

Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies

Douglas Rice
Christopher Zorn

February 20, 2015



McCutcheon v. FEC, Breyer's Dissent

"Today a majority of the Court overrules this holding. It is wrong to do so. Its conclusion rests upon its own, not a record-based, view of the factors. Its legal analysis is faulty; it misconstrues the nature of the competing constitutional interests at stake. It understates the importance of protecting the political integrity of our governmental institutions. It creates a loophole that will allow a single individual to contribute millions of dollars to a political party or to a candidate's campaign. Taken together with *Citizen's United v. FEC*, today's decision eviscerates our Nation's campaign finance laws, leaving a remnant incapable of dealing with the grave problems of democratic legitimacy that those laws were intended to resolve."



McCutcheon v. FEC, McConnell's Press Release

“The Supreme Court has once again reminded Congress that Americans have a Constitutional First Amendment right to speak and associate with political candidates and parties of their choice.”



McCutcheon v. FEC, Democrats on Twitter



Nancy Pelosi ✓

@NancyPelosi

Follow

W/ [#McCutcheon](#), SCOTUS has chosen to pour even more money into our process & politics. We must restore fairness & pass the [#ByThePeople](#) Act.

10:59 AM - 2 Apr 2014

150 RETWEETS 69 FAVORITES



Sen. Robert Menendez ✓

@SenatorMenendez

Follow

Awful decision - \$ shouldn't buy politics!
[#AmericaIncorporated](#) MT [@jonresnickAP](#): SCOTUS strikes down overall campaign contributions limits.

10:37 AM - 2 Apr 2014

15 RETWEETS 5 FAVORITES



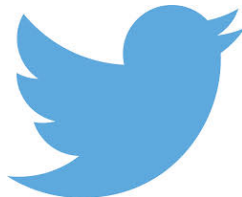
“...[C]omputational study of how opinions, attitudes, emotions, and perspectives are expressed in language...”

– Liu, 2011

Sentiment Analysis

“...[C]omputational study of how opinions, attitudes, emotions, and perspectives are expressed in language...”

– Liu, 2011



Two General Approaches:

- Supervised Machine Learning
 - Classify subset of documents →
 - Train classifier on subset →
 - Score remaining (test) texts.

Two General Approaches:

- Supervised Machine Learning
 - Classify subset of documents →
 - Train classifier on subset →
 - Score remaining (test) texts.
 - Problem? No / expensive “training” data.

Two General Approaches:

- Supervised Machine Learning

- Classify subset of documents →
- Train classifier on subset →
- Score remaining (test) texts.
- Problem? No / expensive “training” data.

- Dictionary-Based

- Identify positive and negative words *ex ante*
- Create document scores based upon those words.

Two General Approaches:

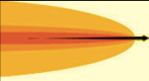
- Supervised Machine Learning

- Classify subset of documents →
- Train classifier on subset →
- Score remaining (test) texts.
- Problem? No / expensive “training” data.

- Dictionary-Based

- Identify positive and negative words *ex ante*
- Create document scores based upon those words.
- Problems? *Domain dependence and validation.*

Standard Sentiment Dictionaries



Linguistic Inquiry and Word Count

- Home
- Dictionaries ▾
- How it Works
- Try Online
- Contact Us

[Click here to buy LIWC](#)


LIWC2007 versus LIWCite7

Compare 2007 and 2001 dictionaries

ANALYZE WORDS

Experience the power of LIWC for twitter personality analysis

Web Hosting Search



What is LIWC?

Linguistic Inquiry and Word Count (LIWC) is a text analysis software program designed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis. LIWC calculates the degree to which people use different categories of words across a wide array of texts, including emails, speeches, poems, or transcribed daily speech. With a click of a button, you can determine the degree any text uses positive or negative emotions, self-references, causal words, and 70 other language dimensions.

