PLSC 597: Modern Measurement

Sentiment Analysis and Other Dictionary-Based Methods

March 29, 2018

Overview: Dictionary-Based Methods

- **Classification** task:
  - *Categorize* documents into classes C, and/or
  - *Score* documents degree of association with those classes.

- Heuristic: **Dictionaries assign weights to words / terms.**

- Formally: For $j \in \{1...J\}$ words in a corpus of $i = \{1...N\}$ documents, the *document score* is:

$$S_i = \frac{\sum_{j=1}^{J} \omega_j X_{ij}}{\sum_{j=1}^{J} X_{ij}}$$

  where
  - X_{ij} is the number of instances of word j in document i, and
  - ω_j is the weight assigned to each word by the dictionary.

General Dictionary-Based Methods: How-To

1. Obtain / preprocess documents (stemming, stop words, etc.)
2. Obtain / create a dictionary
3. Score documents by calculating S_i
   - Weights ω_j can be positive or negative
   - Words in the corpus but not in the dictionary have $\omega_j = 0$
4. (Optional:) Classify documents by mapping $S_i \rightsquigarrow C_i$

Toy Example: "Truthiness"

- Document: {TRUE FALSE TRUE FALSE TRUE}
- Dictionary:

Term	ω_j
TRUE	1.0
FALSE	0.0

- Word counts:

$$\mathbf{x} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- Score:

$$S_i = \frac{(1.0 \times 3) + (0.0 \times 2)}{(3 + 2)} = \frac{3}{5} = 0.6$$

Dictionary-Based Classification Tasks

- Topic(s)
  - What are documents *about*?
  - What thing(s) are *emphasized*?
- Sentiment
  - What is the *emotional valence* of the documents?
  - What are the emotions expressed? (pity, anger, jealousy, etc.)
- Tone / Style
  - Authorship / provenance
  - Specialization of language (e.g., "hold harmless")

Sentiment Analysis

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

– Liu (2011)

Lots of research in computer science and linguistics: Pang and Lee (2004, 2008), Tong (2001), Zhou, Chen and Wang (2010), Das and Chen (2001), Dasgupta and Ng (2009), Pang et al. (2002), Turney (2002), Wiebe (2000), Shanahan, Qu, and Wiebe (2006), Jindal and Liu (2006), Liu (2006), Nigam and Hurst (2005), Hu and Liu (2004), Choi and Cardie (2010), and many, many more...

A good overview is:

Pang, Bo, and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval 2:1-135.

Example...

"Arizona bears the brunt of the country's illegal immigration problem. Its citizens feel themselves under siege by large numbers of illegal immigrants who invade their property, strain their social services, and even place their lives in jeopardy. Federal officials have been unable to remedy the problem, and indeed have recently shown that they are unwilling to do so."

– Justice Scalia, dissenting in *Arizona v. United States* (2012)

Where Do (Sentiment) Dictionaries Come From?

- "Standard" dictionaries
  - Code sentiment in common (contemporary, usually American) English
  - See below; there's a list here
- "By hand"...
  - Requires careful thought / luck / divine help
  - **Validate**. Seriously.
- "Crowdsourced" methods: RAs, MTurk, etc.
  - "On a scale from -10 to 10, how positive is the word...?"
  - Can be made context-specific, etc.
- Statistical approaches
  - Fit a model to some document-level outcome $\rightarrow$ most predictive words = dictionary
  - "Model" = lasso / ridge regression / elastic net, etc.
  - Again, **validation** is key...

Common (English) Sentiment Dictionaries

- General Inquirer (`http://www.wjh.harvard.edu/~inquirer/`)
- AFINN (`http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010`)
- QDAP dictionaries (`https://cran.r-project.org/web/packages/qdap/index.html`)
- WordStat (find it here)
- LIWC (`http://liwc.wpengine.com/`)

Sentiment Dictionary Examples

General Inquirer:

- Words scored either positive (+1) or negative (-1)
- 1637 positive words, 2005 negative words

AFINN (2477 total words, scored [-5,5]):

Term	ω_j
bastard	-5
bitch	-5
⋮	⋮
worn	-1
some kind	0
aboard	1
⋮	⋮
superb	5
thrilled	5

Sentiment Analysis Options in R

- `SentimentAnalysis`
  - Built by finance people → dictionaries, etc.
  - Plays well with `tm`
  - My current favorite (see the vignette)

- `tidyverse`, etc.
  - Requires admission to the cult of Wickham
  - Details here: `https://www.tidytextmining.com/`
  - Tons of tutorials (here, here, here, etc.)

- `RSentiment` (super minimal)

- `sentiment` (deprecated)

`SentimentAnalysis` Details

- Works with character objects, data frames, corpuses / TDMs / DTMs from `tm`
- Built-in dictionaries: General Inquirer, QDAP, two finance-specific (Henry 2008; Loughran and McDonald 2011)
- Can also create dictionaries "by hand" or through predictive power of words vis-a-vis some response (via glm, lasso, etc.)
- `analyzeSentiment` is the workhorse
  - Defaults to using all four built-in dictionaries
  - Stems and removes stop words by default
  - Outputs a `data.frame` with document-level sentiment scores
- Other useful things:
  - Built-in tokenizer / N-gram creator
  - Convert continuous sentiment scores to binary (0/1) or directional (-1/0/1) values
  - Can generate predictions and assess predictive performance...

