tweets (whatsapp is used as the seller's contact method) which are irrelevant to our research goal. So we changed our approach and use hashtags for our queries:

1. #telegram,#feature
2. #telegram,#privacy
3. #telegram
4. #whatsapp,#feature
5. #whatsapp,#feature
6. #whatsapp,#feature

These tweets were collected from December 1st, 2020 00:01AM to January 31st, 2021 11:59PM. The smallest file size is 23MB, and the largest file size is 417MB

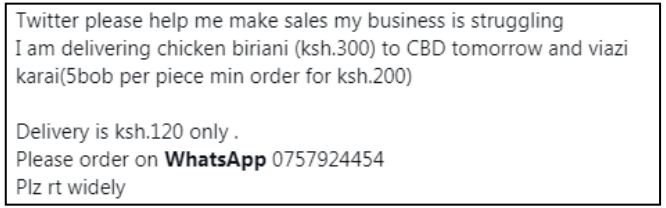


Image: Example of words-based approach

Image: Example of hashtag-based approach

Whatsapp in 2020 against 2021:

- Notice the occurence of words privacy in 2021. On the other hand, the word “feature” occur in 2020 data
- The words are tokenized, so “messenger” becomes “messeng”, and “privacy” becomes “privaci”

Image: wordcloud of WhatsApp in 2020

Image: wordcloud of WhatsApp in 2021

Telegram in 2020 against 2021:

- The Telegram's wordcloud in 2020 is dominated by coronavirus-related words. This is due to its bot integration ability which is widely used for tracking corona-virus information
- Notice that words like whatsapp and privacy started to appear in 2021

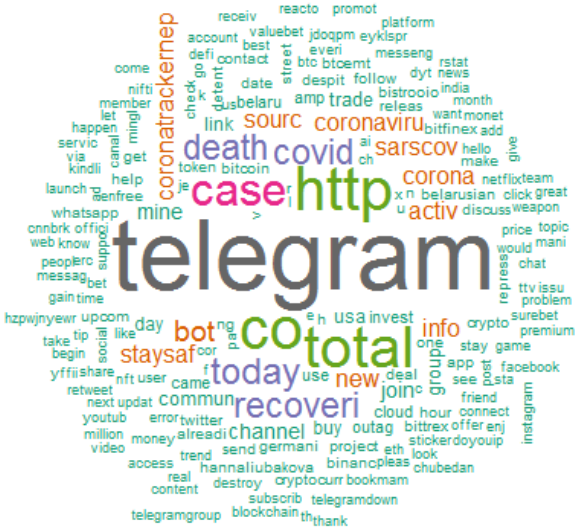


Image: wordcloud of Telegram in 2020

Image: wordcloud of Telegram in 2021

#telegram #feature against #whatsapp #feature in 2021:

- Notice that privacy-related words frequently occured on both wordclouds although the search tag was #feature.
- There are some feature-related words, such as “watusi” in WhatsApp and “bot” in telegram, but we cannot see any of them is dominant in the wordcloud

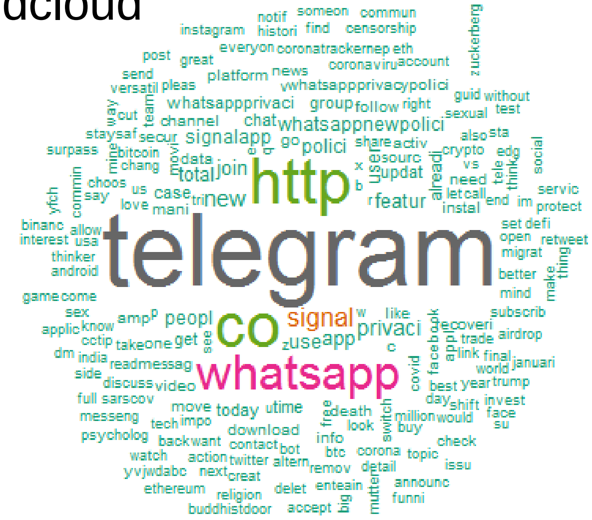


Image: wordcloud of #telegram #feature in 2021

Image: wordcloud of #whatsapp #feature in 2021

#telegram #privacy against #whatsapp #privacy:

- As expected, many of the words are related to the new WhatsApp's privacy policy
- Another interesting word in the WhatsApp's wordcloud is "signalapp", which also gain popularity because of the privacy policy issue.

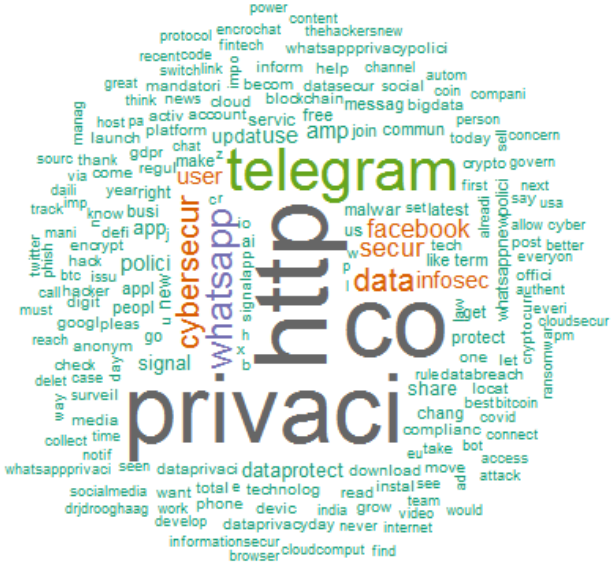


Image: wordcloud of #telegram #privacy in 2021

Image: wordcloud of #whatsapp #privacy in 2021

LDA of Telegram in 2020

- The number on LDA graph shows the occurence of each word in that specific topic, NOT on overall tweets.
- The 1st topic revolves around coronavirus tracking bot, which uses telegram as its technology
- It is important to notice that in the sentiment analysis, these negative words will downgrade the telegram's score although its usage shows the usefulness of this application
- Another topic discuss about trading, cryptocurrency, and investment. Telegram is widely popular for these kinds of forum because of its feature to host a large forum up to 200,000 members

