MBA (BA-304) Text Mining, Social Media & Web Analytics

End-Semester Exam Report

on

Lexicon-based sentiment analysis after Latent Dirichlet Allocation (LDA) from Twitter data

Submitted By:

JAMMALAMADAKA KRISHNA RAVALI (19MBMB12)

(2019 – 2021)

Under the esteemed guidance of
Dr. Varsha Mamidi

INDEX

LIST OF FIGURES i

ABSTRACT ii

1. INTRODUCTION **1**

2. SENTIMENT ANALYSIS **2**

2.1. LEXICONS 3

3. TOPIC MODELING **3**

4. DATA **5**

4.1. ATTRIBUTES 5

4.2. DESCRIPTION 5

4.3. PACKAGES USED IN R 5

4.4. ABOUT HASHTAG 5

5. METHODOLOGY **6**

6. RESULTS AND OBSERVATIONS **8**

7. CONCLUSION **17**

8. FUTURE SCOPE **17**

REFERENCES 17

APPENDIX 17

LIST OF FIGURES

Figure 1. (a) Social media profile percentage in 2008 vs 2020 in USA and **(b)** Number of social media users worldwide from 2017 projected till 2025. 1
Figure 2. Most used social platform as of July 2020 1
Figure 3. Sentiment Analysis Architecture 2
Figure 4. Machine learning-based vs lexicon-based sentiment analysis 3
Figure 5. A flowchart of a text analysis that incorporates topic modelling 4
Figure 6. Topic modelling applications 4
Figure 7: Snippet view of code from twitter sentiment analysis on '#PowerWomen', '#forbesPowerWomen' (as a part of previous assignment) 5
Figure 8. Top 15 most probable terms in two topic model 8
Figure 9. Top 15 most probable terms in 6 topic model 9
Figure 10. Top 15 most probable terms in 10 topic model 9
Figure 11. Highest probability of intial hashtag powerwomen in two topic model 10
Figure 12. Topics with overall highest probability of in two topic model 10
Figure 13. Highest probability of intial hashtags in six topic model 10
Figure 14. Topics with overall highest probability of in six topic model 10
Figure 15. Highest probability of intial hashtags in ten topic model 11
Figure 16. Topics with overall highest probability of in ten topic model 11
Figure 17. Twitter data collected for the new topic 'woman' 11
Figure 18. Visualizing most frequent words under emotion joy and positive (left-woman hashtag, right- powerwoman, forbespowerwoman hashtag) 12
Figure 19. Visualizing most frequent emotion in the data (left- woman hashtag; right- powerwoman, forbespowerwoman hashtag) 12
Figure 20. Date wise sentiment distribution using bing lexicon (top- woman hashtag; bottom- powerwoman, forbespowerwoman hashtag) 13
Figure 21. Lexicon wise sentiment distribution (top- woman hashtag; bottom- powerwoman, forbespowerwoman hashtag) 14
Figure 22. Plot showing most common positive and negative words in the dataset using bing lexicon (top- woman hashtag; bottom- powerwoman, forbespowerwoman hashtag) 15
Figure 23. Word cloud on twitter data after removing stop words ((a)- woman hashtag; (b)- powerwoman, forbespowerwoman hashtag) 16

Abstract:

In this study, topic modelling is performed using Linear Discriminant Analysis is performed on the results of the previous study of sentiment analysis on twitter data fetched based on the trending hastags '#PowerWomen', '#forbesPowerWomen' which is related to the topic of 2020 Powerful Women list released by Forbes with three different factors (k=2,6,10) as test cases and their results are compared in regard to topics having highest probability of occurrence in all the respective factors with the initial terms of study and the topic having highest probability is considered as final term of study to proceed with fetching corresponding twitter data and sentiment analysis is performed.

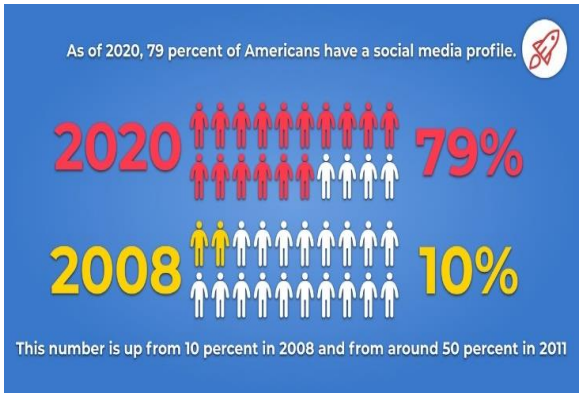
1. Introduction:

In the world of increasing digitization, every person is leaving out some form of digital footprint, which is a trail of data left while using internet on social media platforms. There exist two different types of digital footprint:

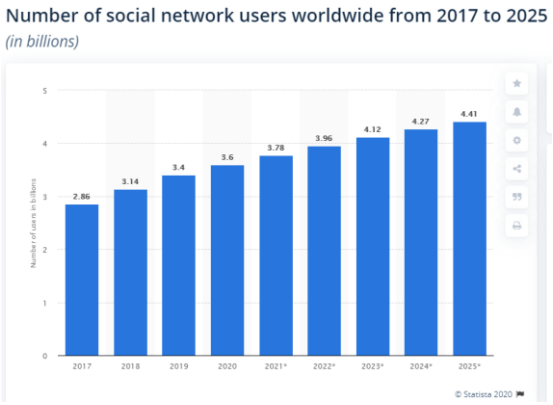
Passive digital footprints may track the user IP address, when it was created, and where they came from; with the footprint later being analyzed. In an offline environment, it may be stored in files, which can be accessed by administrators to view the actions performed on the machine, without seeing who performed them.

Active digital footprints can be stored by a user being logged into a site when making a post or change, with the registered name being connected to the edit. In an offline environment a footprint may be stored in files, when the owner of the computer uses a keylogger, so logs can show the actions performed on the machine, and who performed them [1].

The number of active users of social media platform is increasing with time (Fig. 1).



(a)



(b)

Figure 1. (a) Social media profile percentage in 2008 vs 2020 in USA (source: https://www.broadbandsearch.net/blog/social-media-facts-statistics) and **(b)** Number of social media users worldwide from 2017 projected till 2025.

There are many social media platforms that are available to the users such as Facebook, Twitter, YouTube, Instagram, Snapchat, Whatsapp etc.