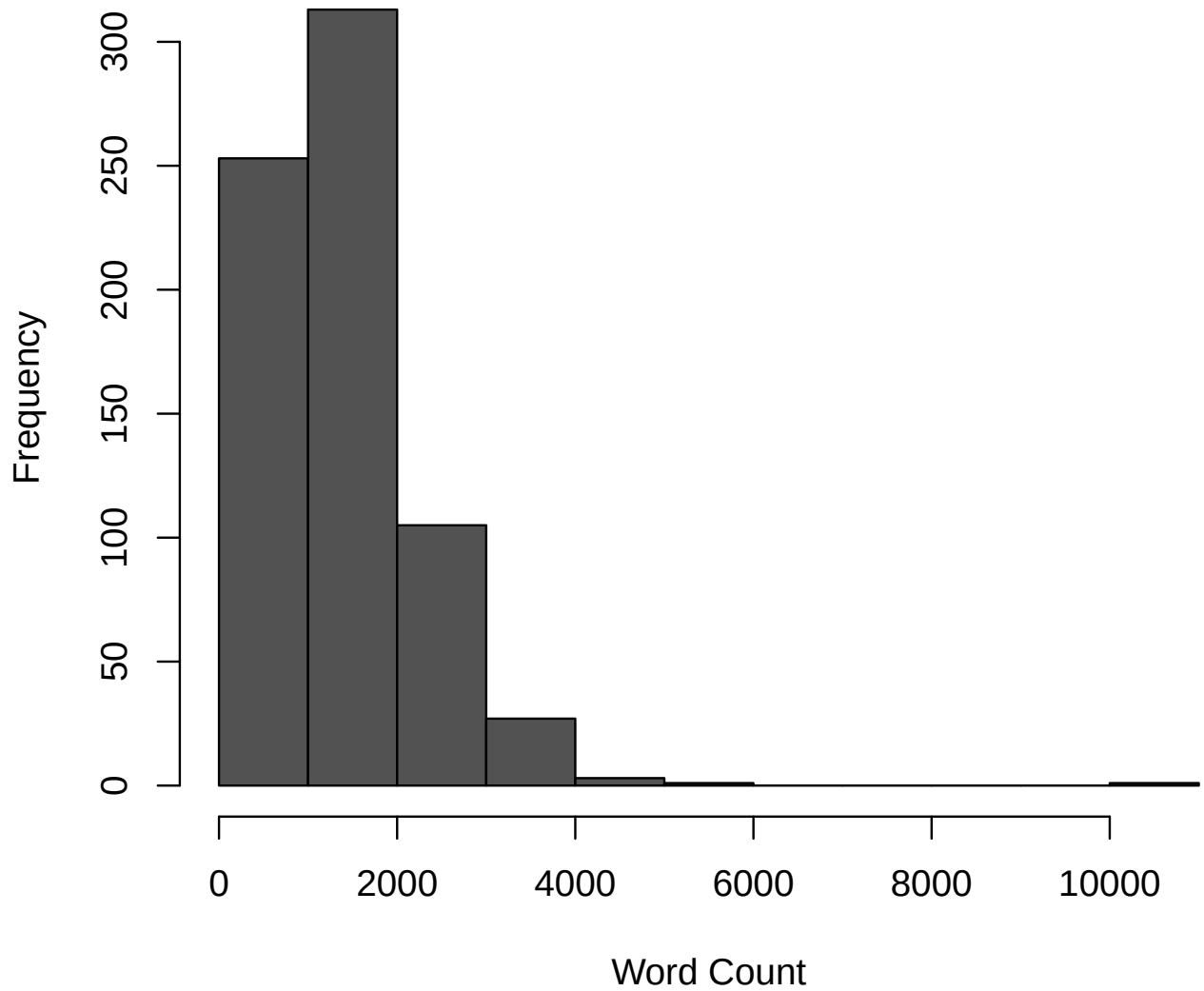
Running Example: UNHCR Speeches

- All speeches made by the High Commissioner of the U.N. Refugee Agency, 1970-2016 ($N = 703$)
- Metadata include ID, speaker, title, and date
- Source: `https://www.kaggle.com/franciscadias/un-refugee-speech-analysis/`

UNHCR Speeches...

```
> UN <- read.csv(text=temp,
+                stringsAsFactors=FALSE,allowEscapes=TRUE)
> rm(temp)
>
> UN$content <- removeNumbers(UN$content) # no numbers
> UN$content <- str_replace_all(UN$content, "[\n]", " ") # line breaks
> UN$content <- removeWords(UN$content,stopwords("en")) # remove stopwords
> UN$Year <- as.numeric(str_sub(UN$by, -4)) # Year of the speech
> UN$foo <- str_extract(UN$by, '\\b[^,]+$')
> UN$Date <- as.Date(UN$foo, format="%d %B %Y") # date of speech
> UN$foo <- NULL
> UN$Author <- "Goedhart"  # Fix names...
.
.
.
> # Corpus:
>
> UN2 <- with(UN, data.frame(doc_id = id,
+                            text = content))
> ds <- DataframeSource(UN2)
> UNC <- Corpus(ds)
> meta(UNC)
data frame with 0 columns and 703 rows
>
> # Some tools in SentimentAnalysis...
>
> UNCount<-countWords(UNC,removeStopwords=FALSE)
> summary(UNCount$WordCount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     50     762    1283    1404    1864   10948
```

