**End Term Exam**

**Text Analytics**

**Name – Souvik Samanta**

**Roll No - 19PGDM064**

**Question1.**

**a) Selection technique –** The dataset we have is a labelled data where the sentiment column has two labels – positive and negative. Thus, the problem can be defined as a classification problem. Here, we can apply supervised classification algorithms such as Naïve-Bayes and Support Vector Machine for classifying the data. As in our case the data is binary in nature and they contain positive and negative classes. Naïve-Bayes is quite suitable in this case.

**b) Steps with R codes and output –**

# Loading Libraries and reading the dataset

# Setting Working Directory  
setwd("D:/Analytics/R/TA/EXAM")  
  
# Loading libraries  
library(tm)

library(SnowballC)  
library(wordcloud)

library(RColorBrewer)  
library(e1071)   
library(caret)

library(dplyr)

# Importing Dataset  
textdata <- read.csv("Question1 (2).csv")  
  
#View the first few lines of the dataset and the structure  
head(textdata)

## review  
## 1 One of my sisters friends lent me this game, and it is too damn hard! It carries the appearance of a kids game, but you have to learn how to do tons of intricate moves that require you to twist and turn your hands into all sorts of awkward positions, and you have to search seemingly endless levels for 100 notes, to improve your 'score'! You also have to find these impossibly hidden jigsaw puzzle pieces, that require you to do almost impossible tasks to get them! AND I AM ONLY UP TO STAGE THREE!!!!! Maybe if you have no life nad can stay home all the time you might get some enjoyment out of this, but otherwise keep away! AND IT IS DEFINATELY NOT RECOMMENDED FOR KIDS - THEY WILL PULL THEIR HAIR OUT WITHIN THE HOUR!  
## 2 Chuck Jones's 'Rabbit Seasoning', the second in the much beloved hunting trilogy, is often considered to be the best of the three. While I find it almost impossible to choose between this trio of fantastic cartoons, I would have to concede that 'Rabbit Seasoning' is the most finely honed script. Here, the emphasis is placed on language as Bugs and Daffy run through a series of complex dialogues in the grand tradition of Abbot and Costello's 'Who's on next' routine. As a long term Daffy fan, I have always been delighted by the hunting trilogy because it is consistently Daffy who gets all the best lines (the famous "Pronoun trouble" being one of the all time classics) and does most of the work. Bugs plays the role of cool manipulator while Elmer, as always, is the befuddled dupe. Part of what makes the hunting trilogy so much fun is that Daffy and Elmer pose so little threat to Bugs that he is basically just kicking back and having some easy laughs. Elmer falls into every trap that is laid for him but it is poor old Daffy who comes off worst, being shot in the face again and again, his beak ending up in more and more ridiculous positions. It all builds to the inevitable climactic declaration "You're despicable". As intricate an example of Chuck Jones's impeccable timing as you'll come across, 'Rabbit Seasoning' is a true classic.  
## 3 A lovely little B picture with all the usual Joe Lewis touches.... people ripping up pillows and auras of lurking fear. Also, alas, an ending that comes out of nowhere, because, apparently, the auteur has lost interest in the movie, or perhaps because as a B picture it has to fit into a slot.  
## 4 This was one of my favorite movies from childhood. I watched it so many times,eventually my tape wore out. I was a huge fan of this show and still am.The thing I love most about this movie is that it appeals to so many people, both young and old. I watch this movie now and laugh just as hard as I did the first time I saw it. I am now able to appreciate all the adult jokes that I never got as a child. My favorite characters are Elmyra and Foulmouth. Almost fifteen years later, my dad (a huge fan of the movie as well) and I are still quoting lines from this movie. I love the part where Foulmouth and Shirley go to the movies. "You save the seats, Shirl and I'll snag the dadgum snacks." I also loved the storyline of Plucky and Hampton and his family going to Happy World Land. Wade Pig reminded me a lot of my dad. I love the part when they finally get to Happy World Land and all they do is ride on the monorail. This movie is hilarious and appeals to children and kids. The animation, jokes and everything about it are top notch. If you have not seen it, rent it. You won't be sorry.  
## 5 Previously, I wrote that I loved "Titanic", cried at its ending (many times over), and I'm a guy in his 60's. I also wondered about why this great movie, which won so many awards and was applauded by so many critics, was given only a 7.0 rating by imdb.com users.<br /><br />Well, I looked at the breakdown of the user ratings. While 29.0% of all votes gave it a 10 rating, 10.7% gave it a 1 rating. These 10.7% of these irrational imdb users, in effect, pulled the overall rating down to 7.0. <br /><br />In my previous comments, I blamed this very unusual voting pattern (a sudden surge in 1 ratings, with a high 10 rating, dropping only gradually and then suddenly reversing course and jumping at the 1 rating level) on only one thing: hatred for Leonardo DiCaprio. Believe me, I've tuned into enough chat rooms to see the banter by young people (young men, mostly), who defame him left and right. They absolutely hate the man, and they will have no part in giving him any credit in "Titanic". (To answer one other user: I am NOT talking about someone who just really doesn't like the movie that much, and gave it a 5 or a 6, etc. Everyone has, and is entitled to, his/her own taste. But no one can convince me that the imdb rating of only 7.0 overall for "Titanic", pulled to that level by an inordinate number of ridiculous 1 ratings, is a fair reflection of the overall motion picture.)<br /><br />Let me demonstrate my point by comparing the imdb user voting pattern of "Titanic" to 5 randomly chosen box office and critical "bombs" (there are many more, but these 5 will prove my point). "Heaven's Gate" (1980) was pulled from the theaters quickly after a very poor box office showing, and imdb voters' ratings were: 23.2% 10 ratings and 9.2% 1 ratings (overall rating of 6.1). "Big Top Pee-wee" (1988) got 4.3% 10 ratings and 9.9% 1 ratings (overall rating of 4.5). "Cat People" (1982) got 6.1% 10 ratings and 2.6% 1 ratings (overall rating of 5.8). "Blind Date" (1987) got 3.0% 10 ratings and 2.8% 1 ratings (overall rating of 5.3). "Jumpin' Jack Flash" (1986) got 4.4% 10 ratings and 3.7% 1 ratings (overall rating of 5.2). WHAT DO ALL OF THESE FILMS HAVE IN COMMON WITH "TITANIC"? ALL OF THE PERCENTAGES OF THEIR 1 RATINGS ARE LOWER !!!! THAN "TITANIC", AND NONE OF THESE STINKERS EVER WAS NOMINATED FOR A SINGLE AWARD. Again, "Titanic" got 10.7% 1 ratings! Compare that to the other 5 movies I just mentioned.<br /><br />Can there be any explanation other than the hatred of Leo factor?<br /><br />  
## 6 This movie probably never made a blip on the radar screen, but it's got quite a bit of quality. It's pretty lifelike, yet you think "It's only a movie." Duvall and Close portray common people, and you'd never even realize they are now big-name actors. It seems that the jerk in this story is a little too old to be chasing Eugene's girlfriend, but I guess it's possible. It seems unlikely that the kid would travel from Montana to Nevada by himself, but I guess it's possible. You might think that the family troubles in this movie would never happen in your own family, but I guess it's possible. I remember Glenn Close saying something like "You think the work you do is the hardest part of your life, but it isn't."  
## sentiment  
## 1 negative  
## 2 positive  
## 3 positive  
## 4 positive  
## 5 positive  
## 6 positive

str(textdata)

