Lekta AI recruitment challenge

by Xavier Sułkowski

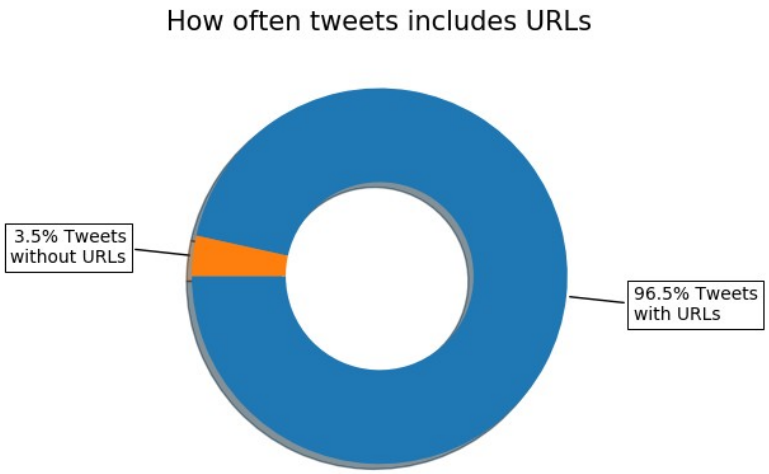
1. Task description

Purpose of the task is to perform an Extensive Data Analysis of text data from *Health News in Twitter Data Set*, focusing on finding the topics of tweets.

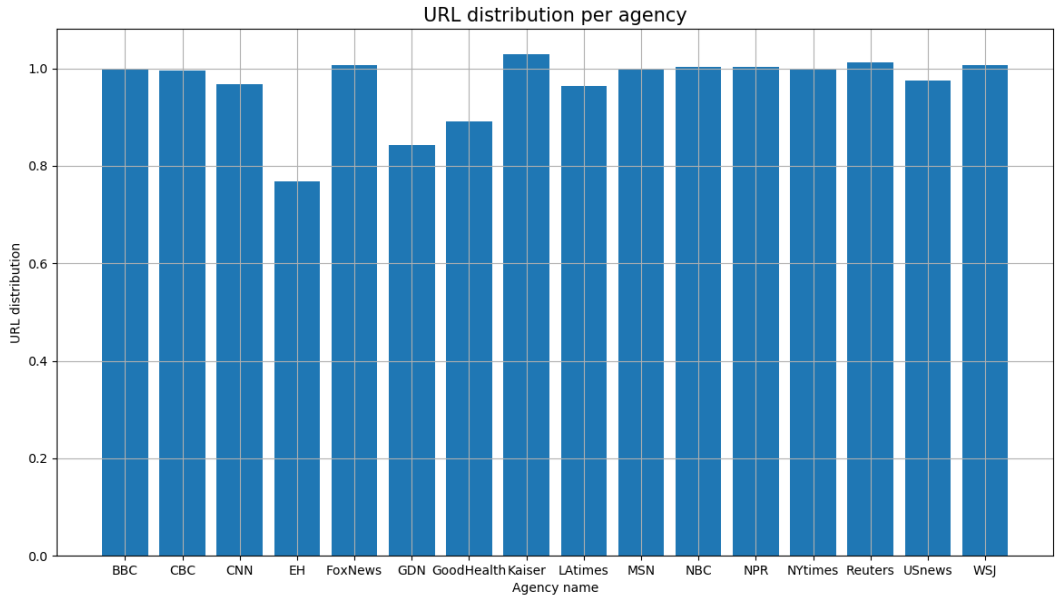
2. Task solution

First of all, it is important to find out what a Twitter is. It is an American online news and social networking service on which users post and interact with short messages known as "tweets". In this case we focusing on tweets posted by **26** health agency between **2011** and **2015** year.

Before we face with topics it is important to get some knowledge about agencies and their tweets. Twitter is often used by news agencies to reach as many people as possible by posting very short messages with URL to agency's website embedded, if post is catchy then user will visit website. Let's check how many posts includes URL links.