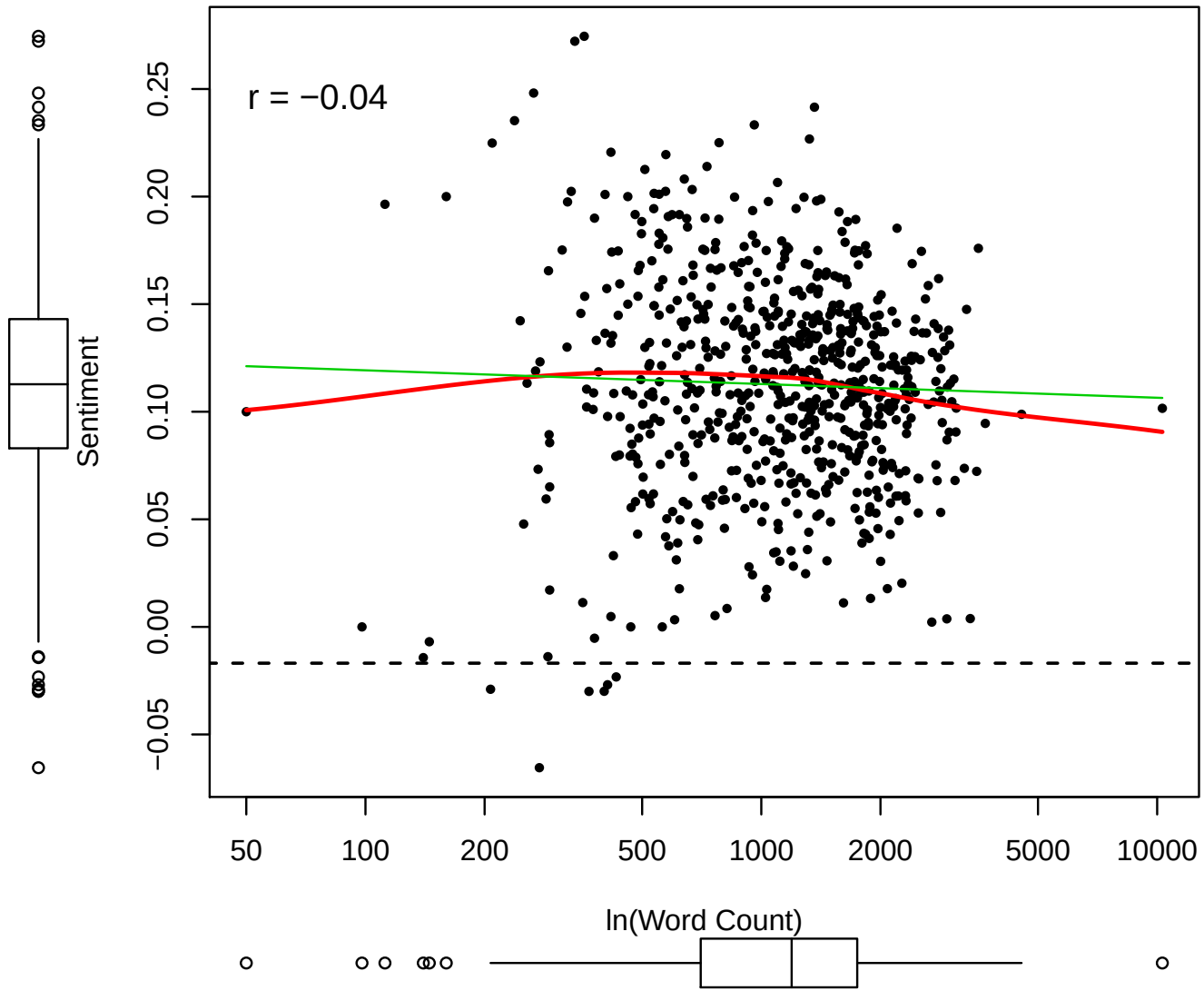
UNHCR Speech Word Counts, 1970-2016



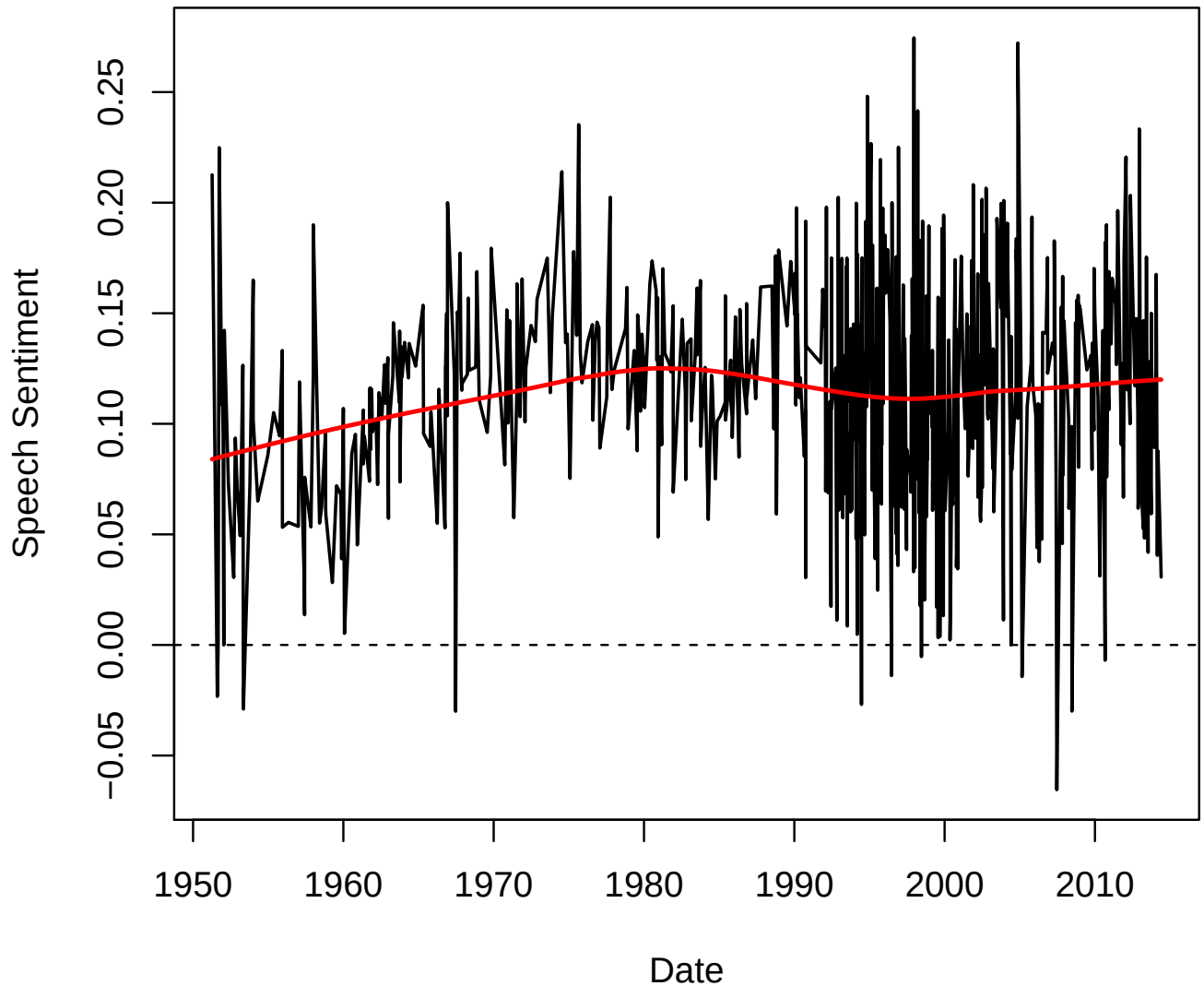
Simple Sentiment Analysis

```
> UNSent <- analyzeSentiment(UNC)

> summary(UNSent)
   WordCount      SentimentGI       NegativityGI     PositivityGI    SentimentHE
 Min.   :   50   Min.   :-0.065   Min.   :0.002   Min.   :0.00   Min.   :-0.011
 1st Qu.:  703   1st Qu.: 0.083   1st Qu.:0.115   1st Qu.:0.23   1st Qu.: 0.011
 Median : 1193   Median : 0.113   Median :0.134   Median :0.25   Median : 0.017
 Mean   : 1299   Mean   : 0.113   Mean   :0.135   Mean   :0.25   Mean   : 0.017
 3rd Qu.: 1747   3rd Qu.: 0.143   3rd Qu.:0.154   3rd Qu.:0.27   3rd Qu.: 0.022
 Max.   :10306   Max.   : 0.275   Max.   :0.237   Max.   :0.36   Max.   : 0.072
  NegativityHE     PositivityHE    SentimentLM      NegativityLM    PositivityLM
 Min.   :0.0000   Min.   :0.000   Min.   :-0.119   Min.   :0.000   Min.   :0.000
 1st Qu.:0.0043   1st Qu.:0.019   1st Qu.:-0.043   1st Qu.:0.045   1st Qu.:0.026
 Median :0.0070   Median :0.024   Median :-0.024   Median :0.057   Median :0.032
 Mean   :0.0075   Mean   :0.025   Mean   :-0.027   Mean   :0.060   Mean   :0.032
 3rd Qu.:0.0101   3rd Qu.:0.029   3rd Qu.:-0.009   3rd Qu.:0.073   3rd Qu.:0.038
 Max.   :0.0249   Max.   :0.072   Max.   : 0.044   Max.   :0.136   Max.   :0.068
 RatioUncertaintyLM SentimentQDAP    NegativityQDAP  PositivityQDAP
 Min.   :0.000      Min.   :-0.066   Min.   :0.000   Min.   :0.003
 1st Qu.:0.011      1st Qu.: 0.064   1st Qu.:0.056   1st Qu.:0.144
 Median :0.014      Median : 0.084   Median :0.075   Median :0.160
 Mean   :0.015      Mean   : 0.084   Mean   :0.076   Mean   :0.161
 3rd Qu.:0.019      3rd Qu.: 0.108   3rd Qu.:0.094   3rd Qu.:0.178
 Max.   :0.044      Max.   : 0.231   Max.   :0.174   Max.   :0.260
```

