Tweet Sentiment Analysis Project

I Zidan

19/06/2020

Mission

- Model/prepare and predict sentiments for the provided tweets datasets.
- The datasets used was extracted from Kaggle

Given

- train.csv file provided by Kaggle for the purpose of training the model
- validation.csv dataset for final prediction. Also provided by kaggle

Requirements

- Provide a model that is able to predict a sentiment for a given tweet
- Apply model on the provided datasets

My thoughts

Sentiment and text analysis and mining is a very challenging area. Many models are out there, but trust that you will always come across scenarios that you have either missed or think of. I am a begginer in this field, but does not stop me from thinking, the challenges are down to the fact that humans are unpredictable and can write anything and use text and notations creatively and differently. I felt sometimes lost and sometimes overwhelmed.I did learn my limitations and the need to take the next step in datascience. The key is to get involved as much as you can.

Methodology

I have read a lot about the topic and decided to run the given data through many algorithms and pick the best one and in turn use for the final prediction. This is really the essence of what I learned in this course. It is always recommended to look at options and pick the best. So, to do this, the below will be done:

1. Load data
2. Clean data
3. Take a peak at the data
4. Partition data

5. Use training data to train model and run through models namely syuzhet,Afinn, bing, nrc, vegnitte, sentimentr/rinker
6. Validate all models against training data
7. Pick and show the best model
8. Again run the test data through all models model
9. Validate all models against test data
10. Pick and show best one
11. Use validation data with the best model
12. Show prediction

References

- Course lecture notes/videos (big help)
- Sentimentr documentation
- vegnitte documentation
- tidytext documentation

Note

Please be aware that much of the work work was done on the preparation and building. kindly bear with me and thank you.

Loading required libraries

The first task in this journey is to load the required libraries. Some of the important libraries for the task are tidytext, textclean, vegnitte, Syizhet, sentimentr. please refer to code to see full list.

Global functions

the below functions will help in cleaning and preparing the model.

```
# This function is responsible for performing the initial cleaning of the tweets text
# Function require cleantext package
# function names are self explanatory.
# for full reference please visit https://github.com/trinker/textclean

cleanText  <- function(text){
  text = replace_url(text)
  text = replace_emoticon(text)
  text = replace_tag(text)
  text = replace_non_ascii(text)
  # Replace contractions with both words
  text = replace_contraction(text)
  text = replace_date(text)
  text = replace_number(text)
  text = replace_html(text)
  text = replace_kern(text)
```

```r
  return(text)
}

# Function to convert text sentiment to integers
# this is done for validation purposes later on
decodeSentiment<- function(sn) {
  case_when(
    sn == "neutral" ~ 0,
    sn == "negative" ~ -1,
    sn  == "positive" ~ 1
  )
}

# Function to convert integer sentiment to text sentiment
# this is used for the final output
encodeSentiment<- function(sn) {
  case_when(
    sn == 0 ~ "neutral",
    sn < 0  ~ "negative",
    sn > 0  ~ "positive"
  )
}

# Function to interpret predicted sentiments to integer form (-1,0,1) for consistency
matchSentiment<- function(sn) {
  case_when(
    sn == 0 ~ 0,
    sn < 0  ~ -1,
    sn > 0  ~ 1
  )
}


# set working Directory mainly to pickup the datafile from a known location
# set working Directory mainly to pickup the datafile from a known location
setEnv <- function(){
  setwd(str_c(getwd(),"/"))
  str_c("Working directory is set. Place data file in this location ",getwd(),"/")
}

# Loading the tweets data provided
# Files are placed on Git to download if required
load.Data <- function(dataFile,method){

  if (method == "remote") {

    if (dataFile=="train.csv") {
       urlfile <- "https://raw.githubusercontent.com/imadzidan/Tweet-Sentiment-Analysis-Project/master/
    }else {
      urlfile <- "https://raw.githubusercontent.com/imadzidan/Tweet-Sentiment-Analysis-Project/master/va
    }

    df <- read_csv(url(urlfile))
  }else {

    df <- read_csv(dataFile)

  }


    return(df)

}

# The check_text function helps as it gives you suggestions on how to clean the text
# This check_text exists in the textClean library
check.Text <- function(df){

  # The check_text function helps as it gives you suggestions on how to clean the text
  x <- as.factor(df$text)
  return(check_text(df$text))

}

# Function to clean the tweets text and drop unwanted columns.
# Function will create a new column "modeled_text" which is the cleaned version of the text.
# The clean text will be used to get the sentiment
clean.Data <- function(df){
  df <- df %>% drop_na() %>% mutate(modeled_text = cleanText(text))
  df <- df[-c(2,3)]
  return(df)
}
```

Partition the data into train set and test set.

the split is 50% for the training and 50% for the test. The aim is to include as many variations as possible.This deemed to be helpfull when validating the result.

```
# partition data to start the text analysis
# 50-50. 50% for training and 50% for test. this is to include as many variation as possible.
# it seems to give a better validation estimates.
partition.Data <- function(df){
  tweets_test_index <- createDataPartition(y = df$sentiment, times = 1,  p = 0.5, list = FALSE)
  tweets_train_set <- df[-tweets_test_index,]
  tweets_test_set <- df[tweets_test_index,]
  return(list(tweets_train_set,tweets_test_set))
  }
```

