Dictionary-Based Text Analysis

Bamberg Summer Institute in Computational Social Science

Carsten Schwemmer, University of Bamberg

2019-08-01

Many thanks to Chris Bail for providing material for this lecture

Dictionary-based methods

- dictionary methods count the number of words that appear in each document that have been assigned a particular meaning or value to the researcher
- such words may or may not be weighted (e.g. for sentiment analysis)
- sometimes they are only used for filtering or labeling documents in preparation for further analysis

Example - representation of immigrant groups

“.. how does the government want to ensure that the Federal Employment Office will bring residents with a migratory background into vocational training in similar proportions in their respective age groups as compared to Germans?” - translated parliamentary question by Mechthild Rawert, SPD, March 2011

- identify questions that address needs of disadvantaged immigrant groups (= substantive representation)
- use indicator in regression approach to examine what factors drive representative behavior

`https://doi.org/10.1080/01402382.2018.1560196`

Example - representation of immigrant groups

abgeschoben, abschiebehaftbedingungen, abschiebestopps, abschiebung, abschiebungen, altübersiedler, aufenthaltstitel, antidiskriminierungsrichtlinie, antidiskriminierungsstelle, arbeitserlaubnis, aslybewerberleistungsbezug, assoziationsrecht, asyl, asylantrag, asylantragstellern, asylanträge, asylbewerber, asylbewerberinnen, asylbewerberleistungsbezug, asylbewerberleistungsgesetz, asylbewerberleistungsgesetzes, asylbewerberleistungsgestz, asylbewerberleisungsgesetz, asylbewerbern, asylbewerbers, asylblg, asylsuchende, asylsuchenden, asylsuchendenzahlen, asylsuchender, asylsystem, asylsystems, asylverfahren, asylverfahrenrichtlinie, asylverfahrensgesetz, asylverfahrensgesetzes, asylverfahrensrecht, asylverfahrensrichtlinie, asylverfahrungsgesetz, aufenthaltgesetz, aufenthaltsstatus, aufenthaltserlaubnis, aufenthaltserlaubnisse, aufenthaltserlaubnis, aufenthaltsgesetz, aufenthaltsgesetze, aufenthaltsgesetzes, aufenthaltsgestaltung, aufenthaltsgewährung, aufenthaltspapiere, aufenthaltsrecht, aufenthaltstitel, ausländer, ausländerbeschäftigungsrecht, ausländerförderung, ausländerjagdschein, ausländerzentralregister, ausländischer, aussiedler, balkanflüchtlinge, bleiberechtsregelung, bleibeberechtigung, bürgerkriegsflüchtlinge, bürgerkriegsflüchtlingen, ...

Example - representation of immigrant groups

…, diskriminierung, doppelstaatlers, drittstaatangehörige, drittstaatsangehörige, drittstaatsangehörigen, dublin-ii, dublinüberstellungsverfahren, ehegattennachzug, einbürgerung, einbürgerungstest, einbürgerungstests, einbürgerungsverhalten, eingebürgert, einreiseerlaubnis, einreisevisum, einwanderern, einwanderungsgruppen, eu-aufnahmerichtlinie, eu-aufnahmerichtlinien, fachkräfteanwerbung, familiennachzug, familienzusammenführung, familienzusammenführungsrichtlinen, familienzusammenführungsrichtlinie, flüchtlinge, flüchtlingen, flüchtlingselend, flüchtlingskonvention, flüchtlingslager, frontex, grenzsicherug, grenzübergangsstellen, herkunftsfamilie, herkunftsland, herkunftsstaaten, integration, integrationsansprüche, integrationsarbeit, integrationscoaching, integrationsfördernd, integrationsförderung, integrationsgipfel, integrationsherausforderungen, integrationskurs, integrationskursbeteiligung, integrationskurse, integrationskursen, integrationsleistung, integrationsleistungen, integrationsministerkonferenz, integrationspolitik, integrationspolitisch, integrationsprogramm, integrationsprogramms, integrationsprojekte, integrationssprachkursleiter, integrationstest, integrationsunwillig, integrationsverordnung, integriert, interkulturelle bildung, intgegrationsprojekte, islam, jugendintegrationskurse, jugendmigrationsdienst, jugendmigrationsdienstes, migranten, migrantinnen, migration, migrationsabkommen, migrationsbiographie, migrationshintergrund, migrationshintergrund, migrationshintergrundes, …