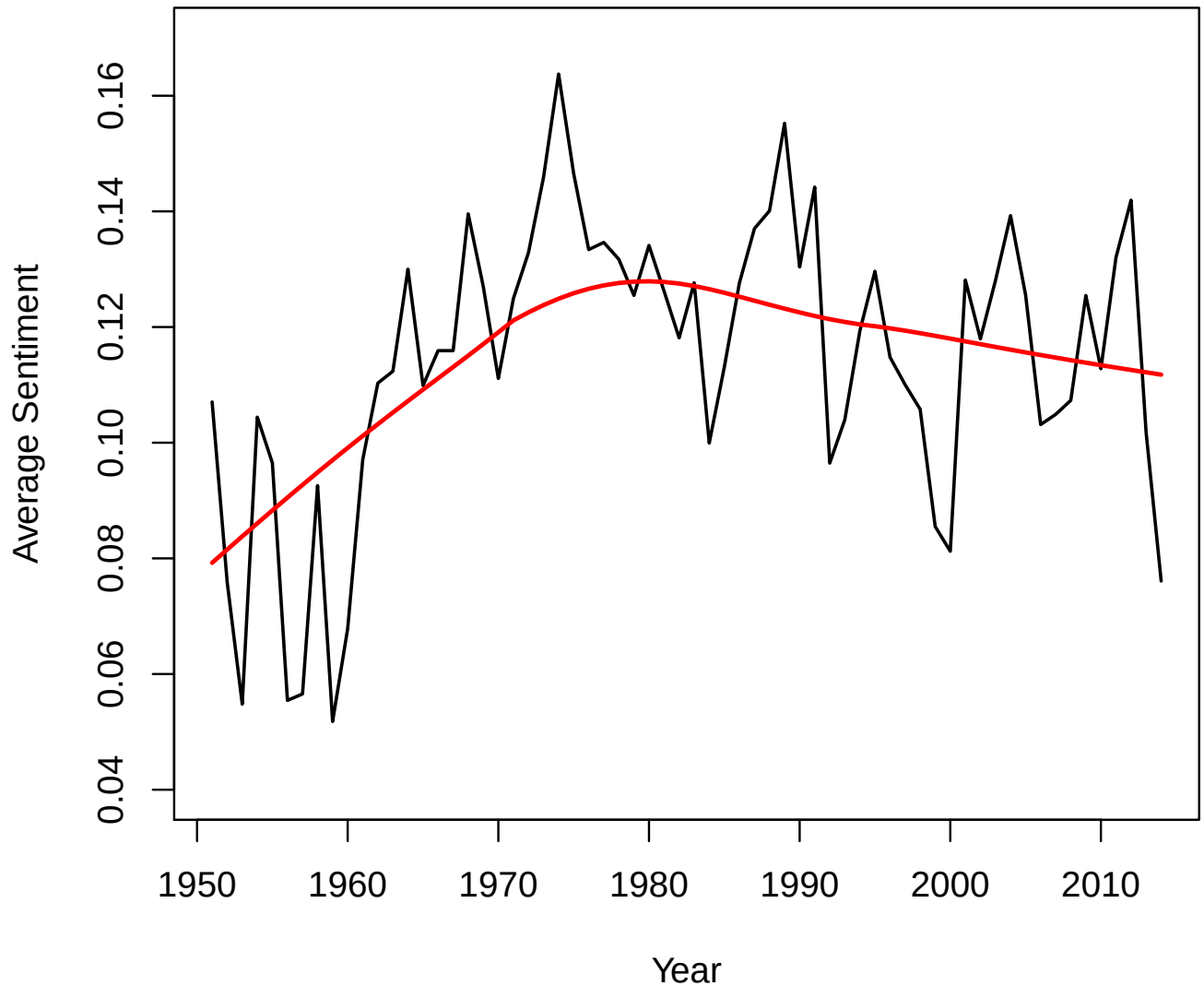
UNHCR: Sentiment vs. Word Count



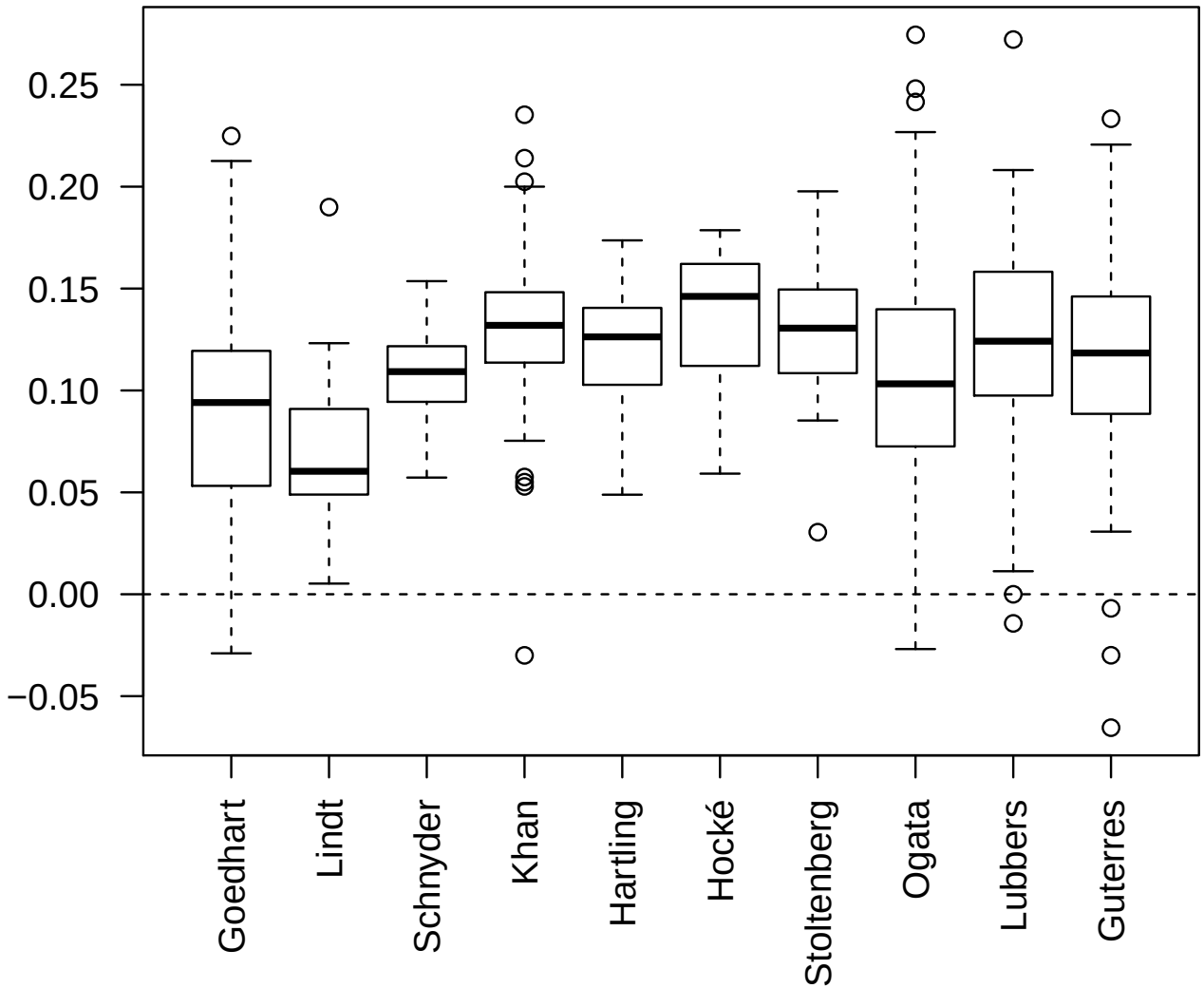
UNHCR: Sentiment Over Time



UNHCR: Annual Sentiment Means



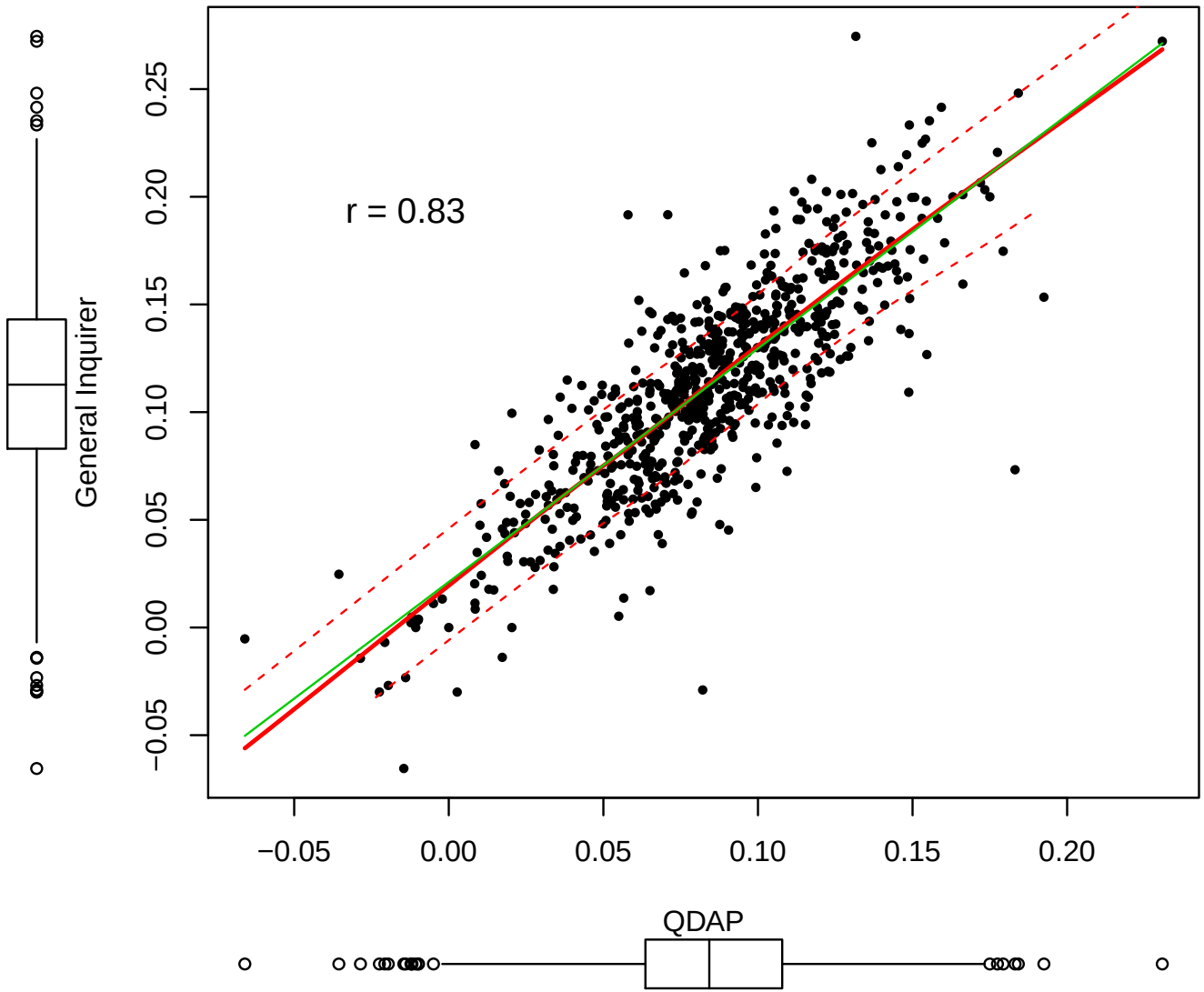
UNHCR: Sentiment By Speaker



Similar Results By Dictionary?

```
> GI<-loadDictionaryGI()
> QD<-loadDictionaryQDAP()
>
> compareDictionaries(GI,QD)
Comparing: binary vs binary

Total unique words: 5100
Matching entries: 2136 (0.42%)
Entries with same classification: 1448 (0.28%)
Entries with different classification: 63 (0.012%)
$totalUniqueWords
[1] 5100

$totalSameWords
[1] 2136

$ratioSameWords
[1] 0.42

$numWordsEqualClass
[1] 1448

$numWordsDifferentClass
[1] 63

$ratioWordsEqualClass
[1] 0.28

$ratioWordsDifferentClass
[1] 0.012
```

Comparing Results w/Different Dictionaries



For Whom Does Dictionary Choice Matter?

```
> DictDiff <- with(UNSent, abs(SentimentGI - SentimentQDAP))

> summary(lm(DictDiff~UN$Author - 1))

Call:
lm(formula = DictDiff ~ UN$Author - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.03758 -0.01658 -0.00173  0.01332  0.10766

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
UN$AuthorGoedhart     0.03151    0.00427    7.39  4.4e-13 ***
UN$AuthorLindt        0.01661    0.00444    3.74   0.0002 ***
UN$AuthorSchnyder     0.02043    0.00340    6.01  3.0e-09 ***
UN$AuthorKhan         0.03012    0.00266   11.33  < 2e-16 ***
UN$AuthorHartling     0.03187    0.00296   10.76  < 2e-16 ***
UN$AuthorHocke        0.04397    0.00487    9.04  < 2e-16 ***
UN$AuthorStoltenberg  0.03973    0.00582    6.83  1.8e-11 ***
UN$AuthorOgata        0.03519    0.00133   26.53  < 2e-16 ***
UN$AuthorLubbers      0.03097    0.00255   12.16  < 2e-16 ***
UN$AuthorGuterres     0.03214    0.00203   15.84  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.022 on 693 degrees of freedom
Multiple R-squared:  0.695,Adjusted R-squared:  0.691
F-statistic:  158 on 10 and 693 DF,  p-value: <2e-16
```

