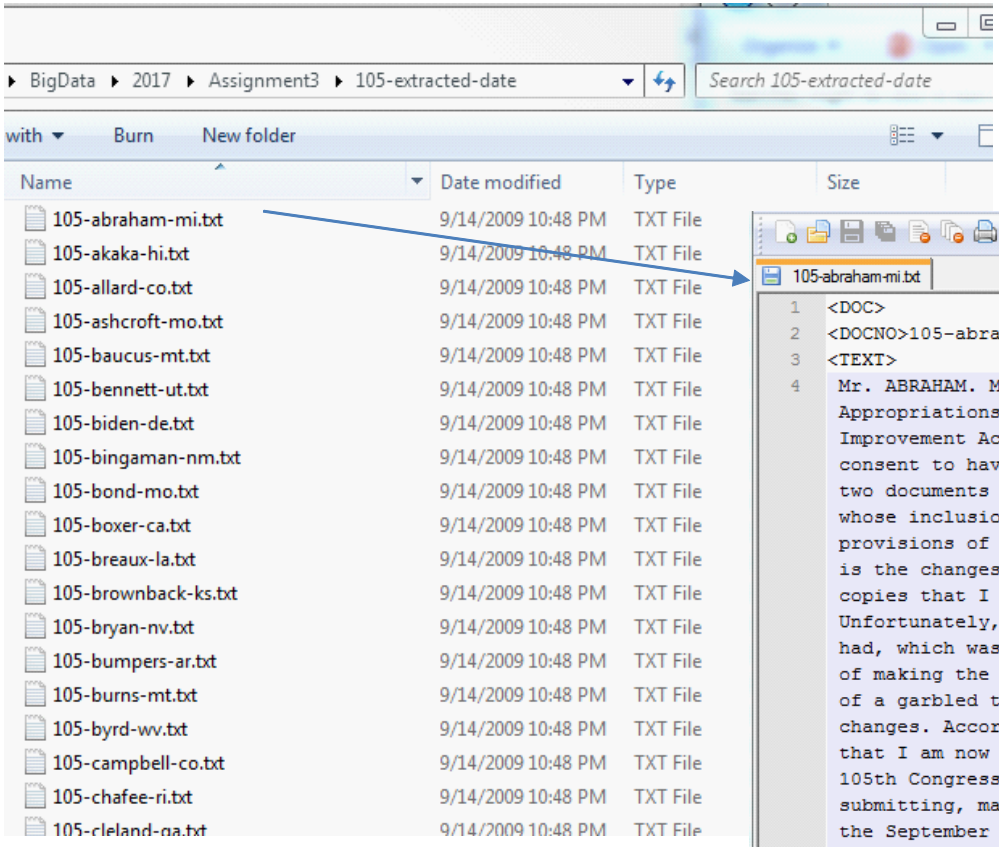
# Text as data

# Text as data

- Text in digital form increases exponentially.
- Commercial applications
  - Advertising (Google, Facebook, Twitter,...)
  - Dominant approach: computational linguistics + machine learning.
- Research applications
  - Stock market sentiment, consumer sentiment unemployment claims, retail sales, ideology of newspapers, deliberations of FOMC,...
- We will discuss three different problems:
  - document representation,
  - classifier construction, and
  - classifier evaluation.

# Text categorization using Machine Learning

- Initial corpus of documents: $\Omega = \left\{ d_1, d_2, ..., d_{|\Omega|} \right\}$.



