Brexit through the Lens of Data Science

Kenneth Benoit
(joint work with Akitaka Matsuo)
Department of Methodology
LSE

Text mining and data/social science

- Treats text as "data" that informs us about the authors of the text - the text itself is only incidental
- Key is some form of comparison, to gain insights about differences
- Involves the application of statistical models to judge differences using probability statements

Analyzing Brexit through Twitter

- EU-funded project
- We collected some 26 million Tweets from January - July 2016
- capture based on #hashtags, @usernames, and search terms

Search terms

Hashtags:

#betterdealforbritain
#betteroffout
#brexit
#euref
#eureferendum
#eusummit
#getoutnow
#leaveeu
#no2eu
#notoeu
#strongerin
#ukineu
#voteleave
#wewantout
#yes2eu
#yestoeu

brexit

Usernames:

@vote_leave
@brexitwatch
@eureferendum
@ukandeu
@notoeu
@leavehq
@ukineu
@leaveeuofficial
@ukleave_eu
@strongerin
@yesforeurope
@grassroots_out
@stronger_in

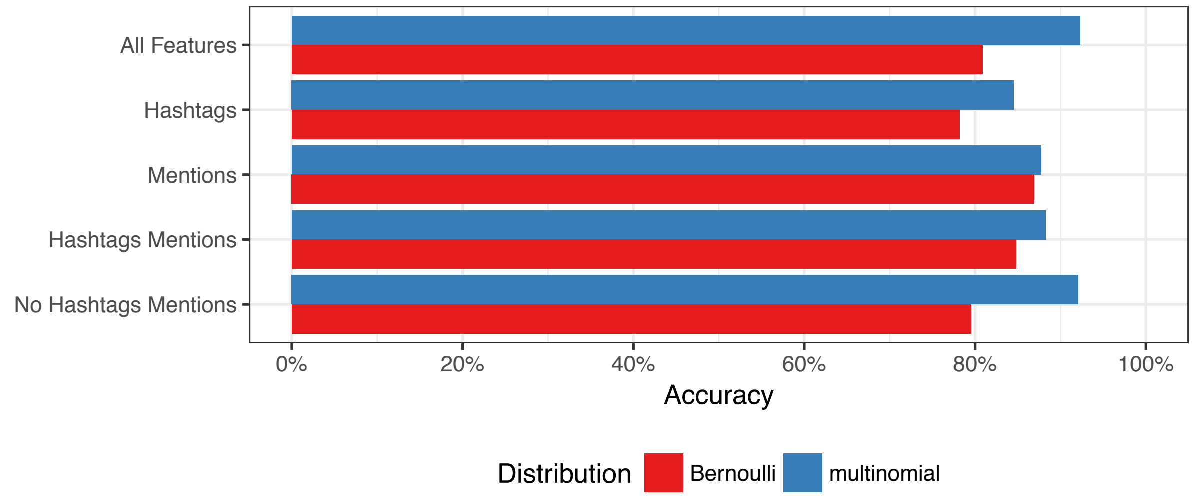
Users in data

- 3.6M unique users
- number of tweets:
  - average: 7.2
  - median: 1 (more than 50% had only one tweet)
  - max: 81.1K

First application: Predicting Leave v. Remain users

- Method: Naive Bayes classifier
- Data source: combined tweet corpus at user level
- Creating training data
    - Select “power-users” (more than 100 tweets in the corpus, 15K users)
    - Check the use of pre-determined set of “leave” and “remain” hashtags
    - Calculate the difference in the use of leave and remain hashtags. Construct training data from top and bottom 10% of power users

Predictive accuracy based on feature types



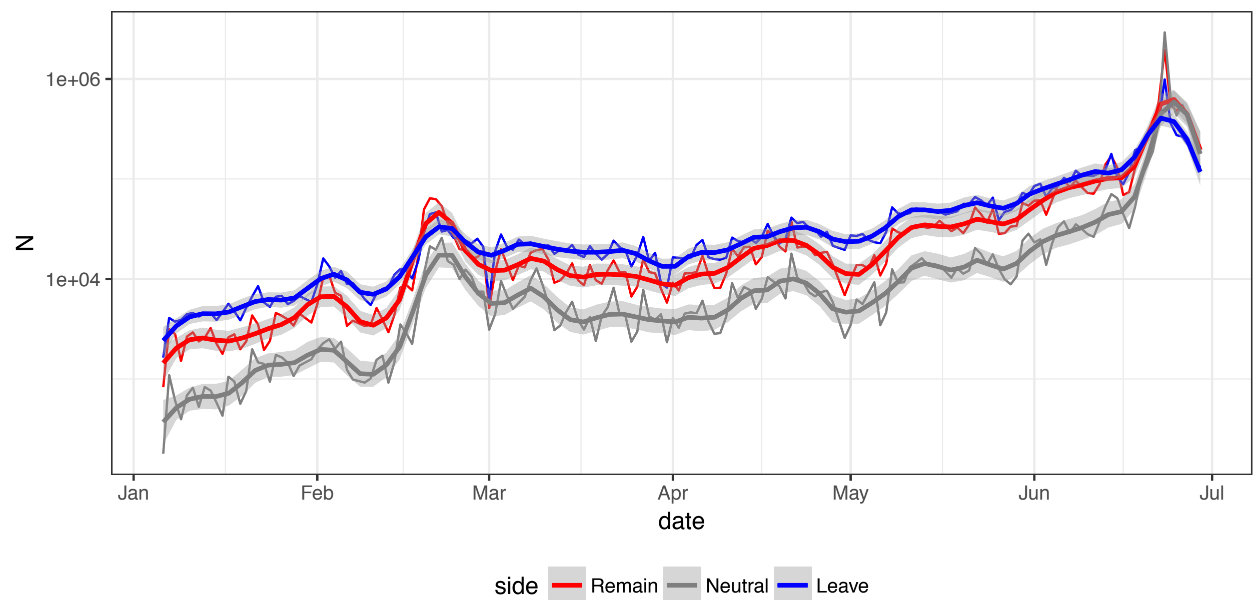
- Use bag of words method (unigrams of #hashtags and @user_mentions, remove features used less than 20 users)