Trianing the model

up to this stage, no output is shown as. sorry but eventually, it will happen.

```r
# This function will take the text through many algorithms.
# Function will provide stats for each algorithm and pick the best one.
# sentimentr/rinker, vegnitta and syuzhet libraries are used.
train.model <- function(which_df){

  #Which dataset to process
  if (which_df == "train") {
  df <- train_set
  }else{
    df <- test_set
  }

   #Remove na if exists
  df <- df %>% drop_na()

  # Run 4 selected algorithms using Syuzhet package
  # Store result in a dataframe
  syuzhet <- setNames(as.data.frame(lapply(c("syuzhet","bing", "afinn", "nrc"),
                        function(x) matchSentiment(get_sentiment(df$modeled_text, method=x)))),
                        c("jockers","bing", "afinn", "nrc"))

  pred_tbl_p1 <- data.frame(
    textID = df$textID,
    modeled_text = df$modeled_text,
    syuzhet,
    #include 5th algorithm from another package vegnitte
    vegnitte= matchSentiment(replace_na(analyzeSentiment(df$modeled_text)$SentimentQDAP,0)),
    actual_sentiment = decodeSentiment(df$sentiment),
    stringsAsFactors = FALSE
  )

  # Run and include 6th algorithm using sentimentr
  (rinker <- with(
    df,
    sentiment_by(
      get_sentences(modeled_text), question.weight = 0,
      averaging.function = average_weighted_mixed_sentiment,list(textID)
    )
  ) %>% filter(!is.na(textID)))

  # Store sentimentr output into a frame
  pred_tbl_p2 <- rinker %>% mutate(sentimentr= matchSentiment(ave_sentiment))
  pred_tbl_p2 <- pred_tbl_p2[,-c(2,3,4)]

  #join the the above two outputs into one data frame (Combining results)
  pred_tbl <- pred_tbl_p1 %>% inner_join(pred_tbl_p2,by="textID")

  return(pred_tbl)

}

# Function to validate predicted sentiment against the actual sentiment.
# sentimentr::validate_sentiment is used
```

```r
validate.Prediction <- function(df){

  nrc_pred <- validate_sentiment(df$nrc, df$actual_sentiment)
  nrc_pred<- nrc_pred %>%
                      mutate(mda=attributes(nrc_pred)$mda,
                      mare=attributes(nrc_pred)$mare,method="nrc")

  bing_pred <- validate_sentiment(df$bing, df$actual_sentiment)
  bing_pred<- bing_pred %>%
                       mutate(mda=attributes(bing_pred)$mda,
                       mare=attributes(bing_pred)$mare,method="bing")

  afinn_pred <- validate_sentiment(df$afinn, df$actual_sentiment)
  afinn_pred <- afinn_pred %>%
                           mutate(mda=attributes(afinn_pred)$mda,
                           mare=attributes(afinn_pred)$mare,method="afinn")

  sentimentr_pred <- validate_sentiment(df$sentimentr, df$actual_sentiment)
  sentimentr_pred <- sentimentr_pred %>%
                                    mutate(mda=attributes(sentimentr_pred)$mda,
                                    mare=attributes(sentimentr_pred)$mare,method="sentimentr")

  jockers_pred <- validate_sentiment(df$jockers, df$actual_sentiment)
  jockers_pred <- jockers_pred %>%
                               mutate(mda=attributes(jockers_pred)$mda,
                               mare=attributes(jockers_pred)$mare,method="jockers")

  vegnitte_pred <- validate_sentiment(df$vegnitte, df$actual_sentiment)

  vegnitte_pred <- vegnitte_pred %>%
                                 mutate(mda=attributes(vegnitte_pred)$mda,
                                 mare=attributes(vegnitte_pred)$mare,method="vegnitte")

  all_pred<- bind_rows(nrc_pred, bing_pred,afinn_pred,sentimentr_pred,jockers_pred)

  # Returning a list of two dataframes. One contains all validation
  # whilst the second contains only the best validation.
  return(list(all_pred,all_pred[which.max(all_pred$accuracy),]))

}

# Function to perform prediction with the best algorithm found in our case afinn
predict.Sentiment<- function(df){

  df <- df %>% drop_na()

  syuzhet <- setNames(as.data.frame(lapply(c("bing"),
             function(x) matchSentiment(get_sentiment(df$modeled_text, method=x)))), c("sentiment"))

  pred <- data.frame(
    textID = df$textID,
    modeled_text = df$modeled_text,
    syuzhet,
    stringsAsFactors = FALSE
  )
  pred <- within(pred, sentiment <- encodeSentiment(sentiment))

  return(pred)
}

#Clean output by dropping and renaming columns.
clean.output <- function(df,alg,n){

  df <- df[c(1,2,n)]

  names(df)[names(df)==alg] <- "sentiment"

  return(within(df, sentiment <- encodeSentiment(sentiment)))

}
```

Running the model

Now, all the building that has been created is called into action.

Load and clean data

```
#----------------------building the model------------------------------
  # Set working directory to local directory.
  # Datafiles should be present
  setEnv()
```

```
## [1] "Working directory is set. Place data file in this location C:/GitWorking/Tweet-Sentiment-Analys
```

```
  # pick up the train.csv file from Git or the location set above
  # remote parameter is set to pick up the data from Git
  # set it to any other value like loc to pickup the file from the current
  # working directory. provided you downloaded the file into the directory
  tweet_train <- load.Data("train.csv","remote")

  # The data looks like this
  head(tweet_train,5)
```

```
## # A tibble: 5 x 4
##   textID    text                                selected_text           sentiment
##   <chr>     <chr>                               <chr>                   <chr>
## 1 cb774db~ I`d have responded, if I were going  I`d have responded, i~ neutral
## 2 549e992~ Sooo SAD I will miss you here in Sa~ Sooo SAD                negative
## 3 088c60f~ my boss is bullying me...            bullying me             negative
## 4 9642c00~ what interview! leave me alone       leave me alone          negative
## 5 358bd9e~ Sons of ****, why couldn`t they put~ Sons of ****,           negative
```

```
# Function from textclean package that gives recommendations
# on where to performs cleanups. very very helpful.
# I will show the output at the end of the report
Text_cleanup_recommendation <- check.Text(tweet_train)


# Take the loaded data through a cleaning process.
# Text column will be prepared to use in the sentiment analysis process
tweet_train <- clean.Data(tweet_train)

# Sample from the cleaned data now looks like this
# modeled_text is the text to source of sentiment analysis
head(tweet_train,5)
```

```
## # A tibble: 5 x 3
##   textID    sentiment modeled_text
##   <chr>     <chr>     <chr>
## 1 cb774db0~ neutral   I`d have responded, if I were going
## 2 549e992a~ negative  Sooo SAD I will miss you here in San Diego!!!
## 3 088c60f1~ negative  my boss is bullying me...
## 4 9642c003~ negative  what interview! leave me alone
## 5 358bd9e8~ negative  Sons of ****, why couldn`t they put them on the releases ~
```

Explore the data

```
# The main conclusion from this exploration is that neutral data seems to occupy most of the dataset.
# The data is very unpredictable. More insights can be drawn but it is very time consuming.
# In my opinion not much exploration we can do about the data itself in this context.
# new texts are very unpredictable and can be anything.
# Exploring the actual data helped in deciding what to clean.


# Show the dataset structure after cleaning
glimpse(tweet_train)
```

```
## Rows: 27,480
## Columns: 3
## $ textID       <chr> "cb774db0d1", "549e992a42", "088c60f138", "9642c003ef"...
## $ sentiment    <chr> "neutral", "negative", "negative", "negative", "negati...
## $ modeled_text <chr> "I`d have responded, if I were going", "Sooo SAD I wil...
```

```
# Records count broken down by sentiment.
# Neutral sentiment seems to have the majority of records
tweet_train %>% group_by(sentiment) %>% summarise(n())
```

```
## # A tibble: 3 x 2
##   sentiment `n()`
##   <chr>     <int>
## 1 negative   7781
## 2 neutral   11117
## 3 positive   8582
```

```
# More information on the breakdown
by(tweet_train, tweet_train$sentiment, summary)
```