The LIWC program can analyze hundreds of standard ASCII text files or Microsoft Word documents in seconds. The LIWC2007 program also allows you to build your own dictionaries to analyze dimensions of language specifically relevant to your interests. The Macintosh version of LIWC2007 has a feature that will highlight in color all the words found in a particular file when it is analyzed. Users can also create dictionaries that include literal phrases (e.g. 'you know') as well as individual words and word stems. These features will soon be available for the Windows LIWC2007 version as well.

The student version of LIWC, LIWCite7, only analyzes plain text files using the LIWC2007 and earlier LIWC2001 dictionaries. LIWCite7 is the student version that is ideal for people with limited text analysis needs.

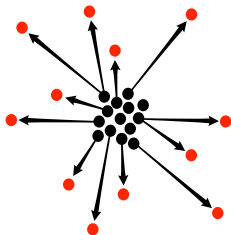
LIWC license

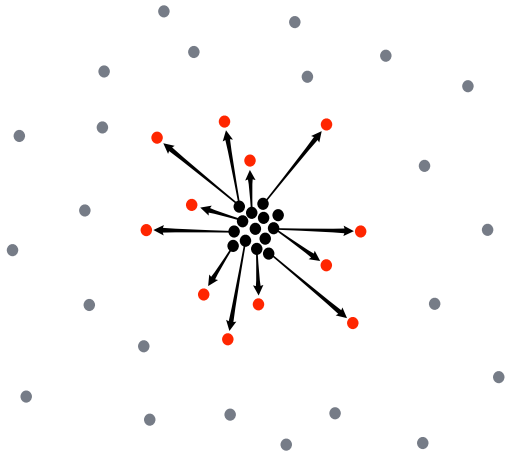
A single-user license for LIWC2007 or LIWCite7 entitles you to install the software on no more than two computers, however discounts available for multi-user versions (see [End User License Agreement here](#)).

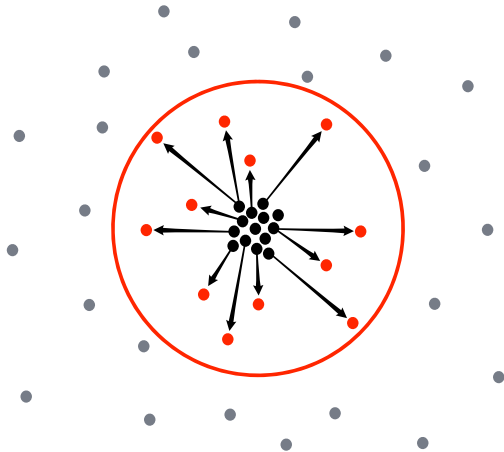
Learn more

To learn more about the development and uses of LIWC, click the 'How it Works' link in the menu above. You can also read more about the categories and dictionary features of the LIWC2007 dictionary by [clicking this link](#).