our version

Predicting Leave v. Remain

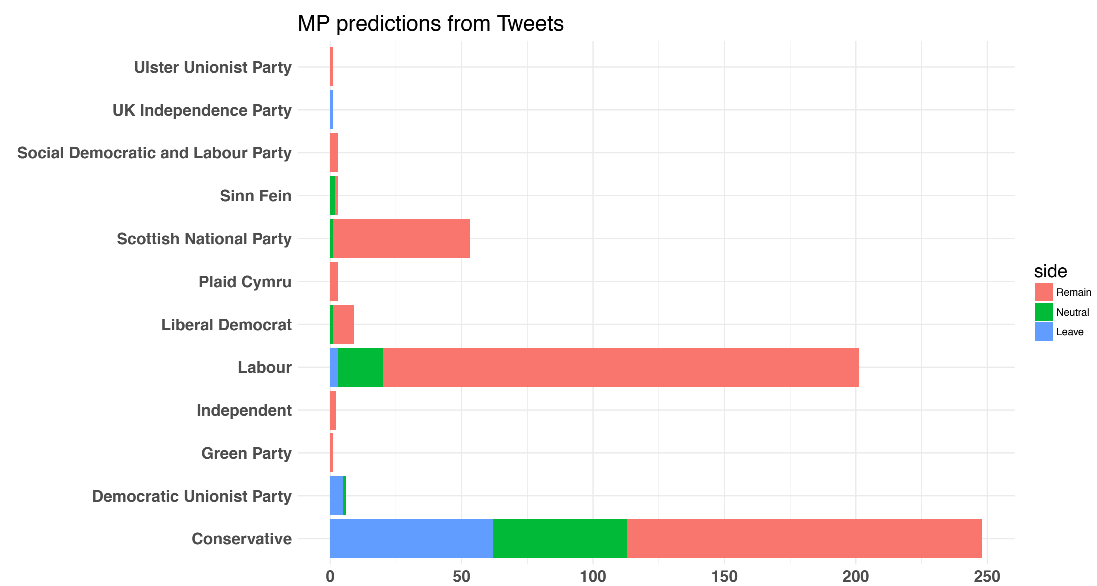
	N	%
Remain	9,780,223	36.93%
Neutral	7,786,297	29.40%
Leave	8,914,207	33.66%

Patterns of posts, Leave v. Remain

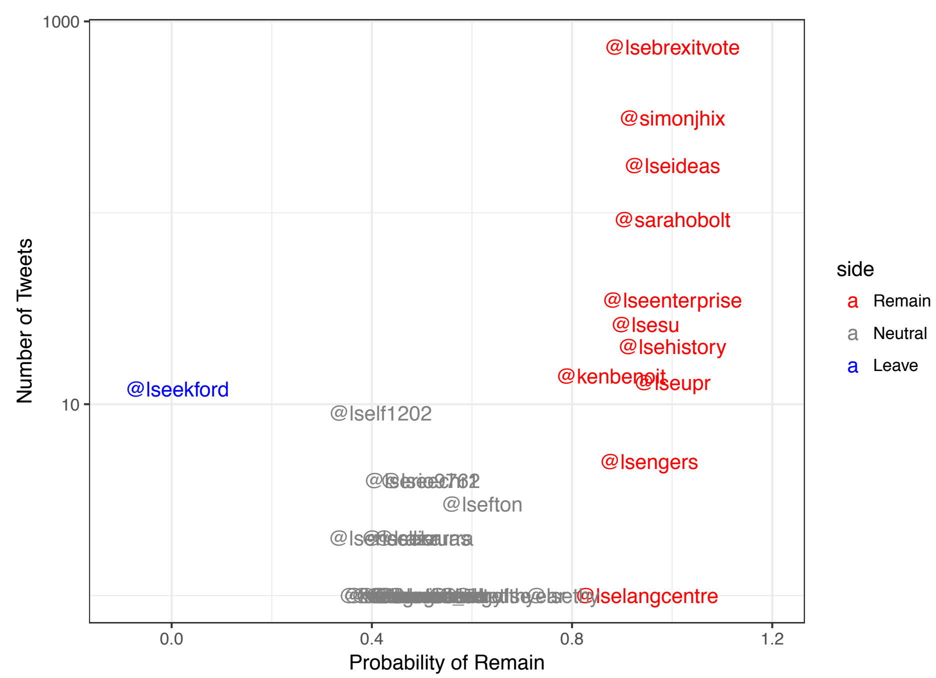


our version

MP side predictions by party



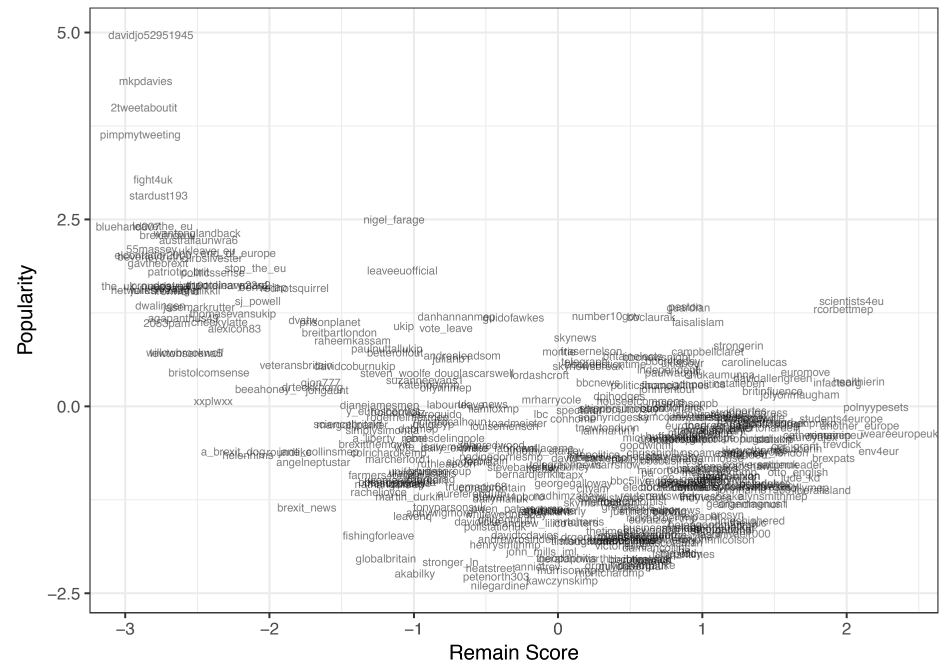
A few well-known Twitter users




linda seekford
@lseekford
Joined October 2012
Tweet to linda seekford
40 Photos and videos
TWEETS 4,078
FOLLOWING 232
FOLLOWERS 123
LIKES 3,376
Follow
Tweets
Tweets & replies
Media
linda seekford @lseekford · Jan 20
Sad liberals spewing hate
AJ+ @ajplus
#Inauguration2017: We're in DC showing you everything going down on this historic day. periscope.tv/w/a1DlKzQzNDk5…
linda seekford @lseekford · Jan 20
Thank God we survived the last 8 yrs
linda seekford @lseekford · Jan 17
John McCain is,a snake. Go away you old goat
linda seekford @lseekford · Jan 20
Mass impeachment for the traitor democracts