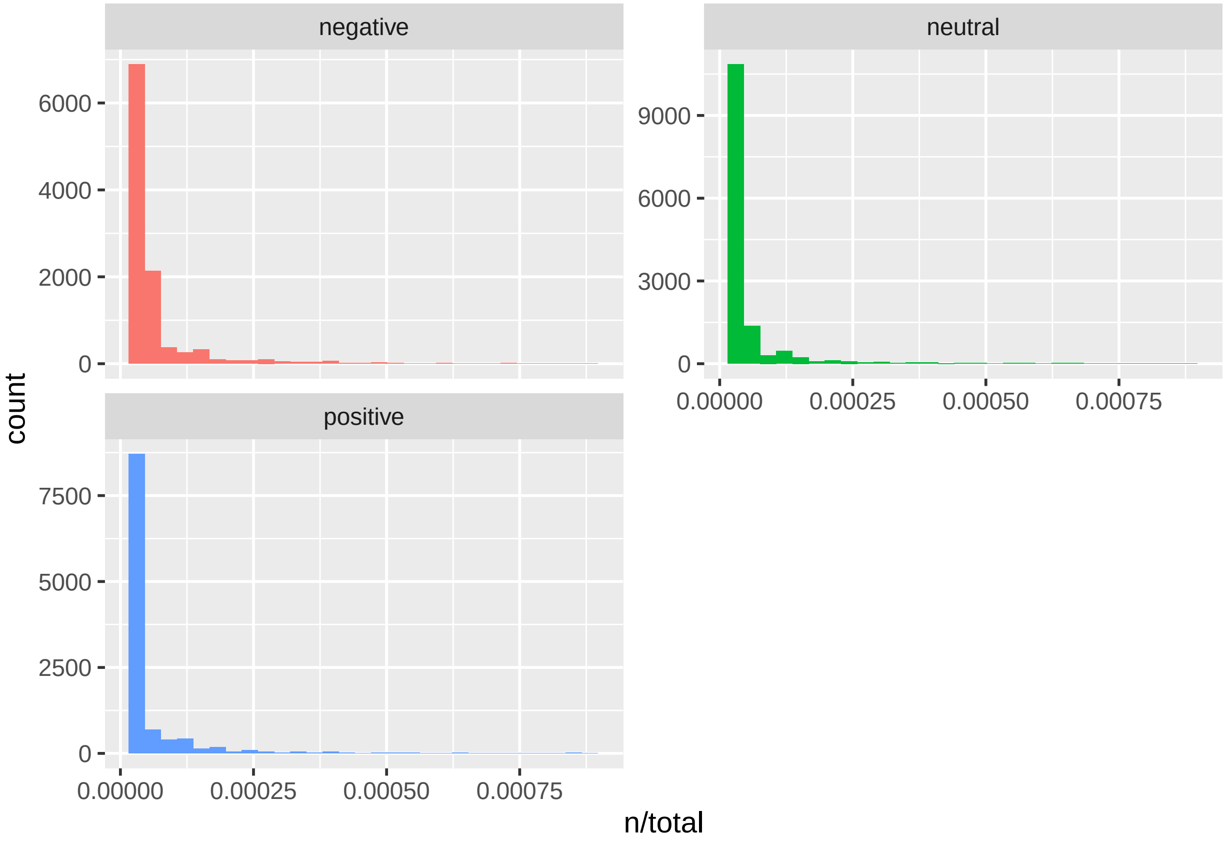
```
## tweet_train$sentiment: negative
##     textID           sentiment          modeled_text      
##  Length:7781        Length:7781        Length:7781       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
## ------------------------------------------------------------ 
## tweet_train$sentiment: neutral
##     textID           sentiment          modeled_text      
##  Length:11117       Length:11117       Length:11117      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
## ------------------------------------------------------------ 
## tweet_train$sentiment: positive
##     textID           sentiment          modeled_text      
##  Length:8582        Length:8582        Length:8582       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character
```

```
# Quick graph showing words count per sentiment
sentiment_words <- tweet_train %>%
unnest_tokens(word, modeled_text) %>% filter(!word %in% stop_words$word) %>%
count(sentiment, word, sort = TRUE)

#Tidy text to show a data split by sentiment.
total_words <- sentiment_words %>%
group_by(sentiment) %>%
summarize(total = sum(n))

sentiment_words <- left_join(sentiment_words, total_words)

# Ploting word counts by sentiment.
ggplot(sentiment_words, aes(n/total, fill = sentiment)) +
geom_histogram(show.legend = FALSE) +
xlim(NA, 0.0009) +
facet_wrap(~sentiment, ncol = 2, scales = "free_y")
```