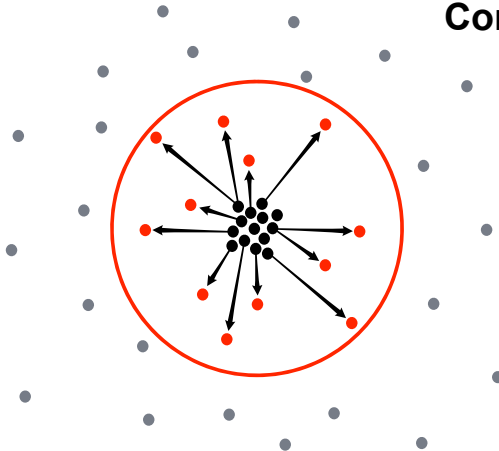


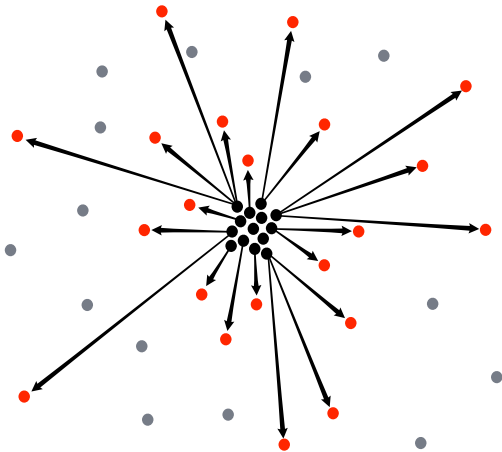


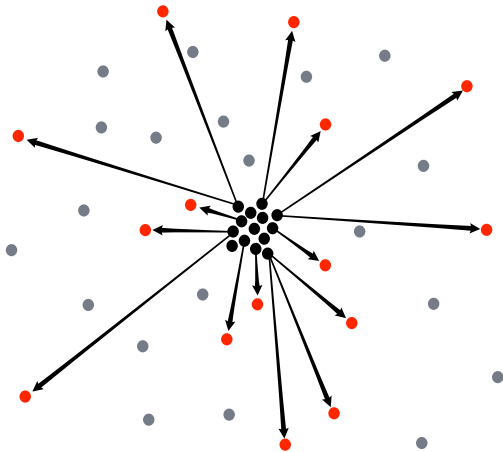




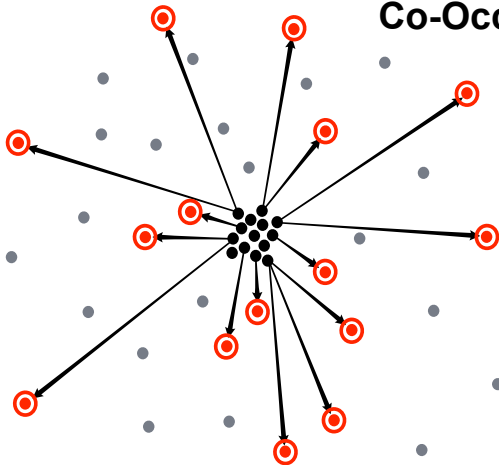
Conjoined







Co-Occurrence

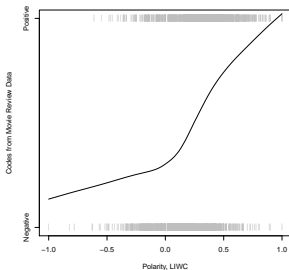


Validation: Movie Reviews

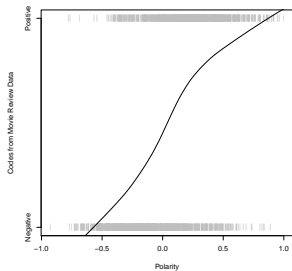


- Source: IMDB (Pang and Lee 2004)
- 1000 positive, 1000 negative
- General vocabulary + benchmarks

Movie Reviews: Comparison With Dictionary



LIWC



Polarity

Movie Reviews: Comparison with Classifiers

Model	Mean	Min	Max
Naive Bayes	79.7	77.0	81.5
Maximum Entropy	79.7	77.4	81.0
Support Vector Machines	79.4	72.8	82.9
LIWC Only	58.5	-	-
Our Combined Measure	72.5	-	-

NOTE: Estimates for naive Bayes, maximum entropy, and support vector machines classifiers are taken from Pang et al. (2002).

Application: U.S. Supreme Court

“The position urged by respondent would upset this carefully drawn approach in a manner that is both unnecessary for the protection of the Fifth Amendment privilege and injurious to legitimate law enforcement.”

“Questions of precedent to one side, we find respondent’s understanding of the Sixth Amendment both practically and theoretically unsound.”

– Justice O’Connor for the majority in *Moran v. Burbine*, 475 U.S. 412 (1986)



Application: U.S. Supreme Court



“It is not only the Court’s ultimate conclusion that is deeply disturbing; it is also its manner of reaching that conclusion.”

“[T]he Court’s truncated analysis . . . is simply untenable.”