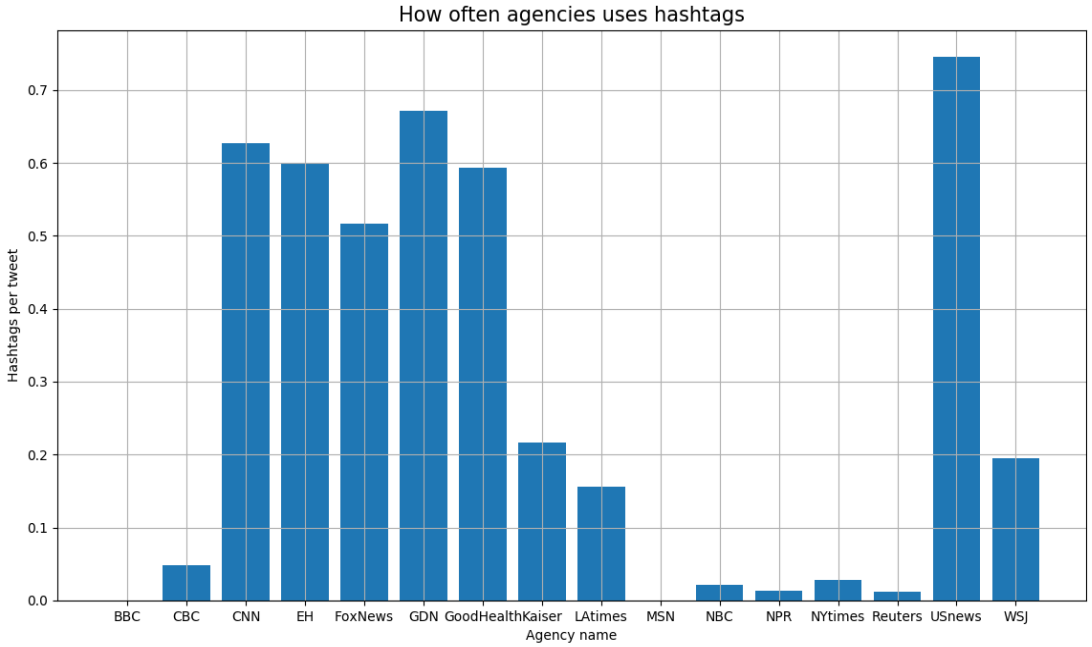


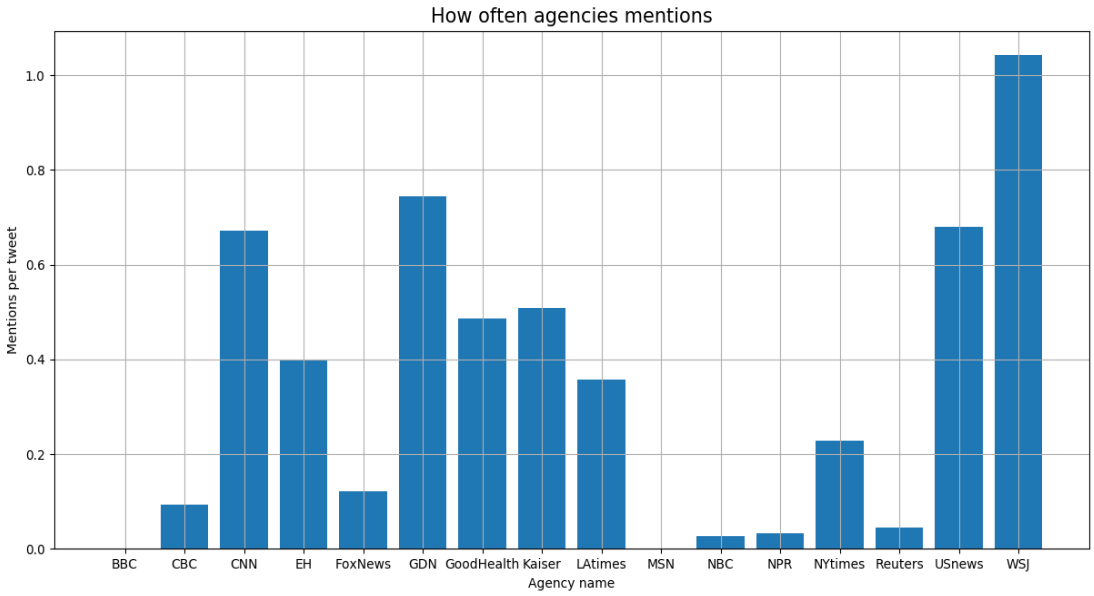
Pie chart clearly shows that almost every tweet includes URL link. Another connected question is about links distribution to a tweet, let's take a look.