## 'data.frame': 621 obs. of 2 variables:  
## $ review : chr "One of my sisters friends lent me this game, and it is too damn hard! It carries the appearance of a kids game,"| \_\_truncated\_\_ "Chuck Jones's 'Rabbit Seasoning', the second in the much beloved hunting trilogy, is often considered to be the"| \_\_truncated\_\_ "A lovely little B picture with all the usual Joe Lewis touches.... people ripping up pillows and auras of lurki"| \_\_truncated\_\_ "This was one of my favorite movies from childhood. I watched it so many times,eventually my tape wore out. I wa"| \_\_truncated\_\_ ...  
## $ sentiment: chr "negative" "positive" "positive" "positive" ...

# Defining the categories as factors  
textdata$sentiment <- factor(textdata$sentiment)  
  
# Check the number of positive and negative sentiments  
table(textdata$sentiment)

##   
## negative positive   
## 330 291

prop.table(table(textdata$sentiment))

##   
## negative positive   
## 0.531401 0.468599

# Create corpus  
txt\_corpus <- VCorpus(x = VectorSource(textdata$review))  
  
# Check the text in some rows  
lapply(txt\_corpus[5:8], as.character)

## $`5`  
## [1] "Previously, I wrote that I loved \"Titanic\", cried at its ending (many times over), and I'm a guy in his 60's. I also wondered about why this great movie, which won so many awards and was applauded by so many critics, was given only a 7.0 rating by imdb.com users.<br /><br />Well, I looked at the breakdown of the user ratings. While 29.0% of all votes gave it a 10 rating, 10.7% gave it a 1 rating. These 10.7% of these irrational imdb users, in effect, pulled the overall rating down to 7.0. <br /><br />In my previous comments, I blamed this very unusual voting pattern (a sudden surge in 1 ratings, with a high 10 rating, dropping only gradually and then suddenly reversing course and jumping at the 1 rating level) on only one thing: hatred for Leonardo DiCaprio. Believe me, I've tuned into enough chat rooms to see the banter by young people (young men, mostly), who defame him left and right. They absolutely hate the man, and they will have no part in giving him any credit in \"Titanic\". (To answer one other user: I am NOT talking about someone who just really doesn't like the movie that much, and gave it a 5 or a 6, etc. Everyone has, and is entitled to, his/her own taste. But no one can convince me that the imdb rating of only 7.0 overall for \"Titanic\", pulled to that level by an inordinate number of ridiculous 1 ratings, is a fair reflection of the overall motion picture.)<br /><br />Let me demonstrate my point by comparing the imdb user voting pattern of \"Titanic\" to 5 randomly chosen box office and critical \"bombs\" (there are many more, but these 5 will prove my point). \"Heaven's Gate\" (1980) was pulled from the theaters quickly after a very poor box office showing, and imdb voters' ratings were: 23.2% 10 ratings and 9.2% 1 ratings (overall rating of 6.1). \"Big Top Pee-wee\" (1988) got 4.3% 10 ratings and 9.9% 1 ratings (overall rating of 4.5). \"Cat People\" (1982) got 6.1% 10 ratings and 2.6% 1 ratings (overall rating of 5.8). \"Blind Date\" (1987) got 3.0% 10 ratings and 2.8% 1 ratings (overall rating of 5.3). \"Jumpin' Jack Flash\" (1986) got 4.4% 10 ratings and 3.7% 1 ratings (overall rating of 5.2). WHAT DO ALL OF THESE FILMS HAVE IN COMMON WITH \"TITANIC\"? ALL OF THE PERCENTAGES OF THEIR 1 RATINGS ARE LOWER !!!! THAN \"TITANIC\", AND NONE OF THESE STINKERS EVER WAS NOMINATED FOR A SINGLE AWARD. Again, \"Titanic\" got 10.7% 1 ratings! Compare that to the other 5 movies I just mentioned.<br /><br />Can there be any explanation other than the hatred of Leo factor?<br /><br />"  
##   
## $`6`  
## [1] "This movie probably never made a blip on the radar screen, but it's got quite a bit of quality. It's pretty lifelike, yet you think \"It's only a movie.\" Duvall and Close portray common people, and you'd never even realize they are now big-name actors. It seems that the jerk in this story is a little too old to be chasing Eugene's girlfriend, but I guess it's possible. It seems unlikely that the kid would travel from Montana to Nevada by himself, but I guess it's possible. You might think that the family troubles in this movie would never happen in your own family, but I guess it's possible. I remember Glenn Close saying something like \"You think the work you do is the hardest part of your life, but it isn't.\""  
##   
## $`7`  
## [1] "The dazzling seventeen-minute dance sequence of George Gershwin's 1928 orchestral piece, \"An American in Paris\", is an indisputable masterwork. Choreographed with precision and unparalleled flair by Gene Kelly, the vibrant combination of color, music and dance is still eye-poppingly startling as the piece is broken down into scenes inspired by selected master artists - Dufy in the opening Place de la Concorde piece, Manet in the flower market, Utrillo in a Paris street, Rousseau at the fair, Vincent Van Gogh in the spectacular Place de l'Opera piece, and Toulouse-Lautrec for the Moulin Rouge where Kelly wears his famous white bodysuit. The 97 minutes that precede this finale are not as exciting, not by a long shot, but there are certain charms to be had in viewing the entire 1951 Oscar-winning musical.<br /><br />Director Vincente Minnelli and screenwriter Alan Jay Lerner have fashioned a surprisingly sophisticated if rather slight romantic story focused on Jerry Mulligan, a former G.I. who has remained in Paris after the end of WWII trying to make a living as a painter. With his braggadocio manner and athletic dancing style, Gene Kelly can be concurrently ingratiating and irritating as a screen personality, but he seems to find his oeuvre as the carefree Jerry. The love-triangle plot is focused on Jerry's involvement with Milo Roberts, a self-proclaimed art patron but a sexual predator when it comes to young artists. On their first date in a crowded Montmartre nightclub, Jerry unapologetically falls for Lise, a young woman who turns out to be the fiancÃ©e of Henri, a professional entertainer and friend of Jerry's pal, Adam, an out-of-work concert pianist. Romantic complications ensue until the inevitable ending but not before several classic Gershwin songs are performed.<br /><br />The best of these is the most imitated - a swooningly romantic song and dance to \"Our Love Is Here to Stay\" along a faux-Seine River in a blue hazy mist with yellow fog lights. The way Kelly and Leslie Caron circle each other is transcendent as they approach each other tentatively at first and then synchronize beautifully to the music leading to the final clinch. Few films have so elegantly and succinctly shown two people falling in love. \"I Got Rhythm\" and \"S'Wonderful\" spotlight Kelly's nimble tap-dancing and agreeable singing, while \"Embraceable You\" is danced impressively by Caron in a five-scene montage of Henri's all-over-the-map description of Lise to Adam. Designed to show off Caron's dancing versatility, the sequence is similar to the one in \"On the Town\" where Vera-Ellen showed off her considerable dancing skills when Kelly's sailor character described his multi-faceted vision of Miss Turnstiles.<br /><br />As Lise, the nineteen year-old Caron (in her first film) dances superbly throughout and handles her role with unformed charm with her acting talent not to bloom for several years. Looking quite glamorous, Nina Foch plays older as the manipulative Milo and manages to be likable enough for us to care about her fate, while Oscar Levant is just his sardonic self as Adam. Performing an elegant \"I'll Build a Stairway to Paradise\", George GuÃ©tary plays Henri so agreeably that you feel bad that he does lose the girl at the end. This is not the best all-around MGM musical, but there is certainly enough movie magic to make this quite worthwhile. The 2000 DVD contains a fairly pristine print but little else in terms of extras."  
##   
## $`8`  
## [1] "This show is amazing! I love each and every episode. Carrie is a spitfire and Doug is a lovable and at times a moron. Arthur, Spence, Danny, Deacon and Carrie's Boss add just a nice end touch to the show, tying up all of the funny, pee in your pants moments. In one of the seasons, Doug tries to get Carrie drunk, because she is nicer when she is drunk. Nice husband right? Carrie isn't much better, when her boss needs a IPS driver to testify in a small case at her job, Carrie hesitates, because she views Doug as a slob and doesn't want him to embarrass her, so she hires Doug's friend instead. Wife of the year. But, who i believe to be funniest is yell-at-random Arthur. He is drop-dead hilarious, and angry. Hey, you would be too if you had to live in a basement where the mold has a funny smell and makes you dizzy. This show is hilarious, and if you haven't seen it yet, then you haven't lived!"