Partition, train and validate training data

```
# create a partition 50% training and 50% test
# As mentioned before, this betters the validation numbers.
# Partition function retuns a list of two frames
dataPartitions <- partition.Data(tweet_train)

# assign the first dataframe in the list to training
train_set <- dataPartitions[[1]]

# training will run multiple algorthms and returns all results
train_pred <- train.model("train")

# Sample from the data output looks like this
head(as_tibble(train_pred),5)
```

```
## # A tibble: 5 x 9
##   textID modeled_text jockers  bing afinn   nrc vegnitte actual_sentiment
##   <chr>  <chr>          <dbl> <dbl> <dbl> <dbl>    <dbl>            <dbl>
## 1 549e9~ Sooo SAD I ~      -1    -1    -1     0       -1               -1
## 2 088c6~ my boss is ~      -1    -1    -1     0       -1               -1
## 3 28b57~ - some sham~       0     0     1    -1        0                0
## 4 50e14~ Soooo high         0     0     0     0        0                0
## 5 e0502~ Both of you        0     0     0     0        0                 0
## # ... with 1 more variable: sentimentr <dbl>
```

```
  # Now validation of the above output will take place.
  # The best result is decided on the accuracy figure.
  # mda (Mean Directional Accuracy) and mare(Mean Absolute Rescaled Error) are
  # also important, hence their inclusion in the output
  # More explanation found on https://rdrr.io/cran/sentimentr/man/validate_sentiment.html
  validation <- validate.Prediction(train_pred)

  # validation process will return a list of two dataframes
  # this is the first dataframe in the list and contains all validations.
  train_pred_validations <- as_tibble(validation[[1]])[,c(1,2,3,4,5,6,7,8)]

  # All validation frame look this this
  train_pred_validations
```

```
## # A tibble: 10 x 8
##    average precision recall accuracy     F   mda  mare method
##    <fct>       <dbl>  <dbl>    <dbl> <dbl> <dbl> <dbl> <chr>
##  1 macro       0.547  0.523    0.689 0.524 0.533 0.272 nrc
##  2 micro       0.533  0.533    0.689 0.533 0.533 0.272 nrc
##  3 macro       0.655  0.643    0.764 0.645 0.646 0.203 bing
##  4 micro       0.646  0.646    0.764 0.646 0.646 0.203 bing
##  5 macro       0.655  0.648    0.760 0.640 0.641 0.213 afinn
##  6 micro       0.641  0.641    0.760 0.641 0.641 0.213 afinn
##  7 macro       0.335  0.336    0.565 0.333 0.347 0.412 sentimentr
##  8 micro       0.347  0.347    0.565 0.347 0.347 0.412 sentimentr
##  9 macro       0.633  0.617    0.733 0.595 0.599 0.248 jockers
## 10 micro       0.599  0.599    0.733 0.599 0.599 0.248 jockers
```

```
  # this is the second dataframe in the list and contains the best validation results
  train_best_prediction <- as_tibble(validation[[2]])[,c(4,6,7,8)]

  # best prediction dataframe looks like this
  train_best_prediction
```

```
## # A tibble: 1 x 4
##   accuracy   mda  mare method
##      <dbl> <dbl> <dbl> <chr>
## 1    0.764 0.646 0.203 bing
```

```
  # Note
```

```
## [1] "the best model for training dataset is bing with accuracy of 76.374%"
```

```
 # Sample of the final output of the training data set
  head(as_tibble(train_pred),5)
```

```
## # A tibble: 5 x 3
##   textID     modeled_text                                                sentiment
##   <chr>      <chr>                                                 <chr>
## 1 549e992a42 Sooo SAD I will miss you here in San Diego!!!         negative
## 2 088c60f138 my boss is bullying me...                             negative
## 3 28b57f3990 - some shameless plugging for the best Rangers forum on ~ neutral
## 4 50e14c0bb8 Soooo high                                            neutral
## 5 e050245fbd Both of you                                           neutral
```

Load, train and validate test dataset

```
#----------------------testing the model-------------------------------

# Repeating steps applied when training the model on the test data set

# assign the second dataframe in the partition list to test
# this will create the test dataset
test_set <- dataPartitions[[2]]

test_pred <- train.model("test")

validation <- validate.Prediction(test_pred)

test_pred_validations <- as_tibble(validation[[1]])[,c(1,2,3,4,5,6,7,8)]
test_pred_validations
```

```
## # A tibble: 10 x 8
##    average precision recall accuracy     F   mda  mare method
##    <fct>       <dbl>  <dbl>    <dbl> <dbl> <dbl> <dbl> <chr>
##  1 macro       0.541  0.518    0.685 0.518 0.528 0.274 nrc
##  2 micro       0.528  0.528    0.685 0.528 0.528 0.274 nrc
##  3 macro       0.649  0.636    0.759 0.637 0.638 0.208 bing
##  4 micro       0.638  0.638    0.759 0.638 0.638 0.208 bing
##  5 macro       0.655  0.646    0.759 0.637 0.639 0.215 afinn
##  6 micro       0.639  0.639    0.759 0.639 0.639 0.215 afinn
##  7 macro       0.325  0.325    0.557 0.323 0.336 0.417 sentimentr
##  8 micro       0.336  0.336    0.557 0.336 0.336 0.417 sentimentr
##  9 macro       0.624  0.611    0.729 0.588 0.594 0.251 jockers
## 10 micro       0.594  0.594    0.729 0.594 0.594 0.251 jockers
```

```
test_best_prediction <- as_tibble(validation[[2]])[,c(4,6,7,8)]
test_best_prediction
```

```
## # A tibble: 1 x 4
##   accuracy   mda  mare method
##      <dbl> <dbl> <dbl> <chr>
## 1    0.759 0.639 0.215 afinn
```

```
# Note
```

```
## [1] "the best model for test dataset is afinn with accuracy of 75.945%"
## # A tibble: 5 x 3
##   textID    modeled_text                                                 sentiment
##   <chr>     <chr>                                                        <chr>
## 1 cb774db0~ I`d have responded, if I were going                          neutral
## 2 9642c003~ what interview! leave me alone                               negative
## 3 358bd9e8~ Sons of ****, why couldn`t they put them on the releases ~ neutral
## 4 6e0c6d75~ 2am feedings for the baby are fun when he is all smiles a~ positive
## 5 fc2cbefa~ Journey!? Wow... u just became cooler. hehe... (is that p~ positive
```

```
# This concludes the training and test process.
# The conclusion is that when running algorithms for training dataset the best score
# came from algorithm bing. However, afinn shows as the best score when running against test data.
# I have decided to go for bing as the accuracy it produced was higher than the afinn.
# Final model will run with bing algorithm.
```

load, clean and predict sentiment

```
#----------------------running the model on validation data set--------
#--------------------- Final prediction Begin --------------------------------

# load final dataset for prediction from Git
# to pick up file from local directory,change parameter
# remote into any other value such as loc
tweet_validation <- load.Data("validation.csv","remote")
```

```
## Parsed with column specification:
## cols(
##   textID = col_character(),
##   text = col_character(),
##   sentiment = col_character()
## )
```

```
# clean data
tweet_validation <- clean.Data(tweet_validation)

# the data to be predicted looks like this
as_tibble(tweet_validation)
```

```
## # A tibble: 3,534 x 2
##    textID     modeled_text
##    <chr>      <chr>
##  1 f87dea47db Last session of the day
##  2 96d74cb729 Shanghai is also really exciting (precisely -- skyscrapers galore~
##  3 eee518ae67 Recession hit Veronique Branquinho, she has to quit her company, ~
##  4 01082688c6 happy bday!
##  5 33987a8ee5 - I like it!!
##  6 726e501993 that`s great!! weee!! visitors!
##  7 261932614e I THINK EVERYONE HATES ME ON HERE lol
##  8 afa11da83f soooooo wish i could, but im in school and myspace is completely ~
##  9 e64208b4ef and within a short time of the last clue all of them
## 10 37bcad24ca What did you get? My day is alright.. haven`t done anything yet. ~
## # ... with 3,524 more rows
```

```r
# Perform prediction
prediction <- predict.Sentiment(tweet_validation)

# output the final prediction data frame
as_tibble(prediction)
```

```
## # A tibble: 3,534 x 3
##    textID    modeled_text                                             sentiment
##    <chr>     <chr>                                                    <chr>
##  1 f87dea47~ Last session of the day                                  neutral
##  2 96d74cb7~ Shanghai is also really exciting (precisely -- skyscrape~ positive
##  3 eee518ae~ Recession hit Veronique Branquinho, she has to quit her ~ negative
##  4 01082688~ happy bday!                                              positive
##  5 33987a8e~ - I like it!!                                            positive
##  6 726e5019~ that`s great!! weee!! visitors!                          positive
##  7 26193261~ I THINK EVERYONE HATES ME ON HERE lol                    negative
##  8 afa11da8~ soooooo wish i could, but im in school and myspace is co~ neutral
##  9 e64208b4~ and within a short time of the last clue all of them      neutral
## 10 37bcad24~ What did you get? My day is alright.. haven`t done anyth~ neutral
## # ... with 3,524 more rows
```

```r
#------------------------- Final prediction END ---------------------------------
```