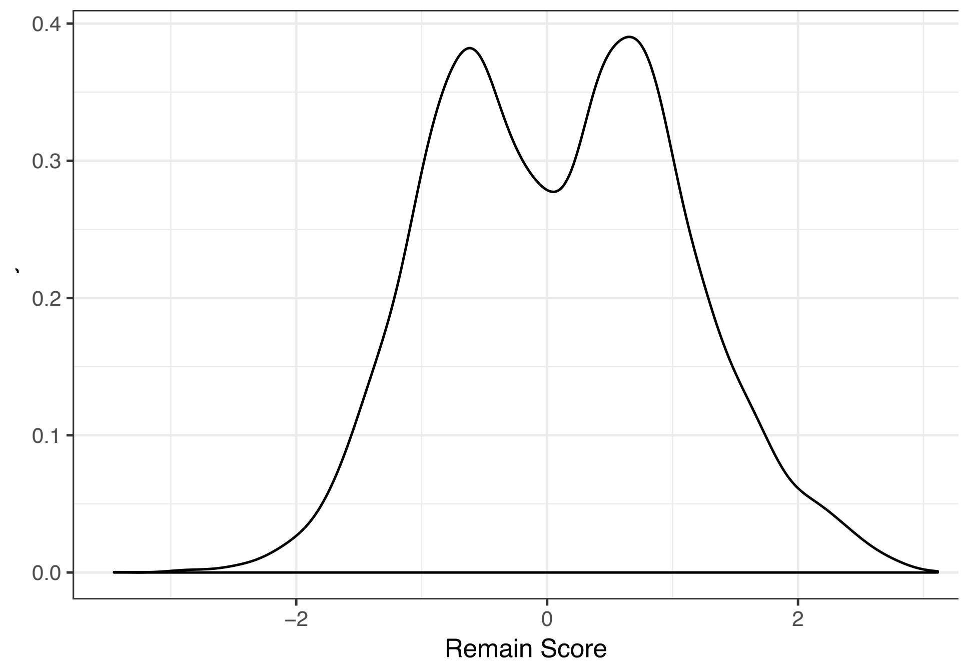
Leave and Remain Hashtags

Followership network analysis, followees' positions (Barberà 2015)



Plot of top 300 hashtags. Cross side edges are highlighted. There are some connections, but the number of edges is smaller than the next figure.

Followership network analysis, followers distribution (Barberà 2015)



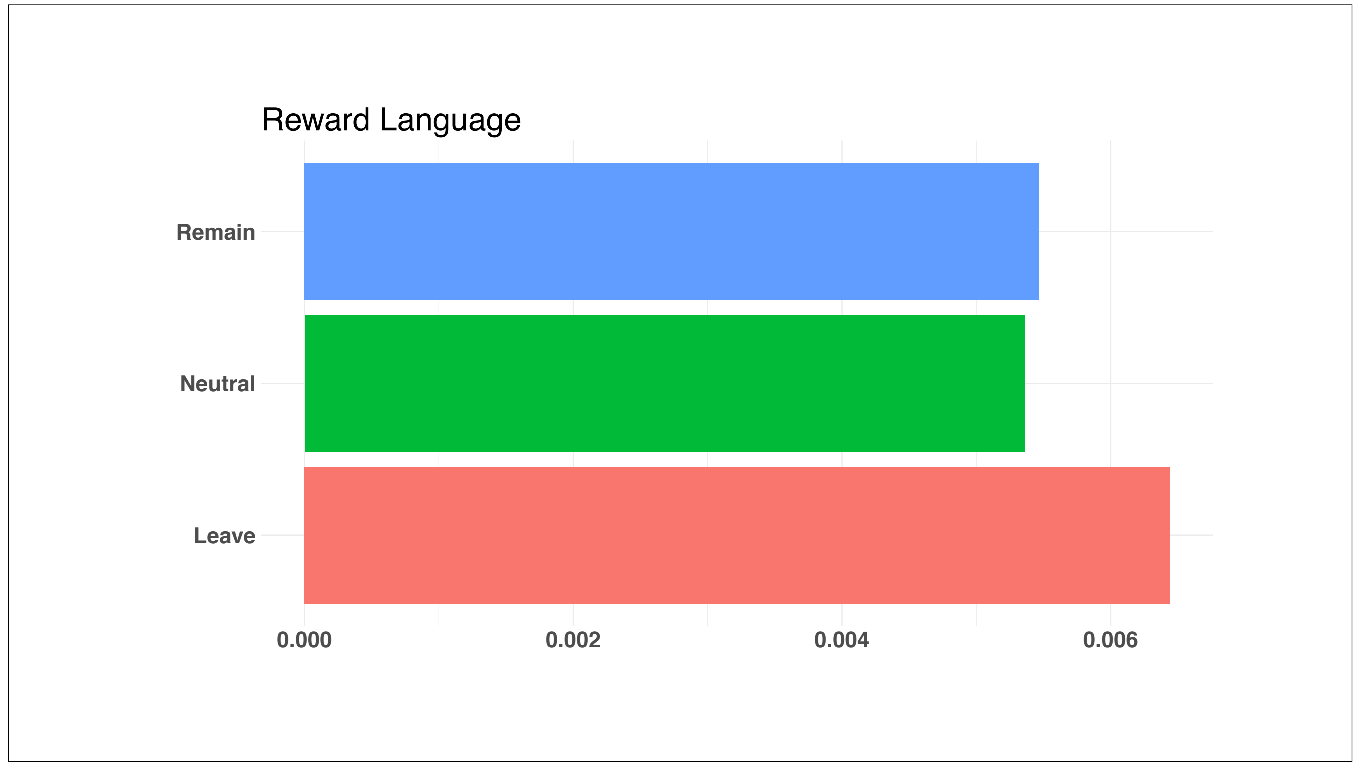
Plot of top 300 hashtags. Cross side edges are highlighted. There are some connections, but the number of edges is smaller than the next figure.

