E-petition popularity: Do linguistic and semantic factors matter?

Loni Hagen
Assistant Professor
School of Information
University of South Florida
Linguistic and Semantic Factors in Written Texts

- Extremity (much more, extremely, very, wonderful)
- Informativeness and novelty of text (number of unique words)
- Repetition
- Request (please, rt, retweet, spread, pls, plz)
- Sentiment
- Internet activity
- Named entities
- Topics
Research Questions

• To understand whether, and to what extent, a number of linguistic and semantic factors are related to the popularity of e-petitions.

80 to 85% of the data in the world is unstructured form, primarily text
Data
E-petition platform “We the People”

E-petition system launched by US federal government in 2011.

- Over 15 million users
- Over 22 million signatures
- Over 400,000 petitions created
- 150 signatures => web presence
- 100,000 signatures => official response
## Title

**Pardon Edward Snowden**

Edward Snowden is a national hero and should be immediately issued a a full, free, and absolute pardon for any crimes he has committed or may have committed related to blowing the whistle on secret NSA surveillance programs.

Created: Jun 09, 2013  
Issues: Civil Rights and Liberties, Government Reform, Human Rights

---

## Total Signatures

**163,318**

---

## Signature Information

<table>
<thead>
<tr>
<th>Creator</th>
<th>Signature Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. D.</td>
<td>Marshall, MO</td>
</tr>
<tr>
<td></td>
<td>November 10, 2014</td>
</tr>
<tr>
<td>K. L.</td>
<td>Chelsea, MA</td>
</tr>
<tr>
<td></td>
<td>November 10, 2014</td>
</tr>
<tr>
<td>K. S.</td>
<td>Daviston, MI</td>
</tr>
<tr>
<td></td>
<td>November 10, 2014</td>
</tr>
</tbody>
</table>
Data Collected (Sept. 2011 - Jan. 2015)

• WtP API
• 3,344 petition texts (title + rationale)
• Signature information
• Petition creation time
• Preprocessing:
  • Converted all words to lower case
  • Removed white space
  • Eliminated punctuation
  • Removed short words of only one or two characters using the Natural Language Toolkit (Bird, Klein, & Loper, 2009)
Before and After Stemming

Preprocessed petition:
stop animal homelessness its roots

Stemmed:
stop anim homeless it root
Title: Deport Justin Bieber and revoke his green card.

Rationale: “We the people of the United States feel that we are being wrongly represented in the world of pop culture. We would like to see the dangerous, reckless, destructive, and drug abusing, Justin Bieber deported and his green card revoked. He is not only threatening the safety of our people but he is also a terrible influence on our nations youth. We the people would like to remove Justin Bieber from our society.”

Signature counts: 273,968
Linguistic and semantic characteristics of persuasive texts

**Linguistic Characteristics of Persuasive Texts**
- Extremity, Urgency, Request, Internet Activity, Novelty and Informativeness, Repetition, and Sentiment

**Semantic Information in E-petition**
- Named entities (person, location, organization), and Topics

**Theory driven**

**Data driven**
Methodology

• Information Extraction
• Variables
• Regression Analysis
Information Extraction
Input: 1,671 petition texts

Output: Table (1,671 rows * 28 columns)

Validation
Validation
Validation

Unstructured texts

Structured

R
Python
Java
Approaches

1. Lexicon generation
   • Extremity: much more, extremely, very, and wonderful (Craig & Blankenship, 2011, p. 295)

2. Dictionary-based approach
   • Urgency: Find synonyms of seed words (“immediately” “immediate” and “urgent”) from WordNet => 53 words in the “Urgency” list

3. Tagging approaches
   • Stanford CoreNLP: Sentiment and named entity recognition tagger

4. Machine learning (MALLET)
   • Topic Modeling
Topic Modeling: unsupervised machine learning to find clusters of words

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analysis to compute known genomes, concluded that today’s organisms can be sustained with just 250 genes; and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions...


