Exploratory Data Analysis

Sentiment Analysis

Other Formats
Jupyter

1 Introduction

I will be performing sentiment analyses using several of the different approaches described on Figure 1 below. I will make some sligth changes that I think reflect a more accurate classification for methods of conducting sentiment analysis.

Figure 1: Thank you for this, Oliver

I will also be providing some background information about each approach, and my own thoughts about it.

The purpose of this document is to provide some insights into each of the different techniques and help with the decision of how to proceed.

2 Lexicon Based Approaches

2.1 Dictionary based approach

Here I think a better sub-classification scheme may be:

1. Single word/ keyword spotting
   1. In this case, a single word (unigram) is matched with a particular sentiment. This is obviously a naive approach with several flaws. Perhaps the most obvious one is that it neglects negation. Another flaw is that it is based on surface features.
2. Lexical affinity
   1. This is an slightly more sophisticated approach compared to keyword spotting and it is based on assigning a probability (rather that a hard label) to a certain word. For example, the word accident instead of receiving a label of either positive or negative, is given a probability (e.g. 0.7 of being negative). Dictionary of lexical affinity are usually derived from *linguistic corpora*. Although an step up from keyword spotting, they still lack the capacity to incorporate negation into the final output and they are usually domain-dependent, so it is hard to find generalisable models.

2.1.1 Keyword spotting

▸ Code

2.1.1.1 AFINN Lexicon

The AFINN lexicon gives a score of -5 to 5 to classify the sentiment from extremely negative to extremely positive. The overall sentiment is given by the sum of these values. See @tbl_afinn_sample below for examples.

▸ Code

Table 1: AFINN Lexicon (sample)

word	value
manipulation	-1
fascist	-2
responsive	2
noob	-2
disorganized	-2
downcast	-2
faithful	3
victimizing	-3
backed	1
fraudster	-4
beautify	3
skepticism	-2
restores	1
drowned	-2
reassures	1
naive	-2
sore	-1
motherfucker	-5
harmful	-2
daring	2

▸ Code

Figure 2: Change of sentiment over time (agg. to week) according to AFINN lexicon

2.1.1.2 Bing Lexicon

Gives a hard label (i.e. positive, negative) to each word. See Table 2 below for examples.

▸ Code

Table 2: Bing Lexicon (sample)

word	sentiment
sorely	negative
concerned	negative
tease	negative
steadiest	positive
bullying	negative
rapt	positive
compliant	positive
god-given	positive
fervent	positive
amply	positive
constructive	positive
oppressive	negative
transparent	positive
temper	negative
jeeringly	negative
cataclysmically	negative
illegally	negative
inspiring	positive
drawback	negative
stampede	negative

▸ Code

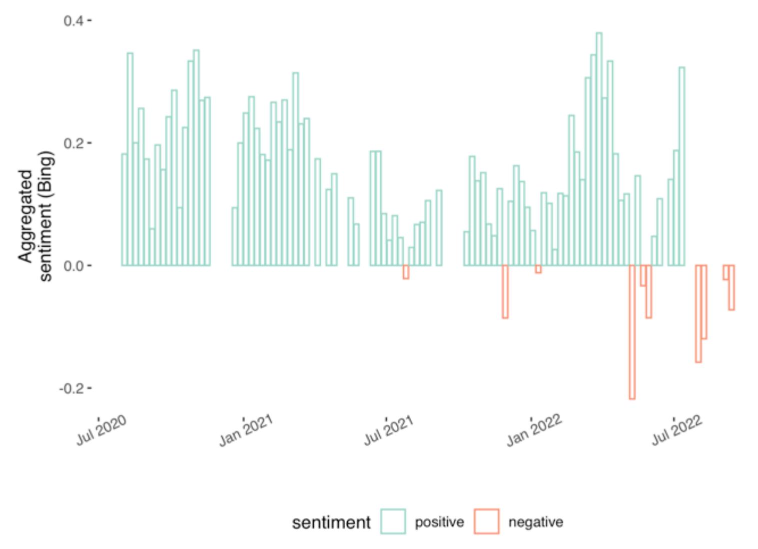

Figure 3: Change of sentiment over time (agg. to week) according to bing lexicon

2.1.1.3 NRC Lexicon

The NRC lexicon assigns words into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. See Table 3 below for examples.