Almost every agency includes URL to post and Top Players as BBC, CBC, MSN or NewYork Times ends their tweet with link, they are using Twitter to pass information form website to word in short, catchy form and encourage users to visit agency website.

Another well known marketing mechanisms from Twitter are mentions and hashtags. Both uses for increase range of recipient.





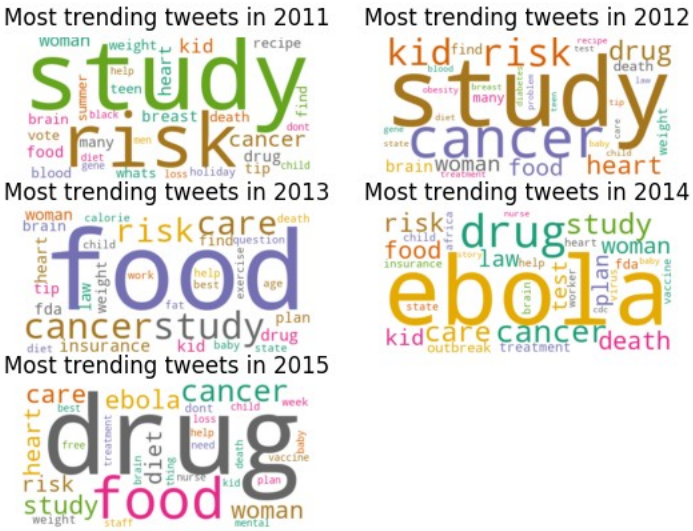
Basing on the two above charts we can see that mentioned before Top Player use less often marketing mechanisms. We can guess that agencies which uses hashtags more often tries to reach more users by including this keywords which users are looking for. We can also presume that agencies which mention other users are not collecting information themselves but they passing information from other agencies or independent journalists.

There we can also mention some most often mentioned persons and agencies.

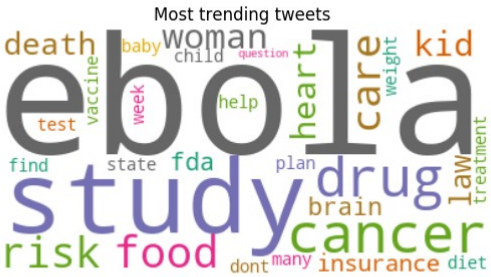
Mention	**Frequency**
@goodhealth	897
@cynthiasass	834
@stefaniei	725
@gdnhealthcare	419
@wsj	313
@pharmalot	240
@cnnhealth	226
@everydayhealth	214
@cslnyt	178
@lauralandrowsj	176

Leaving simple statistics, we can **focus on content**. I decided to take a look on tweets topics year by year.

First of all I construct wordclouds for each year which allows to pick trending topics.



Then I compare it to trending tweets in all data.



There are some most repeating topics in every year and “ebola” topic comes out nowhere. Basing on this wordclouds it’s is possible to pick some trends in health tweets.