Example - representation of immigrant groups

..., minderheitsangehoerige, minderheitsangehörige, immigranten, optionskind, optionskinder, optionspflicht, optionspflichtige, rassismus, resettlement-programms, roma-minderheit, rückführungsabkommen, rückführungsentscheidungen, rücknahmeabkommen, rückübernahmeabkommen, rückübernahmeabkommens, rücküberstellung, sammelunterkünfte, sammelunterkünften, scheineheverdachts, scheineheverdachtsfälle, sprachförderung, sprachkurs, sprachkurse, sprachkursen, sprachtest, spätaussiedler, staatenlose, staatsangehörigkeit, staatsangehörigkeitsgesetz, staatsangehörigkeitsrecht, staatsbürgerschaft, visa, visagebühren, visapflicht, visavergabe, visum, visumantrags, visumanträge, visumbefreiung, visumfreiheit, visumgebühren, visums, visumsanträge, visumsbefreiung, visumsfreiheit, visumsgebühren, visumspflicht, visumverfahren, zugewandert, zuwanderer, zuwanderern, zuwanderung

191 terms in total, identified with qualitative validations using a Shiny app: `https://cschwem2er.github.io/pathways/`

Applying dictionaries - data for this lecture

A data set of tweets by Donald Trump (yay!)

```
library(tidyverse)
load(url(”https://cbail.github.io/Trump_Tweets.Rdata”))
trumptweets$text[1:2]
```

```
## [1] ”Just met with UN Secretary-General António Guterres
who is working hard to ”Make the United Nations Great
Again.” When the UN does more to solve conflicts around the
world, it means the U.S. has less to do and we save money.
@NikkiHaley is doing a fantastic job!
https://t.co/pqUv6cyH2z”
## [2] ”America is a Nation that believes in the power of
redemption. America is a Nation that believes in second
chances - and America is a Nation that believes that the
best is always yet to come! #PrisonReform
https://t.co/Yk5UJUYgHN”
```

Quanteda corpus

```r
library(quanteda)
trump_dfm <- corpus(trumptweets, text_field = 'text',
                    docid_field = 'status_id') %>%  dfm()

head(trump_dfm, 3, 5)
```

```
## Document-feature matrix of: 3 documents, 5 features
(66.7% sparse).
## 3 x 5 sparse Matrix of class ”dfm”
## features
## docs just met with un secretary-general
## 997577906007298048 1 1 1 2 1
## 997573139663028224 0 0 0 0 0
## 997568208369577985 0 0 0 0 0
```

{https://twitter.com/realDonaldTrump/status/997577906007298048}

Quanteda dictionaries

Quantedfa dictionaries consist of lists with one or multiple character vectors. Patterns can for instance be specified used *glob* or *regex* patterns (see `valuetype()`):

```
dict <- dictionary(list(terror = c(”terror*”, ”threat”),
                    economy = c(”jobs”, ”business”, ”grow”, ”work”),
                    immigration = c('immig*', 'migra*')))
dict_match <- dfm_lookup(trump_dfm, dict) # apply dictionary
textstat_frequency(dict_match)
```

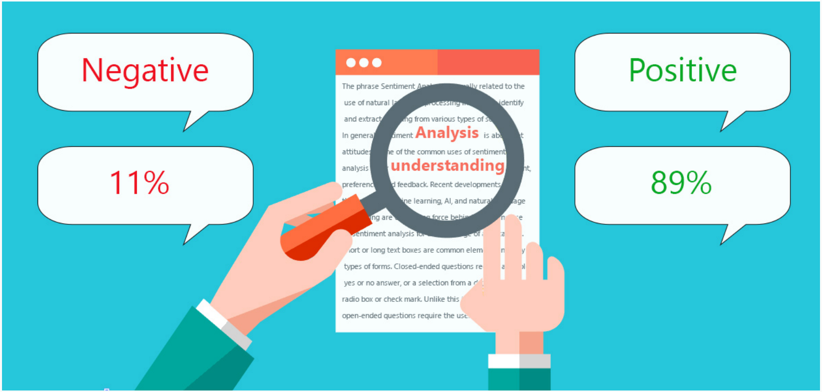
```
##        feature frequency rank docfreq group
## 1     economy       291    1     223   all
## 2 immigration        77    2      62   all
## 3      terror        63    3      55   all
```

Trump tweets related to immigration

```r
trumptweets <- bind_cols(trumptweets, as.data.frame(dict_match))
trumptweets %>% arrange(desc(immigration)) %>%
  head(2) %>% pull(text)
```

```
## [1] ”The Schumer-Rounds-Collins immigration bill would
be a total catastrophe. @DHSgov says it would be ”the end
of immigration enforcement in America.” It creates a giant
amnesty (including for dangerous criminals), doesn't build
the wall, expands chain migration, keeps the visa...”
## [2] ”My Administration has identified three major
priorities for creating a safe, modern and lawful
immigration system: fully securing the border, ending chain
migration, and canceling the visa lottery. Congress must
secure the immigration system and protect Americans.
https://t.co/xV1lgfhjBU”
```