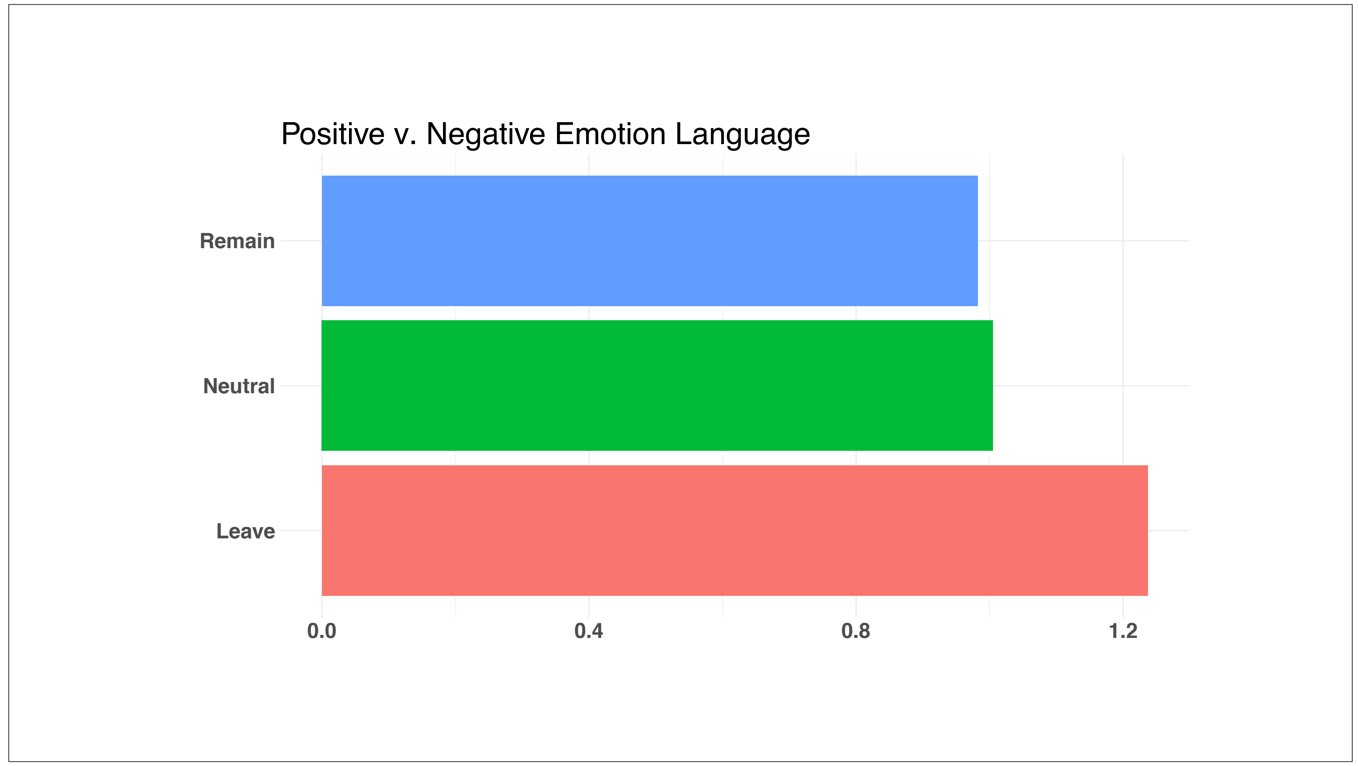
Sentiment Analysis

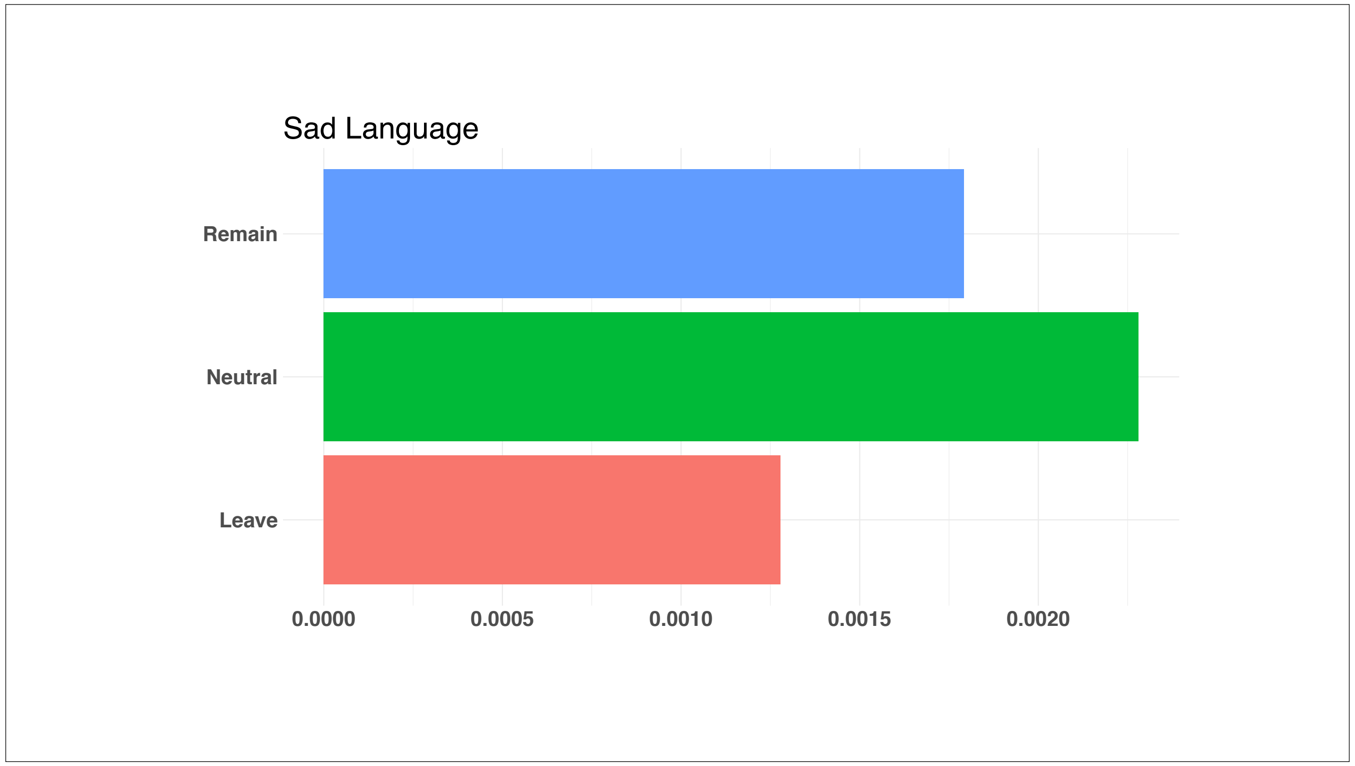
- Looks up terms from the Linguistic Inquiry and Word Count, a psychological dictionary
- Contains categories about:
  - positive and negative emotion
  - politics
  - power
  - quantitative language
  - tentative language
  - sadness
  - future v. past orientation