“[T]he Court’s balancing approach is profoundly misguided.”

– Justice Stevens, dissenting.

Application: U.S. Supreme Court

“Justice Stevens’ apocalyptic suggestion that we have approved any and all forms of police misconduct is demonstrably incorrect.”

“Among its other failings, the dissent declines to follow *Oregon v. Elstad*, a decision that categorically forecloses Justice Stevens’ major premise....Most importantly, the dissent’s misreading of *Miranda* itself is breathtaking in its scope.”

The dissent’s “lengthy exposition” featured an “entirely undefended suggestion” and “incorrectly reads our analysis.”



Comparison of Approaches

Supervised Learning

- Vote splits?
- Dissenting opinions and unanimous majorities.
- Treat class probability as polarity.
- Trained random forest model using SCDB data.

Comparison of Approaches

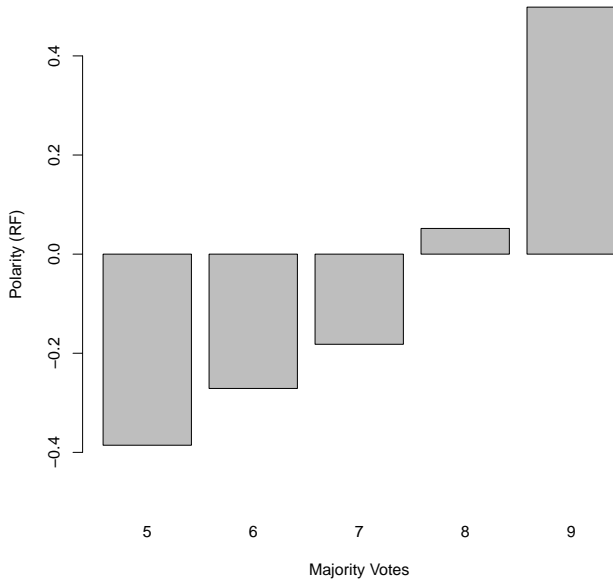
Supervised Learning

- Vote splits?
- Dissenting opinions and unanimous majorities.
- Treat class probability as polarity.
- Trained random forest model using SCDB data.

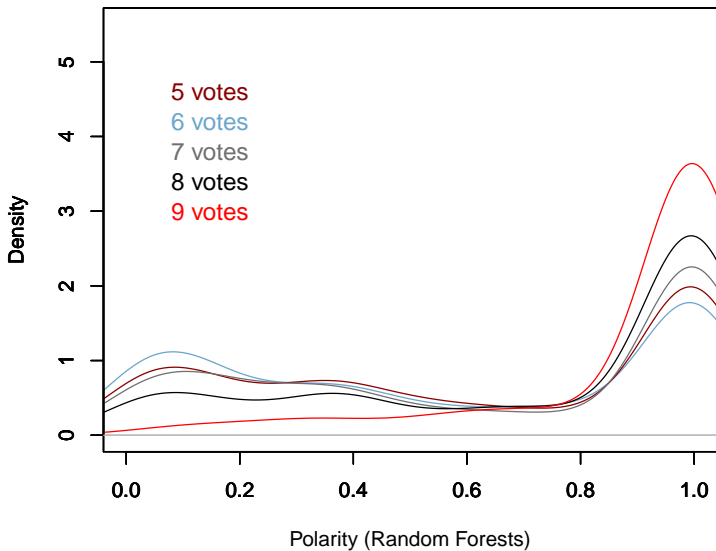
Our Approach

- All U.S. Supreme Court opinions, 1793-2000.
- Estimate polarity for all 34,946 total opinions.
- Compare across SCDB subset.