# Document representation

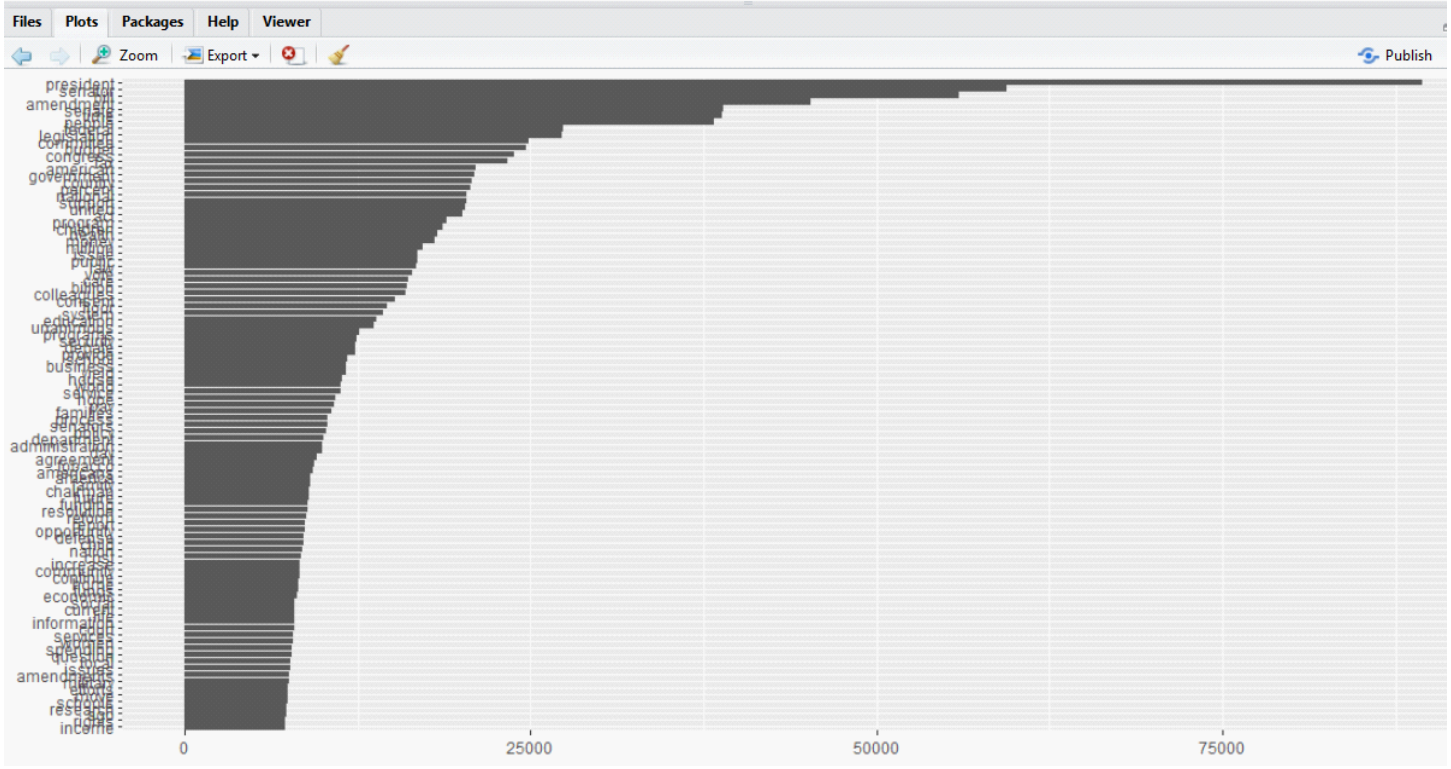
- Indexing maps a text $d_j$ to a compact representation.
  - meaningful units of text (lexical semantics)
  - meaningful natural language rules for the combination of these units (compositional semantics) typically ignored.
- $d_j$ represented by vector of term weights

$$\bar{d}_j = \left( w_{1j}, ..., w_{|T|_j} \right),$$

  where $T$ is the set of terms (features) that occur at least once in the training set.
- Terms are typically words (bag of words approach)
  - more sophisticated representations do not yield significantly better effectiveness
  - phrases (sequences of words) may still be preferred since they carry meaning

# Term weights

- Zipf's law: word frequency approx 1/n.
- Words like president is not very informative since every document contains it.



# Term weights

- Weights are typically based on term-frequency-inverse document frequency (*tfidf*)

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

  - $\#(t_k, d_j)$ = number of times $t_k$ occurs in $d_j$
  - $\#_{Tr}(t_k)$ = number of documents in $Tr$ in which $t_k$ occurs.
  - (i) the more often a term occurs in a document, the more it is representative of its content, and (ii) the more documents a term occurs in, the less discriminating it is.

- Weights are often normalized so that the vectors $\overline{d}_j$ are of equal length:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_s, d_j))}}.$$

# Reducing term space dimensionality

- Pre-processing of words
  - Stopwords (i.e. topic-neutral words such as articles, prepositions, conjunctions, etc.) are typically removed.
  - Stemming (i.e. grouping words that share the same morphological root) are typically revmoved (although suitability to TC is controversial).
- Term selection
  - Keep terms that occur in at least $x$ documents (reduce dimensionality by a factor of 10 with no loss in effectiveness, Yang and Pedersen, 1997)
  - R1 regularization
  - Score functions (what terms are have strongest simple correlation with outcome).
- Term extraction: e.g. latent semantic indexing (pca).

