Lupus Symptoms on Twitter

Sabrina Kent

Introduction

When studying immunology, I was intrigued by the way diseases like rheumatoid arthritis and lupus work. Lupus, in particular, is fascinating - not ony its mechanism of action, but all the factors leading to difficulty in diagnosing patients correctly. A simple google on lupus hits on the basics:

- It can take years for a patient to be diagnosed with lupus
- Lupus can "masquerade" as other diseases and cause great frustration (and worse)
- A huge part of the issue is the wide variety - and therefore the patient profile inconsistency - of symptoms

Digging deeper, I went to lupus.org and, from there, to their lupus message board. Searching for posts about patients exhibiting unusual symptoms led to something that floored me - according to many posts, lupus patients struggle with getting their doctors to acknowledge that their less typical symptoms are, indeed, related to lupus. However, many of these same posts had an astounding number of responses from other patients confirming that the "atypical" symptom being discussed was actually experienced by many members of the online lupus community.

That seems like an important disconnection between what lupus patients experience and what established medical knowledge says lupus does to a patient.

I'd like very much to collect data on patient-reported symptoms (specifically, symptoms that the patients feel are connected to their lupus diagnosis) and connect that with medical papers on lupus. However, data mining directly from a message board like the one I used to collect anecdotal evidence of an interesting problem would not be ethical.

So I went to twitter with my favorite r package for this kind of work, rtweet. (Kearney MW (2019). rtweet: Collecting Twitter Data. R package version 0.6.9, https://cran.r-project.org/package=rtweet.)

My official starting question is: is Twitter a good source for data on patient-reported lupus symptoms?

The more specific question, of course is about patient-reported lupus symptoms that are not considered to be standard symptoms of lupus by the medical community, but I'm going to start by casting a wide net and doing an exploratory analysis.

```
library(ggplot2)
library(tidyr)
library(rtweet)
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

##Collect data from twitter

To download a data frame of tweets including "lupus" as a hashtag requires a few minutes work with rtweet. Because of twitter's terms of use, I won't be including the actual tweets I collected, but will be able to share the resulting statistical information.

```
og_df <- search_tweets(
  "#lupus", n = 18000, include_rts = FALSE
)
```

Remove unneeded columns

Of the 90 variables downloaded automatically, only a few are potentially useful. At this point I want to keep information on the actual tweet and its hashtags, the account it came from, information on popularity - and if there is an URL attached to the tweet.

```
lupus_df <- select(og_df, screen_name, text, favorite_count, hashtags, urls_expanded_url)
```

Remove tweets from organizations linking to articles

I kept the variable containing linked URLs, because many of the tweets I first downloaded seemed to be generated by organizations and served as advertisements back to articles on other websites about lupus. Those tweets are not relevant here, and removing them will clean up my data nicely.

I want to keep tweets that don't have links in them.

So I simply filtered my dataframe for only tweets for which the URL variable is "NA," easily increasing the proportion of tweets from individuals rather than organizations and, particularly, removing tweets that don't discuss lupus directly.

```
no_orgs <- filter(lupus_df, is.na(urls_expanded_url))
```

Create new columns

I'm going to start my exploration by testing for the proportion of my tweets that contain interesting words:

- "fibromyalgia" is another difficult-to-diagnose disease that is often comorbid with lupus
- "flare" is the term used by both patients and medical professionals for times when disease activity increases and more symptoms are seen
- "symptom" - is the magic word!

```
no_orgs <- no_orgs %>%
  mutate(contains_flare = grepl("flare", text, fixed = TRUE))

no_orgs <- no_orgs %>%
  mutate(contains_fibro = grepl("fibro", text, fixed = TRUE))

no_orgs <- no_orgs %>%
  mutate(contains_symptom = grepl("symptom", text, fixed = TRUE))
```

Summarize data

What percent of tweets contain lupus + fibro, lupus + flare, lupus + symptoms - fibro, lupus + flare + symptoms, lupus - flare + symptoms

```
num_flare = sum(no_orgs$contains_flare)
num_fibro = sum(no_orgs$contains_fibro)
num_symptom = sum(no_orgs$contains_symptom)

flare_and_symptom = sum(no_orgs$contains_symptom & no_orgs$contains_flare)
```

```
no_flare_yes_symptom = sum(no_orgs$contains_symptom & !no_orgs$contains_flare)

num_flare
```

```
## [1] 16
```

```
num_fibro
```

```
## [1] 23
```

```
num_symptom
```

```
## [1] 11
```

```
flare_and_symptom
```

```
## [1] 0
```

```
no_flare_yes_symptom
```

```
## [1] 11
```

So only 17 tweets in my data frame contain the word “flare” and only 19 contain “symptom”? That’s not very helpful - unless my dataframe is somehow very tiny? I should have checked the dimensions immediately. It’s never too late!