text cleaning recommendation

```r
# Sample of the cleanup recommendations (as promised)
Text_cleanup_recommendation
```

```
##
##
## ===========
## CONTRACTION
## ===========
##
## The following observations contain contractions:
##
## 10440, 15127, 17778, 19642, 26589
##
## This issue affected the following text:
##
## 10440: How I met your mother and Scrubs in role! YEAH! 'Cause I`m FLY!
## 15127: OK its official I'M OLD! at least I feel likewise OLD & TIREDD & WASTED!!
## 17778: i dnt think i can ever get tired of'The Climb'its just 1 of those sngs u`ll always remember
## 19642: _Peabody I`m pretty sure we got sent home a couple of times too. `tis the week to remember Ma
## 26589: **** girl I`m so down but ya gotta let me know so I can get my kit together & I got a flyer
##
## *Suggestion: Consider running `replace_contraction`
##
##
## ====
## DATE
## ====
##
## The following observations contain dates:
##
## 1266, 1899, 2721, 5773, 6752, 7859, 12421, 16008, 16235, 19667...[truncated]...
##
## This issue affected the following text:
##
## 1266: http://naturalismo.files.wordpress.com/2008/01/elliott10.jpg my hero
## ...[truncated]...
## 1899: Went out to get groceries...prices are inflating  Gas went up another 10 cents to hit $2.49...
## ...[truncated]...
## 2721: IN $RF .94 - target $5.30.  OUT $DNDN @$21.85 near days low
## ...[truncated]...
## 5773: 08.05.09 partying at the Pineforest  http://tinyurl.com/ojugsb
## ...[truncated]...
## 6752: see you..... 08.08.09
## ...[truncated]...
## 7859: check out q100 right now..99.7
## ...[truncated]...
## 12421: self-portrait week http://unbecominglily.blogspot.com/2009/05/announcing.html  would you like
## ...[truncated]...
## 16008: Check this video out -- ScriptGirl Report 05.08.09 http://bit.ly/hclXP ... I can officially n
## ...[truncated]...
## 16235: http://79.170.44.101/buma.ro/ temporary address not working either, m8. They must have done s
## ...[truncated]...
## 19667: there it is. postieeee  http://andshehopes.blogspot.com/2009/05/kewpie.html
## ...[truncated]...
##
## *Suggestion: Consider running `replace date`
##
##
## =====
## DIGIT
## =====
##
## The following observations contain digits/numbers:
##
## 7, 15, 18, 21, 24, 25, 33, 36, 37, 44...[truncated]...
##
## This issue affected the following text:
##
## 7: 2am feedings for the baby are fun when he is all smiles and coos
## ...[truncated]...
## 15: test test from the LG enV2
## ...[truncated]...
## 18: i`ve been sick for the past few days  and thus, my hair looks wierd.  if i didnt have a hat on i
## ...[truncated]...
## 21: oh Marly, I`m so sorry!!  I hope you find her soon!! <3 <3
## ...[truncated]...
## 24: gotta restart my computer .. I thought Win7 was supposed to put an end to the constant rebootine
## ...[truncated]...
## 25: SEe waT I Mean bOuT FoLLOw fRiiDaYs... It`S cALLed LoSe fOLloWeRs FridAy... smH
## ...[truncated]...
## 33: If it is any consolation I got my BMI tested hahaha it says I am obesed  well so much for being u
## ...[truncated]...
## 36: Thats it, its the end. Tears for Fears vs Eric Prydz, DJ Hero   http://bit.ly/2Hpbg4
## ...[truncated]...
## 37: Born and raised in NYC and living in Texas for the past 10 years!  I still miss NY
## ...[truncated]...
## 44: RATT ROCKED NASHVILLE TONITE..ONE THING SUCKED, NO ENCORE!  LIKE IN THE 80`S THEY STILL HAVE A FU
## ...[truncated]...
##
## *Suggestion: Consider using `replace_number`
##
##
## ========
## EMOTICON
## ========
##
## The following observations contain emoticons:
##
## 6, 9, 11, 18, 21, 36, 44, 51, 53, 58...[truncated]...
##
## This issue affected the following text:
##
## 6: http://www.dothebouncy.com/smf - some shameless plugging for the best Rangers forum on earth
## ...[truncated]...
## 9: Both of you
## ...[truncated]...
## 11: as much as i love to be hopeful, i reckon the chances are minimal =P i`m never gonna get my cake
## ...[truncated]...
## 18: i`ve been sick for the past few days  and thus, my hair looks wierd.  if i didnt have a hat on it
## ...[truncated]...
## 21: oh Marly, I`m so sorry!!  I hope you find her soon!! <3 <3
## ...[truncated]...
## 36: Thats it, its the end. Tears for Fears vs Eric Prydz, DJ Hero   http://bit.ly/2Hpbg4
## ...[truncated]...
## 44: RATT ROCKED NASHVILLE TONITE..ONE THING SUCKED, NO ENCORE!  LIKE IN THE 80`S THEY STILL HAVE A FU
## ...[truncated]...
## 51: Then you should check out http://twittersucks.com and connect with other tweeple who hate twitter
## ...[truncated]...
## 53: hm... Both of us I guess...
## ...[truncated]...
## 58: will be back later.  http://plurk.com/p/rp3k7
## ...[truncated]...
##
## *Suggestion: Consider using `replace_emoticons`
##
##
## =======
## ESCAPED
## =======
##
## The following observations contain escaped back spaced characters:
##
## 658, 964, 1613, 3633, 9764, 10350, 10847, 11482, 13910, 25121...[truncated]...
```

```