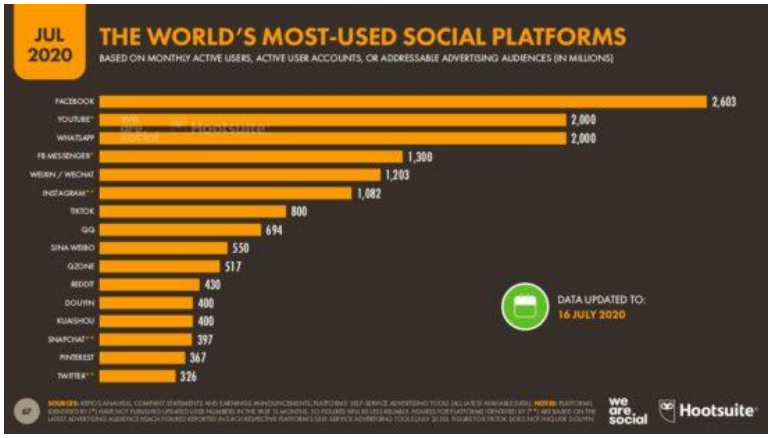


Figure 2. Most used social platform as of July 2020 (source: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/)

Marketers are also using social media as platform to improve their businesses (particularly for B2B, B2C modes, as in fig. 2) by engaging with customers as a part of post-sales support, to collect feedback, to know customer sentiment etc. using different analytical techniques. This process of collecting, analyzing information from social networks to track the online conversations of people about their product or services is called social media analytics. It is particularly used to get in-depth idea about social customers. In order to fetch this information, the stack of tools and techniques used is called social media intelligence [2]. Even though social media applications like Facebook, Twitter, Instagram provides users with their usage statistics and analytics information, it is of limited scope for analysis. There are several trending tools available in the market to support extensive analytic requirements such as:

Google analytics, Socialbakers, Hootsuite, Sprout Social, Awario, Talkwalker etc. [3].

2. Sentiment Analysis: It is also known as opinion mining or emotion AI. It uses Natural Language Processing and text mining, linguistics to identify, extract, quantify and to study the information related to a particular subject area of interest [4]. It classifies the opinions as positive, negative or neutral. The architecture is depicted in fig. 3.

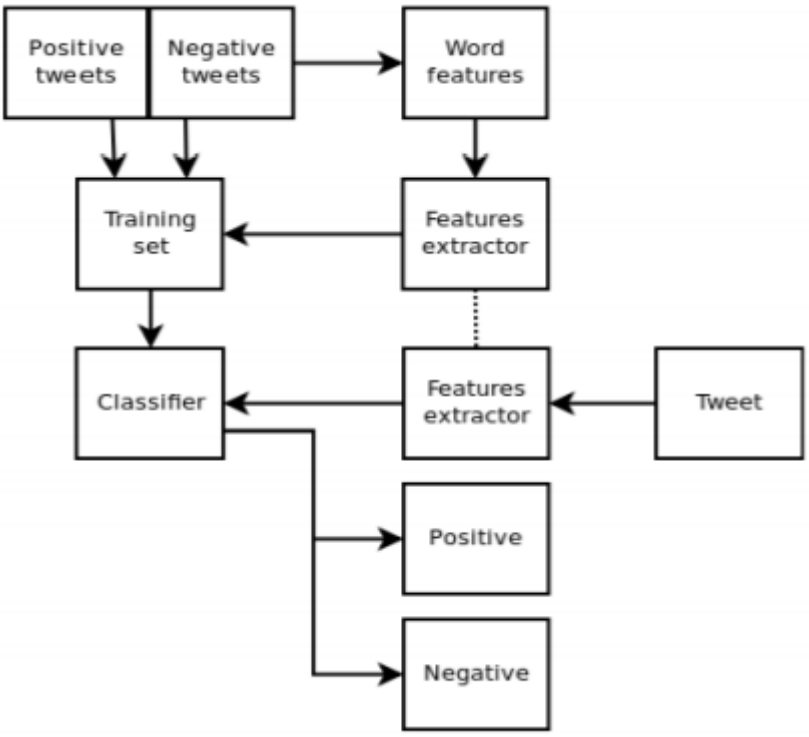


Figure 3. Sentiment Analysis Architecture [5]

The process of sentiment analysis can be handled in two ways: Machine learning based, Lexicon based (fig. 4).

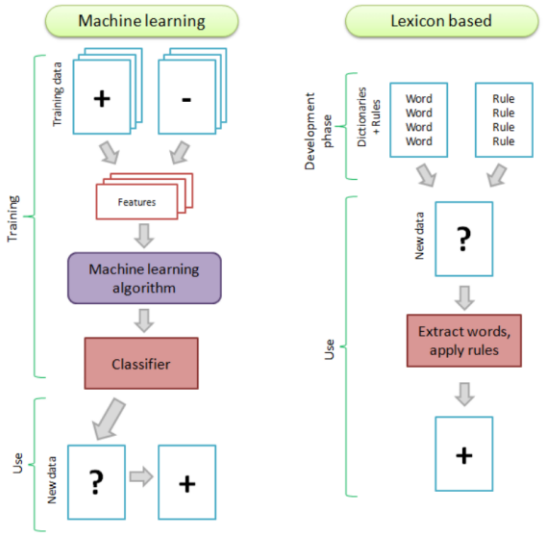


Figure 4. Machine learning-based vs lexicon-based sentiment analysis

Challenges in sentiment analysis:

Entity detection, Sarcasm detection, Negation detection, Handling comparison, subject detection [5].

2.1. Lexicons: In this study, three lexicons were used,

The AFINN lexicon assigns words with a score between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. AFINN includes 2,476 words in total.

The NRC lexicon categorizes words into positive or negative sentiments as well as 8 different emotions including anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. In total, 6,468 words were rated.

The BING lexicon categorizes words in a binary fashion into positive and negative categories. In total, 6,788 words are rated [6].

3. Topic Modeling:

It is a technique to group a collection of documents or texts. It is one of the unsupervised learning classification techniques used on documents to find natural groups (fig. 5). There are different ways to perform topic modelling and Latent Dirichlet Allocation (LDA) is most popular among them. Along with this there are GUI tools like GTMT (GUI Topic Modeling Tool), MALLET a topic modeling routine of MALLET NLP tool kit, STMT (Stanford topic modeling toolbox) [7].

It is mainly used in the are of computer science with prime focus on text mining and information retrieval. It is one such area which gained huge attention since it was introduced from researchers in various domains [8].