```
dim(og_df)
```

```
## [1] 1195   90
```

```
dim(no_orgs)
```

```
## [1] 515   8
```

```
100 - dim(no_orgs)/dim(og_df) * 100
```

```
## [1] 56.90377 91.11111
```

So almost 60% of my lupus-related tweets contained hyperlinks in them.

Interesting Note: Over half of the sampled lupus-related tweets were from organizations, not individuals

That’s not very encouraging as far as my original question, but it does make sense. I collected a sampling of tweets based only on whether they contained #lupus. And twitter contains constant conversations of every imaginable type. I should be expecting a small proportion of the original dataset to contain tweets about symptoms.

Collecting a larger data set over time will increase the number of tweets pertinent to my question, but I’m going to keep exploring to set up a reusably analysis.

Examine hashtags in tweet collection

My next step is to collect all the hashtags in the dataset and rank them by frequency, creating a dataframe of tags.

```
#create a vector from every element in the hashtags column of no_orgs
all_tags <- c(no_orgs$hashtags, recursive = TRUE)
all_tags <- tolower(all_tags)

#creates a vector of length all_tags with FALSE at
#indices of first occurrence of a tag, TRUE else
dups_index <- duplicated(all_tags)
```

```r
#create a dataframe holding each tag in column 1, number
#of times that tag occurs in twitter data in column 2
tag_freq_df <- as.data.frame(table(all_tags))
popular_tag_freq_df <- tag_freq_df[tag_freq_df$Freq > 1, ]

#sort by tag frequency
popular_tag_freq_df <- popular_tag_freq_df %>%
  arrange(desc(Freq))
```

```r
#the first row will be for the most frequent tag - lupus. Remove it.
popular_tag_freq_df <- popular_tag_freq_df[-c(1),]
names(popular_tag_freq_df)
```

```
## [1] "all_tags" "Freq"
```

```r
dim(popular_tag_freq_df)
```

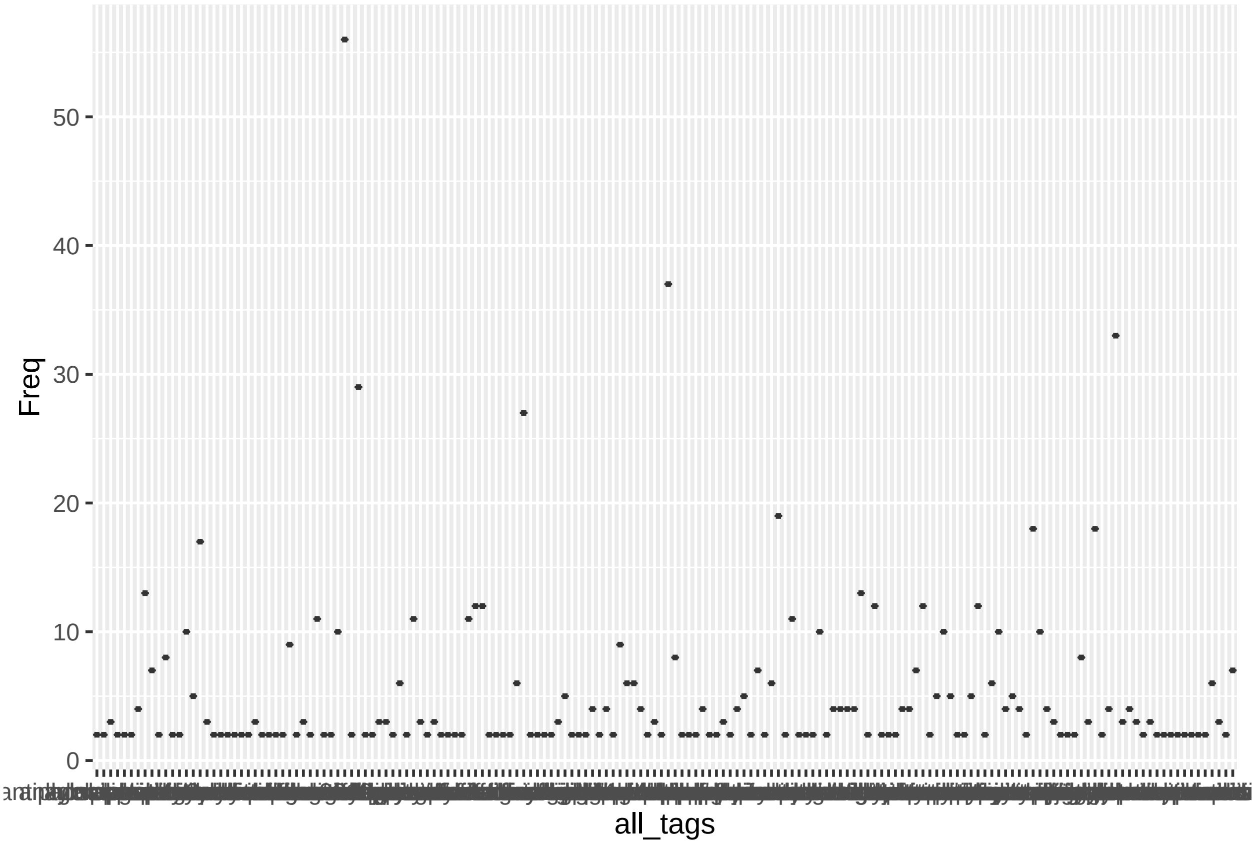
```
## [1] 166   2
```

I now have a data frame containing under 200 individual hashtags (after removing #lupus, of course), and the number of times each occurred.

##Create a graph to show which hashtags are most popular

It's high time for some graphics. When creatting my data frame of hashtags, I left out any tag that only occurred once. It will be very interesting to see if most of the tags left occur many times, or if the reverse is true - many tags occur relatively infrequently.