##
## This issue affected the following text:
##
## 658: lol hi emmy, latin would help me study for the aptitude tests to get into grad school ;\ thats
## ...[truncated]...
## 964: mad the rain got me...now i cant go see jaiden   *|)|/-\|\|/-\*
## ...[truncated]...
## 1613: is hungryyyyyyy!! going to eat traditional indian food...the pakistani way. woowoo!  hahaha! >
## ...[truncated]...
## 3633: word to yer mother!!   \m/
## ...[truncated]...
## 9764: tres rude VICTOR!  :\
## ...[truncated]...
## 10350: I would if I was drivin :\ hahaha. but get me a Carol C. Special, yeah?
## ...[truncated]...
## 10847: Ahhhh \you are soo smart  Thanks for this schooling of thoughts
## ...[truncated]...
## 11482: Ahhhh \you are soo smart  Thanks for this schooling of thoughts  Have you taught before?
## ...[truncated]...
## 13910: how bad has life gotten where u  werecounting on the church50\50 raffle? the answer is real ba
## ...[truncated]...
## 25121: I had a horrible dream  it had to do with a scary face, now im awake for the rest of the nigh
## ...[truncated]...
##
## *Suggestion: Consider using `replace_white`
##
##
## ====
## HASH
## ====
##
## The following observations contain Twitter style hash tags (e.g., #rstats):
##
## 59, 71, 167, 180, 252, 253, 258, 264, 316, 560...[truncated]...
##
## This issue affected the following text:
##
## 59: Aw. Torn ace of hearts  #Hunchback
## ...[truncated]...
## 71: I still smell of smoke  #kitchenfire
## ...[truncated]...
## 167: #lichfield #tweetup sounds like fun  Hope to see you and everyone else there!
## ...[truncated]...
## 180: All the cool people I want to find for following today are #English, and I guess the English do
## ...[truncated]...
## 252: #powerblog What is this powerblog challenge you keep talking about?  I`m a newbie follower
## ...[truncated]...
## 253: almost died. Laptop screen was set to 100% brightness after I reinstalled Windows Vista. Got a
## ...[truncated]...
## 258: discovered cause of a bug in the new #NetPLAYER 4 build. Publishing bug fix now, hopefully new
## ...[truncated]...
## 264: good news: finally finished my #EASactive workout that has been paused for 6 hours. bad news: my
## ...[truncated]...
## 316: It looks like the office TV DOES get MLB Network... and it looks like MLBN will NOT be televisi
## ...[truncated]...
## 560: there were attempts to somehow extend inner classes, which would be close to #closure, can`t fi
## ...[truncated]...
##
## *Suggestion: Consider using `qdapRegex::ex_tag' (to capture meta-data) and/or replace_hash
##
##
## ====
## HTML
## ====
##
## The following observations contain HTML markup:
##
## 1784, 9225, 12251, 15295, 17358, 18684, 19889, 21154, 23646, 25222...[truncated]...
##
## This issue affected the following text:
##
## 1784: on jacksonville beach walking in the cold **** water   but have to work in the morning   ily <
## ...[truncated]...
## 9225: Allergies suck ducks nuts.     <=====8=====>
## ...[truncated]...
## 12251: Whoishonorsociety <never wear your pajama pants to school  >
## ...[truncated]...
## 15295: Good nite everybody!  <:Baby Boy:>
## ...[truncated]...
## 17358: ) no more Chemistry!!! I`m gonna choose English. I find it (Chem.) kinda boring in the end! g
## ...[truncated]...
## 18684: Welll my folkiesss(; im offf to dream land;work in the mornin;ugh;ewwy. Talkkk to me;but tomo
## ...[truncated]...
## 19889: I like fridays generally, but class is extended today  and I`m starving :X haha </whine> O:-P
## ...[truncated]...
## 21154: I never get them and the hubby is due in next week.. <cries> hope its gone by then
## ...[truncated]...
## 23646: fireworks @ KBOOM concert... second best I`ve ever seen...preceded only by last year`s show 2
## ...[truncated]...
## 25222: Oh, ouch. That hurt.  **** washing machine is out to get me!  <.<    >.>
## ...[truncated]...
##
## *Suggestion: Consider running `replace_html`
##
##
## ==========
## INCOMPLETE
## ==========
##
## The following observations contain incomplete sentences (e.g., uses ending punctuation like '...'):
##
## 3, 10, 18, 23, 24, 25, 28, 31, 39, 44...[truncated]...
##
## This issue affected the following text:
##
## 3: my boss is bullying me...
## ...[truncated]...
## 10: Journey!? Wow... u just became cooler.  hehe... (is that possible!?)
## ...[truncated]...
## 18: i`ve been sick for the past few days  and thus, my hair looks wierd.  if i didnt have a hat on i
## ...[truncated]...
## 23: is cleaning the house for her family who is comming later today..
## ...[truncated]...
## 24: gotta restart my computer .. I thought Win7 was supposed to put an end to the constant rebootines
## ...[truncated]...
## 25: SEe waT I Mean bOuT FoLLOw fRiiDaYs... It`S cALLed LoSe fOLloWeRs FridAy... smH
## ...[truncated]...
## 28: On the way to Malaysia...no internet access to Twit
## ...[truncated]...
## 31: I`m going home now. Have you seen my new twitter design? Quite....heavenly isn`****?
## ...[truncated]...
## 39: i`m soooooo sleeeeepy!!! the last day o` school was today....sniffle....
## ...[truncated]...
## 44: RATT ROCKED NASHVILLE TONITE..ONE THING SUCKED, NO ENCORE!  LIKE IN THE 80`S THEY STILL HAVE A FU
## ...[truncated]...
##
## *Suggestion: Consider using `replace_incomplete`
##
##
## ====
## KERN
## ====
##
## The following observations contain kerning (e.g., 'The B O M B!'):
##
## 25, 44, 205, 677, 787, 996, 1394, 1620, 1925, 2785...[truncated]...
##
## This issue affected the following text:
##
## 25: SEe waT I Mean bOuT FoLLOw fRiiDaYs... It`S cALLed LoSe fOLloWeRs FridAy... smH
## ...[truncated]...
## 44: RATT ROCKED NASHVILLE TONITE..ONE THING SUCKED, NO ENCORE!  LIKE IN THE 80`S THEY STILL HAVE A FU
## ...[truncated]...
## 205: I AM SUCH A CREEPER  I feel disappointed because of it. **** my cyberstalking skills   the inter
## ...[truncated]...
## 677: _xo they were so pretty and took like an hour to do  CAN I DO URSSSSS!
## ...[truncated]...
## 787: OH NEVERMIND I THINK THIS THING IS UNSALVAGEABLE
## ...[truncated]...
## 996: I am twittering, LIKE A BOSS. Thanks Savvv
## ...[truncated]...
## 1394: her son is 7 and captured it outside...THANK GOD I HAVE A LITTLE GIRL
## ...[truncated]...
## 1620: WiSHiNG ALL THE MOTHERS A HAPPY MOTHER`S DAY
## ...[truncated]...
## 1925: IM SOWWIE I WAS A LIL LATE  LOL it looked good though ;)
## ...[truncated]...
## 2785: aww , its ok,we ended up getting in later than expected and didnt go...I would of called you i
## ...[truncated]...
##
## *Suggestion: Consider using `replace_kern`
##
```