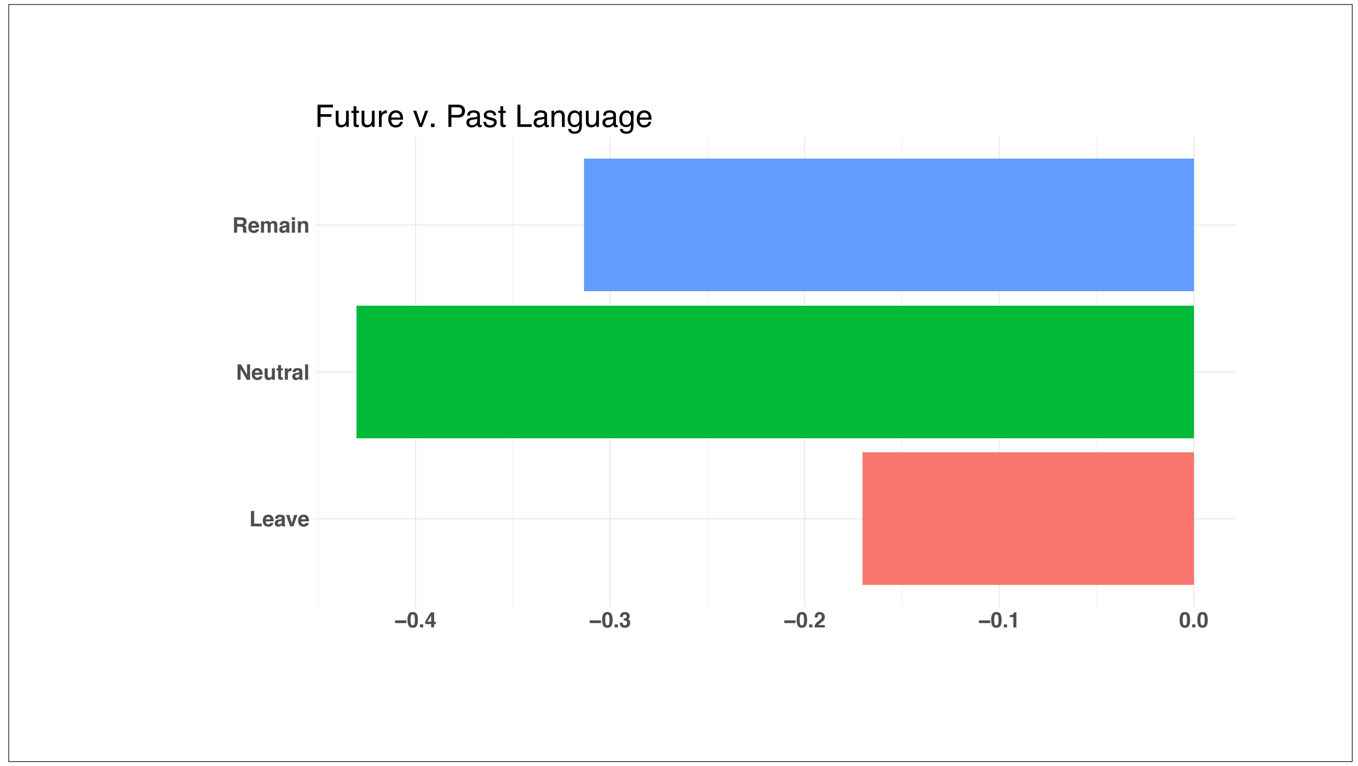
example: “reward” language

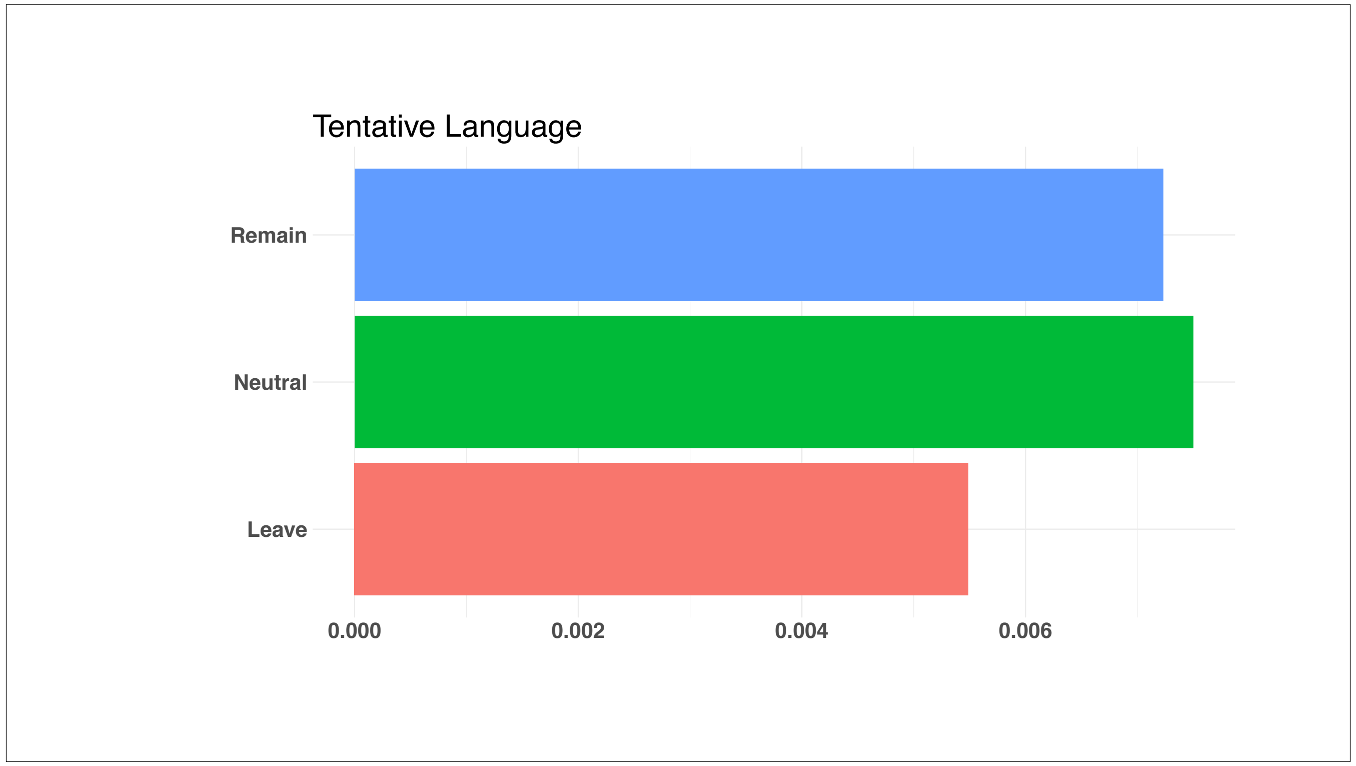
```
> data_dictionary_liwc[["reward"]]
  [1] "access*"       "accrue*"       "accumul*"      "achievable"    "achieve*"      "achievi*"
  [7] "acquir*"       "add"           "added"         "adding"        "adds"          "advanc*"
 [13] "advantag*"     "adventur*"     "amass*"        "approach"      "approached"    "approaches"
 [19] "approaching"   "award*"        "benefit"       "benefits"      "best"          "bet"
 [25] "bets"          "better"        "betting"       "bold"          "bonus*"        "confidence"
 [31] "confident"     "confidently"   "crave"         "craving"       "dare"          "dared"
 [37] "dares"         "daring"        "desir*"        "eager"         "eagerly"       "eagerness"
 [43] "earn"          "earned"        "earning"       "earnings"      "earns"         "enthus*"
 [49] "excite"        "excited"       "excitedly"     "excitement"    "exciting"      "fearless*"
 [55] "fulfill*"      "gain*"         "get"           "gets"          "getting"       "goal*"
 [61] "good"          "got"           "gotten"        "great"         "greed*"        "invigor*"
 [67] "jackpot*"      "luck"          "lucky"         "obtain"        "obtainable"    "obtained"
 [73] "obtaining"     "obtains"       "opportun*"     "optimal*"      "optimism"      "optimistic"
 [79] "perfect"       "perfected"     "perfecting"    "perfection"    "perfectly"     "plus"
 [85] "positive"      "positively"    "positives"     "positivi*"     "prize*"        "profit*"
 [91] "promot*"       "reward*"       "score*"        "scoring"       "seize*"        "snag*"
 [97] "steal*"        "stole"         "succeed*"      "success"       "successes"     "successful"
[103] "successfully"  "surpass*"      "take"          "taken"         "takes"         "taking"
[109] "took"          "triumph*"      "victor*"       "wager"         "wagered"       "wagering"
[115] "wagers"        "willing"       "win"           "winn*"         "wins"          "won"
```