# Score functions

| Function | Denoted by | Mathematical form |
|---|---|---|
| *Document frequency* | $\#(t_k, c_i)$ | $P(t_k \mid c_i)$ |
| *DIA association factor* | $z(t_k, c_i)$ | $P(c_i \mid t_k)$ |
| *Information gain* | $IG(t_k, c_i)$ | $\sum_{c \in \{c_i, \overline{c}_i\}} \sum_{t \in \{t_k, \overline{t}_k\}} P(t, c) \cdot \log \frac{P(t, c)}{P(t) \cdot P(c)}$ |
| *Mutual information* | $MI(t_k, c_i)$ | $\log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$ |
| *Chi-square* | $\chi^2(t_k, c_i)$ | $\frac{\lvert Tr \rvert \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]^2}{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}$ |
| *NGL coefficient* | $NGL(t_k, c_i)$ | $\frac{\sqrt{\lvert Tr \rvert} \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}}$ |
| *Relevancy score* | $RS(t_k, c_i)$ | $\log \frac{P(t_k \mid c_i) + d}{P(\overline{t}_k \mid \overline{c}_i) + d}$ |
| *Odds Ratio* | $OR(t_k, c_i)$ | $\frac{P(t_k \mid c_i) \cdot (1 - P(t_k \mid \overline{c}_i))}{(1 - P(t_k \mid c_i)) \cdot P(t_k \mid \overline{c}_i)}$ |
| *GSS coefficient* | $GSS(t_k, c_i)$ | $P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)$ |

# Chi-square feature selection

- Suppose $U$ is a random variable that takes values
  - $e_t = 1$ (the document contains term $t$) and
  - $e_t = 0$ (the document does not contain $t$).
- Suppose $C$ is a random variable that takes values
  - $e_c = 1$ (the document is in class $c$) and
  - $e_c = 0$ (the document is not in class $c$).
- $\chi^2$ is applied to the independence of $U$ and $C$.

$$\chi^2(t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} =,$$

  - $N_{e_t e_c}$ = observed # occurences of term $t$ in document of class $c$
  - $E_{e_t e_c}$ = expected # such occurences, assuming that term and class are independent.

# Chi-square feature selection

For example,

$$E_{11} = N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N}.$$

An arithmetically simpler way to compute the chi-squared is

$$\chi^2(t,c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})},$$

which is used in Gentzkow and Shapiro (2010)

# Language and Ideology in Congress, Diermeier et al. (2008)

Predict party label based on words in speeches in US House and the Senate 2005.

- Classifier SVM
- Term weights
  1. bool: dummy for word in document
  2. ntf: normalized word frequency.
  3. tfidf.
- Terms
  1. word
  2. stemmed words
  3. nouns
  4. verbs
  5. adjectives
  6. adverbs.

# Data

Download speech data from thomas.gov

> Mrs. MURRAY. Madam President, here we are, once again debating this issue. Since we began debating how to criminalize women's health choices yesterday, the Dow Jones has dropped 170 points; we are 1 day closer to a war in Iraq; we have done nothing to stimulate the economy or create any new jobs or provide any more health coverage. But here we are, debating abortion in a time of national crisis. (Senator Murray, 2003, p. S3422)

# Data

Pre-processing: tokenization, stemming and part-of-speech tagging.
Term selection: minimum term frequency of 50 and document frequency of 10 for a word to be selected as a feature.
removed the top fifty most frequent words as 'stopwords'

TABLE 1 *Feature Set Sizes*

| | Feature set | | | | | |
|---|---|---|---|---|---|---|
| | word | stem | noun | verb | adj. | adv. |
| Size | 19,459 | 11,395 | 8,831 | 6918 | 3665 | 890 |

# Model selection:

TABLE 2 *'Leave-one-out' Cross Validation on the Training Set Representations*

| | Feature sets | | | | | | |
|---|---|---|---|---|---|---|---|
| | word | stem | noun | verb | adj. | adv. | Average acc. |
| Boolean | 88.9 | 88.0 | 89.5 | 82.9 | 89.2 | 77.2 | 86.1 |
| Normalized frequency | 88.6 | 89.7 | 84.6 | 66.7 | 84.6 | 61.8 | 90.7 |
| *tf*idf* | 95.4 | 94.0 | 95.4 | 93.4 | 93.4 | 88.9 | 93.3 |