Stripping down, computer analysis yields an estimate of the minimum modern and ancient genomes.

<blei (2012), p.78>
Topic Modeling: unsupervised machine learning to find clusters of words

<Topic proportions of a petition>

Stop Animal Homelessness at Its Roots

Every year in the United States, an estimated 6 to 8 million lost, abandoned, or unwanted dogs and cats enter animal shelters and nearly half of these animals many of them healthy, young, and adoptable must be euthanized because there are too many animals and not enough good homes. This tragedy occurs because people don't spay and neuter their animals and because greedy breeders continue to churn out more puppies. Because all dogs and cats are precious and because no more animals need to be bred when so many others go without hope of being adopted. PETA is calling for a mandatory spay-and-neuter law until all dogs and cats in the United States have a home to call their own. Sign the petition calling for a mandatory spay-and-neuter law to help end the animal overpopulation crisis.
15 topics extracted

• Example topics (top 8 most frequent words in topics)
  • marijuana, legal, drug, cannabis, medical, substances, schedule, alcohol
  • sex, legal, marriage, families, couples, file, court, provide
  • police, office, killed, law, murder, men, shot, death
  • ...
Variables
<table>
<thead>
<tr>
<th>Variables</th>
<th>Development Strategy</th>
<th>Evaluation Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremity</td>
<td>Manual lexicon generation</td>
<td>.</td>
</tr>
<tr>
<td>Urgency</td>
<td>Dictionary-based lexicon generation</td>
<td>.</td>
</tr>
<tr>
<td>Informativeness</td>
<td>Frequency counting</td>
<td>.</td>
</tr>
<tr>
<td>Repetition</td>
<td>Frequency counting</td>
<td>.</td>
</tr>
<tr>
<td>Request</td>
<td>Manual lexicon generation</td>
<td>.</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Tagging</td>
<td>.</td>
</tr>
<tr>
<td>Internet Activity</td>
<td>Manual lexicon generation</td>
<td>.</td>
</tr>
<tr>
<td>Named Entity</td>
<td>Tagging</td>
<td>F-measure</td>
</tr>
<tr>
<td>Topic</td>
<td>Unsupervised learning</td>
<td>10-fold cross validation</td>
</tr>
</tbody>
</table>
Variables

• Dependent variable: Logarithm of signature counts

• Control variables
  • Logarithm of numbers of signatures gathered on first 24 hours
  • Logarithm of the number of petitions started on same day

• Independent variables
  • Linguistic style variables
  • Semantic variables
### Linguistic Style Variables

<table>
<thead>
<tr>
<th>Lexicon generation</th>
<th>1. Extremity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary-based Tagging</td>
<td>2. Request</td>
</tr>
<tr>
<td>Frequency counting</td>
<td>3. Internet Activity</td>
</tr>
<tr>
<td></td>
<td>4. Urgency</td>
</tr>
<tr>
<td></td>
<td>5. Sentiment</td>
</tr>
<tr>
<td></td>
<td>6. Informativeness</td>
</tr>
<tr>
<td></td>
<td>7. Repetition</td>
</tr>
</tbody>
</table>
1. Named entities (Stanford CoreNLP)
   - Person
   - Location
   - Organization
2. 15 Topics (MALLET)
Regression Analysis
## Hierarchical Ordinary Least Squares Regression

<table>
<thead>
<tr>
<th>Model</th>
<th>Dependent variable</th>
<th>Control variable block</th>
<th>Independent variable blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LogSigCount</td>
<td>Control Variable Block</td>
<td>Linguistic Style Variable Block</td>
</tr>
<tr>
<td>2</td>
<td>LogSigCount</td>
<td>Control Variable Block</td>
<td>Linguistic Style Variable Block</td>
</tr>
<tr>
<td>3</td>
<td>LogSigCount</td>
<td>Control Variable Block</td>
<td>Linguistic Style Variable Block</td>
</tr>
<tr>
<td>4</td>
<td>LogSigCount</td>
<td>Control Variable Block</td>
<td>Topic Variable Block</td>
</tr>
</tbody>
</table>