```
##
## =============
## MISSING VALUE
## =============
##
## The following observations contain missing values:
##
## 315
##
## *Suggestion: Consider running `drop_NA`
##
##
## ==========
## MISSPELLED
## ==========
##
## The following observations contain potentially misspelled words:
##
## 2, 5, 6, 8, 10, 13, 15, 18, 21, 22...[truncated]...
##
## This issue affected the following text:
##
## 2: <<<<S<<oo>>>>o>> SAD I <<wi>>ll m<<i<<ss>>>> you here in <<Sa>>n Di<<eg>>o!!!
## ...[truncated]...
## 5: S<<o<<ns>>>> of ****, <<wh>>y <<c<<ould>>n>>`t <<th>>ey p<<ut>> <<th>>em on <<th>>e r<<<<el>>e>>a
## ...[truncated]...
## 6: <<ht<<tp>>>>://<<www>>.<<d<<<<ot>>h>><<eb>>o<<un>>cy>>.com/<<s<<mf>>>> - <<som>>e sha<<m<<el>>>>e
## ...[truncated]...
## 8: <<S<<<<<<oo>>o>>o>>>> hi<<gh>>
## ...[truncated]...
## 10: <<Jo<<ur>>>><<ne>>y!? Wow... u j<<ust>> <<b<<ec>>>><<ame>> c<<oo>><<le>>r.  <<<<heh>>e>>... (is
## ...[truncated]...
## 13: My S<<har>>p<<ie>> is <<r<<<<un>><<ni>>>>n>>g <<DANGE<<Ro>>u<<sl>>y>> low on i<<nk>>
## ...[truncated]...
## 15: <<te>>st <<te>>st f<<ro>>m <<th>>e LG <<enV>>2
## ...[truncated]...
## 18: i`<<ve>> b<<ee>>n s<<<<ic>>k>> <<fo>>r <<th>>e past f<<ew>> d<<ay>>s  a<<nd>> t<<hus>>, my <<h<<
## ...[truncated]...
## 21: oh <<Ma<<r<<ly>>>>>>, I`m so s<<o<<rr>>>>y!!  I ho<<pe>> you <<fi>><<nd>> her <<s<<oo>>>>n!! <3
## ...[truncated]...
## 22: <<Pla<<yi>>n>>g Gh<<os>>t On<<li>><<ne>> is r<<ea<<l<<ly>>>>>> i<<<<nt>>e>>re<<sti>><<ng>>. The
## ...[truncated]...
##
## *Suggestion: Consider running `hunspell::hunspell_find` & `hunspell::hunspell_suggest`
##
##
## ========
## NO ALPHA
## ========
##
## The following observations contain elements with no alphabetic (a-z) letters:
##
## 8121, 26006
##
```

```
## This issue affected the following text:
##
## 8121: ****
## 26006: ?
##
## *Suggestion: Consider cleaning the raw text or running `filter_row`
##
##
## ==========
## NO ENDMARK
## ==========
##
## The following observations contain elements with missing ending punctuation:
##
## 1, 4, 5, 6, 7, 8, 9, 10, 11, 12...[truncated]...
##
## This issue affected the following text:
##
## 1: I`d have responded, if I were going
## ...[truncated]...
## 4: what interview! leave me alone
## ...[truncated]...
## 5: Sons of ****, why couldn`t they put them on the releases we already bought
## ...[truncated]...
## 6: http://www.dothebouncy.com/smf - some shameless plugging for the best Rangers forum on earth
## ...[truncated]...
## 7: 2am feedings for the baby are fun when he is all smiles and coos
## ...[truncated]...
## 8: Soooo high
## ...[truncated]...
## 9: Both of you
## ...[truncated]...
## 10: Journey!? Wow... u just became cooler.  hehe... (is that possible!?)
## ...[truncated]...
## 11: as much as i love to be hopeful, i reckon the chances are minimal =P i`m never gonna get my cake
## ...[truncated]...
## 12: I really really like the song Love Story by Taylor Swift
## ...[truncated]...
##
## *Suggestion: Consider cleaning the raw text or running `add_missing_endmark`
##
##
## ====================
## NO SPACE AFTER COMMA
## ====================
##
## The following observations contain commas with no space afterwards:
##
## 246, 295, 472, 500, 570, 583, 868, 943, 1132, 1278...[truncated]...
##
## This issue affected the following text:
##
## 246: yeah I was thinking about that ,ahaha
## ...[truncated]...
## 295: Family is here,hanging with them
## ...[truncated]...
## 472: The birds are out,, oh man... That`s NOT cool && I didn`t sleep yet for the night!!!
## ...[truncated]...
## 500: Woke Up,I Wanna Stay In My bed
## ...[truncated]...
## 570: I don`t think I`ve ever been so tierd in my life.Ugh,goodnight.So sleeping in tomorrow
## ...[truncated]...
## 583: http://twitpic.com/4v0vr - cool! you and your mother have awesome hair styles,! Wish her a happ
## ...[truncated]...
## 868: Morning tweeple,way to early again
## ...[truncated]...
## 943: _babe woo! im getting mine on monday,cant wait  x
## ...[truncated]...
## 1132: Why are young people attracted to trouble? this makes me sad!  ,<3 kMv
## ...[truncated]...
## 1278: http://twitpic.com/4vuuy - thats the most colorful thing ive seen all day,wow.
## ...[truncated]...
##
## *Suggestion: Consider running `add_comma_space`
##
##
## =========
## NON ASCII
## =========
##
## The following observations contain non-ASCII text:
##
## 45, 193, 433, 646, 855, 951, 961, 1087, 1157, 1497...[truncated]...
##
## This issue affected the following text:
##
## 45: I love to! But I`m only available from 5pm.  and where dear? Would love to help  convert her vid
## ...[truncated]...
## 193: *phew*  Will make a note in case anyone else runs into the same issueï¿½
## ...[truncated]...
## 433: I love mine, too . happy motherï¿½s day to your mom , John Taylor  . much love to you, too .
## ...[truncated]...
## 646: meeting just in time that iï¿½m trying to win something  prize`s friday!
## ...[truncated]...
## 855: Just got confirmed that itï¿½s pizza-time with some ex co-workers on friday...looking forward t
## ...[truncated]...
## 951: Well thatï¿½s disappointing to hear.
## ...[truncated]...
## 961: today Jon Doe plays at the Moho. ia m excited  itï¿½s gonna be funny.. but before i have to car
## ...[truncated]...
## 1087: _Guy Would luv to hear music too but iï¿½m out of batteries - the tv plays besides but i think
## ...[truncated]...
## 1157: CCï¿½s video for Long Gone premiers today on yahoo, dont miss it: http://new.music.yahoo.com/v
## ...[truncated]...
## 1497: vocï¿½ que sumiu forever do msn.
## ...[truncated]...
##
## *Suggestion: Consider running `replace_non_ascii`
```