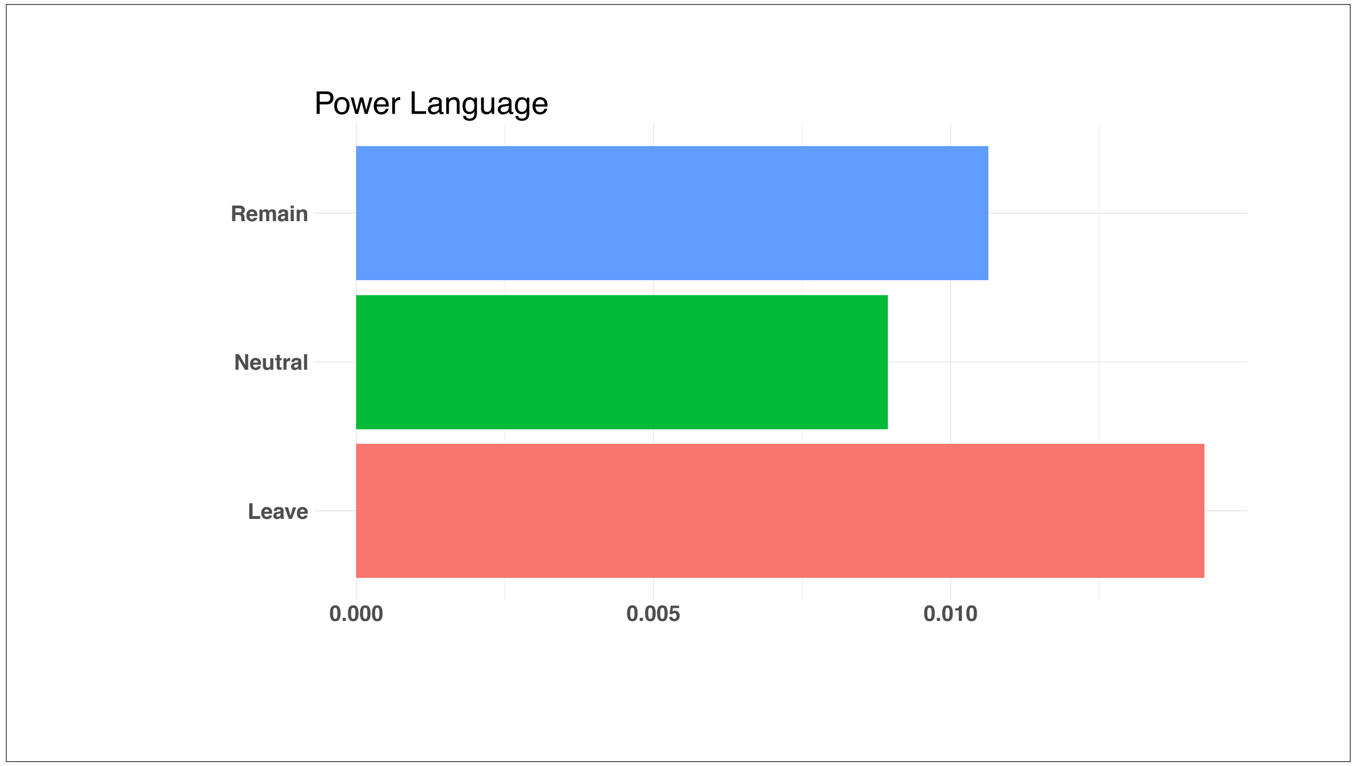

Reward Language
Remain
Neutral
Leave
0.000
0.002
0.004
0.006


Positive v. Negative Emotion Language
Remain
Neutral
Leave
0.0
0.4
0.8
1.2

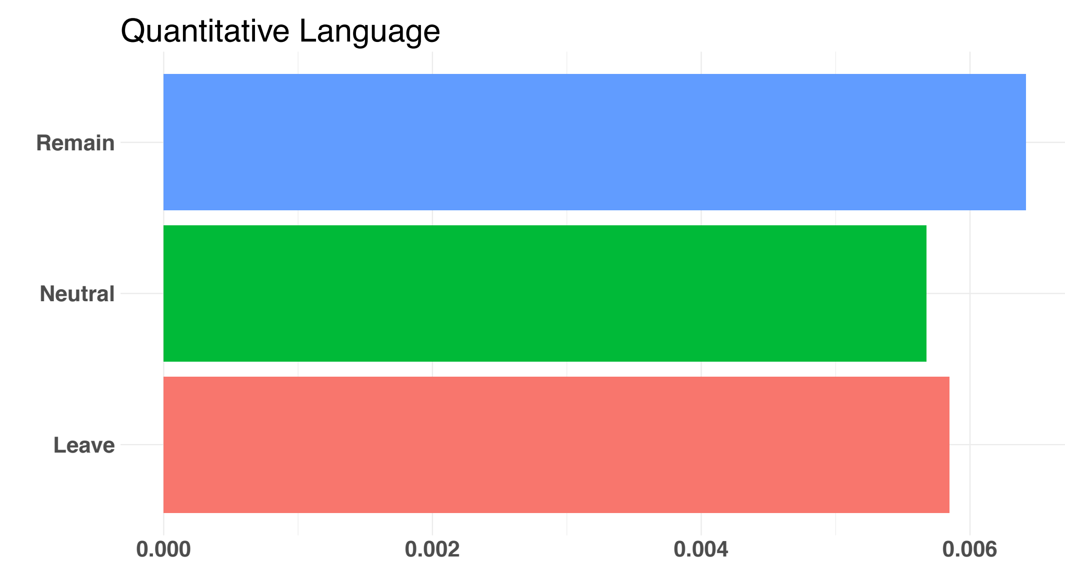

Sad Language
Remain
Neutral
Leave
0.0000
0.0005
0.0010
0.0015
0.0020


Future v. Past Language
Remain
Neutral
Leave
–0.4
–0.3
–0.2
–0.1
0.0


Tentative Language
Remain
Neutral
Leave
0.000
0.002
0.004
0.006

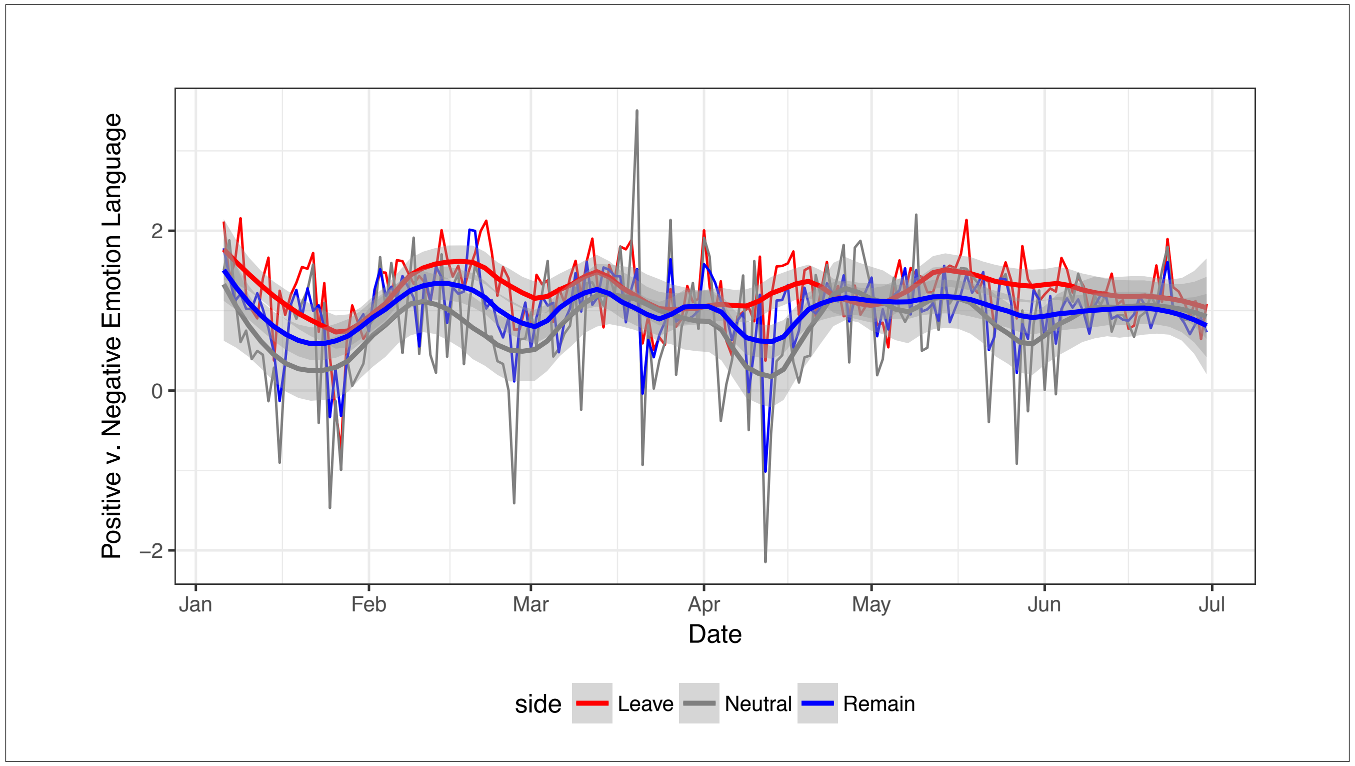