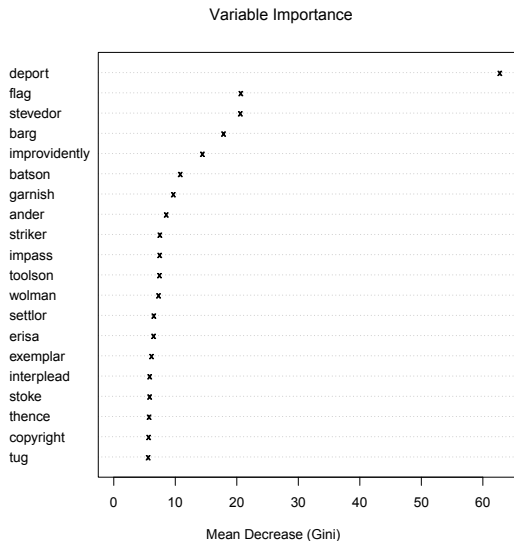
Results: Random Forest



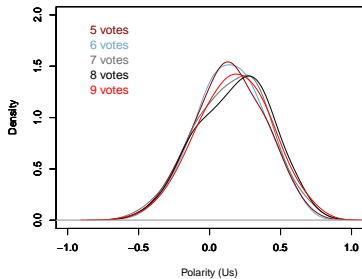
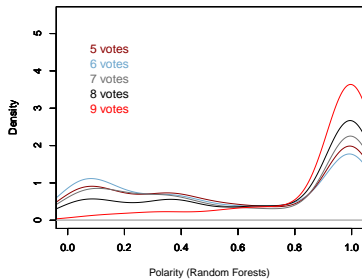
Results: Random Forest



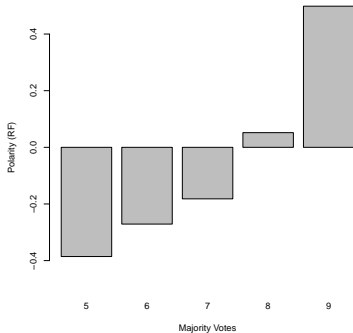
Random Forest: Feature Importance



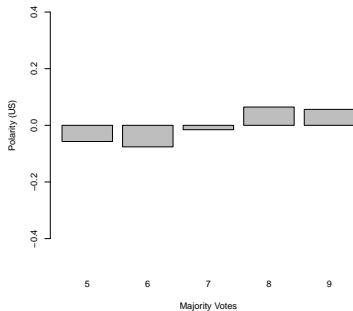
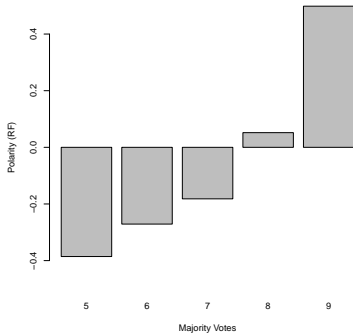
Comparison



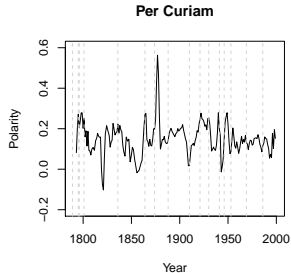
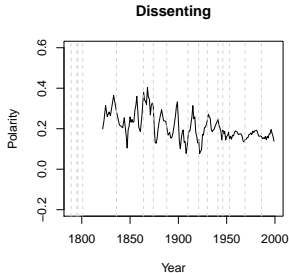
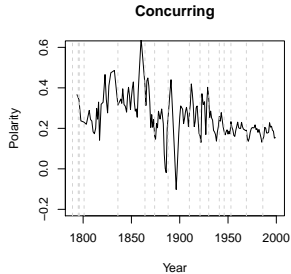
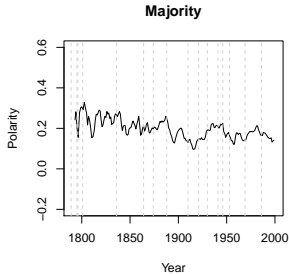
Comparison: Machine Learning