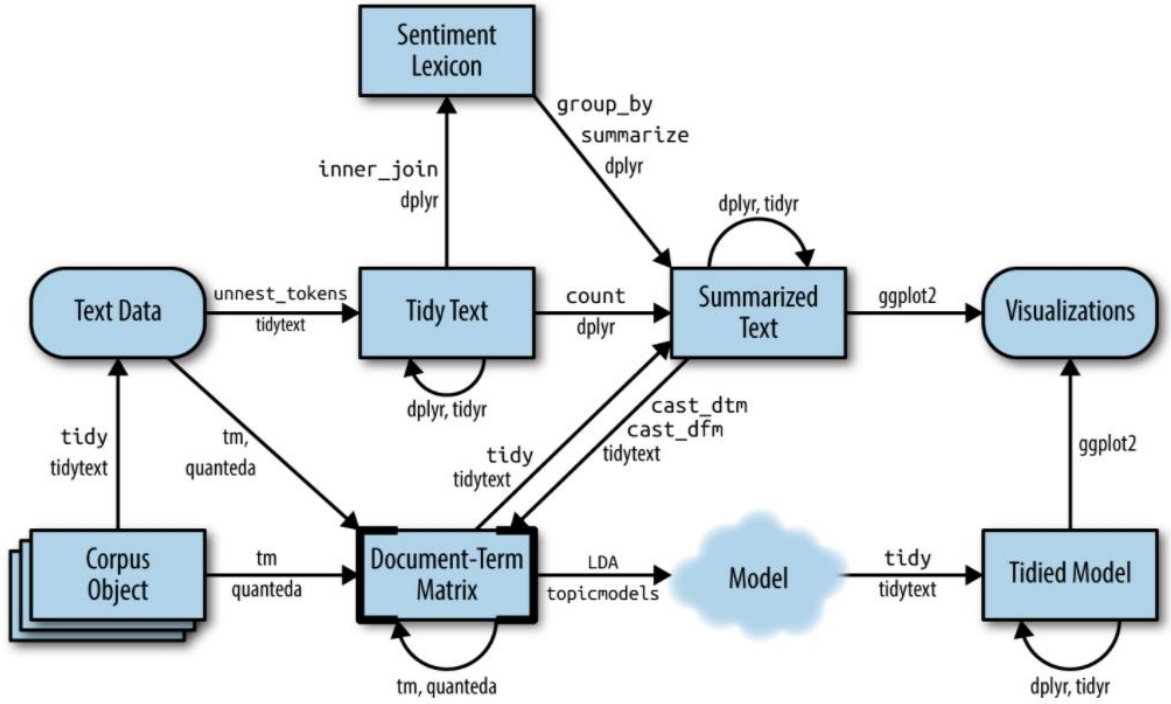


Figure 5. A flowchart of a text analysis that incorporates topic modelling [9]

Areas of application of Topic modelling (fig.6):

Figure 6. Topic modelling applications. https://medium.com/@fatmafatma/industrial-applications-of-topic-model-100e48a15ce4

4. Data:

The data from twitter using hashtags '#PowerWomen', '#forbesPowerWomen' from Dec 1st, 2020 till Dec 14th, 2020 which are in English excluding retweets with maximum limit on tweets fetched set at 1500 are fetched as a part of sentiment analysis on twitter data project using twitter developer authentication is loaded in to a csv file. Fig. 7 presents the snippet view of the code.

```
power_tweets <- searchTwitter("#PowerWomen OR  #forbesPowerWomen -RT", n =
1500,lang='en',since='2020-12-01',until='2020-12-14')
typeof(power_tweets)
#coverting list to dataframe
df_forbes_women <- twListToDF(power_tweets)
#viewing data
View(df_forbes_women)
#viewing created dates in data
View(unique(df_forbes_women$created))
#number of rows in data
nrow(df_forbes_women)
#writing data to csv file in local
write.csv(df_forbes_women, "C:/Users/krish/Downloads/forbes_powerfulwomen_data.csv")
```

Figure 7: Snippet view of code from twitter sentiment analysis on '#PowerWomen', '#forbesPowerWomen' (as a part of previous assignment).

4.1. Attributes:

"X", "text",
"favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","replyToUID",
"statusSource","screenname","retweetCount","isRetweet","retweeted","longitude","latitude"

4.2. Description:

text: The text of the status
screenName: Screen name of the user who posted this status
id: ID of this status
replyToSN: Screen name of the user this is in reply to
replyToUID: ID of the user this was in reply
statusSource: Source user agent for this tweet
created: When this status was created
truncated: Whether this status was truncated
favorited: Whether this status has been
favorited retweeted: TRUE if this status has been retweeted
retweetCount: The number of times this status has been retweeted
longitude, latitude: location details [10]
rest are binary variables like isRetweet, retweeted etc which are self-explanatory.
Tools used: R, MS Excel.

4.3 Packages used in R: rtweet, dplyr, stringr, qdapRegex, tidytext, tidyr, ggplot2, wordcloud, ROAuth, twitteR, RCurl, topicmodels, qdap, tm, snowballC.

4.4. About hashtag:

Since 2004, Forbes has been compiling the list of most powerful women in the world, which is edited by the reputed Forbes journalists which include women from different categories like politics, business, fashion, technology, media and entertainment, finance, philanthropy etc.

Angela Merkel (Chancellor of Germany) remaining in top spot since 2006. The list of 2020 was released on Dec 8th which includes names of Christine Lagarde (Politics & Policy - Head of European central bank), Kamala Harris (Politics- USA vice president elect), Melinda Gates (Philanthropy- cochair of Bill and Melinda Gates foundation), Mary Barra (Business- GM's CEO) [11].

5. Methodology:

Steps involved:

1. Reading Data:

- Reading the data from csv file which already had twitter data of the topic under study '#PowerWomen', '#forbesPowerWomen'
- Slicing fields, 'text','created' as separate fields to perform sentiment analysis w.r.to date of tweet

```
#loading text,created fields data to a seperate variable
pw_text1<- pw_data[,c("text","created")]
#type
typeof(pw_text1)
#unique values of created field
View(unique(pw_text1$created))
#splitting 'created' field data which is in timestamp format to date format
library(stringr)
pw_text1$created_date<- stringr::str_split_fixed(pw_text1$created," ",2)[,1]
#view data
View(pw_text1)
#view date data which is stored in to created time
View(pw_text1$created_date)
pw_text_data<- pw_text1[,c("text","created_date")]
head(pw_text_data)
```

- Pre-processing data to remove special characters, digits, punctuations, white spaces etc.
- Converting the resultant data to a dataframe and writing that processed data to a csv file
- Converting the dataframe in to Volatile Corpus
- Pre-Processing the corpus
- Converting the resultant data to Document Term Matrix

```
# Create Document Term Matrix
```{r echo=TRUE}
DTM <- DocumentTermMatrix(x=pw_corpus_pp)
nrow(DTM)
#1346
```

2. Topic Modelling:

- Topic Modelling using LDA() function with different factor/topic levels

```
library(topicmodels) #library to perform LDA

#applying lda on the dataset with different values of k (2,6,10)
#setting seed for reproducibility of the results


#a) k=2
pw_lda_2 <- LDA(DTM, k = 2, control = list(seed = 1234))
pw_lda_2 #A LDA_VEM topic model with 2 topics.