Power Language
Remain
Neutral
Leave
0.000
0.005
0.010

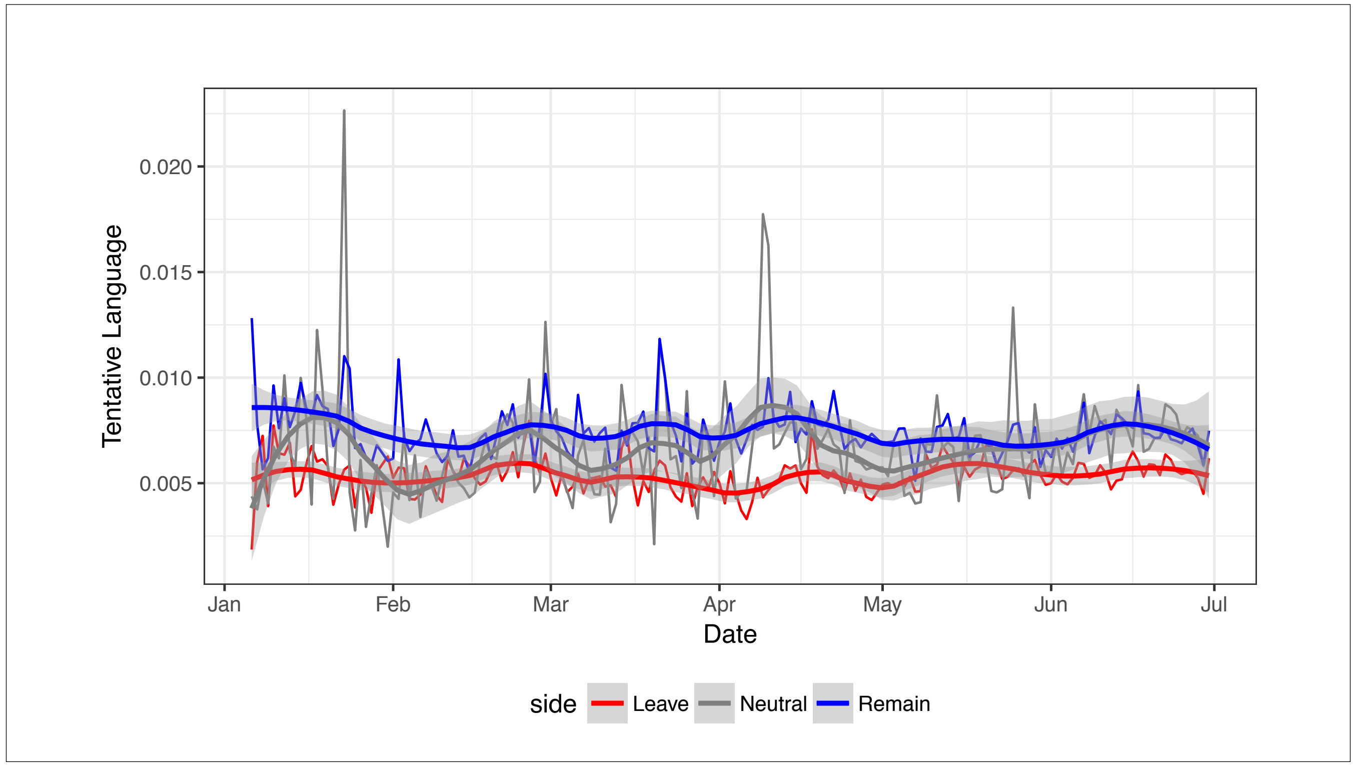

Quantitative Language
Remain
Neutral
Leave
0.000
0.002
0.004
0.006

Sentiment Analysis conclusions: Leave was more

- reward-oriented
- positive
- assertive of power
- less quantitative
- less tentative
- less sadness
- future-, versus past-, oriented



Positive v. Negative Emotion Language
2
0
−2
Jan
Feb
Mar
Apr
May
Jun
Jul
Date
side
Leave
Neutral
Remain