# Creating new corpus for cleaning  
corpus\_clean <- txt\_corpus  
  
# Remove Numbers  
corpus\_clean <- tm\_map(x = corpus\_clean, FUN = removeNumbers)  
  
# Transform all letters to lower case  
corpus\_clean <- tm\_map(x = corpus\_clean, FUN = content\_transformer(tolower))  
  
# Remove punctuation  
corpus\_clean <- tm\_map(x = corpus\_clean, FUN = removePunctuation)  
  
# Remove stop words  
corpus\_clean <- tm\_map(x = corpus\_clean, FUN = removeWords, stopwords())  
  
# add extra stop words: 'fy' within the inbuilt list  
myStopwords <- c(stopwords("english"), "the", "one")  
  
# remove stopwords from corpus  
corpus\_clean <- tm\_map(corpus\_clean, removeWords, myStopwords)  
  
# Remove extra white spaces  
corpus\_clean <- tm\_map(x = corpus\_clean, FUN = stripWhitespace)  
  
# Rechecking the corpus  
for (i in c(5:8)){  
 writeLines(as.character(corpus\_clean[[i]]))  
}

## previously wrote loved titanic cried ending many times im guy s also wondered great movie won many awards applauded many critics given rating imdbcom usersbr br well looked breakdown user ratings votes gave rating gave rating irrational imdb users effect pulled overall rating br br previous comments blamed unusual voting pattern sudden surge ratings high rating dropping gradually suddenly reversing course jumping rating level thing hatred leonardo dicaprio believe ive tuned enough chat rooms see banter young people young men mostly defame left right absolutely hate man will part giving credit titanic answer user talking someone just really doesnt like movie much gave etc everyone entitled hisher taste can convince imdb rating overall titanic pulled level inordinate number ridiculous ratings fair reflection overall motion picturebr br let demonstrate point comparing imdb user voting pattern titanic randomly chosen box office critical bombs many will prove point heavens gate pulled theaters quickly poor box office showing imdb voters ratings ratings ratings overall rating big top peewee got ratings ratings overall rating cat people got ratings ratings overall rating blind date got ratings ratings overall rating jumpin jack flash got ratings ratings overall rating films common titanic percentages ratings lower titanic none stinkers ever nominated single award titanic got ratings compare movies just mentionedbr br can explanation hatred leo factorbr br   
## movie probably never made blip radar screen got quite bit quality pretty lifelike yet think movie duvall close portray common people youd never even realize now bigname actors seems jerk story little old chasing eugenes girlfriend guess possible seems unlikely kid travel montana nevada guess possible might think family troubles movie never happen family guess possible remember glenn close saying something like think work hardest part life isnt  
## dazzling seventeenminute dance sequence george gershwins orchestral piece american paris indisputable masterwork choreographed precision unparalleled flair gene kelly vibrant combination color music dance still eyepoppingly startling piece broken scenes inspired selected master artists dufy opening place de la concorde piece manet flower market utrillo paris street rousseau fair vincent van gogh spectacular place de lopera piece toulouselautrec moulin rouge kelly wears famous white bodysuit minutes precede finale exciting long shot certain charms viewing entire oscarwinning musicalbr br director vincente minnelli screenwriter alan jay lerner fashioned surprisingly sophisticated rather slight romantic story focused jerry mulligan former gi remained paris end wwii trying make living painter braggadocio manner athletic dancing style gene kelly can concurrently ingratiating irritating screen personality seems find oeuvre carefree jerry lovetriangle plot focused jerrys involvement milo roberts selfproclaimed art patron sexual predator comes young artists first date crowded montmartre nightclub jerry unapologetically falls lise young woman turns fiancã©e henri professional entertainer friend jerrys pal adam outofwork concert pianist romantic complications ensue inevitable ending several classic gershwin songs performedbr br best imitated swooningly romantic song dance love stay along fauxseine river blue hazy mist yellow fog lights way kelly leslie caron circle transcendent approach tentatively first synchronize beautifully music leading final clinch films elegantly succinctly shown two people falling love got rhythm swonderful spotlight kellys nimble tapdancing agreeable singing embraceable danced impressively caron fivescene montage henris alloverthemap description lise adam designed show carons dancing versatility sequence similar town veraellen showed considerable dancing skills kellys sailor character described multifaceted vision miss turnstilesbr br lise nineteen yearold caron first film dances superbly throughout handles role unformed charm acting talent bloom several years looking quite glamorous nina foch plays older manipulative milo manages likable enough us care fate oscar levant just sardonic self adam performing elegant ill build stairway paradise george guã©tary plays henri agreeably feel bad lose girl end best allaround mgm musical certainly enough movie magic make quite worthwhile dvd contains fairly pristine print little else terms extras  
## show amazing love every episode carrie spitfire doug lovable times moron arthur spence danny deacon carries boss add just nice end touch show tying funny pee pants moments seasons doug tries get carrie drunk nicer drunk nice husband right carrie isnt much better boss needs ips driver testify small case job carrie hesitates views doug slob doesnt want embarrass hires dougs friend instead wife year believe funniest yellatrandom arthur dropdead hilarious angry hey live basement mold funny smell makes dizzy show hilarious havent seen yet havent lived