#b) k=6
pw_lda_6 <- LDA(DTM, k = 6, control = list(seed = 1234))
pw_lda_6 #A LDA_VEM topic model with 6 topics.

#c) k=10
pw_lda_10 <- LDA(DTM, k = 10, control = list(seed = 1234))
pw_lda_10 #A LDA_VEM topic model with 10 topics.
```

- Tidying the modelled data using tidy() function in tidytext package
```

```
library(tidytext)
#a) for two topic lda model case
pw_topics_2 <- tidy(pw_lda_2, matrix = "beta")
View(pw_topics_2)

#b) for six topic lda model case
pw_topics_6 <- tidy(pw_lda_6, matrix = "beta")
View(pw_topics_6)

#c) for ten topic lda model case
pw_topics_10 <- tidy(pw_lda_10, matrix = "beta")
View(pw_topics_10)
```

- Fetching top 15 most probable terms/topics from each case of k (2,6,10)
- Visualizing the above results and analysing the differentiation in topic under each k
- Obtaining topics/terms with highest probability and its corresponding probability value under factors of each k value
- Comparing these results with the probability of our initial topic of consideration '#PowerWomen', '#forbesPowerWomen' and identifying topic/term with overall highest probability in all cases (which happened to be a word different from the initially considered topics – 'woman').

3. Sentiment Analysis:

- Now, repeating the process of sentiment analysis on the data pulled from twitter with this new topic of highest probability of occurrence, with the steps given as follows:
    a. Fetching data for the hashtags #woman from twitter from Dec 13th 2020 till Dec 15th 2020 with tweet language set to English and maximum limit on tweets set to 800 without retweets

```
power_tweets <- searchTwitter("#woman -RT", n =
800,lang='en',since='2020-12-13',until='2020-12-15')
```

    b. Writing the data to csv file in local

```
#writing data to csv file in local
write.csv(df_forbes_women, "C:/Users/krish/Downloads/woman.csv")
```

    c. Reading the same csv file again as dataframe in to R environment
    d. Slicing fields, 'text','created' as separate fields to perform sentiment analysis w.r.to date of tweet
    e. Pre-processing data to remove special characters, digits, punctuations, white spaces etc.
    f. Loading all 3 lexicons nrc, bing, afinn

```
library(tidytext)
sentiments
get_sentiments("afinn") #afinn gives scores in -5 to 5 range
get_sentiments("bing") #bing lexicon gives only positive or negative sentiments no
emotions
get_sentiments("nrc") #gives emotion as well as sentiment
```

    g. Converting text to tokens

```
library(qdap)
#tokenizing data grouped on created_date
pw_tokens<- pw %>% group_by(created_date) %>% mutate(linenumber = row_number()) %>%
as.vector() %>% ungroup() %>% unnest_tokens(word, text)
```

    h. Selecting few emotions out of the list of nrc emotions (positive, joy in this case), identifying words in the dataset that matches with these emotions and visualizing result

```
nrc_joy_pos <- get_sentiments("nrc") %>%
  filter(sentiment %in% c("joy","positive"))
head(nrc_joy_pos)
```

    i. Identifying most prominent emotion in the dataset and visualizing the results

```
ggplot(top_emo)+ggtitle("Most frequent emotion in nrc for selected twitter data")+geom_bar(aes(reorder(sentiment,n),n,fill=sentiment),stat='identity')+xlab("sentiment")+ylab("frequency")
```

j. Using bing lexicon to visualize the date wise sentiment variations in data

k. Obtaining sentiment scores in the twitter data using all 3 lexicons and visualizing them together

```
#uniting all lexicons
```{r}
#appending all rows from 3 lexicons in to single entity and using it to plot
bind_rows(bing_pw,afinn_pw,nrc_pw) %>%
  ggplot(aes(index, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")+ggtitle("Lexicon wise sentiment")
```
```

l. Plotting a word cloud using wordcloud() function from wordcloud package for the data setting maximum words limit to 150 and minimum frequency to 5

```
library(wordcloud)
pw_tokens %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 150,scale=c(4,0.5),min.freq=5,random.order = FALSE,colors=brewer.pal(n=6,name="Dark2")))#setting max words limit to 150 , minimum freq to 5
```

m. Analysing the results from all the visualizations obtained

6. Results and observations:

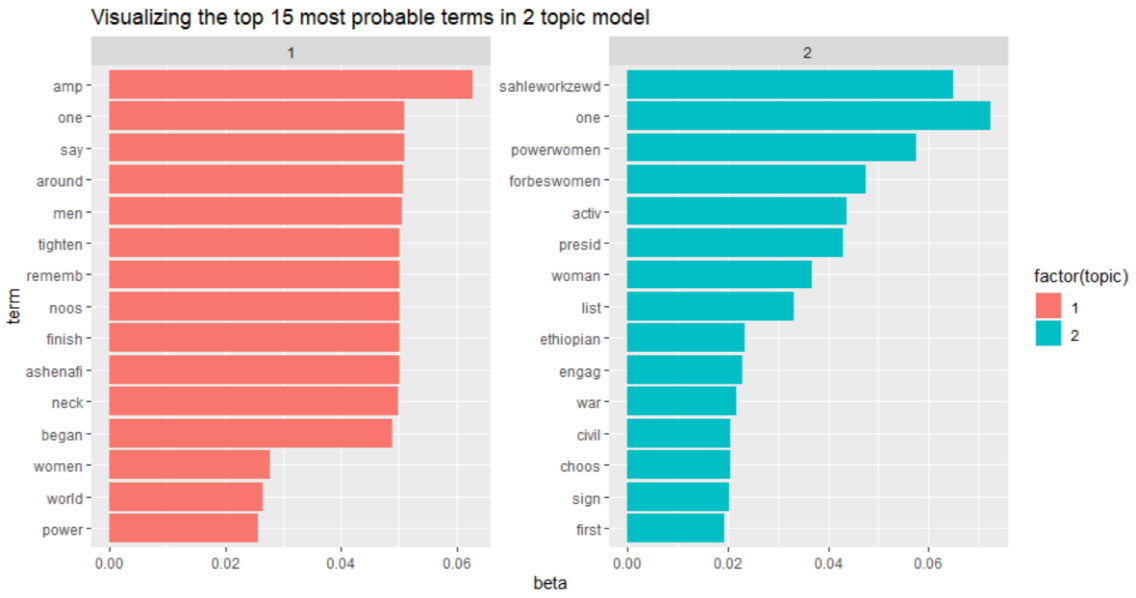


Figure 8. Top 15 most probable terms in two topic model

From the visualization above (fig. 8), we can see that there is good differentiation of terms between topic1 and topic2 with not severe visible overlap of words. Topic 1 is talking more about power, tighten, around, say, men etc, it seems to describe more verb/action terms. Topic 2 appears to have more content about the hashtag terms forbeswomen, powerwomen, sahleworkzewd (current president of ethiopia) etc.