TABLE 3 *tf*idf*–SVM *Extreme Prediction Results (with Extreme Training Set)*

| | Feature sets | | | | | |
|---|---|---|---|---|---|---|
| *tf*idf*–SVM | word | stem | noun | verb | adj. | adv. |
| Extreme | 92.0 | 86.0 | 84.0 | 88.0 | 94.0 | 52.0 |

# Word weights

TABLE 4 *tf*idf*–SVM *Feature Set Analysis for All Vocabulary*

| Words | | | |
|---|---|---|---|
| Liberal | | Conservative | |
| FAS: −199.49 | SBA: −113.10 | habeas: 193.55 | homosexual: 103.07 |
| Ethanol: −198.92 | Nursing: −109.38 | CFTC: 187.16 | everglades: 102.87 |
| Wealthiest: −159.74 | Providence: −108.73 | surtax: 151.81 | tower: 101.67 |
| Collider: −142.28 | Arctic: −108.30 | marriage: 145.79 | tripartisan: 101.23 |
| WIC: −140.14 | Orange: −107.98 | cloning: 141.71 | PRC: 102.90 |
| ILO: −139.89 | Glaxo: −107.81 | tritium: 133.49 | scouts: 97.55 |
| Handgun: −129.01 | Libraries: −107.70 | ranchers: 132.95 | nashua: 99.32 |
| Lobbyists: −128.95 | Disabilities:−106.44 | BTU: 121.92 | ballistic: 97.22 |
| Enron: −127.71 | Prescription: −106.31 | grazing: 121.59 | salting: 94.28 |
| Fishery: −127.30 | NIH: −105.52 | unfunded: 120.82 | abortion: 91.94 |
| Hydrogen: −122.59 | Lobbying: −105.35 | catfish: 120.82 | NTSB: 93.81 |
| Souter: −121.40 | NRA: −105.20 | IRS: 114.91 | Haiti: 97.28 |
| PTSD: −119.87 | Trident: −104.15 | unborn: 111.88 | PAC: 92.85 |
| Gun: −119.52 | RNC: −103.46 | Taiwan: 111.13 | taxing: 90.39 |
| Firestone: −117.90 | Lobbyist: −99.38 | PLO: 106.56 | nonseverability: 89.26 |
| Lakes: −114.84 | Homelessness: −95.68 | EMS: 103.99 | embryonic: 88.83 |

# Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards

- 1.5 million messages posted on Yahoo! Finance and Raging Bull about the 45 companies in the Dow Jones Industrial Average and the Dow Jones Internet Index
  - time and title
  - text often contains a predicted price change and some explanation
  - number of words typically 20 to 50.

```
---------------------
FROM YF
COMP ETYS
MGID 13639
NAME CaptainLihai
LINK 1
DATE 2000/01/25 04:11
SKIP
TITL ETYS will surprise all pt II
SKIP
TEXT ETYS will surprise all when it drops to below 15$ a pop, and even then
TEXT it will be too expensive.
TEXT
TEXT If the DOJ report is real, there will definately be a backlash against
TEXT the stock. Watch your asses. Get out while you can.
---------------------
FROM YF
COMP IBM
MGID 43653
NAME plainfielder
LINK 1
DATE 2000/03/29 11:39
SKIP
TITL BUY ON DIPS - This is the opportunity
SKIP
```

# Naive Bayes Classifier

- Term selection: top 1,000 by information gain.
- Labels: buy, hold, sell. Manually classify training data set of 1,000 messages

**Table I**

**Naive Bayes Classification Accuracy within Sample and Overall Classification Distribution**

The first percentage column shows the actual shares of 1,000 hand-coded messages that were classified as buy (B), hold (H), or sell (S). The buy-hold-sell matrix entries show the in-sample prediction accuracy of the classification algorithm with respect to the learned samples, which were classified by the authors (Us).

| Classified: by Us | % | By Algorithm | | |
|---|---|---|---|---|
| | | Buy | Hold | Sell |
| Buy | 25.2 | 18.1 | 7.1 | 0.0 |
| Hold | 69.3 | 3.4 | 65.9 | 0.0 |
| Sell | 5.5 | 0.2 | 1.2 | 4.1 |
| 1,000 messages[a] | | 21.7 | 74.2 | 4.1 |
| All messages[b] | | 20.0 | 78.8 | 1.3 |

[a]These are the 1,000 messages contained in the training data set.
[b]This line provides summary statistics for the out-of-sample classification of all 1,559,621 messages.