# Create Document Term Matrix  
DTM <- DocumentTermMatrix(x = corpus\_clean)  
  
dtm.matrix <- as.matrix(DTM)  
wordcount <- colSums(dtm.matrix)  
topten <- head(sort(wordcount, decreasing=TRUE), 10)  
topten # Top ten words with counts

## movie film like just good even see really story get   
## 1058 929 467 458 328 317 290 282 271 257

## Create training and test set  
  
# Create Training Set  
DTM\_train <- DTM[1:round(nrow(DTM)\*0.80, 0), ]  
  
# Create Test Set  
DTM\_test <- DTM[(round(nrow(DTM)\*0.80, 0)+1):nrow(DTM), ]  
  
# Create vectors with labels for the training and test set  
train\_labels <- textdata[1:round(nrow(textdata)\*0.80, 0), ]$sentiment  
test\_labels <- textdata[(round(nrow(textdata)\*0.80, 0)+1):nrow(DTM), ]$sentiment  
  
# Check proportion of ham and spam is similar on the training and test set  
prop.table(table(train\_labels))

## train\_labels  
## negative positive   
## 0.5291751 0.4708249

prop.table(table(test\_labels))

## test\_labels  
## negative positive   
## 0.5403226 0.4596774

library(wordcloud)  
  
# Create wordcloud for the whole dataset  
wordcloud(words = corpus\_clean,  
 min.freq = 100, # minimum number of times a word must be present before it appears  
 random.order = FALSE, # Arrange most frequent words to be in the center of the word cloud  
 color = (colors = c("#4575b4","#74add1","#abd9e9","#e0f3f8","#fee090","#fdae61","#f46d43","#d73027")) # Colour gradient for the font  
)



##  
threshold <- 0.1  
min\_freq = round(DTM$nrow\*(threshold/100),0) # calculate minimum frequency  
min\_freq

## [1] 1

# Create vector of most frequent words  
frequent\_words <- findFreqTerms(x = DTM, lowfreq = min\_freq)  
  
str(frequent\_words)

## chr [1:15600] "â„comedy" "â„mechanics" "â\200“" "aaliyah" "aaron" ...

# Filter DTM to only have most frequent words  
DTM\_train\_most\_frequent <- DTM\_train[, frequent\_words]  
DTM\_test\_most\_frequent <- DTM\_test[, frequent\_words]  
  
# Check dimension of DTM  
dim(DTM\_train\_most\_frequent)

## [1] 497 15600

# Create function that converts numeric values to "Yes" or "No" if word is present or absent in document  
is\_present <- function(x) {  
 x <- ifelse(test = x > 0, yes = "Yes", no = "No")  
}  
  
# Apply is\_present() function to training and test DTM  
DTM\_train\_most\_frequent <- apply(X = DTM\_train\_most\_frequent,  
 MARGIN = 2, # Apply function to columns  
 FUN = is\_present) # Specify function to be used  
  
DTM\_test\_most\_frequent <- apply(X = DTM\_test\_most\_frequent,  
 MARGIN = 2, # Apply function to columns  
 FUN = is\_present) # Specify function to be used  
  
  
library(e1071)  
  
# Create model from the training dataset  
spam\_classifier <- naiveBayes(x = DTM\_train\_most\_frequent, y = train\_labels)  
train\_predictions <- predict(object = spam\_classifier, newdata = DTM\_train\_most\_frequent)  
  
# Create confusion matrix  
confusionMatrix(data = train\_predictions, reference = train\_labels, positive = "positive", dnn = c("Prediction", "Actual"))

## Confusion Matrix and Statistics  
##   
## Actual  
## Prediction negative positive  
## negative 263 3  
## positive 0 231  
##   
## Accuracy : 0.994   
## 95% CI : (0.9825, 0.9988)  
## No Information Rate : 0.5292   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9879   
##   
## Mcnemar's Test P-Value : 0.2482   
##   
## Sensitivity : 0.9872   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 0.9887   
## Prevalence : 0.4708   
## Detection Rate : 0.4648   
## Detection Prevalence : 0.4648   
## Balanced Accuracy : 0.9936   
##   
## 'Positive' Class : positive   
##

## Make predictions on test set  
test\_predictions <- predict(object = spam\_classifier, newdata = DTM\_test\_most\_frequent)  
  
## Create confusion matrix  
  
# install.packages("caret")  
  
library(caret)  
  
# Create confusion matrix  
confusionMatrix(data = test\_predictions, reference = test\_labels, positive = "positive", dnn = c("Prediction", "Actual"))

## Confusion Matrix and Statistics  
##   
## Actual  
## Prediction negative positive  
## negative 56 30  
## positive 11 27  
##   
## Accuracy : 0.6694   
## 95% CI : (0.5792, 0.7512)  
## No Information Rate : 0.5403   
## P-Value [Acc > NIR] : 0.002376   
##   
## Kappa : 0.3174   
##   
## Mcnemar's Test P-Value : 0.004937   
##   
## Sensitivity : 0.4737   
## Specificity : 0.8358   
## Pos Pred Value : 0.7105   
## Neg Pred Value : 0.6512   
## Prevalence : 0.4597   
## Detection Rate : 0.2177   
## Detection Prevalence : 0.3065   
## Balanced Accuracy : 0.6548   
##   
## 'Positive' Class : positive   
##

**c) Results and assessment of model –**

From the above output, we can see that, within the training data, the model gives 99.4% accuracy and in validation data the accuracy is close to 67% - which is a good result. For further development we can try other classification models also along with some modifications made in corpus.