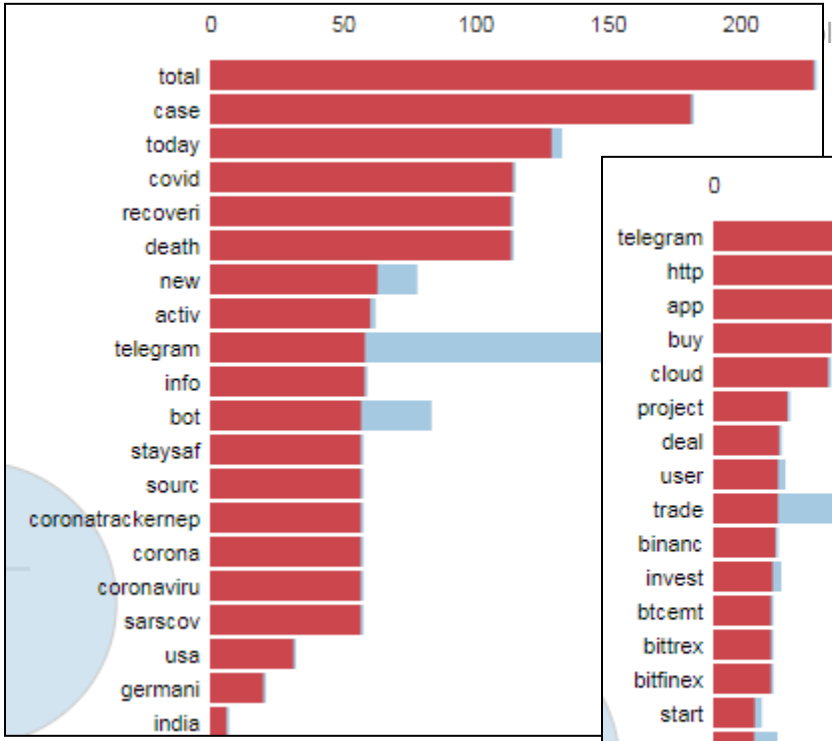


Image: LDA of telegram in 2020

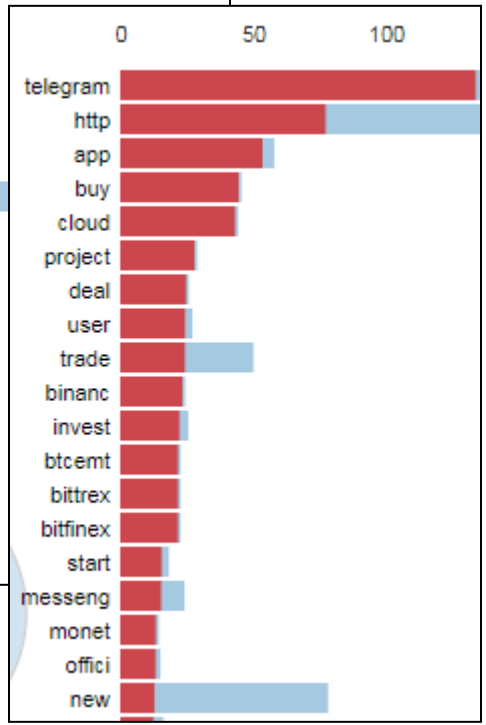


Image: The 2nd topic in LDA of telegram in 2020

LDA of Telegram in 2021

- The first topic revolves around the new WhatsApp's privacy policy, with a specific phrase "whatsappprivacypolici" was mentioned.
- LDA automatically discovered the second topics, but it turns out that the topic is still discussing about the privacy policy. It shows that this topic is very dominant in 2021.

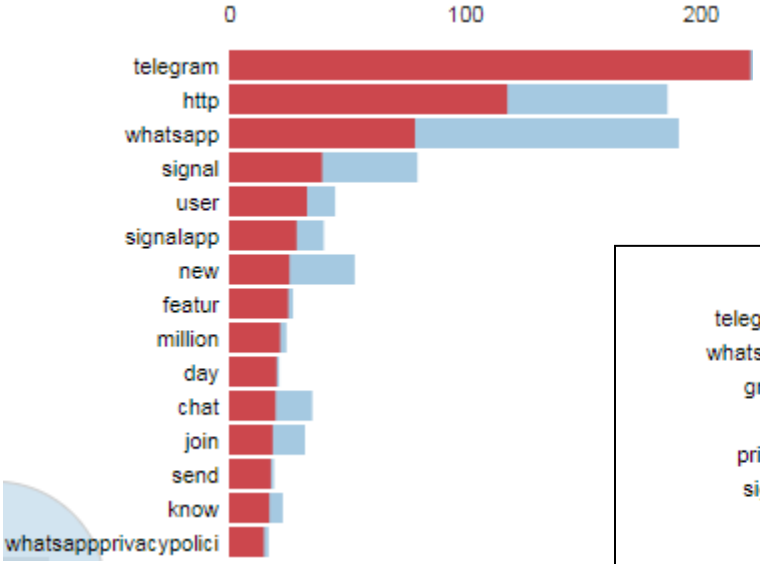


Image: LDA of telegram in 2021

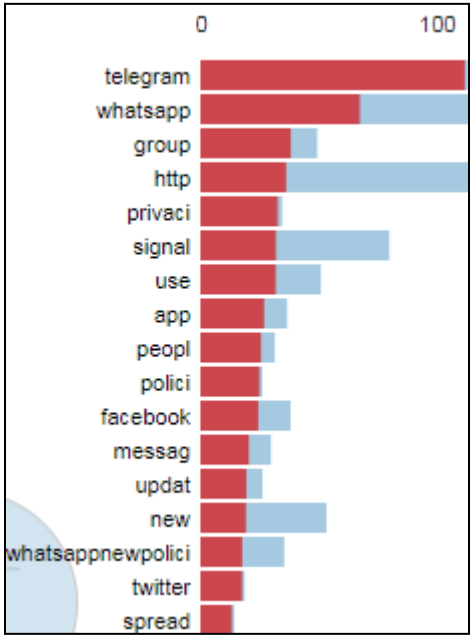


Image: The 2nd topic in LDA of telegram in 2021

LDA of WhatsApp in 2020

- The 1st topic revolves around Fouad Raheb's watusi application, which is a third party application to enhance WhatsApp's feature.
- The second topic is about different social medias, such as "snapchat" "socialmedia", etc.

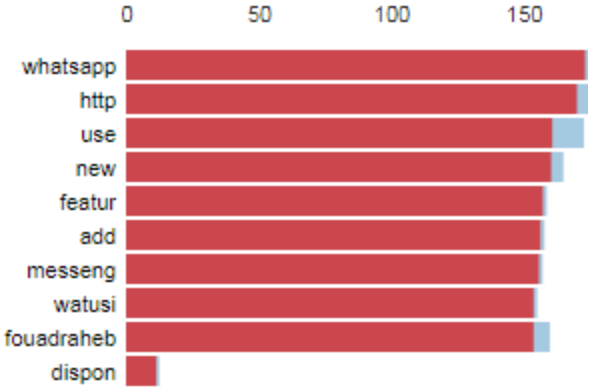


Image: LDA of WhatsApp in 2020

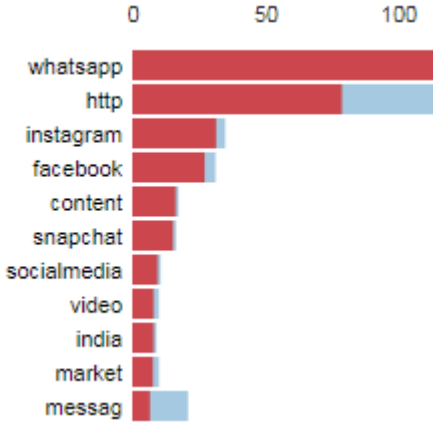


Image: The 2nd topic in LDA of WhatsApp in 2020

LDA of WhatsApp in 2021

- The first topic revolves around the change of privacy policy. Negative words started to appear, such as “complain” and “switch”
- The second topic still discuss about features, such as “watusi” and “featur”

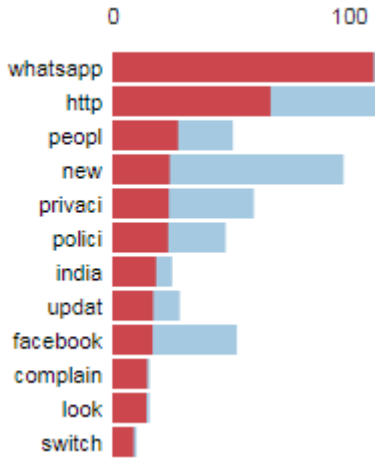


Image: LDA of WhatsApp in 2021

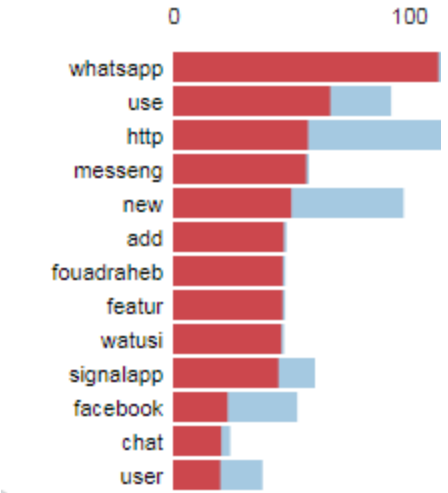


Image: The 2nd topic in LDA of WhatsApp in 2021

LDA of #whatsapp #feature in 2021

- Just as implied in the #whatsapp LDA before, privacy issue has become a dominant issue, even when #feature hashtag has been added