```r
#visualize the frequencies of tags that occur more than once
ggplot(popular_tag_freq_df, aes(x = all_tags, y = Freq)) +
  geom_boxplot()
```



That graph is much too dense to read - but is it useless? It does show that:

- There are a LOT of tags
- Many of these tags occur 10 times or less (in 10 or fewer tweets)
- It would be helpful to know exactly how many tweets and how many tags we have now

```
#How many rows (each row representing a tweet) are in the processed source dataframe?
dim(no_orgs)
```

```
## [1] 515   8
```

```
#How many rows (each row representing a hashtag) are in our dataframe?
dim(popular_tag_freq_df)
```

```
## [1] 166   2
```

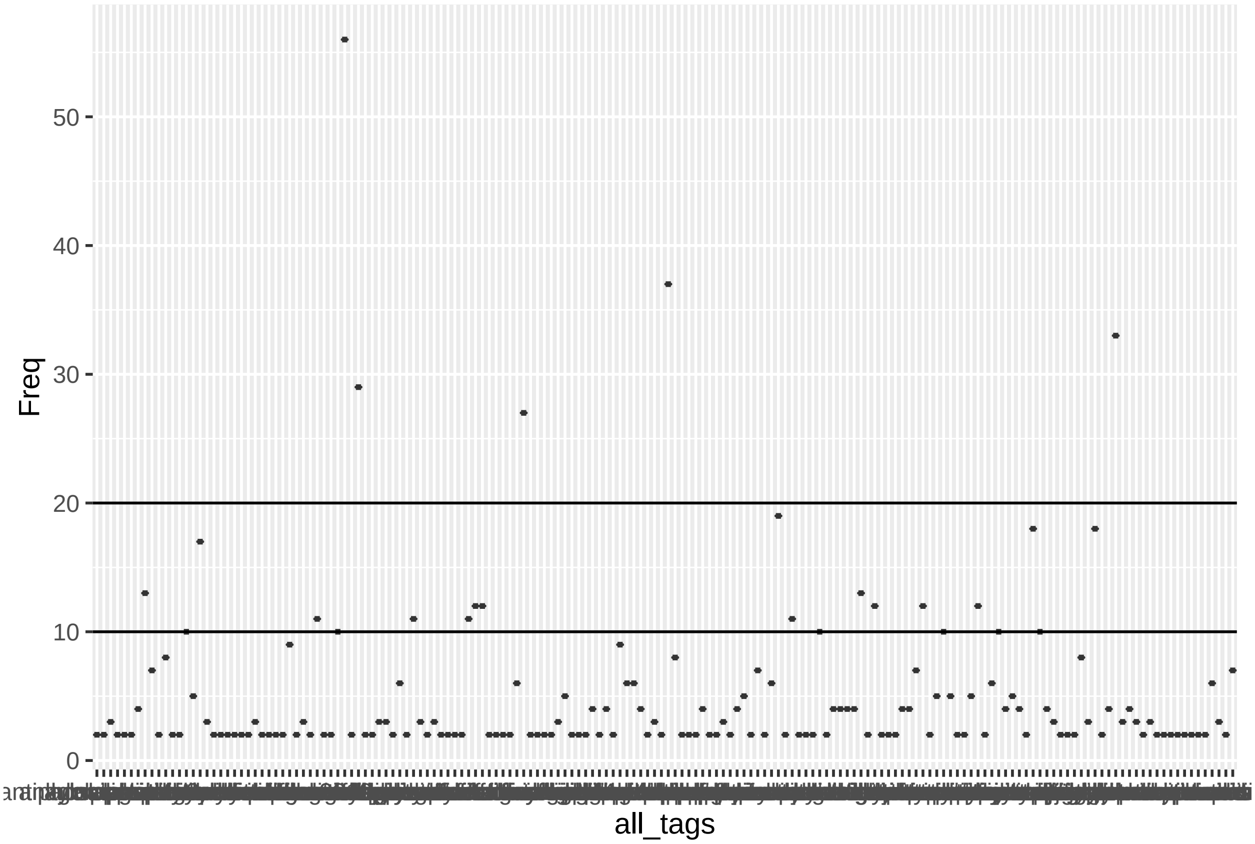
```
#If a tweet occurs 10 times (in 10 tweets), what percent of collected tweets is it appearing in?
100 * 10/dim(no_orgs)
```