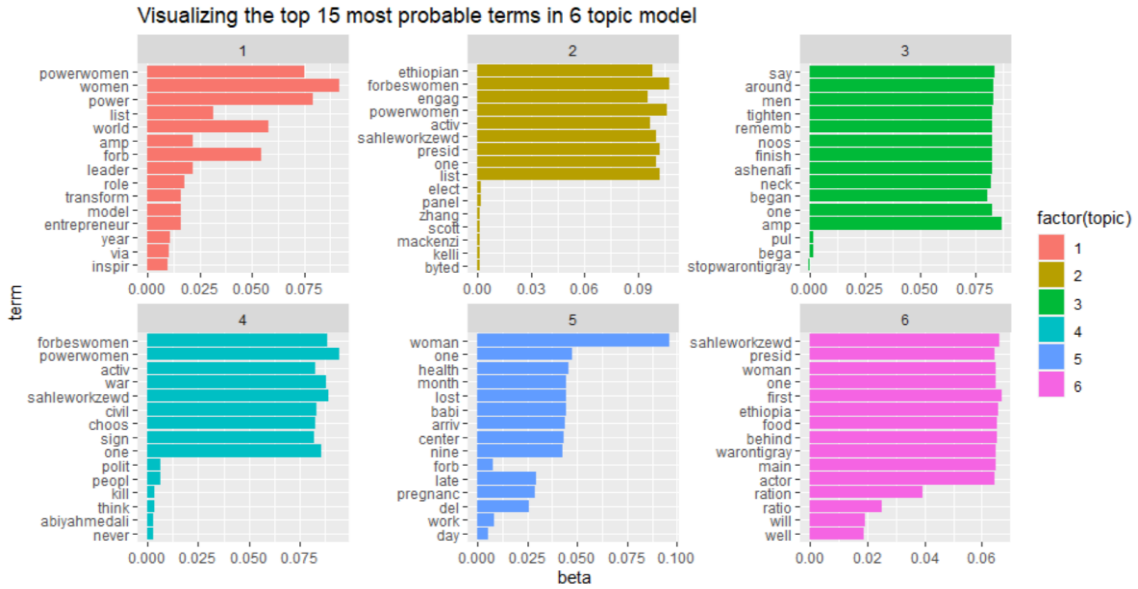


Figure 9. Top 15 most probable terms in 6 topic model

From the second visualization (fig. 9), we see that some of the terms overlap between few topics like ethiopian, Ethiopia in topic 2 & 6 ; women, woman in topic 1 & 5; one in topic 5 & 6; list in topic 1 & 2; powerwomen in topic 1,2 & 4. But still there is some good differentiation each topic wise in terms of the context.

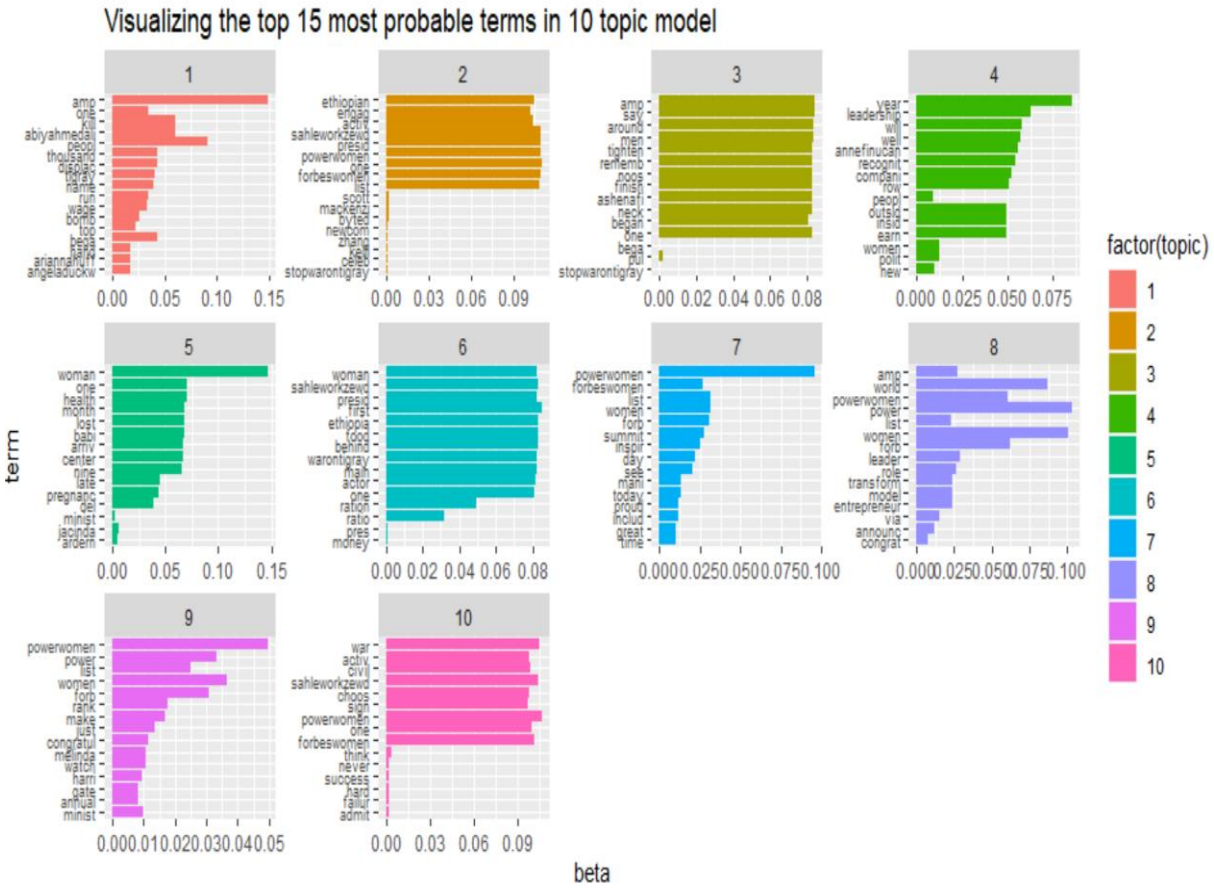


Figure 10. Top 15 most probable terms in 10 topic model

From the third visualization (fig. 10), we see some term overlap between few topics like Stopwarontigray in topic 2 & 3;women,woman in topic 5,7,8 & 9; sahleworkzewd in topic 2&6;list in topic 2,7,8 & 9;powerwomen etc. But still, there is only slight differentiation each topic wise in terms of the context.