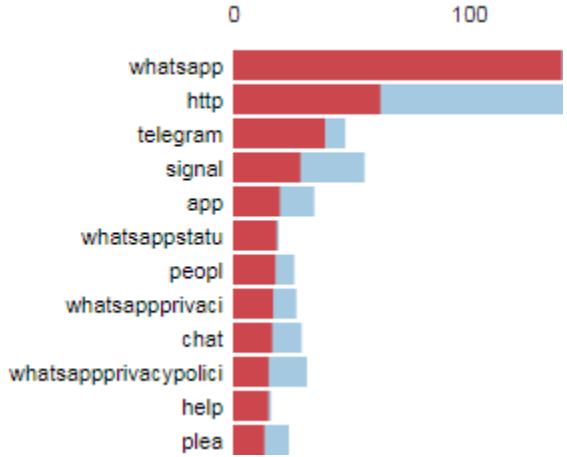


Image: LDA of #whatsapp #feature in 2021

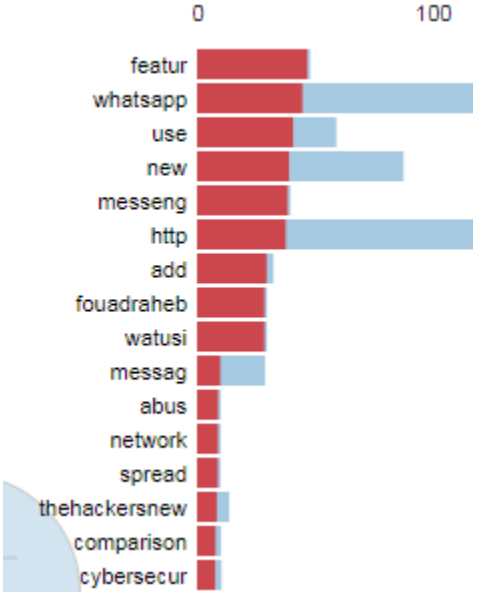


Image: The 2nd topic in LDA of #whatsapp #feature in 2021

LDA of #telegram #feature in 2021

- Even with #telegram #feature hashtags, the privacy issue is still the dominant issue
- The second topic discuss some words unrelated to the privacy issue, such as “bitcoin” and “update”

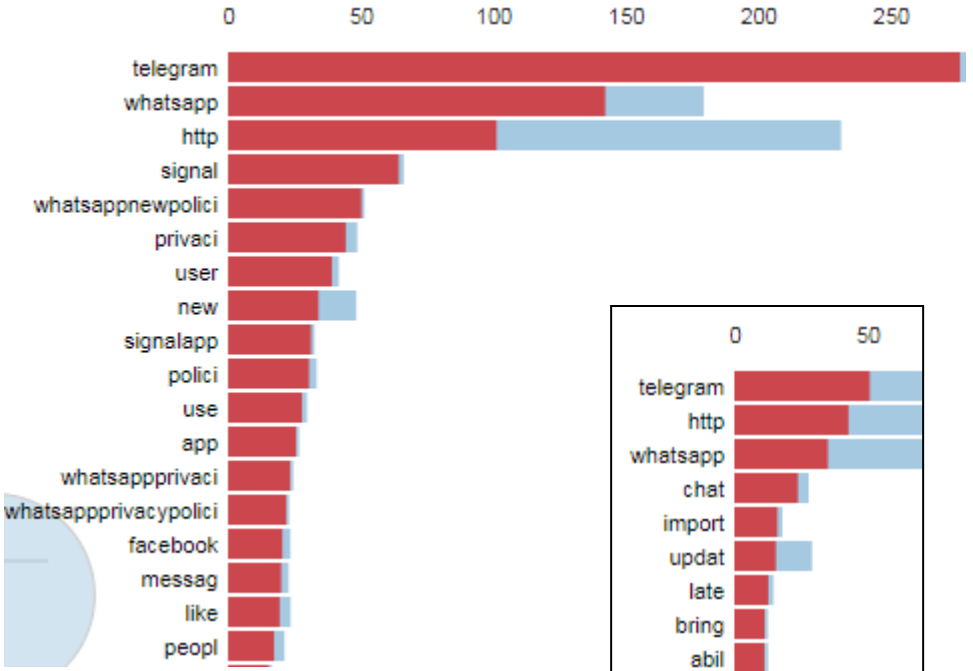


Image: LDA of #telegram #feature in 2021

Image: The 2nd topic in LDA of #telegram #feature in 2021

LDA of #whatsapp #privacy in 2021

- Unsurprisingly, the privacy policy issue become a dominant topic
- Another interesting topic is when Snowden revealed that he used Signal as his messaging application, which shows the occurence of words “snowden” and “signalapp” in the second topic

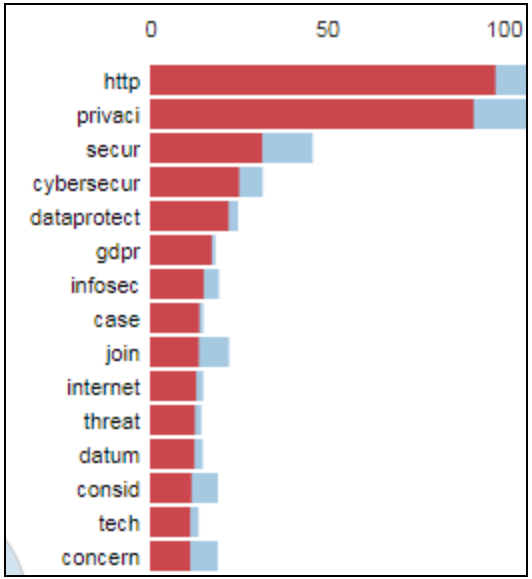


Image: LDA of #whatsapp #privacy in 2021

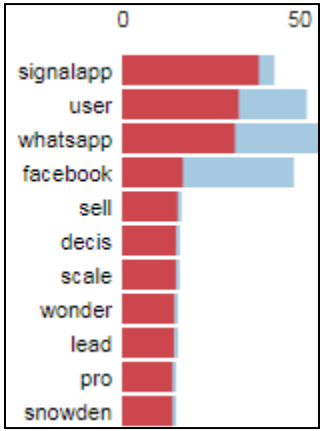


Image: 2nd topic of LDA of #whatsapp #privacy in 2021

LDA of #telegram #privacy in 2021

- The first topic clearly mentioned “whatsappnewpolicy”
- While the second topic discuss about “data privacy” and “security”