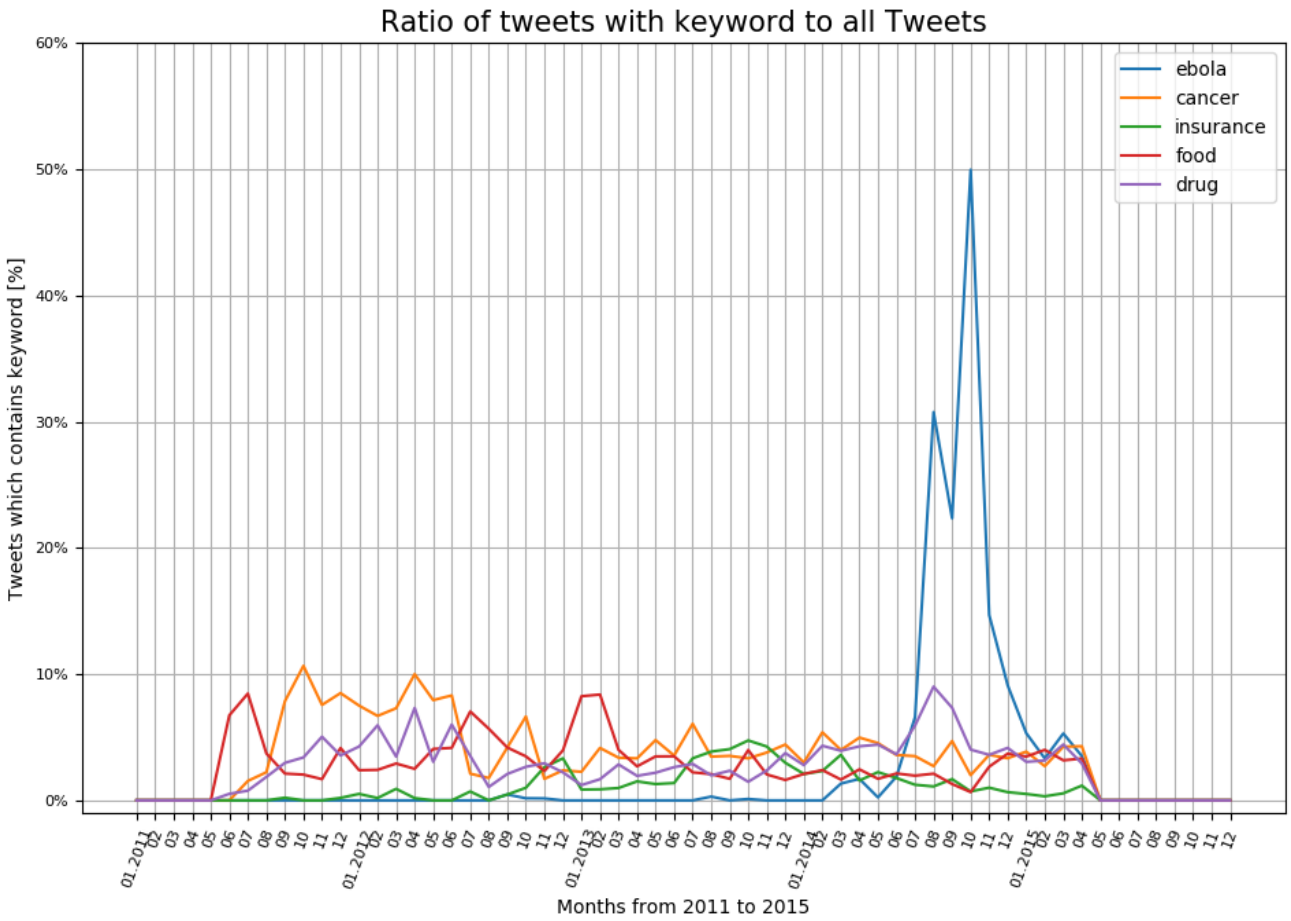
Obviously before creating this models it is necessary to preprocess data. I decided to remove from tweets punctuation, hashtags, mentions, url, english “stopwords” like *'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under'* and some other topic related words like *‘health’, ‘doctor’. I also used lemmatization method to group together the inflected forms of a word so they can be analyzed as a single item.*

In my opinion it is always important to compare collected data in this way to mathematical models, therefore I used TFIDF analysis to model topics from whole data and compare them with wordcloud.

Topic	Words vector
0	0.006*"weight" + 0.006*"food" + 0.006*"tip" + 0.005*"calorie" + 0.005*"fat"
1	0.005*"insurance" + 0.005*"state" + 0.005*"pound" + 0.004*"law" + 0.003*"obamacare"
2	0.008*"recipe" + 0.006*"food" + 0.006*"easy" + 0.006*"happy" + 0.005*"eat"
3	0.010*"ebola" + 0.003*"death" + 0.003*"virus" + 0.003*"vaccine" + 0.003*"africa"
4	0.008*"study" + 0.008*"cancer" + 0.008*"risk" + 0.006*"woman" + 0.006*"age"

After topics comparison I decided to focus on that, how topics popularity has changing in time.

I focused on 5 topics: 'ebola', 'cancer', 'insurance', 'food', 'drug' and plot changing ratio for each topic to all tweets in piece of time.