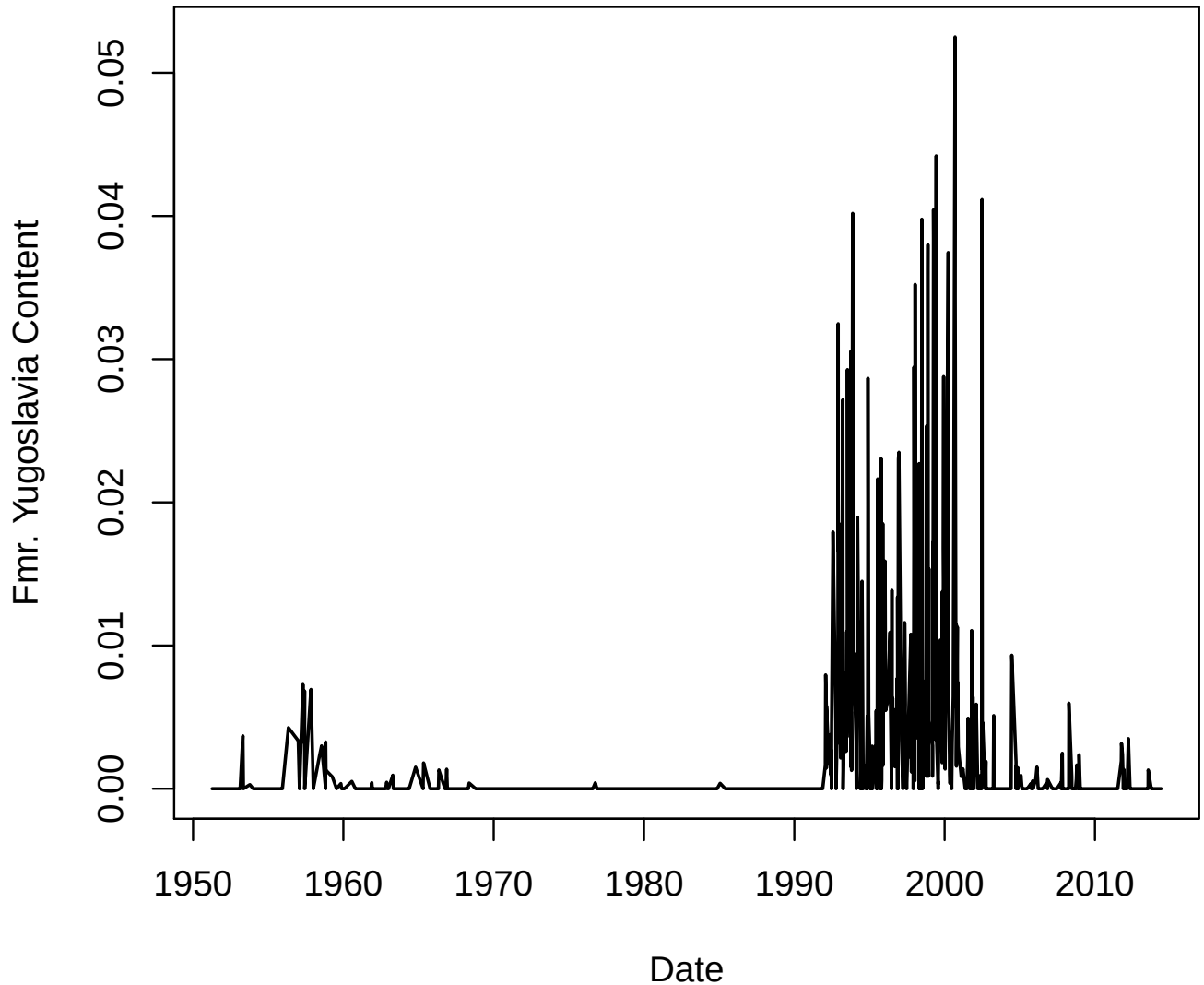
Custom Dictionaries "By Hand"

- Conflict in the former Yugoslavia, 1991-1999
- $\approx$ 2.3 million refugees
- "Europe's biggest refugee crisis since World War II"
- Machine code speeches for content about the former Yugoslavia...

Create and Use a Custom Dictionary

```
> YugoWords <- c("yugoslavia","serbia","bosnia","herzegovina",
+                "kosovo","montenegro","macedonia","croatia",
+                "vojvodina","balkans")

> FmrYugo <- SentimentDictionaryWordlist(YugoWords)

> UNHCRYugo <- analyzeSentiment(UNC,
+                   rules=list("YugoTopic"=list(
+                     ruleRatio,FmrYugo)))

> summary(UNHCRYugo$YugoTopic)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.003   0.003   0.053
```

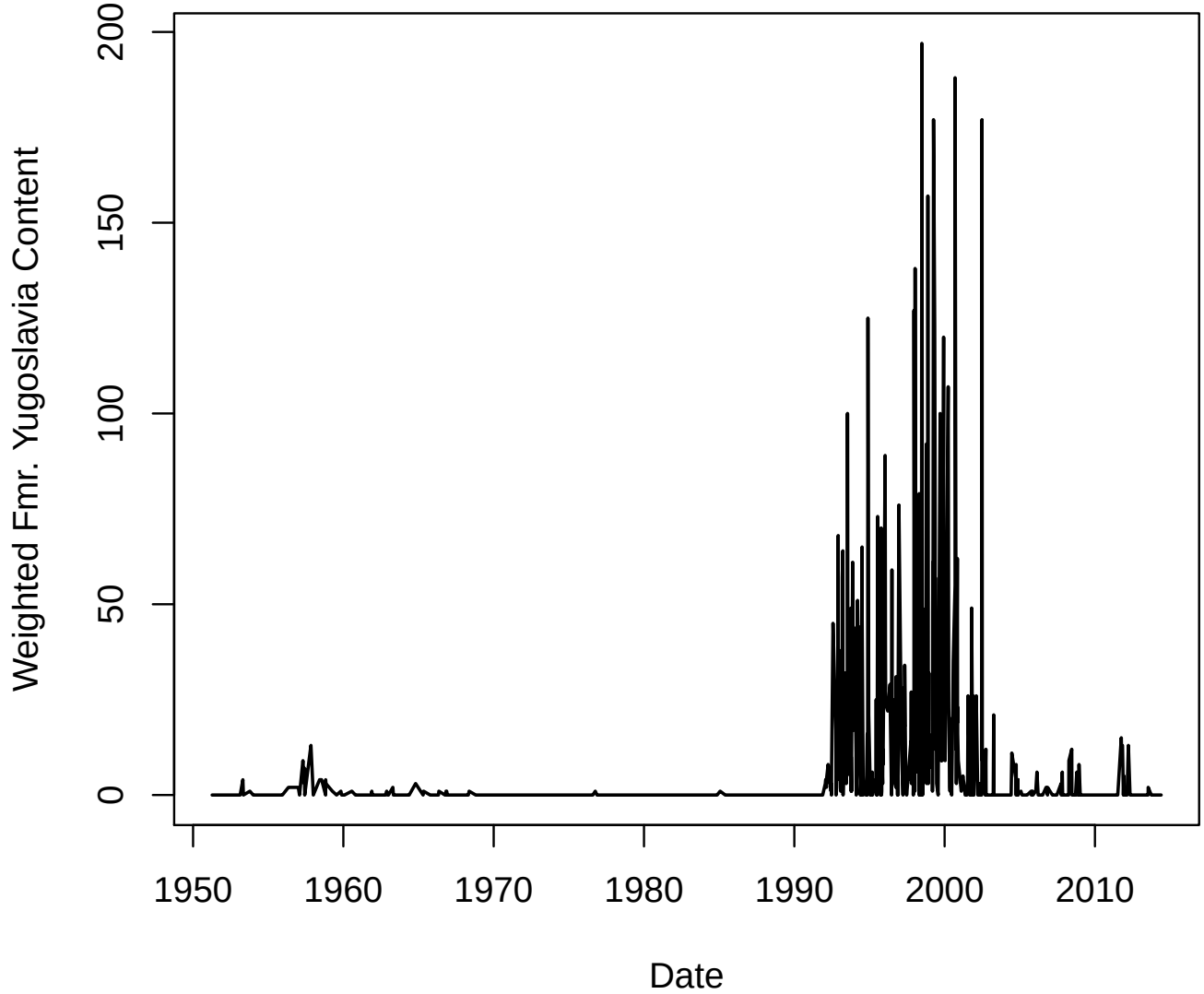
"Former Yugoslavia" Scores Over Time



Adding Weights

```
> YugoWords
 [1] "yugoslavia"  "serbia"      "bosnia"      "herzegovina" "kosovo"
 [6] "montenegro"  "macedonia"   "croatia"     "vojvodina"   "balkans"

> YugoScores <- c(1,3,3,3,3,2,2,2,2,1)

> FmrYugo2 <- SentimentDictionaryWeighted(YugoWords,YugoScores)

> UNHCRYugo2 <- analyzeSentiment(UNC,
+                       rules=list("YugoTopic"=list(
+                       ruleLinearModel,FmrYugo2)))

> summary(UNHCRYugo2)
   YugoTopic
 Min.   :  0
 1st Qu.:  0
 Median :  0
 Mean   : 10
 3rd Qu.:  8
 Max.   :197
```

Weighted "Former Yugoslavia" Scores Over Time



Training A Dictionary

- Dr. Sagato Ogata, U.N. High Commissioner for Refugees, 1991-2000.
- Ph.D. in political science (Berkeley '63)
- Replaced Thorvald Stoltenberg (resigned after only ten months)
- 269 of 703 total speeches
- Can we identify her speeches?