Sentiment analysis



Simple approach

- we’ll be using term weights from the AFINN word list
- this simple approach does not consider valence shifters, e.g. “not nice” (see `sentimentr` package for alternatives)
- other approaches try to identify emotions (e.g. anger, sadness) instead of “positive” vs. “negative”

```r
library(textdata) # contains several sentiment word lists
afinn <- lexicon_afinn() # press 1 to download
sentiment <- c(afinn$value) %>% set_names(afinn$word)
sentiment['sad']
```

```
## sad
##  -2
```

Computing tweet sentiments

```r
sentiment_dfm <- dfm_keep(trump_dfm, names(sentiment)) %>%
  dfm_weight(weights = sentiment) # apply sentiment weights
head(sentiment_dfm, 3, 5)
```

```
## Document-feature matrix of: 3 documents, 5 features (66.7% sparse).
## 3 x 5 sparse Matrix of class ”dfm”
##                      features
## docs                 hard united great solve conflicts
##   997577906007298048   -1      1     3     1        -2
##   997573139663028224    0      0     0     0         0
##   997568208369577985    0      0     0     0         0
```

Merge datasets

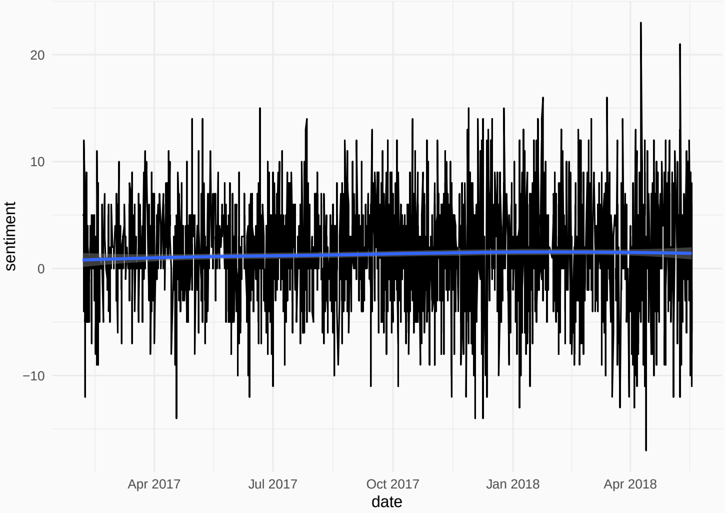
```
trumptweets <- trumptweets %>%  # merge data, create variables
  mutate(date = as.Date(created_at, format = ”%Y-%m-%d %x”),
         sentiment = rowSums(sentiment_dfm))

trumptweets %>% arrange(sentiment) %>% head(1)  %>%
  pull(text) %>% cat() # tweet with most negative sentiment
```

```
## James Comey is a proven LEAKER &amp; LIAR. Virtually
everyone in Washington thought he should be fired for the
terrible job he did-until he was, in fact, fired. He leaked
CLASSIFIED information, for which he should be prosecuted.
He lied to Congress under OATH. He is a weak and.....
```

Sentiment over time

```
trumptweets %>% ggplot(aes(x = date, y = sentiment)) +
  geom_line() + geom_smooth(method = 'loess') + theme_minimal()
```



Many sentiment dictionaries, very different results

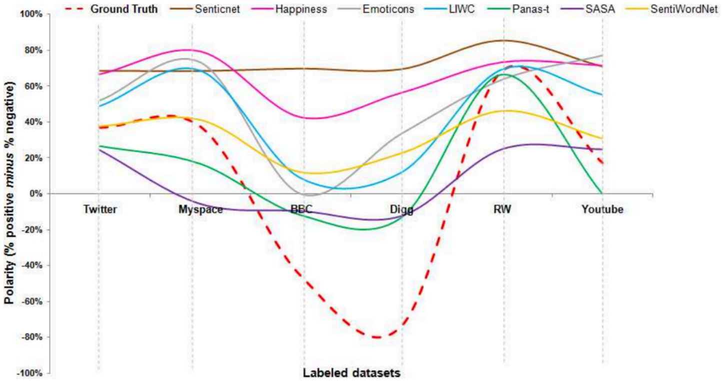


Figure 2: Polarity of the eight sentiment methods across the labeled datasets, indicating that existing methods vary widely in their agreement.

https://doi.org/10.1145/2512938.2512951

When to use dictionary-based approaches?

- quality of dictionary-based methods depends heavily upon the match between learning-corpus and corpus of interest
- creating your own dictionaries might often be the best option, but is time sensitive
- if you are looking for specific things rather than for categorizing documents, dictionary methods often perform better than more sophisticated techniques (e.g. topic modeling)
- computers-assisted methods can be helpful: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12291`

Questions?