# Aggregate measures

Let

$$M_t^c = \sum_i x_i^c,\ c \in \{BUY, HOLD, SELL\},$$

Bullishness,

$$B_t = \frac{M_t^{BUY} - M_t^{SELL}}{M_t^{BUY} + M_t^{SELL}},$$

and log transformations.

$$\sigma_t^2 = \frac{1}{N_t} \sum_i (x_i - B_t)^2, \qquad x_i = x_i^{BUY} - x_i^{SELL} \in \{0, 1\}.$$

Agreement index: $A_t = 1 - \sigma_t$

# Results

**Table IV**

**Contemporaneous Regressions**

The units of observation are the 15-minute intervals between 09:30 and 16:00 on trading days. All regressions use company fixed effects. A coefficient that is significant at the 95% level is indicated with superscript a, while superscript b and superscript c denote significance at the 99% level and 99.9% level, respectively. Absolute $t$-statistics are shown in parentheses. The regressors were obtained from the message boards: the log transformation $\ln(1 + M_t)$ of the number of messages; the bullishness measure $B_t^*$ and the agreement index $A_t \in [0, 1]$. The seven financial dependent variables are the log difference in the bid–ask midpoint from the end of the previous 15-minute interval to the current 15-minute interval (return); the percentage ratio of 15-minute price volatility relative to the interval's average price; the log number of small (<\$100k), medium (\$100k–\$1m) and large (> \$1m) trades; the log number of traded shares (volume); and the daily average of the bid–ask spread. The log number of trades and volume are calculated as $\ln(1 + x)$. Market denotes the log price of the S & P 500 tracking fund (SPY), except in the case of the return regressions, where it denotes the return (difference of the log price) of the SPY.

| | Log of Messages | | Bullishness Index | | Agreement Index | | Market | | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Return | −0.331 | (1.382) | 1.747 | (3.208) | −0.240 | (0.455) | 0.716$^c$ | (120.7) | 0.049 |
| Volatililty | 0.041$^c$ | (35.7) | 0.033$^c$ | (12.74) | −0.029$^c$ | (11.41) | −1.178$^c$ | (81.85) | 0.538 |
| Log small trades | 0.225$^c$ | (102.1) | 0.181$^c$ | (36) | −0.123$^c$ | (25.3) | −1.541$^c$ | (55.88) | (0.984) |
| Log medium trades | 0.119$^c$ | (43.53) | 0.161$^c$ | (25.82) | −0.096$^c$ | (15.84) | −0.464$^c$ | (13.55) | (0.931) |
| Log large trades | 0.082$^c$ | (37.29) | 0.052$^c$ | (10.39) | −0.021$^c$ | (4.382) | −0.222$^c$ | (8.073) | (0.642) |
| Log trading volume | 0.259$^c$ | (82.37) | 0.170$^c$ | (23.81) | −0.109$^c$ | (15.72) | −2.417$^c$ | (61.55) | (0.995) |
| Spread | 0.001 | (0.766) | 0.009$^b$ | (2.861) | −0.004 | (1.369) | −0.047$^b$ | (2.763) | 0.245 |

# Results

- Level of posting predicts .
  - Negative returns on the next day.
  - Subsequent trading volume
  - Volatility both at daily frequencies and also within the trading day
- Disagreement among messages predict volatility.

# Media sentiment and the stock market

- Finance papers use text analysis to examine sentiment of corporate 10-K reports, newspaper articles, press releases, and investor message boards.
- Giving Content to Investor Sentiment: The Role of Media in the Stock Market (Tetlock, JoF, 2007)
  - Proposed measure is based on the linguistic tone of a popular daily WSJ column called "Abreast of the Market" (AM) from 1984 through 1999
  - Computes the relative frequencies of AM words in 77 predetermined categories from the Harvard psychosocial dictionary, such as Strong, Weak, Active, and Passive words.
  - Negative words in the AM column are associated with lower same-day stock returns and predict lower returns the following day
- Garcia (2013) builds on these results in a study of positive and negative words from two NYT columns spanning 1905 to 2005

# Transparency and Deliberation within the FOMC (Hansen et al., QJE 2017)

- What happens to policy deliberations informativeness when these become public?
- Studies minutes of the Federal Open Market Committee (FOMC) deliberations.
- Nov 1993, the Fed agreed to publish the past transcripts and all future transcripts with a five-year lag.
    - Since the 1970s, FOMC meetings were tape recorded to help prepare minutes. Unknown to committee members, these tapes were transcribed and stored in archives.
- Communication measures based on basic text counts and on topic models
    - Estimates the fraction of time each speaker in each meeting spends on each topics.
- Concludes that transparency increases informativeness.