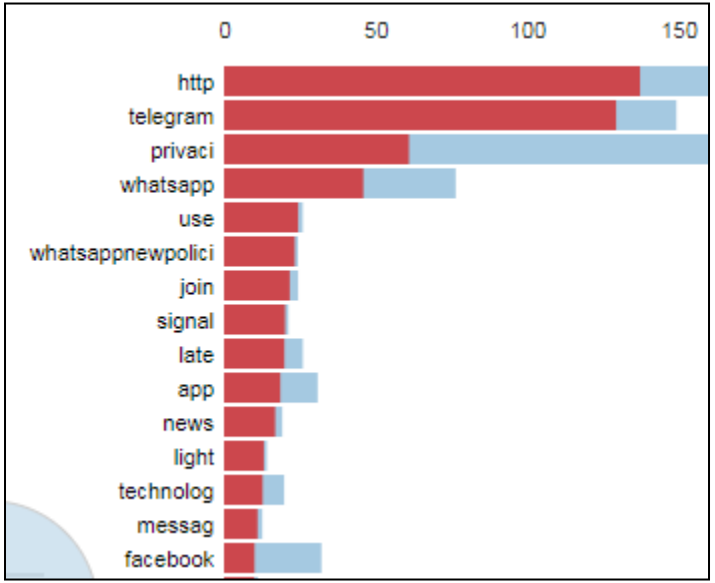


Image: LDA of #telegram #privacy in 2021

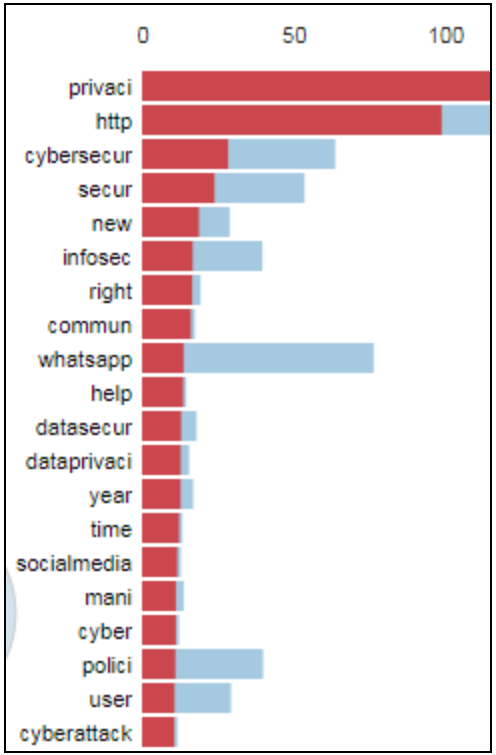


Image: 2nd topic of LDA of #telegram #privacy in 2021

SECTION 4

Exploratory Analysis and Descriptive Statistics

- WhatsApp has a higher mean of sentiment polarity in 2020 compared to 2021
- The histogram shows that VADER classified most of the tweets as neutral, although the negativity in 2021 is more apparent than 2020

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.06660381	0	-0.7798	0.9531	0.2750339

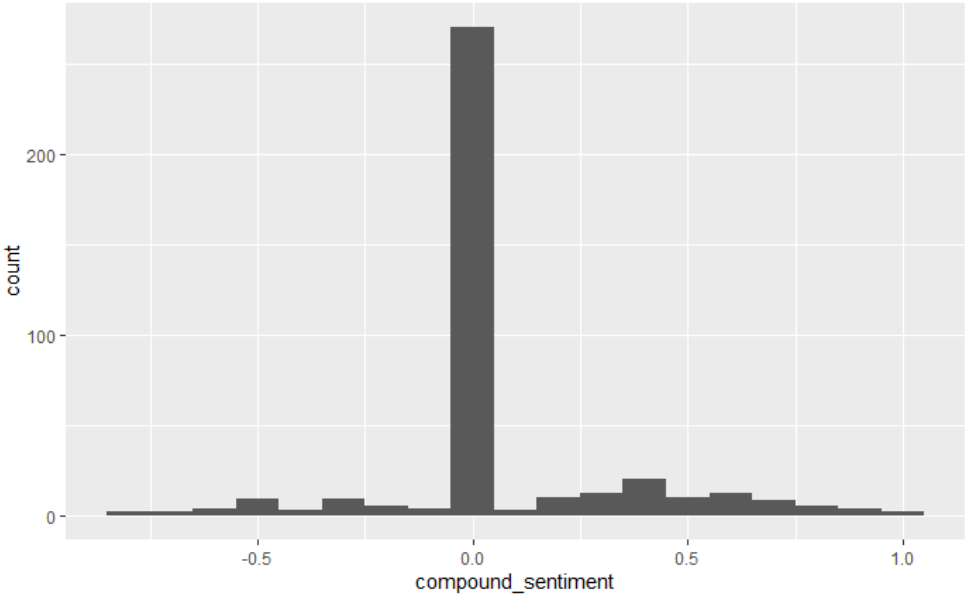


Image: Histogram and descriptive statistics of WhatsApp's sentiment in 2020

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.05931041	0	-0.8519	0.9485	0.3177321

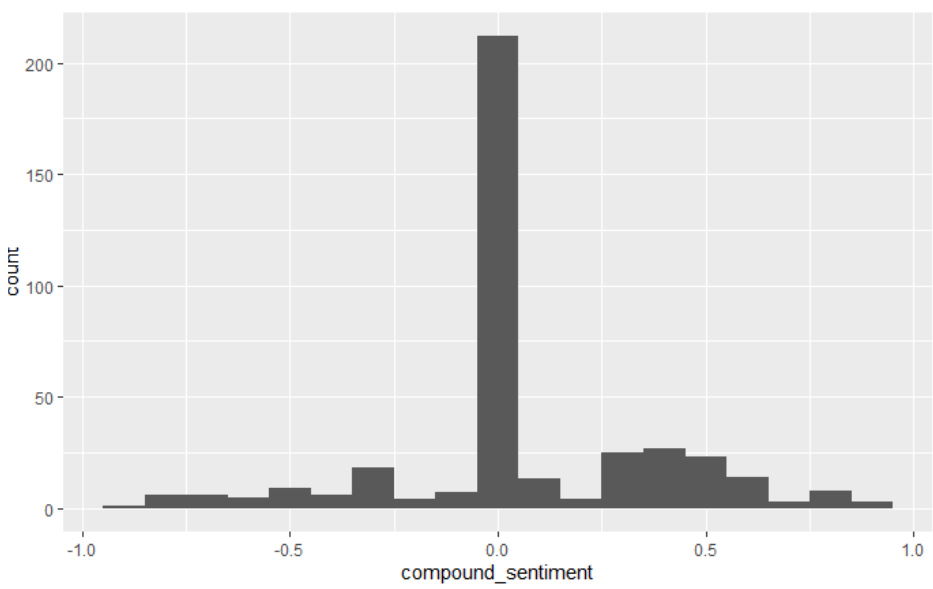


Image: Histogram and descriptive statistics of WhatsApp's sentiment in 2021

- Telegram has a higher mean of sentiment polarity in 2021 than 2020
- Surprisingly, the histogram shows that the negative sentiment is very apparent for Telegram in 2020

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.0101231	0	-0.8885	0.9538	0.5253267

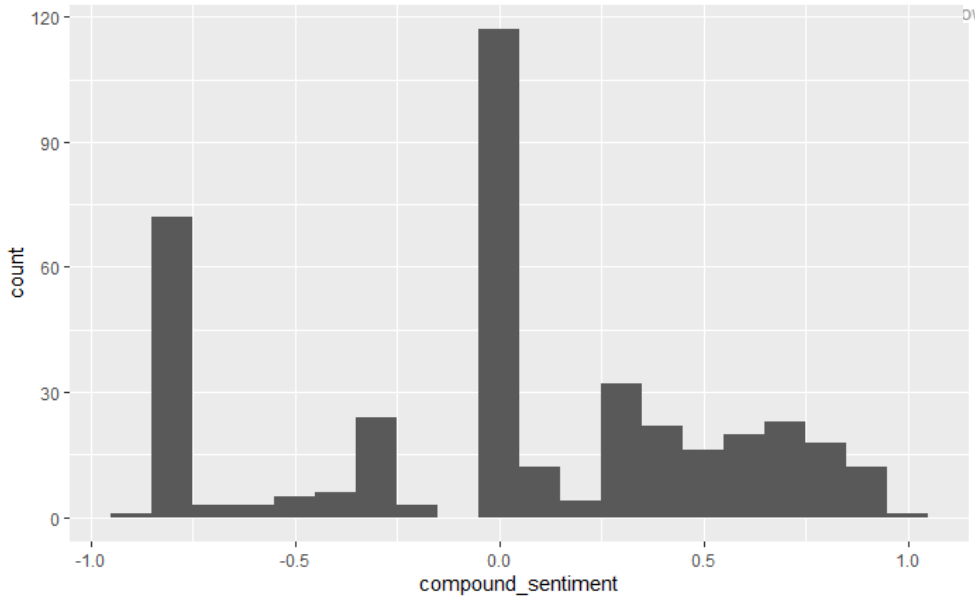


Image: Histogram and descriptive statistics of Telegram's sentiment in 2020

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.1607008	0	-0.8779	0.9552	0.3969986