Generate an "Ogata Dictionary"

```
> OgataDict <- generateDictionary(UNC,UN$Ogata,
+                     modelType="lasso",
+                     control=list(family="binomial"))
>
> summary(OgataDict)
Dictionary type:  weighted (words with individual scores)
Total entries:    38
Positive entries: 24 (63%)
Negative entries: 14 (37%)
Neutral entries:  0 (0%)

Details
Average score:      0.1
Median:             0.017
Min:                -1.1
Max:                5
Standard deviation: 0.86
Skewness:           5.2
```

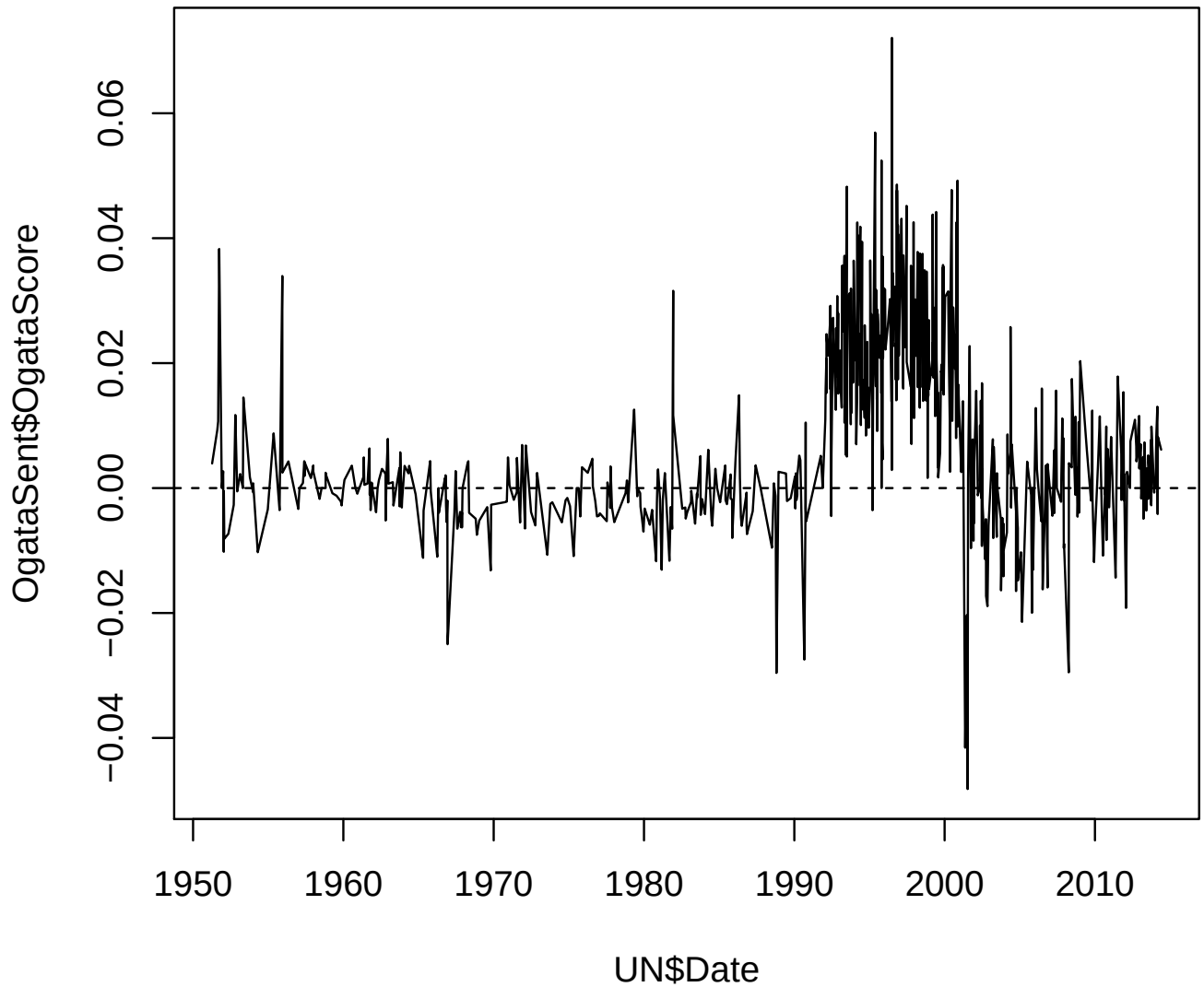
Ogata Dictionary (continued)

```
> OgataDict
Type: weighted (words with individual scores)
Intercept: -4.5
-1.06 ruud
-0.40 antnio
-0.22 check
-0.21 simpli
-0.20 develop
-0.20 forward
-0.18 peacebuild
-0.11 outcom
-0.07 qualiti
-0.04 latin
.
.
.
 0.02 rwanda
 0.03 strategi
 0.03 courag
 0.04 prevent
 0.05 war
 0.05 ethnic
 0.05 lake
 0.06 peacekeep
 0.07 danger
 0.08 flee
 0.14 crucial
 0.14 quick
 0.74 mrs
 5.05 sadako
```