	topic	term	beta
	All	powerwomen	All
1	2	powerwomen	5.752626e-02
3	1	powerwomen	2.025080e-02

	topic	term	beta
	All	forbesw	All
1	2	forbeswomen	4.751519e-02
11	1	forbeswomen	1.348846e-20

Figure 11. Highest probability of intial hashtag powerwomen in two topic model

	topic	term	beta
	All	All	All
1	2	one	0.072291113
2	2	sahleworkzewd	0.065039357

Figure 12. Topics with overall highest probability of in two topic model

From the above results (figs. 11 & 12), the topic chosen powerwomen has highest probability of 0.057 in topic 2 and forbespowerwomen are having their highest probability as 0.04 in topic 2 which is not the overall highest probability of the two topic model which is at 0.072 for the topic/term 'one'.

	topic	term	beta
	All	powerwomen	All
1	2	powerwomen	1.066618e-01
4	4	powerwomen	9.444739e-02
7	1	powerwomen	7.546347e-02

	topic	term	beta
	All	forbeswo	All
1	2	forbeswomen	1.078298e-01
2	4	forbeswomen	8.829815e-02
27	1	forbeswomen	1.050251e-17
34	5	forbeswomen	4.482091e-31
49	6	forbeswomen	4.105301e-77
51	3	forbeswomen	1.398358e-89

Figure 13. Highest probability of intial hashtags in six topic model

	topic	term	beta
	All	All	All
1	2	forbeswomen	0.107829837
2	2	powerwomen	0.106661823

Figure 14. Topics with overall highest probability of in six topic model

From the above results (figs. 13 & 14), the initial topics chosen, powerwomen has highest probability of 0.106 in topic 2 and forbespowerwomen has the highest probability as 0.107 in topic 2 which infact are the overall highest probabilites in 6 topic model.

	topic	term	beta
	All	powerwomen	All
1	2	powerwomen	1.099954e-01
2	10	powerwomen	1.067625e-01
4	7	powerwomen	9.610490e-02
5	8	powerwomen	6.030647e-02
6	9	powerwomen	4.958417e-02

	topic	term	beta
	All	forbeswo	All
1	2	forbeswomen	1.097196e-01
2	10	forbeswomen	1.014391e-01
3	7	forbeswomen	2.631530e-02
4	9	forbeswomen	8.520809e-18
5	1	forbeswomen	7.197592e-49

Figure 15. Highest probability of intial hashtags in ten topic model

	topic	term	beta
	All	All	All
1	1	amp	0.14899874
2	5	woman	0.14669544

Figure 16. Topics with overall highest probability of in ten topic model

From the above results (figs. 15 & 16), it is observed that, the initial topics chosen, powerwomen has highest probability of 0.109 in topic 2 and forbespowerwomen are having their highest probability as 0.109 in topic 2 which is not the overall highest probability in 10 topic model since highest probable words are ‘amp’ with 0.149 and ‘woman’ with 0.146 probability of occurrence.

Hence out of all the three cases of k(2,6,10), the word ‘amp’ had highest overall probability, since it is not a valid word, instead representation of special character, let us proceed with considering the word ‘woman’ as our new topic of consideration to carry out the sentiment analysis (fig. 17).

	text	created_date
1	Everyone ready for drone deliveries from Amazon U F Credit...	2020-12-14
2	A gift for every woman and mom Check out the essential oil...	2020-12-14
3	womensday WomensEqualityDay Feminism st book of its ki...	2020-12-14
4	Congratulations To our First woman Vice President Elect Ka...	2020-12-14
5	Israeli interrogators physically psychologically tortured Pales...	2020-12-14
6	Studio portrait If you d like to learn how to use studio lighti...	2020-12-14
7	ASMR Ear Licker Sings A Beautiful Song Plus Shows Off Her ...	2020-12-14
8	U Shaped Style Cotton Maternity Panties We are nice collect...	2020-12-14

Figure 17. Twitter data collected for the new topic ‘woman’

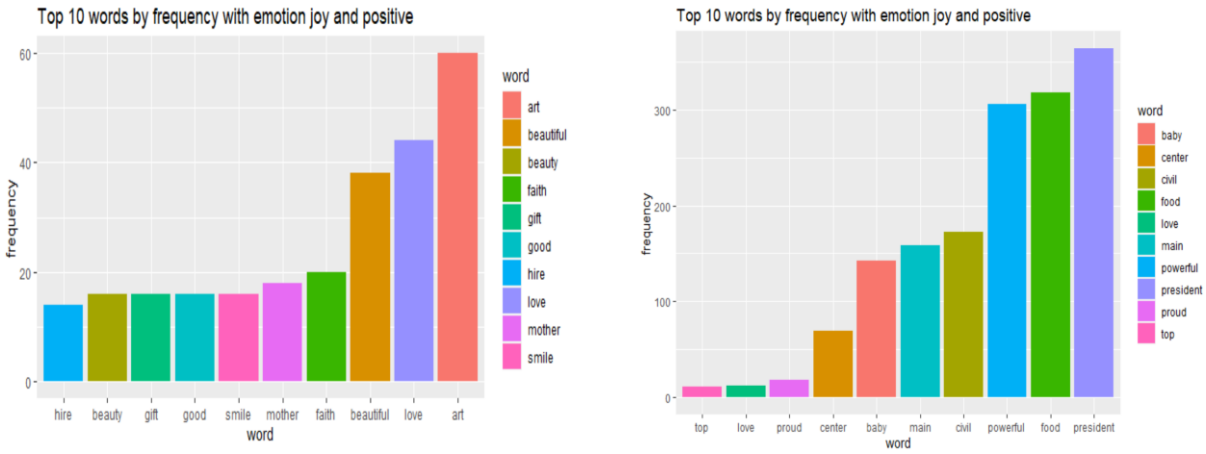


Figure 18. Visualizing most frequent words under emotion joy and positive (left- woman hashtag; right- powerwoman, forbespowerwoman hashtag)

From the above barplots (fig. 18), it is observed that with data from new term 'woman' the most frequent word in the collection of words with emotion 'joy' and 'positive' is 'art' where as the previous study on 'powerwoman', 'forbespowerwoman', it was the word 'president'.