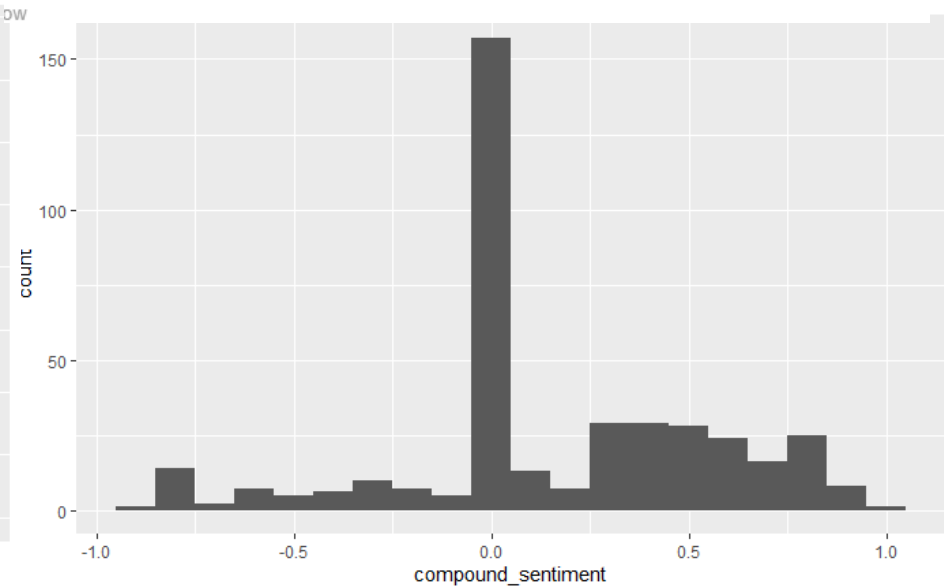


Image: Histogram and descriptive statistics of Telegram's sentiment in 2021

- Telegram has a higher mean of sentiment polarity in 2021 than WhatsApp
- The histogram shows that VADER classified most of the tweets as neutral, although the positivity in is more apparent in Telegram than in WhatsApp

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.05931041	0	-0.8519	0.9485	0.3177321

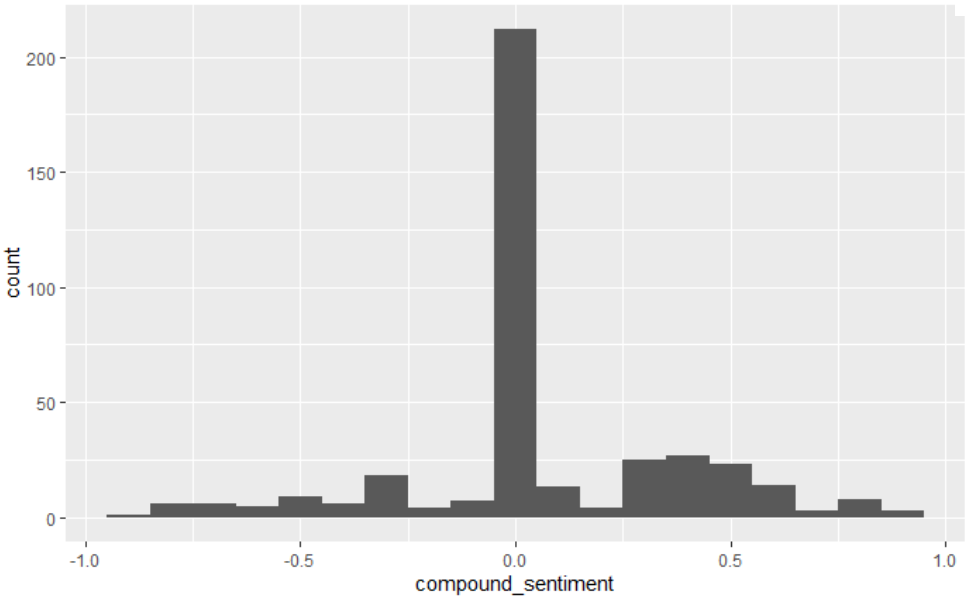


Image: Histogram and descriptive statistics of #whatsapp sentiment in 2021

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.1607008	0	-0.8779	0.9552	0.3969986

1 row

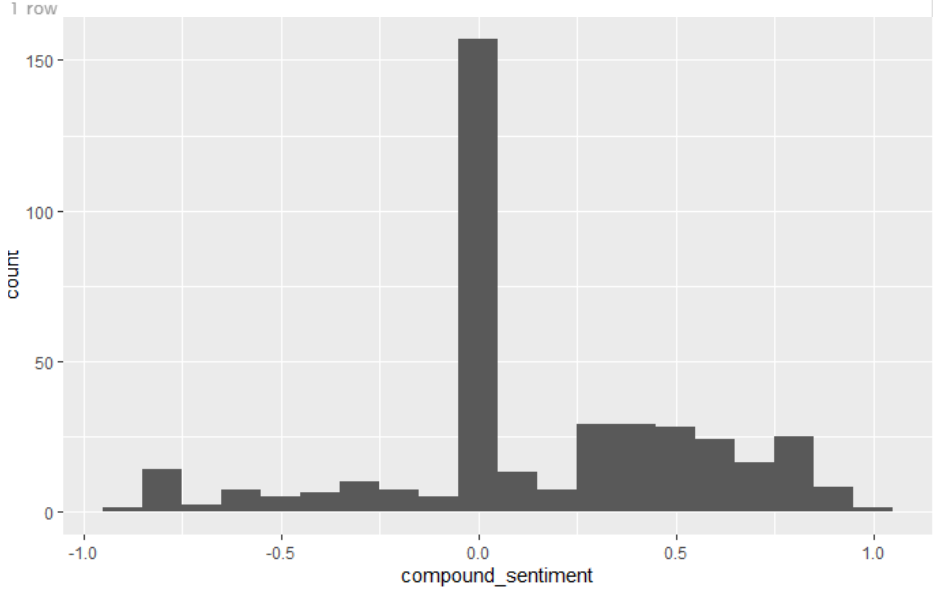


Image: Histogram and descriptive statistics of #telegram sentiment in 2021

- #telegram #privacy has a higher mean of sentiment polarity than #whatsApp #privacy
- The histogram shows that VADER classified most of the tweets as neutral, although the positivity in is more apparent in Telegram than in WhatsApp

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.08892893	0	-0.8481	0.926	0.3147789