Generate Ogata Sentiment Scores

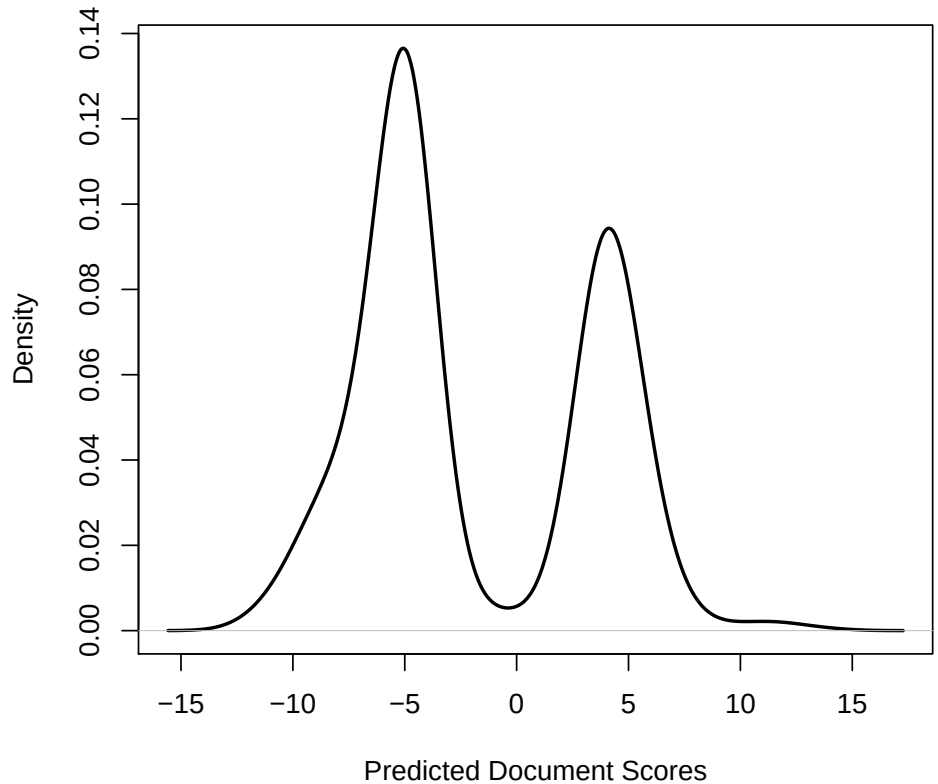
```
> OgataD <- SentimentDictionaryBinary(OgataDict$words[OgataDict$scores>0],
+                                     OgataDict$words[OgataDict$scores<0])
>

> # now, "sentiment":
>
> OgataSent <- analyzeSentiment(UNC,
+                    rules=list("OgataScore"=list(
+                    ruleSentiment,OgataD)))

> summary(OgataSent)
   OgataScore
 Min.   :-0.048
 1st Qu.:-0.002
 Median : 0.004
 Mean   : 0.009
 3rd Qu.: 0.019
 Max.   : 0.072
```

"Ogata Scores" Over Time



Generate In-Sample Predictions...

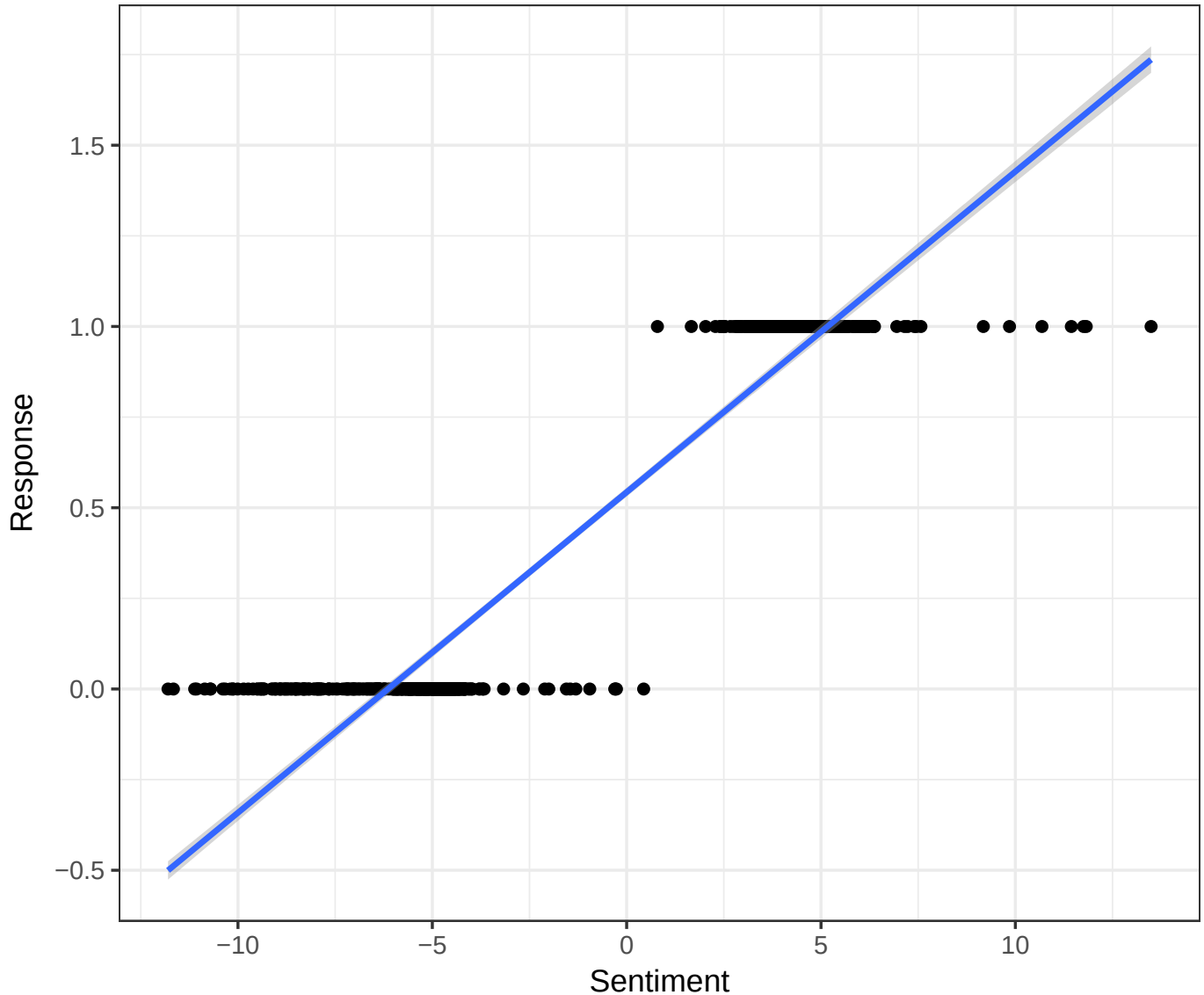
```
> OgataHat <- predict(OgataDict,UNC)
> summary(OgataHat$Dictionary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -11.8    -5.4    -4.5    -1.8     3.8    13.5
```



In-Sample Predictive Power

```
> compareToResponse(OgataHat,UN$Ogata)
                            Dictionary
cor                               0.95
cor.t.statistic                  77.12
cor.p.value                      77.12
lm.t.value                       77.12
r.squared                         0.89
RMSE                              5.23
MAE                               4.82
Accuracy                          0.38
Precision                         0.00
Sensitivity                        NaN
Specificity                       0.38
F1                                0.00
BalancedAccuracy                   NaN
avg.sentiment.pos.response       -1.81
avg.sentiment.neg.response         NaN
```

Predicted vs. Actual Plot



What Should We Actually Do?

Best practice:

- Create dictionary...
- Score a training set of text...
- **Validate!**
  - Assess predictive validity on a test set of text
  - OR: Cross-validate...
  - Compare to human coding / classification!
- Especially important when context matters...

Example: Loughran & McDonald (2011)

- The Harvard IV dictionary assigns negative valence to words that are not negative in accounting/finance (`tax`, `cost`, etc.)
- Also does the reverse (e.g., `litigation`, `misstate`, etc.)

Wrap-Up: Extensions / Challenges / etc.

- **Linguistic complexity**
  - Irony, sarcasm, tone, etc.
  - Complex / subtle negation ("I don't have one guitar; I have many.")
- **Dictionaries**...
  - Specialized vocabularies $\rightarrow$ standard sentiment dictionaries break down (e.g., "love" in tennis)
  - *Minimally-supervised* dictionary creation (Rice & Zorn)
  - Bleeding edge: *Unsupervised* dictionary creation (via negations...)
- **Change over time**
  - Word meanings...
  - Word usage...