```
##
##
## ==================
## NON SPLIT SENTENCE
## ==================
##
## The following observations contain unsplit sentences (more than one sentence per element):
##
## 4, 10, 18, 21, 22, 24, 25, 28, 29, 31...[truncated]...
##
## This issue affected the following text:
##
## 4: what interview! leave me alone
## ...[truncated]...
## 10: Journey!? Wow... u just became cooler.  hehe... (is that possible!?)
## ...[truncated]...
## 18: i`ve been sick for the past few days  and thus, my hair looks wierd.  if i didnt have a hat on i
## ...[truncated]...
## 21: oh Marly, I`m so sorry!!  I hope you find her soon!! <3 <3
## ...[truncated]...
## 22: Playing Ghost Online is really interesting. The new updates are Kirin pet and Metamorph for thir
## ...[truncated]...
## 24: gotta restart my computer .. I thought Win7 was supposed to put an end to the constant rebootine
## ...[truncated]...
## 25: SEe waT I Mean bOuT FoLLOw fRiiDaYs... It`S cALLed LoSe fOLloWeRs FridAy... smH
## ...[truncated]...
## 28: On the way to Malaysia...no internet access to Twit
## ...[truncated]...
## 29: juss came backk from Berkeleyy ; omg its madd fun out there  havent been out there in a minute .
## ...[truncated]...
## 31: I`m going home now. Have you seen my new twitter design? Quite....heavenly isn`****?
## ...[truncated]...
##
## *Suggestion: Consider running `textshape::split_sentence`
##
##
## ===
## TAG
## ===
##
## The following observations contain Twitter style handle tags (e.g., @trinker):
##
## 989, 2473, 2937, 3116, 3217, 3858, 4539, 5124, 5385, 6342...[truncated]...
##
## This issue affected the following text:
##
## 989: nite nite twitts i wish u all a happy sunday i already have my major gift my my 2kids-my bro-n
## ...[truncated]...
## 2473: @_josh_thomas are you coming to sydney?! cool, where can i meet you? id love to meet you, you`
## ...[truncated]...
## 2937: @_katieedwards I can`t yet back, I`ve run out of texts!  I`ll ring you laters xoxo
## ...[truncated]...
## 3116: @_catchfire Happy Birthday Chip`s sister
## ...[truncated]...
## 3217: @_supernatural_ more Demon Sam!! I need it to numb the pain  ****
## ...[truncated]...
## 3858: @_handz_ well you know those 'kind of guys' are just idiots
## ...[truncated]...
## 4539: @_refugee_ /me gets 'Your video will start in 15 seconds', Exiting to watch ... for minutes  #
## ...[truncated]...
## 5124: @_elj OK nice one, cheers boss.  Am liking the lack of FCS today.
## ...[truncated]...
## 5385: @_enzo blech... thats a fail when you`re receiving dollars. I quoted a job in USD last month.
## ...[truncated]...
## 6342: @_finn_ Except I dropped him on my break and now he`s got a ding in his side.  At least it was
## ...[truncated]...
##
## *Suggestion: Consider using `qdapRegex::ex_tag' (to capture meta-data) and/or `replace_tag`
##
##
## ====
## TIME
## ====
##
## The following observations contain timestamps:
##
## 414, 596, 647, 889, 990, 1009, 1096, 1378, 1751, 2418...[truncated]...
##
## This issue affected the following text:
##
## 414: I sure do hope it becomes 4:20 this afternoon ...
## ...[truncated]...
## 596: Just got back from working out. I`m feeling pretty good. work at 4:30
## ...[truncated]...
## 647: Waiting for 5:00 & having cramps
## ...[truncated]...
## 889: Woke up at 7:50 then fell back to sleep. Woke up at 8:50 and back to sleep again. Woke up at 9:
## ...[truncated]...
## 990: Omg. Its 1:47 am and Kim Possible is on Disney Channel right now. I am glued to the screen
## ...[truncated]...
## 1009: Laughing for no reason...maybe its because its 2:27 and I`m tired, haha. Maybe i should go to
## ...[truncated]...
## 1096: Drink #2: And at 12:45pm when leaving the shops I had a Medium Light Coffee Frappuccino. Nom n
## ...[truncated]...
## 1378: Its 11:11...make a wish!
## ...[truncated]...
## 1751: Yep, tomorrow night, 10:30! Just saw the ad *squeals* LOVED this season
## ...[truncated]...
## 2418: Ugh soooo much work to do today while trying to make the 6:10 train to the game... looking lik
## ...[truncated]...
##
## *Suggestion: Consider using `replace_time`
##
##
## ===
## URL
## ===
##
```

```
## The following observations contain URLs:
##
## 6, 18, 36, 51, 58, 65, 173, 202, 216, 229...[truncated]...
##
## This issue affected the following text:
##
## 6: http://www.dothebouncy.com/smf - some shameless plugging for the best Rangers forum on earth
## ...[truncated]...
## 18: i`ve been sick for the past few days  and thus, my hair looks wierd.  if i didnt have a hat on i
## ...[truncated]...
## 36: Thats it, its the end. Tears for Fears vs Eric Prydz, DJ Hero   http://bit.ly/2Hpbg4
## ...[truncated]...
## 51: Then you should check out http://twittersucks.com and connect with other tweeple who hate twitte
## ...[truncated]...
## 58: will be back later.  http://plurk.com/p/rp3k7
## ...[truncated]...
## 65: mannnn..... _ got an iphone!!! im jealous....  http://bit.ly/NgnaR
## ...[truncated]...
## 173: URL in previous post (to timer job) should be http://bit.ly/a4Fdb. I`d removed space which mess
## ...[truncated]...
## 202: http://twitpic.com/66xlm -  hate when my PARKED car gets hit
## ...[truncated]...
## 216: yellow for   ? http://blip.fm/~5z05g
## ...[truncated]...
## 229: 35mins through the 1hr 20mins Google Wave demo, that looks a lot of fun, would love to test it
## ...[truncated]...
##
## *Suggestion: Consider using `replace_url`
```