**Question2.**

**a) Scraping 5 articles on India-China-Standoff**

# Loading Libraries and reading the dataset

library(dplyr)

library(rvest)

library(stringr)  
  
func <- function(htm, xp){  
 article\_html <- read\_html(htm)  
 article\_body <- html\_nodes(article\_html,xpath=xp)  
   
 article\_body\_text <- html\_text(article\_body)  
   
 textdata = data.frame(article\_body\_text)  
 names(textdata) = "text"  
   
 textdata <- data.frame(textdata)  
 return(textdata)  
}  
  
htm1 <- "https://www.ndtv.com/india-news/amid-standoff-with-china-video-shows-army-tanks-near-lac-2301703"  
xp1 <- '//p'  
  
htm2 <- "https://www.news18.com/news/india/india-china-standoff-with-no-thaw-as-winter-sets-in-why-hitting-pause-was-the-only-logical-option-2902225.html"  
xp2 <- '//p'  
  
htm3 <- "https://www.thehindu.com/news/national/ladakh-standoff-new-army-talks-positive-but-no-breakthrough-on-de-escalation/article32667313.ece"  
xp3 <- '//p'  
  
htm4 <- "https://www.financialexpress.com/defence/india-china-standoff-no-loss-of-ground-says-indian-army/2085460/"  
xp4 <- '//p'  
  
htm5 <- "https://www.ndtv.com/india-news/india-will-retaliate-if-its-defences-on-heights-are-breached-top-sources-2292908"  
xp5 <- '//p'  
  
df <- data.frame()  
  
# Keeping relevant rows for final data frame  
df1 <- data.frame(func(htm1, xp1)[1:13,1])  
df2 <- data.frame(func(htm2, xp2))  
df3 <- data.frame(func(htm3, xp3)[1:16,1])  
df4 <- data.frame(func(htm4, xp4)[1:25,1])  
df5 <- data.frame(func(htm5, xp5)[1:16,1])  
  
# Changing the names  
names(df1) <- "text"  
names(df2) <- "text"  
names(df3) <- "text"  
names(df4) <- "text"  
names(df5) <- "text"  
  
df <- rbind(df, df1, df2, df3, df4, df5)  
names(df) <- "text"  
  
  
# Create corpus  
library(tm)

## Loading required package: NLP

myCorpus <- VCorpus(VectorSource(df$text))  
  