1 row

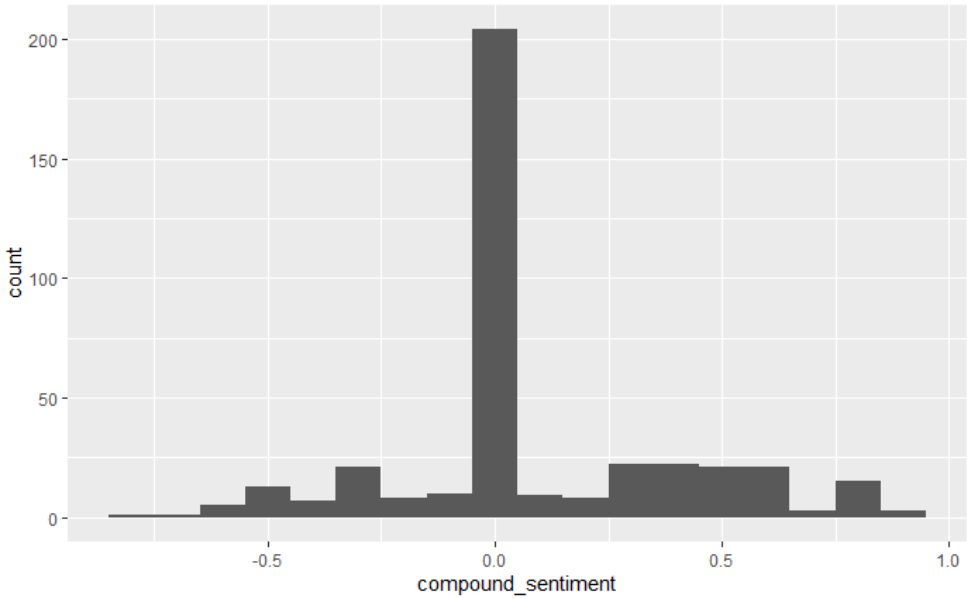


Image: Histogram and descriptive statistics of #whatsapp #privacy sentiment in 2021

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.1400749	0	-0.8625	0.926	0.3766441

1 row

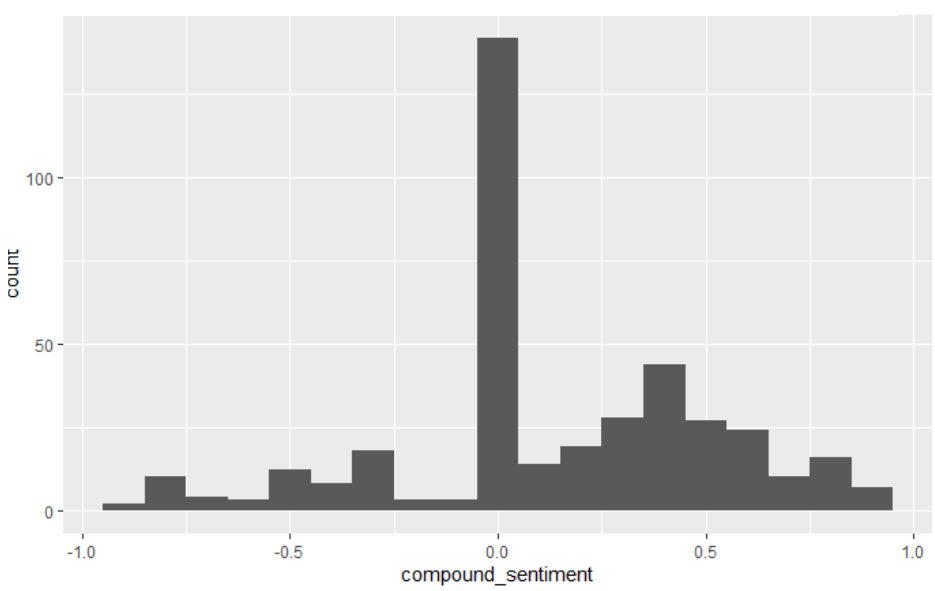


Image: Histogram and descriptive statistics of #telegram #privacy sentiment in 2021

- #telegram #feature has a higher mean of sentiment polarity than #whatsApp #feature
- The histogram shows that VADER classified most of the tweets as neutral

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.1017652	0	-0.8074	0.9274	0.3272712

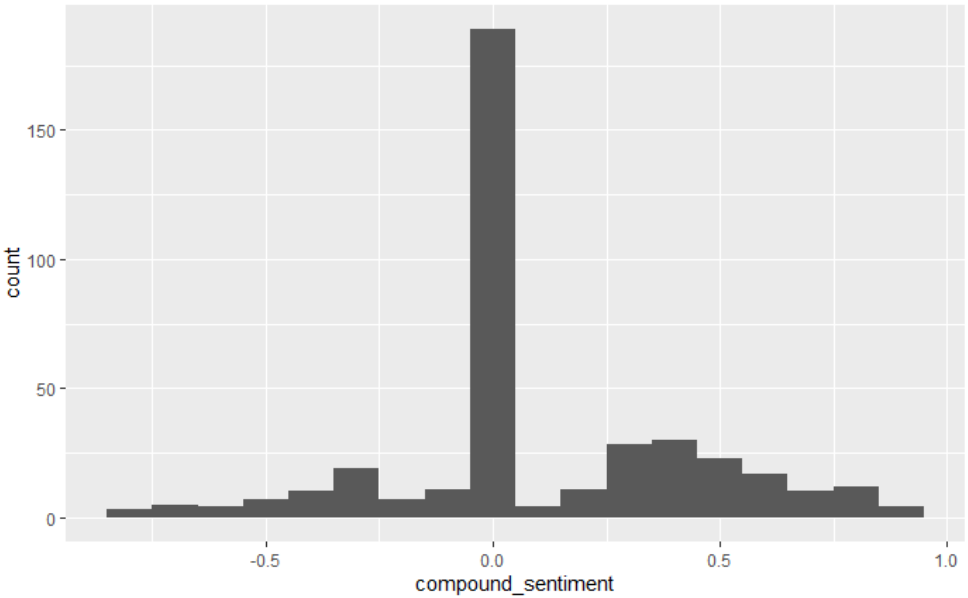


Image: Histogram and descriptive statistics of #whatsapp #feature sentiment in 2021

sentiment_mean <dbl>	sentiment_median <dbl>	sentiment_min <dbl>	sentiment_max <dbl>	sentiment_sd <dbl>
0.1333959	0	-0.9493	0.9349	0.3756577

1 row

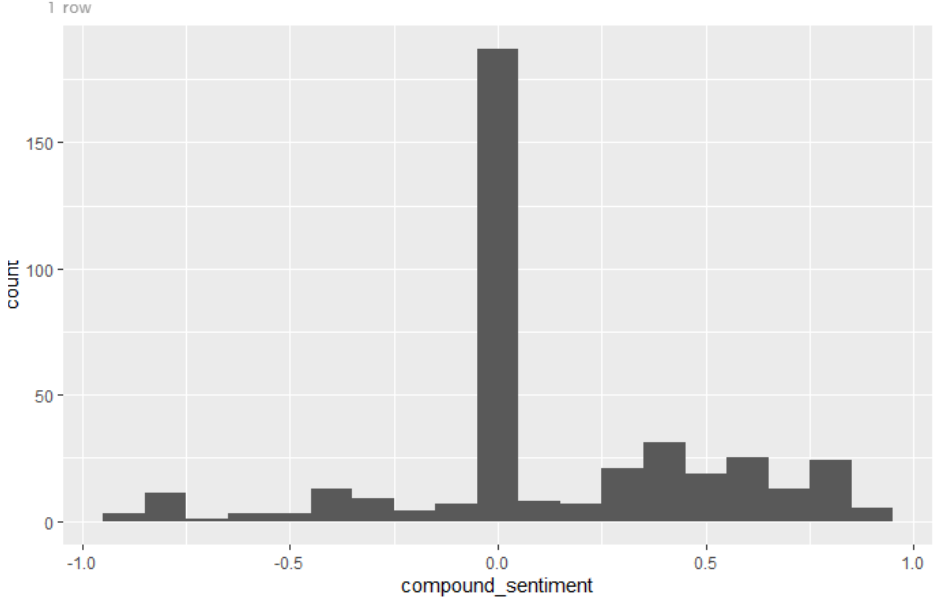


Image: Histogram and descriptive statistics of #telegram #feature sentiment in 2021