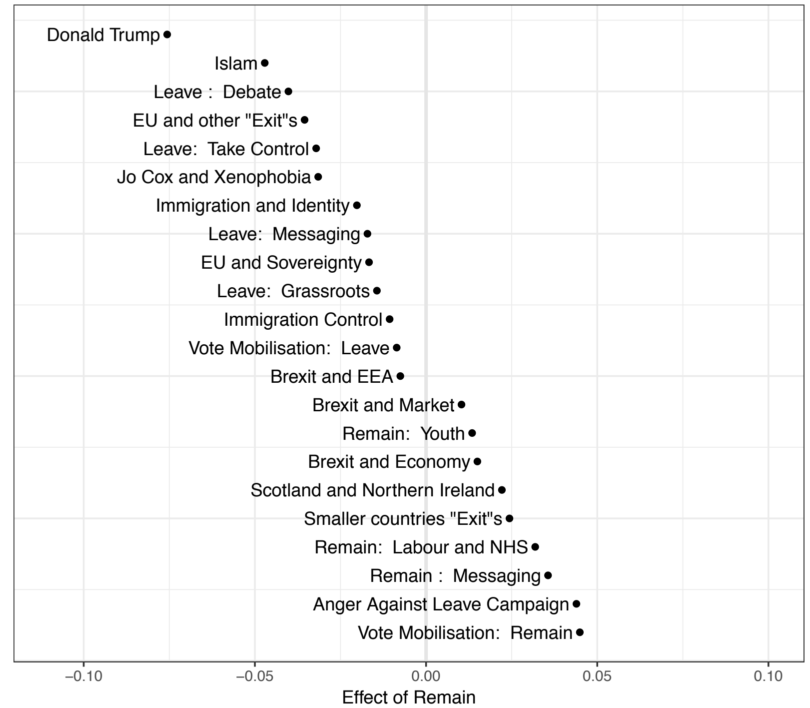
Tentative Language
0.020
0.015
0.010
0.005
Jan
Feb
Mar
Apr
May
Jun
Jul
Date
side
Leave
Neutral
Remain

Topic models

- Using unsupervised machine learning
  - detect topics in twitter conversation
  - topic distributions across sides
- Data
  - tweets from Leave and Remain accounts identified from NB classification
  - combine tweets from each account
  - 700,000 accounts are included
- Method
  - Structural Topic Model (STM) by Roberts, Stewart, and Tingley (2014)
  - Estimate models with 10-40 topics (incremented by 5)
  - covariate in topic prevalence: predicted sides of accounts

STM Results

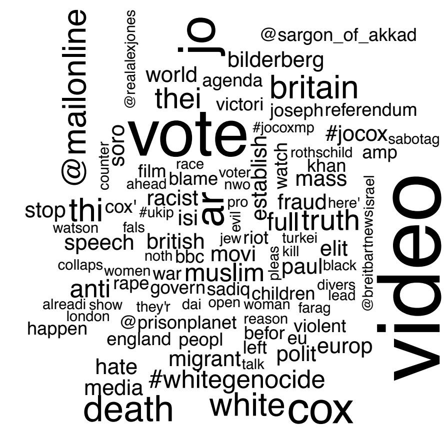
- Selected topics from 40 topics model



STM Results (example of leave topics)

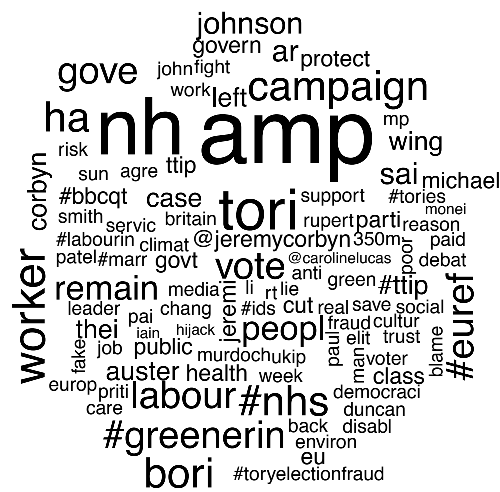
Immigration, Islam

Jo Cox and Xenophobia



STM Results (example of remain topics)

Labour, NHS



Scotland, Northern Ireland

STM Results (examples of remain economic topics)

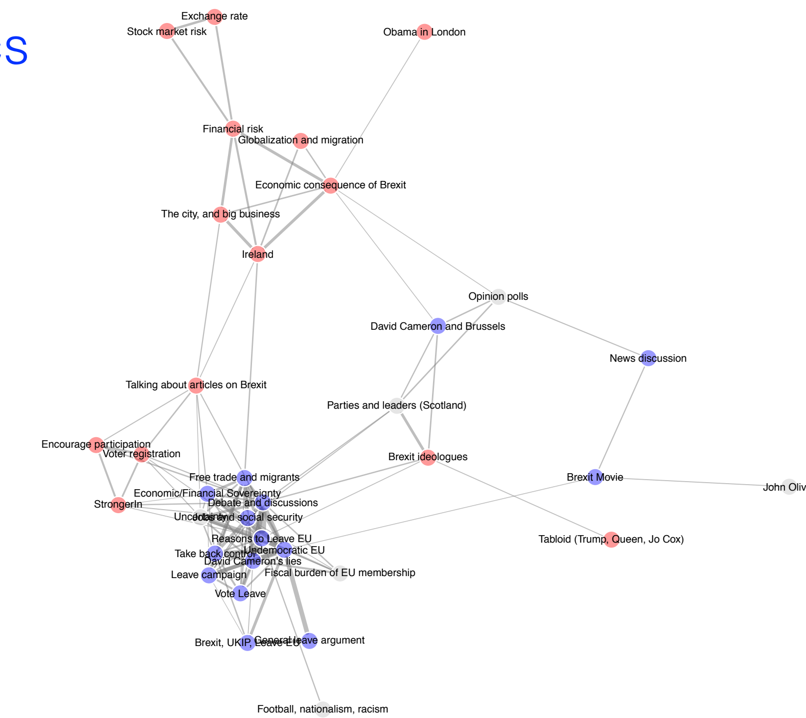
Brexit and Market

Brexit and Economy

Topics by side

Remain	Brexit ideologues	Leave	Brexit Movie
	Economic consequence of Brexit		Brexit, UKIP, Leave EU
	Encourage participation		David Cameron and Brussels
	Exchange rate		David Camerons lies
	Financial risk		Debate and discussions
	Globalization and migration		Economic/Financial Sovereignty
	Ireland		Free trade and migrants
	Obama in London		General leave argument
	Stock market risk		Jobs and social security
	StrongerIn		Leave campaign
	Tabloid (Trump, Queen, Jo Cox)		News discussion
	Talking about articles on Brexit		Reasons to Leave EU
	The city, and big business		Take back control
	Voter registration		Undemocratic EU
			Vote Leave

Networks of topics



Summary: Text analysis was used to

- predict the side of the user
- determining the most common hashtags by side
- using followership networks to estimate “ideology”
- measuring sentiment using dictionaries
- analyzing the topics that were discussed
- mapping connections between topics via networks

How? Using software written in R (and C++ and Python)

quanteda
Quantitative Analysis of Textual Data

CRAN 1.3.0 | downloads 8055/month | downloads 136K | build passing | build passing | codecov 88% | DOI 10.5281/zenodo.1219063 | JOSS Under Review

Aug 12, 2012 – Jun 14, 2018

Contributions: Commits

Contributions to master, excluding merge commits