▸ Code

Table 3: NRC Lexicon (sample)

word	sentiment
wound	anger
pedigree	trust
haven	trust
arrogant	anger
fanaticism	fear
avoiding	fear
memorable	trust
coward	negative
reverie	joy
murderer	anger
perjury	surprise
encroachment	fear
hilarious	joy
leakage	negative
insurrection	anger
hazard	fear
leukemia	anger
sterling	positive
overwhelm	negative
deceitful	disgust

▸ Code

```
Warning: Removed 1 rows containing missing values (`position_stack()`).
```

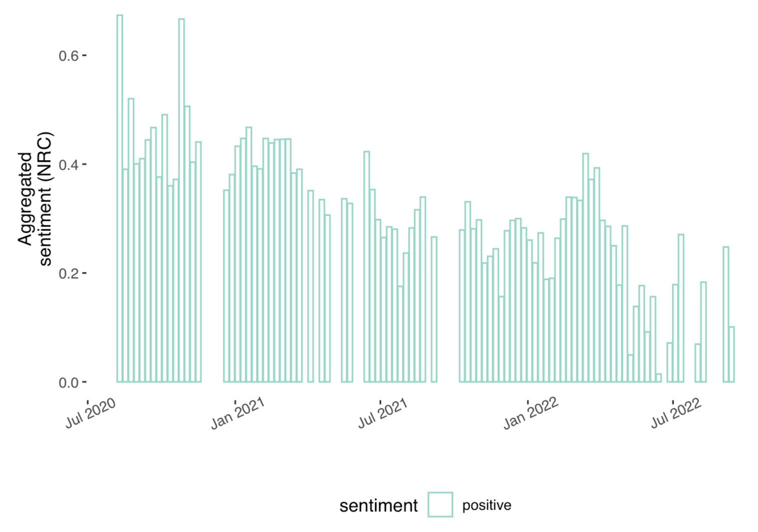

Figure 4: Change of sentiment over time (agg. to week) according to NRC lexicon

2.1.2 Conclusions

Potential issues:

- Because lexicons have only a limited number of words, not all tweets and not all of the words in a given tweet's text are going to be scored
- Some of the words lexicons score as positive (e.g. free) are commonly used by anti-vaxxers groups
- We would be finding the overall sentiment of the tweet, NOT the sentiment towards vaccination, which is what we actually want. . .
- Loss of data (see Table 4 below):
  - All three lexicons drop between a third and a fourth of the data. See below.
  - Likewise, not all words in a tweet's text are used.

▸ Code

Table 4: Tweets kept by using each lexicon dictionary

lexicon	retention
AFINN	65,136 out of 99,997 (65.14%)
Bing	63,011 out of 99,997 (63.01%)
NRC	79,369 out of 99,997 (79.37%)

▸ Code

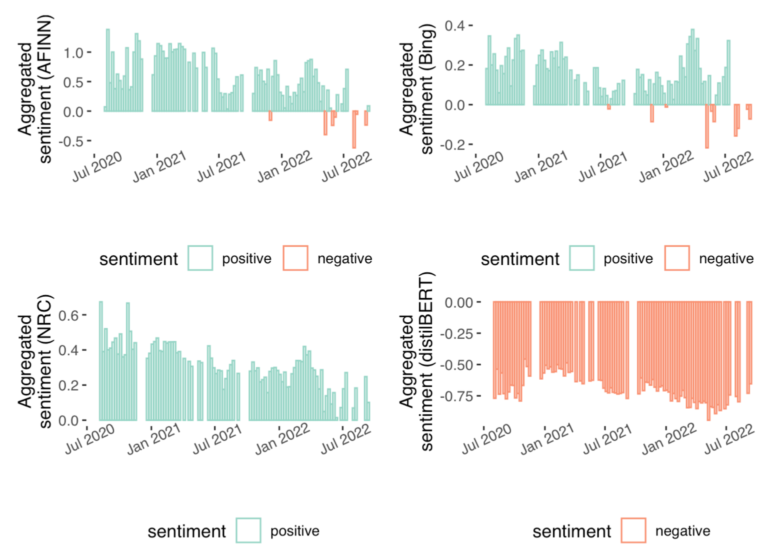

Figure 5: A comparison between the three lexicons and the result from distilBERT