Comparison: Machine Learning



Application: SCOTUS Polarity Over Time



Potential Advantages

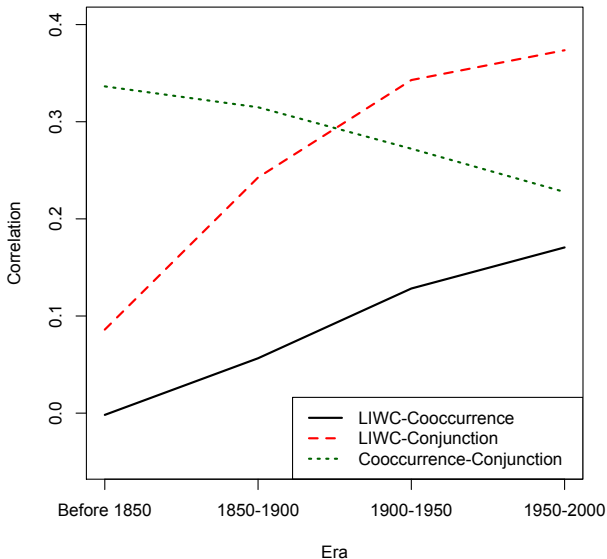
- Context-Specific
- “Minimally Supervised”
- Simple / Intuitive
- Ensemble-able

- Validation / Benchmarking
- Sensitivity Analyses
- Breadth of Applicability
- Application to Temporal Change

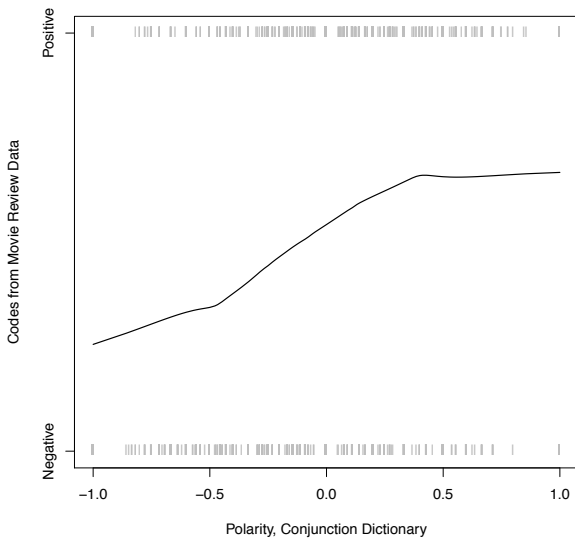
Questions / Comments?

- Eliminate capitalization, punctuation, stop words
- Tokenize negations (“no,” “not,” etc.)
- Tag TOS; retain adverbs, adjectives, nouns (Toutanova et al. 2003)
- Also test against a general dictionary (LIWC)

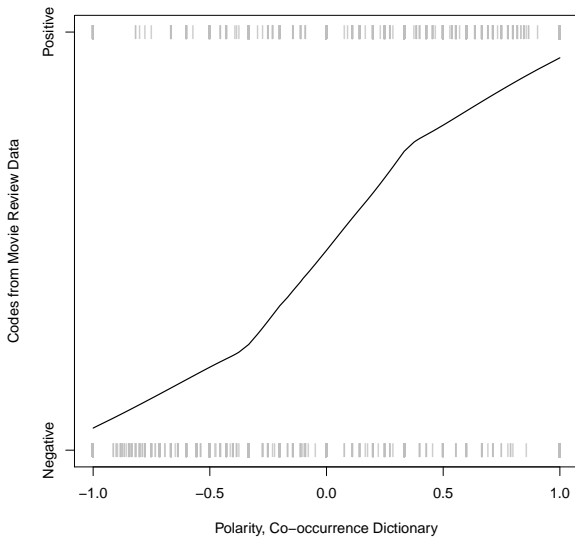
Inter-Measure Correlations Over Time



Polarity by Assigned Rating: Conjunction



Polarity by Assigned Rating: Co-Occurrence



Subjectivity and Polarity by Justice