Comparison of Sentiment Polarities in 2020 vs 2021 for Both Apps

- While on the boxplot it seems that WhatsApp's sentiment is higher in 2021 than 2020, we know from the descriptive statistics that the mean is higher in 2020 than in 2021

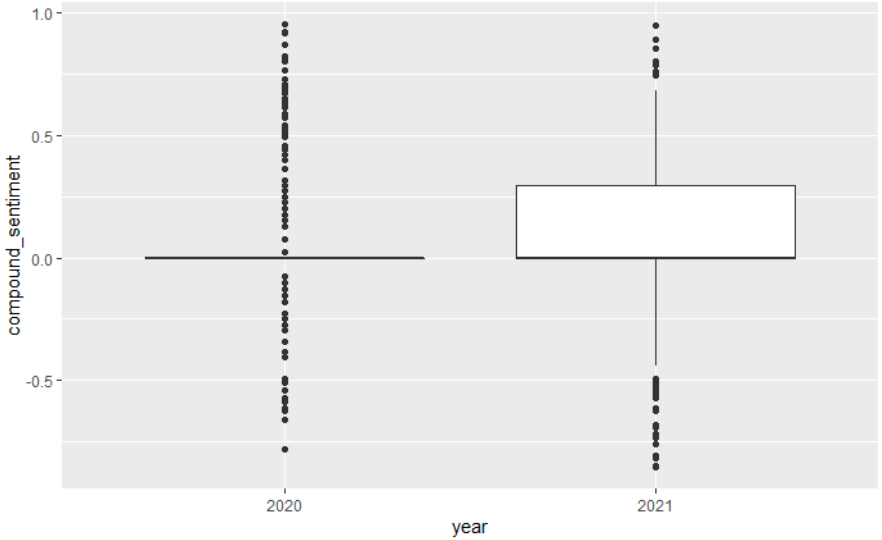


Image: Boxplot of #whatsapp sentiment in 2020 vs 2021

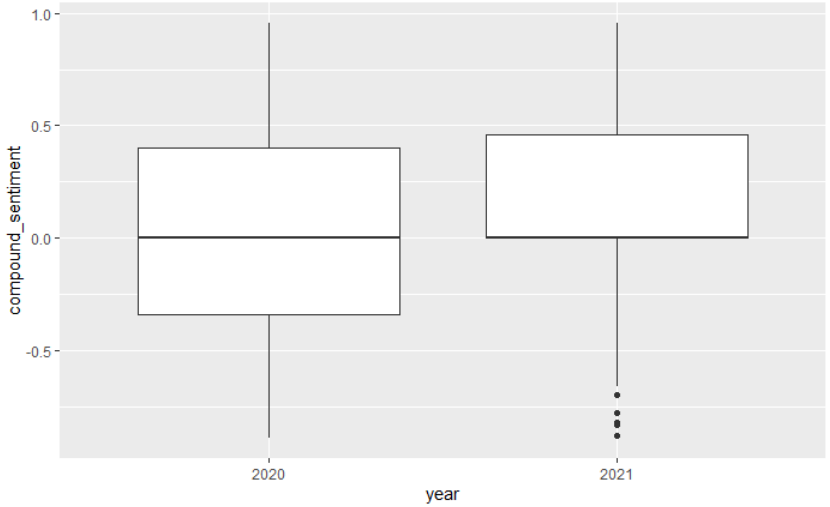


Image: Boxplot of #telegram sentiment in 2020 vs 2021

Comparison of Both Apps with #feature and #privacy tags

- It seems Telegram has the higher sentiment than WhatsApp overall without any additional hashtags

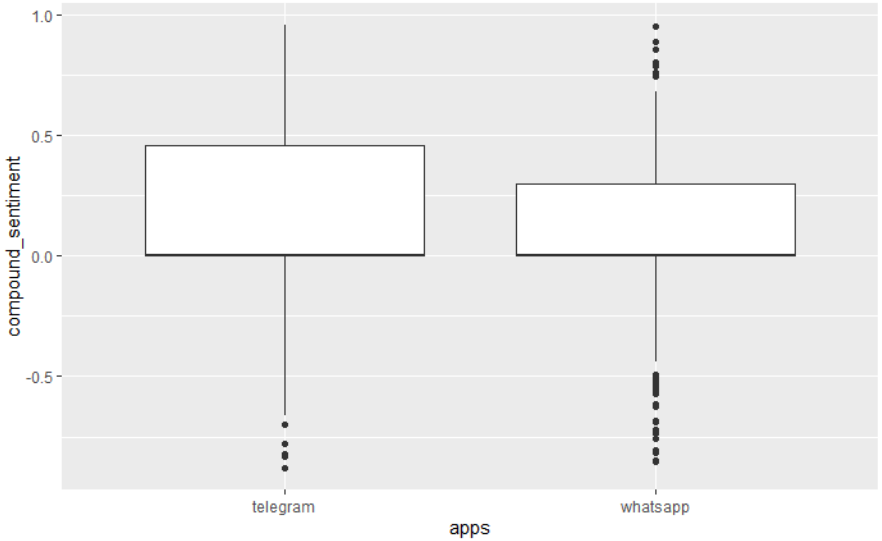


Image: Boxplot of #telegram vs #whatsapp in 2021

Comparison of #whatsapp and #telegram

- It seems Telegram has the higher sentiment than WhatsApp in both hashtags

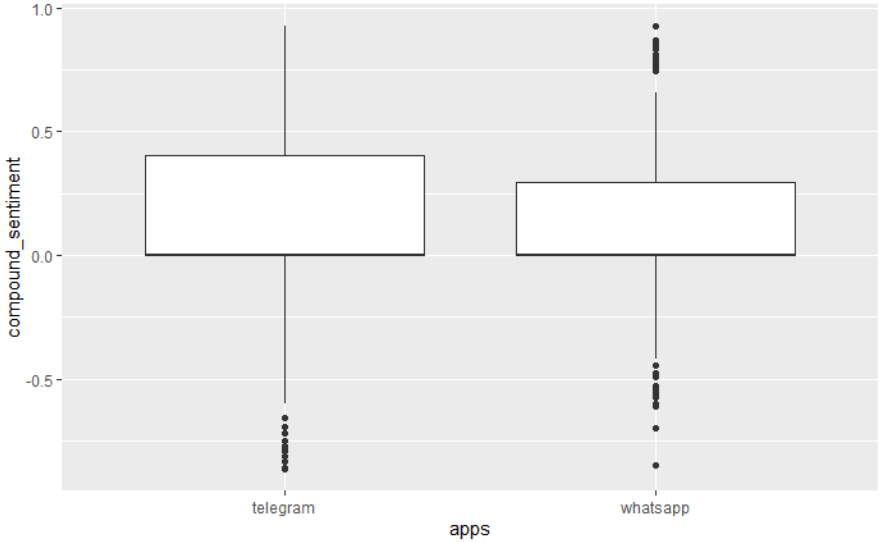


Image: Boxplot of #telegram #privacy vs #whatsapp #privacy in 2021

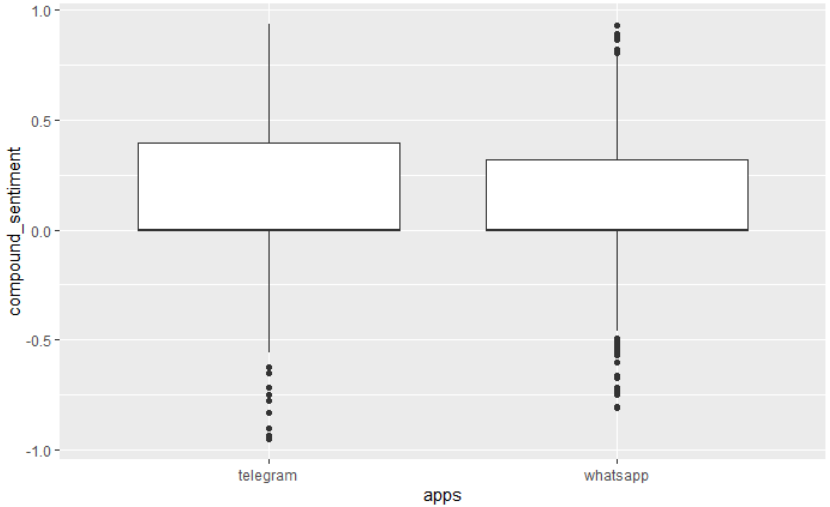


Image: Boxplot of #telegram #feature vs #whatsapp #feature in 2021

SECTION 5

Inferential Tests

Telegram in 2020 vs 2021

We found that Telegram in 2021 (M=0.16, SD=0.40) was on average have a significantly (t(786)=-4.54, p<0.05) higher sentiment polarity than in 2020 (M=0.01, SD=0.53).