# Viewing the corpus  
for(i in 1:83){  
 writeLines(as.character(myCorpus[[i]]))  
}

## India-China Standoff: Army tanks and armoured personnel carriers in forward locations in Eastern Ladakh  
## A video showing Indian army tanks and armoured personnel carriers in forward locations in eastern Ladakh -- the theatre of the face-off with China since April-May -- has been released by the Centre. The videos show rows of T-90 tanks and BMP vehicles in Chumar-Demchock -- possibly the highest deployed tank formation anywhere in the world.    
## India has repeatedly stressed on the necessity to ensure "stability on the ground" after the latest round of Chinese provocation on August 31. The Chinese soldiers had tried to move in on the Indian side of the Line of Actual Control near the south bank of Ladakh's Pangong Lake.  
## On the night of August 29 and August 30, the People's Liberation Army had moved their tanks and troops at night, in contravention of confidence building measures that ban military movements at night. On August 31, more troop movements took pace as ground commanders of the two sides were in discussions to de-escalate the situation.  
## The army is now determined to strengthen security in the whole area..  
## #WATCH Indian Army deploys T-90 & T-72 tanks along with BMP-2 Infantry Combat Vehicles that can operate at temperatures up to minus 40 degree Celsius, near Line of Actual Control in Chumar-Demchok area in Eastern Ladakh.Note: All visuals cleared by competent authority on ground pic.twitter.com/RiRBv4sMud  
## ''Fire and Fury Corps is the only formation of Indian Army and also in the world to have actually deployed mechanised forces in such harsh terrain.  Maintenance of tanks, infantry combat vehicles and heavy guns is a challenge in this terrain,'' said Major General Arvind Kapoor, the Chief of Staff of 14 Corps.  
## The BMP vehicles can function in temperatures of up to minus 40 degrees Celsius.  
## China has also augmented its fire-power in high-altitude areas of Ladakh. Earlier this month, a video of a military exercise at the high-altitude Tibet Autonomous Region was tweeted by the editor of  Global Times. The video, which showed an array of multi-barrel rocket launchers in action, came hours ahead of talks between foreign ministers of the two nations.    
##   
## In his recent statement to the Lok Sabha, Union Defence Minister Rajnath Singh had issued what was seen as a clear warning to China, saying India is "very serious about issues of sovereignty" and the country is prepared for "all contingencies" to ensure that it is maintained.  
## India, however, has maintained that it is ready to resolve the border issue in a peaceful manner and wants the Chinese side "to work together with us".  
## Track Coronavirus pandemic in India and get the latest COVID-19 news from around the world on ndtv.com/coronavirus  
## File photo of an Indian army convoy moving on the Srinagar-Ladakh highway at Gagangeer.  
## After the sixth round of Corps Commander-level talks with China, the first response from government sources was an honest admission – that there was no apparent result. The explanation was that the issue is complex and will need more rounds of talks.  
## This made one aspect quite clear; that the 5-point consensus arrived at in Moscow was not being implemented immediately. This consensus was discussed and mentioned by the two sides in their respective statements, but not in the joint press release issued around 3:30am IST on September 11.  
## The joint press release had mentioned “dialogue, quickly disengage, maintain proper distance and ease tensions” as points “agreed” upon. “The two Foreign Ministers agreed that the current situation in the border areas is not in the interest of either side,” it said.  
## Meanwhile, another joint press release was issued on Tuesday which came 18 hours after the Corps Commander talks ended in Moldo. Significantly, this military meeting also had a diplomat providing the continuity between Moscow and Moldo. Joint secretary (East Asia) Naveen Srivastava has been handling the diplomatic meetings with China under the Working Mechanism for Consultation and Coordination (WMCC) on border affairs and was also part of External Affairs Minister S Jaishankar’s team in Moscow when the latter met his Chinese counterpart Wang Yi.  
## After the Corps Commanders’ meeting, there was only one joint press release, unlike after the Moscow meeting where two separate and one joint statement was issued. The press release issued after the 14-hour Moldo meeting said: The two sides had candid and in-depth exchanges of views on stabilizing the situation along the LAC in the India-China border areas.”  
## Stabilising the situation at the LAC was the crucial point here leading up to the only new point in the release - that both sides will “stop sending more troops to the frontline”. This indicated that for the present, with no agreement on disengagement at the friction points, the two sides had come to terms with the fact that maintaining the current situation without further aggravation was the best case scenario.  
## Former Indian diplomat Rajiv Dogra explains. “A cardinal principle of diplomacy is to keep trying. So it is a positive if the tensions have not escalated further. But diplomatic optimism does not mean shutting your eyes to the reality on ground.”  
## So, Ambassador Dogra warns - “it suits China if this situation prolongs”.  
## Around 2013, during the Depsang stand-off, a senior and seasoned diplomat who has since retired had once explained with regards to China that “their motive” will continue to be to change ground realities. He had said that China’s objective is to create new realities with every stand-off so that whenever maps for the crucial Western and Eastern sectors are exchanged, they can claim newer territories by citing incidences of the past, which are actually blatant transgressions in reality.  
## In this context, the two recent claims of China are significant. First, with regards to the Galwan Valley. In several statements after the bloody June 15 clashes, it claimed that it was India not China that crossed the LAC into their territory. India refuted these claims several times over. "The position with regard to the Galwan Valley area has been historically clear. Attempts by the Chinese side to now advance exaggerated and untenable claims with regard to Line of Actual Control (LAC) there are not acceptable. They are not in accordance with China’s own position in the past,” New Delhi said.  
## On the transgressions in northern bank of Pangong Lake, the Chinese ambassador to India, Sun Weidong, claimed during a webinar by Institute of Chinese Studies that “on the northern bank of Pangong Lake, China’s customary boundary line is in accordance with the LAC. China has not expanded its territorial claim”.  
## The Indian side is cognizant of China’s plan for a long time. However, with several friction points -- in northern bank of Pangong Lake, Depsang, Gogra and now south bank of Pangong Lake -- and little headway during the summer, there was seemingly one choice as the harsh winter sets – the commitment to not aggravate the situation by pausing on sending more troops. Perhaps this was the reason that even though the joint press release talks of the “consensus” between leaders to be implemented, it stops short of spelling out disengagement and de-escalation this time.  
##   
## An Army convoy in Leh on September 22, 2020.   | Photo Credit: ANI   
## In a joint statement issued on Tuesday on the sixth round of Corps Commanders talks, India and China said both sides had agreed to “stop sending more troops to the frontline” and “refrain from unilaterally changing the situation on the ground” while agreeing to hold more talks to resolve the standoff.  
## “They agreed to earnestly implement the important consensus reached by the leaders of the two countries, strengthen communication on the ground, avoid misunderstandings and misjudgements, stop sending more troops to the frontline, refrain from unilaterally changing the situation on the ground, and avoid taking any actions that may complicate the situation,” the joint statement said on the senior military commander level held on Monday. The two sides had candid and in-depth exchanges of views on stabilising the situation along the Line of Actual Control (LAC)in the India-China border areas, it stated.  
## A defence source said while the disengagement is yet to be agreed upon, the recent talks will ensure there won't be any is no further build up along the LAC. c, “take practical measures to properly solve problems on the ground, and jointly safeguard peace and tranquillity in the border area,” the statement added.  
## The talks, which went on for 14 hours, were quite positive despite lack of apparent result, the source said. The talks were held in line with the five point plan agreed between the two foreign ministers in Moscow early this month, another source said.  
## India has pressed for a road map for complete disengagement and de-induction of Chinese troops from all friction points and along the Line of Actual Control (LAC), the source added. For the first time, a Joint Secretary from the Ministry of External Affairs (MEA) was present at the talks.  
## The situation on the ground remains quiet since the foreign ministers meeting, the second source added. Officials had stated that for India, mutual disengagement would not be acceptable and China would have to pull back its troops from the standoff areas.  
## A new strategic report has said that China has doubled air bases and defensive positions near Indian border in three years.  
## The 2017 Doklam crisis appears to have shifted Beijing’s “strategic objectives”, with China more than doubling its air bases, air defence positions and heliports near the Indian border over the past three years, according to the latest report from geopolitical intelligence platform, Stratfor.  
## Since the Doklam crisis, China has started constructing at least 13 entirely new military positions near its borders with India, including three air bases, five permanent air defence positions and five heliports,the report said. “Construction on four of those new heliports started only after the onset of the current Ladakh crisis in May,” the report authored by Sim Tack, Senior Analyst at Stratfor said.  
## Stating that once finished, this infrastructure will provide support for an even greater intensity of Chinese operations, the report observed, “The timing of the Chinese build-up of military facilities along the border with India just prior to the ongoing Ladakh standoff suggests these border tensions are part of a much larger effort by China to assert control over its border regions.”  