```
## [1]   1.941748 125.000000
```

Answer: not many at all. Looking again at that graph, we have a choice to make about our priorities - do we want to see tweets that occur 10 times or more, or tweets that occur 20 times or more (a stricter criterion would give us only tags that are relatively very popular, but there are so few of those that we might lose information)?

Adding horizontal lines at y = 10 and y = 20 to visualize this might help.

```
ggplot(popular_tag_freq_df, aes(x = all_tags, y = Freq)) +
  geom_boxplot() + geom_hline(yintercept = 10) + geom_hline(yintercept = 20)
```



Removing only tags that occur 20 or more times would show us the most popular few, but there's a good chance those could be, by their popularity, very generic:

```
#peeking at the top few tags
head(popular_tag_freq_df)
```

```
##          all_tags Freq
## 2 chronicillness   56
## 3 lupusawareness   37
## 4         spoonie   33
## 5     chronicpain   29
## 6    fibromyalgia   27
## 7    lupuswarrior   19
```

Yes. #chronicillness, #chronicpain, #lupusawareness, and #spoonie are really only useful as other identifiers in the same way #lupus is. An argument could be made for discarding not only tweets with frequency of <10, but also discarding those with frequency of >20 for their over-generalizations.

However, #fibromyalgia and #antiphosphlipidsyndrome are autoimmune diseases and may connect us to interesting tweets for questions of dual-diagnosis or mistaken-diagnosis issues. Since this project is to explore the nature of tweets containing #lupus, it wouldn't be wise to prune away possibly useful data.

Once again, the high variability in the specific subject/purpose of tweets on lupus means that repeating this analysis with a much bigger dataset would be very interesting.

Let's go back to the tag dataframe and remove any with a frequency lower than 10.

Final tag selection

```
final_tag_freq_df <- popular_tag_freq_df[popular_tag_freq_df$Freq >= 10, ]
dim(final_tag_freq_df)
```

```
## [1] 26  2
```

Our final data set contains only about 30 tags. That's small enough to list right here:

```
final_tag_freq_df
```

```
##                 all_tags Freq
## 2         chronicillness   56
## 3         lupusawareness   37
## 4                spoonie   33
## 5            chronicpain   29
## 6           fibromyalgia   27
## 7           lupuswarrior   19
## 8    rheumatoidarthritis   18
## 9                    sle   18
## 10     autoimmunedisease   17
## 11               anxiety   13
## 12 mentalhealthawareness   13
## 13               educate   12
## 14         endometriosis   12
## 15              migraine   12
## 16     multiplesclerosis   12
## 17                  pots   12
## 18                celiac   11
## 19               disease   11
## 20                   eds   11
## 21           lymedisease   11
## 22            autoimmune   10
## 23        chronicfatigue   10
## 24                  mcas   10
## 25            narcolepsy   10
## 26           putonpurple   10
## 27          rheumatology   10
```

Conclusion

Examining these tags reveals some general patterns in the subject of the connected tweets, but shows very few possible symptoms, much less unusual symptoms. The final tag in this set, narcolepsy, is perhaps the most interesting one. By going back to the tweet dataframe and selecting only tweets using #narcolepsy, I could read those tweets (there are only ten, after all) and see if narcolepsy is being reported as a lupus-related disorder.

Discussion

It's not surprising to discover that the wide variety in reasons people tweet about lupus has made it difficult to find information on a very specific question by scraping Twitter. However, beginning again with rtweet's features for collecting tweets over time and building a much, much larger dataset that way could, theoretically, help me. Current computing resources don't allow this, but I do look forward to returning to this data collection and analysis file.