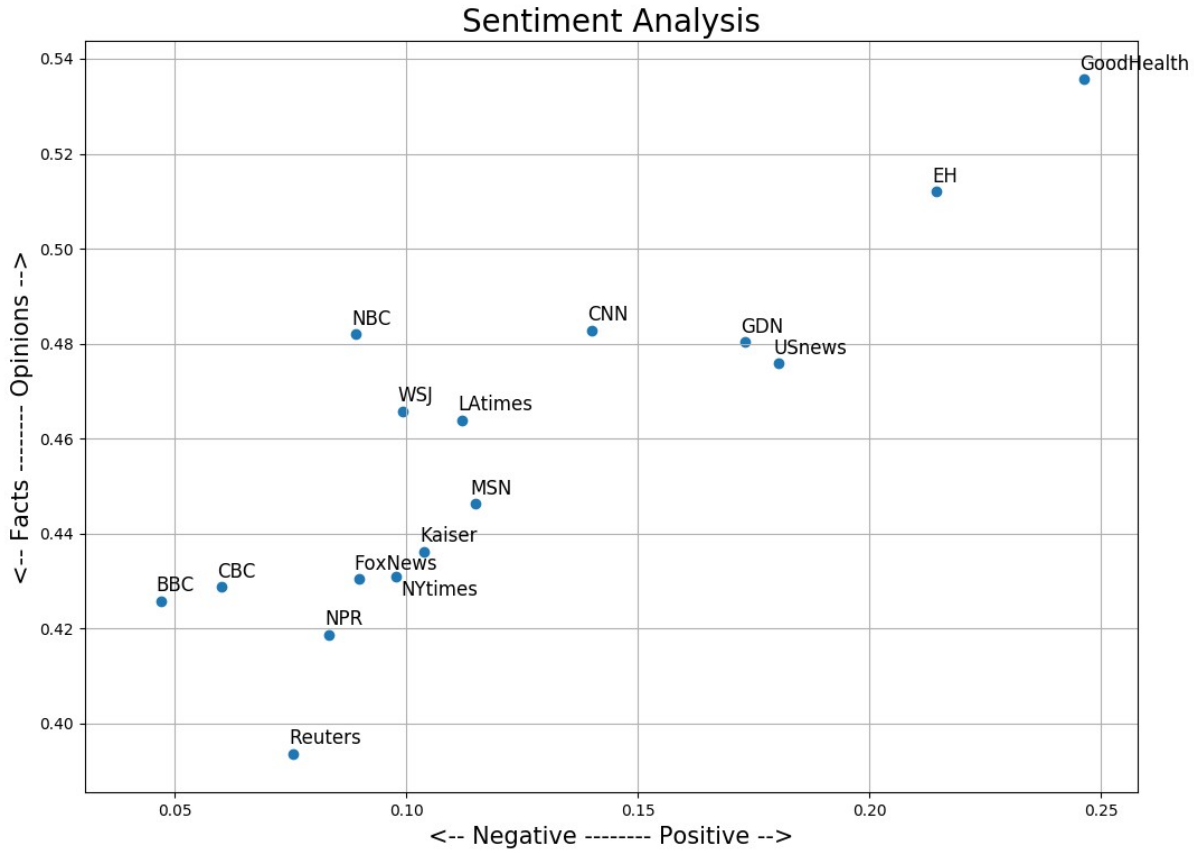


Lets see that topics like ‘cancer’ or ‘food’ are almost almost always have the same popularity. Peak for ‘ebola’ is connected with epidemic at the turn of 2014 and 2015. There is also little peak for insurance keyword before 2014 and its may comes from Obamacare project which starts in 2014.

Then we can take a look on most popular hashtags and compare it with received topics. It may be figure out that a lot of tags comes from fit lifestyle but there are some tags that we can connect with topics from wordclouds and *TFIDF* analysis like ebola, obamacare (which could be connected with insurance)

Hashtag	**Frequency**
#healthtalk	882
#nhs	769
#ebola	411
#getfit	261
#latfit	260
#obamacare	253
#weightloss	235
#health	232
#fitness	211
#recipe	211

At the end lets take a look at sentiment analysis which refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis focus on two parameters – polarity and subjectivity. The polarity score is a number within the range [-1.0, 1.0] where -1.0 is very negative and 1.0 is very positive. The subjectivity is a number within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.



Basing on this scatter plot it is possible to point which agencies delivers facts eg. BBC, Reuters etc.

3. Appendix: third parts

Analyze were made with Python language, with usage of third party libraries:

- pandas
- numpy
- nltk
- textblob
- matplotlib
- and their dependencies

All are free to commercial use.