## The report observed that this approach is similar to China's strategy in the South China Sea, where a build up of permanent defence facilities supports Chinese localised military superiority and significantly raises the “potential cost of military opposition” to Beijing's maritime claims in the region.  
## By applying the same strategy along its land frontier with India, “China aims to discourage Indian resistance or military action during future border disputes by ostentatiously demonstrating its ability and intent to engage in military confrontations,” it stated.  
## China’s strategy aims to confront Indian with an insurmountable challenge in territorial disputes by leaning on broad support capabilities that provide Beijing with a tremendous ability to mobilise forces into disputed border areas, the report added.  
## Observing that New Delhi will also continue to upgrade its overall military capabilities, particularly where gaps exist, the report said efforts by India and China to “translate these capabilities into dominance during border disputes will increase the possibility of direct confrontation.”  
## With inputs from Ananth Krishnan  
## Amidst the fear of tensions escalating between India and China along the Line of Actual Control (LAC), a senior officer has clarified that there has been no loss of ground this time around. What has actually happened is that there is a large a chunk of territory on their side of LAC to which Indian troops have not had access “more than 10 to 15 years” now, Indian Express reported. The officer has also stated that there has been no troops mobilization by Pakistan side, barring the regular ceasefire violations from that side along the Line of Control (LOC) or the Siachen area — located 80 km west of the Depsang Plains.  
## What has happened is that the Pakistan side has been keeping a big number of Indian troops engaged on the Western front. And this has forced the Indian army to increase its presence in Ladakh by pulling out troops from other commands.  
## As has been reported by Financial Express Online, the Indian Army is in a state of operational readiness as the Chinese troops can attempt to ingress from any other location along the LAC.  
## The PLA troops have been aggressively trying to ingress from different points along the LAC. “The Chinese are always pushing and provoking the Indian troops and have been blocked access to four patrolling points since April this year.  
## “It is not only the location in eastern Ladakh, the Indian Army is looking at the Eastern Command too,” explained a senior officer.  
## Story so far –  
## Financial Express Online has reported earlier that though after several rounds of diplomatic as well as military-level talks, during the disengagement, the Chinese troops pulled out from Patrolling Point-14 in Galwan to its side of the LAC. However, there is still a small number of PLA troops who are sitting on the Indian side of the LAC at PP-15 located in Hot Springs, and PP-17A in the nearby Gogra Post.  
## And, the Chinese troops have never left the Finger 4 ridgeline on the Pangong Tso’s north bank, and are less than 500 metres away from Indian troops where the ridge meets Finger 3.  
## Update  
## The Chinese side has positioned two brigades and in the process has cut off India’s access to PP10-13, which is India’s traditional patrolling points in the Depsang Plains area.  
## Importance of Depsang Plains  
## This is located 30 km southeast of the strategically important Daulat Beg Oldie post near the Karakoram Pass in the north.  
## Right in the middle of the mountainous terrain, there is a flat surface which can be used by either side to launch a military offensive, which is similar to the Spanggur Gap in the Chushul sub-sector.  
## In this area, the Line of Patrolling is located beyond India’s perception of the LAC, and this is the area where India has not been able to reach in the last decade and a half and is almost around 972 sq km.  
## “The Indian troops are being stopped by the Chinese from patrolling in these areas. And, India has not lost anything in this round of encounters with the Chinese,” explained the official.  
## So where exactly is the LAC  
## It is `quite inside’ from the International Boundary, and the limit of patrolling is even further inside from the LAC, in Depsang Plains. And the bulk of 972 sq km has been with the PLA for a long time.  
## Since April/May this year the Chinese side started blocking the Indian troops at the Bottleneck – and this is the access to PPs 10-13. And this is the same location where the Chinese troops were positioned in 2013 standoff.  
## What action has the Indian Army taken?  
## The blocking of the Indian troops by the Chinese has been going on for a long time. Indian Army does not want to create another flashpoint right now.  
## The access which has been blocked since April 2020, is around 10 km vertical and 4- 5 km horizontal between those patrolling points.  
## Enough troops have been positioned to handle any misadventure or aggression from the Chinese side.  
## Any re-enforcement at the world’s highest base?  
## No, there has been no –enforcement on the Siachen positions, as there has been no mobilization from Pakistan side, barring the regular ceasefire violations and infiltrations.  
## Indian Army is tied down dealing with counter-terrorism and counter-infiltration – means India is facing two fronts.  
## India China Border Tensions: For over 2 weeks, Chinese action has been focused on south of Pangong lake  
## The present standoff between India and China ''can take any trajectory,'' top government officials have told NDTV, indicating that the possibility of a conflict, localised or across the region, cannot be ruled out.  The room for dialogue and engagement at diplomatic and military levels, however, has not been ruled out and another round of Corps Commander-level talks has been agreed upon, though the date is yet to be decided. While expressing concern at the ground situation in South Pangong, officials said they believed that a war-like situation has not emerged as yet.  
## At the same time, it is believed that China's posturing along the frontier in eastern Ladakh ''is being controlled by the top, not by some exuberant local commanders".  
## NDTV has also learnt of some additional moves made by the Indian army to secure heights along Finger 4 on the north bank of Pangong lake in the last few days.  Through this sizeable deployment, Indian soldiers presently occupy ''heights overlooking the Chinese deployment on Finger 4". Any Chinese move to counter this would mean a Chinese deployment to an altitude of approximately 6,000 metres, or nearly 19,700 feet, extremely challenging from a military stand-point. Indian and Chinese soldiers on Finger 4 ridges are presently separated by a few hundred metres.  
## Significantly, Indian soldiers deployed in South Pangong, where the Indian army pre-empted a Chinese effort to unilaterally alter the Line of Actual Control, have been given robust rules of engagement where they can respond with deadly force in the event that Chinese infantry soldiers close-in to their positions.      
## After 200 Chinese soldiers moved to dominate the heights in South Pangong on the night on August 29-30, Indian soldiers were deployed widely across the region dominating key positions such as Mukhpari, Gurung Hill, Magar Hill, Rezang La and Richin La.  These soldiers have been told that they can engage Chinese forces if the perimeter defences of their positions, composed primarily of barbed wire, is breached.  ''If they try to cross that, it will be a red line,'' sources said.  
## Any decision to engage Chinese forces with lethal forces will, however, be decided by local military commanders, who are monitoring the ground situation constantly.  
## On Monday, nearly 20 Chinese soldiers armed with spears attempted to evict Indian soldiers from their position on the LAC near Mukhpari. Chinese soldiers also fired warning shots. NDTV has now learnt that such attempts at intimidation by the Chinse are happening ''almost on a daily basis".  
## Government sources have explained that China does hold a few heights near the southern banks of the Pangong lake and that the Indian army does not dominate either Black Top or Helmet -- two features which are the Chinese side of the LAC.  
## China is believed to have installed a camera atop Black Top to monitor nearby areas.  After the Chinese move in South Pangong was detected by Indian observation, there was ''a race for heights".    
## On the southern bank of Pangong, ''it was quits'' -- in other words, India and China dominate approximately the same number of key heights.  However, India dominates the Spanggur Gap to the West and all Chinese positions along the banks of the Spanggur Tso, a lake in the region.    
## The ground situation in South Pangong remains tense.  As many as 5,000 to 7,000 Chinese soldiers are deployed here with as many as 50,000 Chinese soldiers deployed across the LAC in Eastern Ladakh.  ''We are 1-1 across Eastern Ladakh,'' said the top government official.  ''If he expands [deploys more numbers], we will expand.  It will not be business as usual,'' he added.  
## The Indian Army and the Air Force are working on a constant basis to monitor the air situation in Eastern Ladakh as China ''has built up sizeable number of air assets". 150 aircraft - fighters, helicopters, electronic warning assets and tankers -- have been deployed on the Chinese side of the LAC in the Ladakh theatre.    
## This presence is being matched by the Indian Air Force. Part of the process of identifying de-escalation in the region will be when the Chinese move back their air assets from the region including surface-to-air missile units, which have been widely deployed.  
## Despite the ground situation, there is hope that ''both sides are trying to stop the actual shooting from happening". There have been multiple rounds of talks at diplomatic and military levels which have resulted in partial Chinese withdrawal in the Galwan and North Pangong region, nowhere close to India's expectations after the Chinese moved into areas stretching from the Depsang Plains, Galwan, Hot Springs, Gogra, North and South Pangong in May this year.    
## With Indian and Chinese Foreign Ministers set to meet in Moscow later this week, and there is a sense of anticipation and hope that the potential for a clash between India and China is taken out of the equation.  ''We are not in any position of weakness,'' asserted the top official who spoke to NDTV.  If there is a conflict, ''there will be a price to pay on both sides".