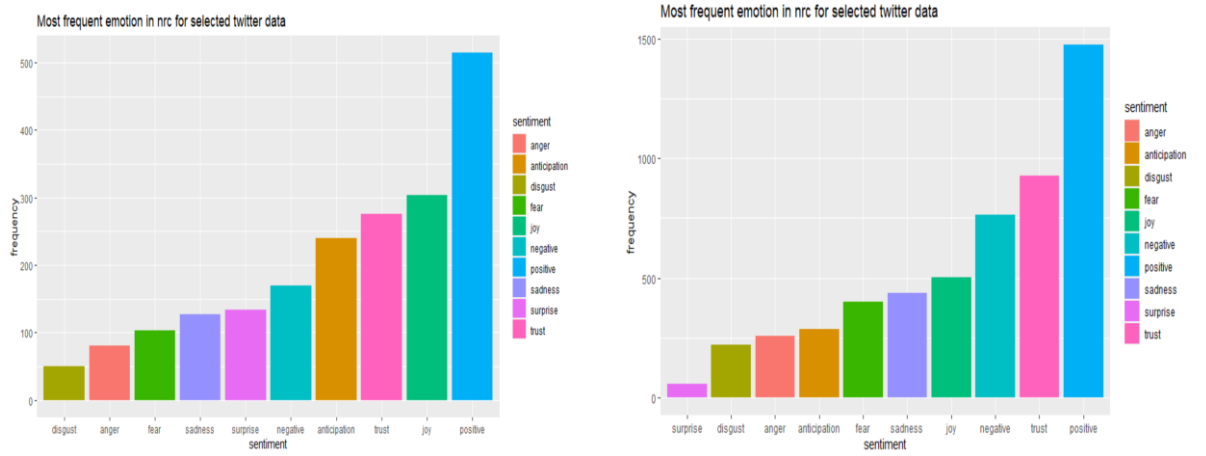


Figure 19. Visualizing most frequent emotion in the data (left- woman hashtag; right- powerwoman, forbespowerwoman hashtag)

From the above barplots (fig. 19), it is observed that with data from new term 'woman' the most frequent emotion in the collection of words with 'NRC' lexicon is 'positive' and same is the case with the results from previous study on 'powerwoman', 'forbespowerwoman' hashtags.

Figure 20. Date wise sentiment distribution using bing lexicon (top- woman hashtag; bottom- powerwoman, forbespowerwoman hashtag)

From the above results (fig. 20), it is observed that, with bing lexicon, extreme variations are observed on Dec14th than Dec 13th for the word 'woman'. In the case of initial topics, twitter interactions are observed to be triggered from Dec 8th since the Forbes powerful woman 2020 list was released on that day.

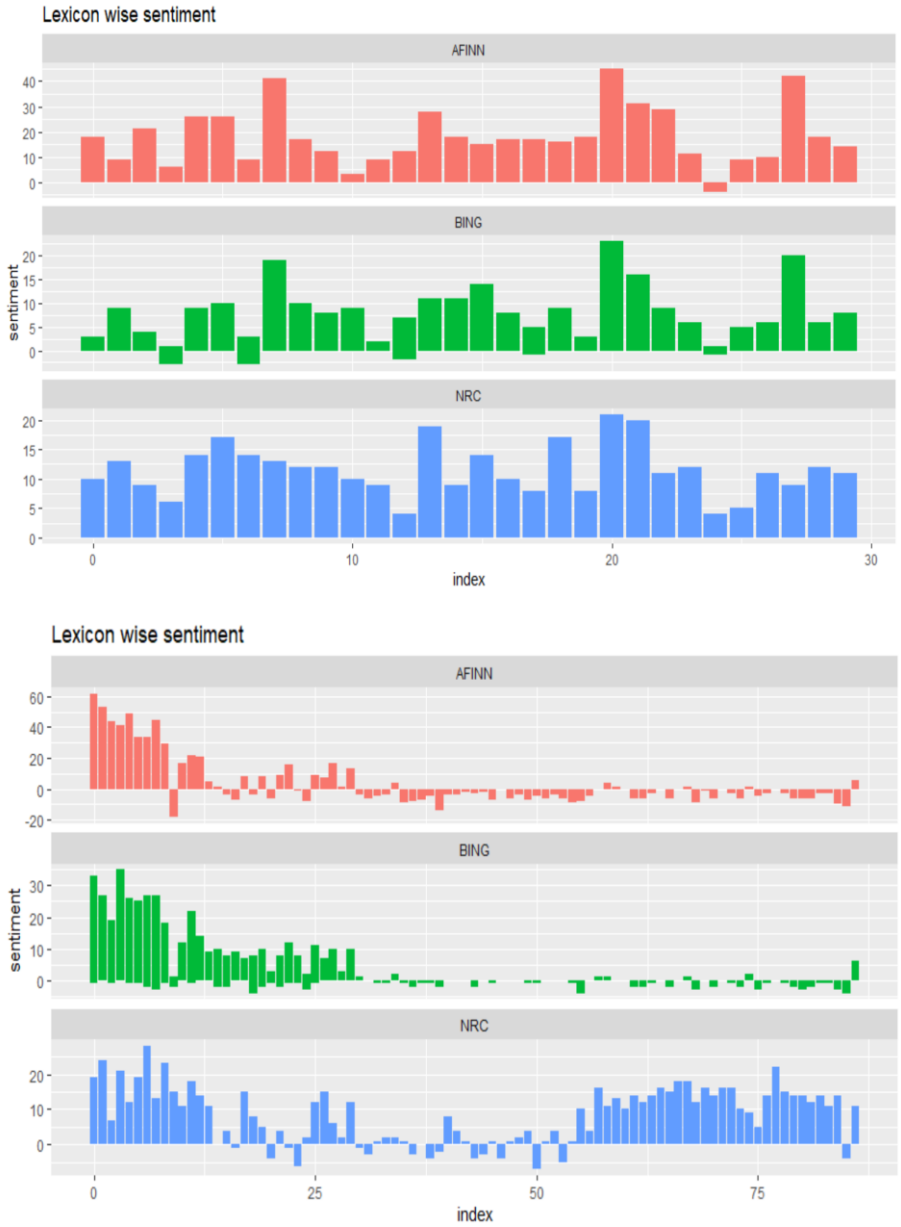


Figure 21. Lexicon wise sentiment distribution (top- woman hashtag; bottom- powerwoman, forbespowerwoman hashtag)

From the above results (fig. 21), higher sentiment values are observed in afinn lexicon compared to other two lexicons in both the topic cases.