The effect size according to Cohen's d is 0.32, which is small

```
        Two Sample t-test

data:  tele_2020 and tele_2021
t = -4.5392, df = 786, p-value = 6.532e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.21569579 -0.08545954
sample estimates:
mean of x mean of y
0.0101231 0.1607008
```

	.y. <chr>	group1 <chr>	group2 <chr>	effsize <dbl>	n1 <int>	n2 <int>	magnitude <ord>
1	compound_sentiment	2020	2021	-0.3234022	394	394	small

1 row

WhatsApp in 2020 vs 2021

We did NOT found that WhatsApp in 2021 (M=0.06,SD=0.32) has significantly different (t(786)=0.34, p>0.05) sentiment polarity from 2020 (M=0.07 SD=0.28).

Thus, we fail to reject the null hypothesis

The effect size according to Cohen's d is 0.02, which is negligible

```
        Two Sample t-test

data:  wa_2020 and wa_2021
t = 0.3445, df = 786, p-value = 0.7306
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03426524  0.04885204
sample estimates:
 mean of x  mean of y
0.06660381 0.05931041
```

	.y. <chr>	group1 <chr>	group2 <chr>	effsize <dbl>	n1 <int>	n2 <int>	magnitude <ord>
1	compound_sentiment	2020	2021	0.02454443	394	394	negligible

1 row

#whatsapp #feature vs #telegram #feature

We did NOT found that tweets with hashtags #whatsapp #feature (M=0.13 SD=0.38) has significantly different (t(786)=1.26, p>0.05) sentiment polarity from #telegram #feature (M=0.10 SD=0.33).

Thus, we fail to reject the null hypothesis

The effect size according to Cohen's d is 0.09, which is negligible

```
        Two Sample t-test

data:  whatsapp and telegram
t = -1.2602, df = 786, p-value = 0.208
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08090179  0.01764037
sample estimates:
mean of x mean of y
0.1017652 0.1333959
```

	.y. <chr>	group1 <chr>	group2 <chr>	effsize <dbl>	n1 <int>	n2 <int>	magnitude <ord>
1	compound_sentiment	telegram	whatsapp	0.08978444	394	394	negligible

1 row

#whatsapp #privacy vs #telegram #privacy

We found that tweets with hashtags #telegram #privacy in 2021 (M=0.14 SD=0.38) was on average have a significantly (t(786)=2.07, p<0.05) higher sentiment polarity than #whatsapp #privacy (M=0.09 SD=0.31).

The effect size according to Cohen's d is 0.15, which is negligible

```
        Two Sample t-test

data:  whatsapp and telegram
t = -2.0682, df = 786, p-value = 0.03894
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.099689274 -0.002602604
sample estimates:
 mean of x  mean of y
0.08892893 0.14007487
```

	.y. <chr>	group1 <chr>	group2 <chr>	effsize <dbl>	n1 <int>	n2 <int>	magnitude <ord>
1	compound_sentiment	telegram	whatsapp	0.1473553	394	394	negligible

1 row

#whatsapp vs #telegram

We found that tweets with hashtags #telegram in 2021 (M=0.16 SD=0.40) was on average have a significantly (t(786)=3.96, p<0.05) higher sentiment polarity than #whatsapp (M=0.06 SD=0.32).

The effect size according to Cohen's d is 0.28, which is small

```
        Two Sample t-test

data:  whatsapp and telegram
t = -3.9579, df = 786, p-value = 8.246e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.15167683 -0.05110388
sample estimates:
 mean of x  mean of y
0.05931041 0.16070076
```

	.y. <chr>	group1 <chr>	group2 <chr>	effsize <dbl>	n1 <int>	n2 <int>	magnitude <ord>
1	compound_sentiment	telegram	whatsapp	0.2819874	394	394	small

1 row

SECTION 6

Summary of Key Findings

Summary of Key Findings

1. Privacy has become the dominant topic in January 2021 for both WhatsApp and Telegram. From the topic models, we can see different topics in 2020, such as the covid-19 tracker application in Telegram and the watusi application in WhatsApp. But in 2021, the privacy topic becomes so dominant even when we add the hashtag #feature on our query.
2. Telegram enjoyed a statistically significant sentiment boost in January 2021 compared to December 2020.
3. Telegram overall has a higher sentiment polarity than WhatsApp in January 2021 due to WhatsApp's change in the privacy policy.
4. Other than Telegram, Signal also enjoyed a popularity boost due to WhatsApp's change in the privacy policy.

SECTION 7

Key Takeaways for Design

Design Implications

1. Privacy policy is a very important thing. Our study showed that the changes in WhatsApp's privacy policy significantly boosted the popularity of Telegram against WhatsApp. A well-designed privacy policy should make the users feel safe to entrust their data.

2. Our LDA model showed that some of the Telegram's features were widely adopted by communities, such as bot integration to track the Covid-19 statistics, which became a dominant topic in December 2020.

3. On the other hand, watusi application was a dominant topic for WhatsApp in 2020. It is an unofficial application for customizing WhatsApp's feature, which was available in a Cydia store for jailbroken devices. Unlike Telegram, WhatsApp is not very flexible in integration with a third party application, so some users resorted to use this unofficial application.

SECTION 7

Recommendations

Recommendations

The topic about privacy was very dominant in the tweets in January 2021, so we did not have enough information to analyze the impact of the features in each application to their sentiment polarity. Further study can be conducted in this topic by tweaking the search query.

Others

This project is not related to any other class’ assignment or capstone project.

INDIANA UNIVERSITY BLOOMINGTON

References

Abbas, S. (2021, January 30). 11 reasons why you should use Telegram instead of Whatsapp. Retrieved March 22, 2021, from https://sirajea.medium.com/11-reasons-why-you-should-use-telegram-instead-of-whatsapp-ab0f80fbfa79#:~:text=There%20is%20no%20significant%20difference.&text=Users%20can%20send%20any%20kind,a%20good%20number%20of%20users.&text=Users%20on%20telegram%20can%20log,receive%20messages%20on%20all%20devices.

Baloch, H., Artashyan, A., Abdullah, Nick, Lancaster, M., Singh, S., . . . Blogger, A. (2021, February 06). Telegram became the most popular app in the world. Retrieved March 21, 2021, from https://www.gizchina.com/2021/02/06/telegram-became-the-most-popular-app-in-the-world/

Brownlee, J. (2020, April 23). A gentle introduction to statistical power and power analysis in python. Retrieved March 22, 2021, from https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/

Bucher, B. (2020, December 02). Messaging app usage statistics around the world. Retrieved March 21, 2021, from https://www.messengerpeople.com/global-messenger-usage-statistics/

Kaleru, S., & Dhanikonda, S. R. (2018). Exploratory Data Analysis and Latent Dirichlet Allocation on Yelp Database. *International Journal of Applied Engineering Research, 13*.