**b) Preparing the corpus and cleaning it**

## data pre-processing  
  
# convert to lower case  
myCorpus <- tm\_map(myCorpus, content\_transformer(tolower))  
  
# remove punctuation  
myCorpus <- tm\_map(myCorpus, removePunctuation)   
  
# remove numbers  
myCorpus <- tm\_map(myCorpus, removeNumbers)  
  
# Removing stop-words  
myStopwords <- c(stopwords("english"), "india", "china", "indian", "chinese")  
myCorpus <- tm\_map(myCorpus, removeWords, myStopwords)  
  
# Removing white spaces (extra spaces)  
myCorpus <- tm\_map(myCorpus,stripWhitespace)  
  
  
# Term document matrix  
dtm <- DocumentTermMatrix(myCorpus)  
dtm.matrix <- as.matrix(dtm)  
wordcount <- colSums(dtm.matrix)  
topten <- head(sort(wordcount, decreasing=TRUE), 10)  
topten # Top ten words with counts

## troops lac army pangong said side situation military   
## 23 20 18 17 17 17 17 16   
## border along   
## 15 14

**c) Text analysis –** We will mainly use n-grams analysis where mostly unigram and bi-grams have been used. Thus, we can check which words or pair of words are being used largely and making some sense. We will not go for tri-gram as the data of 5 articles are very limited and tri-grams are not making that sense, hence ignored.

# Plotting the data -  
# For top 10 -  
dtm <- DocumentTermMatrix(myCorpus)  
  
dtm.matrix <- as.matrix(dtm)  
wordcount <- colSums(dtm.matrix)  
topten <- head(sort(wordcount, decreasing=TRUE), 10)  
library(reshape2)  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':  
##   
## annotate

dfplot <- as.data.frame(melt(topten))  
dfplot$word <- dimnames(dfplot)[[1]]  
dfplot$word <- factor(dfplot$word,  
 levels=dfplot$word[order(dfplot$value,  
 decreasing=TRUE)])  
  
fig <- ggplot(dfplot, aes(x=word, y=value)) + geom\_bar(stat="identity")  
fig <- fig + xlab("Word in Corpus")  
fig <- fig + ylab("Count")  
print(fig)



# N-gram analysis -------------  
  
textdata <- data.frame(text = sapply(myCorpus, as.character), stringsAsFactors = FALSE)  
  
library(tidytext)  
library(dplyr)  
uni\_gram <- as.data.frame(textdata%>%  
 unnest\_tokens(word,text) %>%  
 count(word,sort=TRUE))  
uni\_gram[1:50,]

## word n  
## 1 troops 23  
## 2 lac 20  
## 3 army 18  
## 4 side 18  
## 5 situation 18  
## 6 pangong 17  
## 7 said 17  
## 8 military 16  
## 9 border 15  
## 10 ground 15  
## 11 along 14  
## 12 talks 14  
## 13 two 14  
## 14 soldiers 13  
## 15 will 13  
## 16 ladakh 12  
## 17 line 12  
## 18 air 10  
## 19 eastern 10  
## 20 joint 10  
## 21 south 10  
## 22 areas 9  
## 23 control 9  
## 24 lake 9  
## 25 positions 9  
## 26 s 9  
## 27 sides 9  
## 28 actual 8  
## 29 also 8  
## 30 area 8  
## 31 bank 8  
## 32 can 8  
## 33 deployed 8  
## 34 near 8  
## 35 points 8  
## 36 region 8  
## 37 report 8  
## 38 standoff 8  
## 39 agreed 7  
## 40 corps 7  
## 41 depsang 7  
## 42 new 7  
## 43 patrolling 7  
## 44 release 7  
## 45 since 7  
## 46 time 7  
## 47 around 6  
## 48 disengagement 6  
## 49 heights 6  
## 50 however 6

str(uni\_gram)

## 'data.frame': 884 obs. of 2 variables:  
## $ word: chr "troops" "lac" "army" "side" ...  
## $ n : int 23 20 18 18 18 17 17 16 15 15 ...

data\_bigrams <- textdata %>%  
 unnest\_tokens(bigram, text, token = "ngrams", n = 2)%>%  
 count(bigram,sort=TRUE)  
  
data\_bigrams[1:50,]

## bigram n  
## 1 pangong lake 8  
## 2 actual control 7  
## 3 eastern ladakh 7  
## 4 line actual 7  
## 5 south pangong 7  
## 6 bank pangong 6  
## 7 press release 6  
## 8 depsang plains 5  
## 9 foreign ministers 5  
## 10 joint press 5  
## 11 side lac 5  
## 12 two sides 5  
## 13 along lac 4  
## 14 along line 4  
## 15 border areas 4  
## 16 ground situation 4  
## 17 sending troops 4  
## 18 soldiers deployed 4  
## 19 agreed upon 3  
## 20 air bases 3  
## 21 control lac 3  
## 22 friction points 3  
## 23 joint statement 3  
## 24 long time 3  
## 25 night august 3  
## 26 northern bank 3  
## 27 pakistan side 3  
## 28 patrolling points 3  
## 29 release issued 3  
## 30 report said 3  
## 31 round corps 3  
## 32 situation ground 3  
## 33 source said 3  
## 34 stop sending 3  
## 35 troops frontline 3  
## 36 across region 2  
## 37 air assets 2  
## 38 air defence 2  
## 39 air force 2  
## 40 apparent result 2  
## 41 armoured personnel 2  
## 42 army convoy 2  
## 43 army tanks 2  
## 44 august soldiers 2  
## 45 barring regular 2  
## 46 black top 2  
## 47 blocking troops 2  
## 48 bmp vehicles 2  
## 49 border disputes 2  
## 50 border tensions 2

**d) Text analysis insights–** As we can infer from the bigram analysis, we can see that “Pangong lake” is under high consideration. This indicates a great attention to that place. Also, “press release”, “control” these are the keywords which are attention seeking. Thus the analysis depicts that there are some critical issues on this topic on several news.