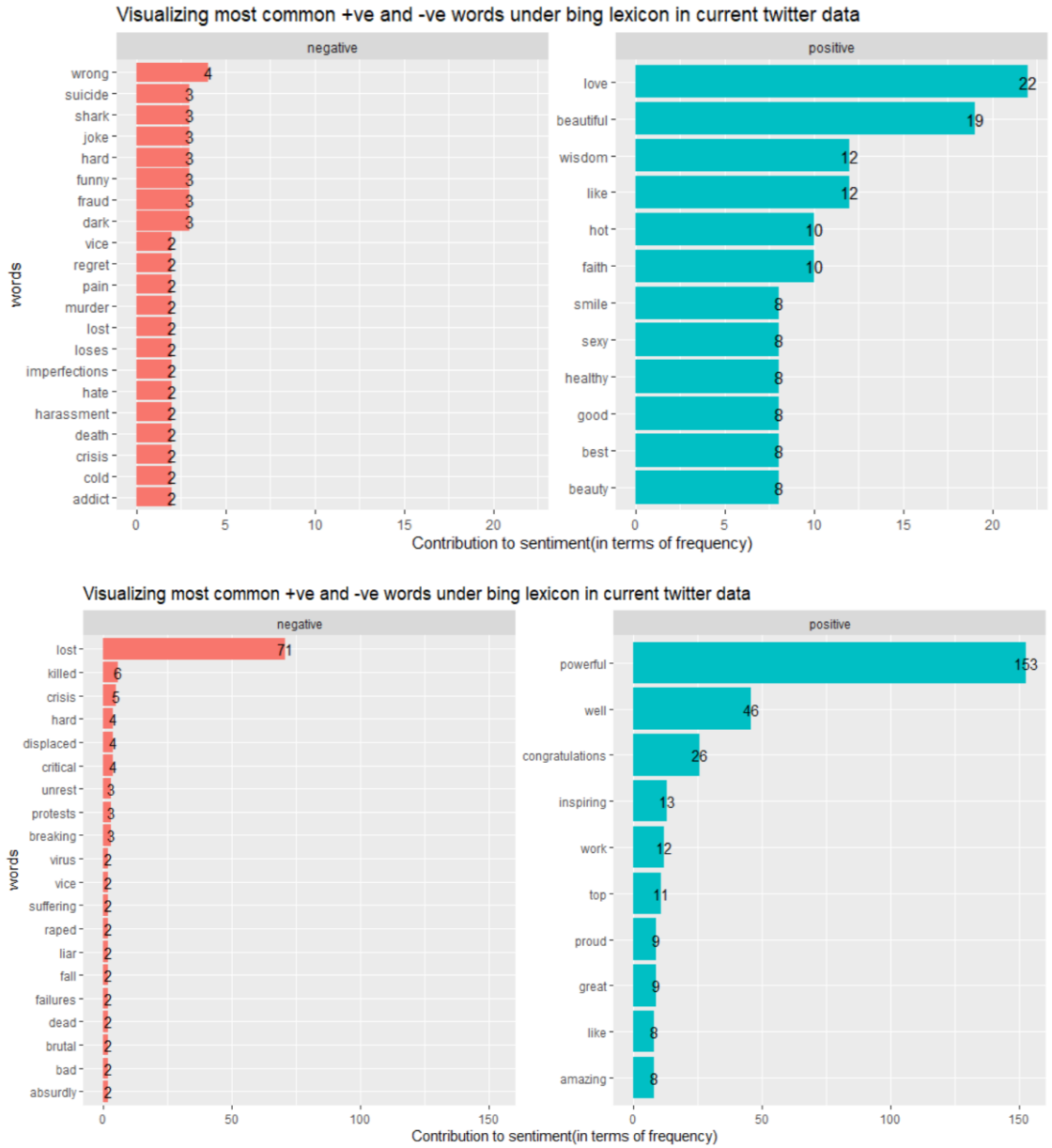


Figure 22. Plot showing most common positive and negative words in the dataset using bing lexicon (top- woman hashtag; bottom- powerwoman, forbespowerwoman hashtag)

From the horizontal barplots above (fig. 22), it is observed that the most common positive sentiment is ‘love’ with frequency 22 followed by ‘beautiful’ with frequency 19 and the most common negative term is ‘wrong’ with frequency 4 followed by ‘suicide’ with frequency 3. In the case of initial hashtags, the most common positive sentiment is ‘powerful’ with frequency 153 and most negative sentiment is ‘lost’ with frequency 71. Since the maximum word limit set for initial hashtags ‘powerwoman’ & ‘forbespowerwoman’ is 1500 whereas for latest hashtag ‘woman’ is 800, difference in frequencies is observed.

(a)

(b)

Figure 23. Word cloud on twitter data after removing stop words ((a)- woman hashtag; (b)- powerwoman, forbespowerwoman hashtag)

From the wordclouds above (fig. 23), it is observed that wordcloud generated from the new term ‘woman’ has words with almost similar frequency whereas is there is some varied frequency of terms is observed in wordcloud on ‘powerwoman’, ‘forbespowerwoman’ hashtag data. The most frequent term from the latest word cloud is 'woman' which is our key word, whereas in the initial study it is ‘powerwoman’ and ‘sahleworkzewde’.

7. Conclusion:

From the results, it is concluded that the process of LDA helped in identifying topics, based on highest probability of occurrence rather than directly picking a random trending topic/hashtag from huge corpus related to forbes power women. In this study, LDA helped to identify highest probable topic by comparing results of three cases (k = 2, 6, 10) where 'woman' is found to be most probable(~0.14%) than our actual/initial topic of consideration 'forbespowerwomen' and 'powerwomen'(~0.10%).

8. Future scope:

- Implementing machine learning based sentiment analysis
- Execute GUI based topic models such as STMT, GTMT, MALLET
- Deploying the model into cloud for real-time sentiment analysis

References:

[1] https://en.wikipedia.org/wiki/Digital_footprint

[2] https://netbasequid.com/blog/what-is-social-media-analytics-why-is-it-important/

[3] https://www.socialbakers.com/blog/social-media-analytics-tools

[4]https://en.wikipedia.org/wiki/Sentiment_analysis#:~:text=Sentiment%20analysis%20(also%20known%20as,affective%20states%20and%20subjective%20information.

[5] https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf

[6] https://books.psychstat.org/textmining/sentiment-analysis.html

[7] Graham, S., Weingart, S., Milligan, I. (2015). Exploring Big Historical Data: The Historian's Macroscope. Singapore: World Scientific Publishing Company.

[8] Liu, L., Tang, L., Dong, W. *et al.* An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **5,** 1608 (2016).

[9] https://www.tidytextmining.com/topicmodeling.html

[10] https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

[11] https://www.forbes.com/power-women/list/

Appendix:

[1] End exam code https://github.com/KrishnaRJ422/Twitter-Sentiment-Analysis/blob/main/Text%20mining%20using%20LDA%20and%20sentiment%20analysis%20on%20twitter%20data.Rmd

[2] Internal 3 Sentiment Analysis code https://github.com/KrishnaRJ422/Twitter-Sentiment-Analysis/blob/main/Text%20mining%20using%20LDA%20and%20sentiment%20analysis%20on%20twitter%20data.Rmd