Model 1, Model 2, Model 3, and Model 4 include the same dependent variable (LogSigCount) and control variable block (Control Variable Block). Model 1 also includes the independent variable block Linguistic Style Variable Block, Model 2 includes Linguistic Style Variable Block, Model 3 includes Linguistic Style Variable Block, and Model 4 includes Topic Variable Block.
Results
## Hierarchical Ordinary Least Squares Regression

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td>LogSigCount</td>
<td>LogSigCount</td>
<td>LogSigCount</td>
<td>LogSigCount</td>
</tr>
<tr>
<td><strong>Control variable block</strong></td>
<td>Control Variable Block</td>
<td>Control Variable Block</td>
<td>Control Variable Block</td>
<td>Control Variable Block</td>
</tr>
<tr>
<td><strong>Independent variable blocks</strong></td>
<td></td>
<td>Linguistic Style Variable Block</td>
<td>Linguistic Style Variable Block</td>
<td>Linguistic Style Variable Block</td>
</tr>
<tr>
<td><strong>NER Variable Block</strong></td>
<td>NER Variable Block</td>
<td>NER Variable Block</td>
<td>NER Variable Block</td>
<td>Topic Variable Block</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>1,671</td>
<td>1,671</td>
<td>1,671</td>
<td>1,671</td>
</tr>
<tr>
<td><strong>R² Change</strong></td>
<td></td>
<td>0.01***</td>
<td>0.004**</td>
<td>0.08***</td>
</tr>
<tr>
<td><strong>Adjusted R²</strong></td>
<td>0.24</td>
<td>0.25</td>
<td>0.26</td>
<td>0.32</td>
</tr>
</tbody>
</table>

**Note:** * p ≤ .1, ** p ≤ .05, *** p ≤ .001
Findings

Statistically Significant Variables

- **Extremity** (-0.17, p-value: 0.027)
- **Person** (-0.11, p-value: 0.032)
- **Topics**
  - **Positive:**
    - religion_gay (1.17, p-value: 0.008)
    - secession (1.18, p-value: 0.000)
    - gun (1.09, p-value: 0.024)
  - **Negative:**
    - children (-1.67, p-value: 0.001)
    - china (-2.12, p-value: 0.000)
    - awareness (-1.30, p-value: 0.062)
    - student_visa (-1.27, p-value: 0.001)
    - white_genocide (-2.13, p-value: 0.000)
Linguistic and Semantic Variables are Significant Predictors of Petition Popularity

• Extremity:
  • Negatively correlated with petition popularity in the Petition setting, in contrast to previous studies.
  • Attributable to the discrepancies in the study settings (Lab; well-written).
  • The analysis of large volumes of texts led us to findings seemingly contradictory to previous findings that used small sets of texts in laboratory settings for human experimentation.

• Person names:
  • Specific and particular presentation of policy
  • Problem of being too specific
  • But specificity helps acquiring more support when an issue is familiar such as “gun” or “religion and gay”
Familiar topics are positively correlated and unfamiliar topics are negatively correlated with petition popularity

- Popular topics
  - Religion_gay:
    - Remove “In God we trust” from money, or to remove the words “One nation under God” from the pledge of allegiance
  - Gun: oppose or support gun control policy
  - Secession:
    - November 6, 2012 right after the re-election of President Obama

- Unpopular topics: children, china, student Visa, white_genocide
Post hoc analyses

1,671 petitions

Model A: 25,000 signatures required for White House response

Model B: 100,000 signatures required for White House response
**Result: Post Hoc Analysis**

<table>
<thead>
<tr>
<th></th>
<th>Linguistic style</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremity</td>
<td>Negative</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>Repetition</strong></td>
<td>-</td>
<td><strong>Negative</strong></td>
<td>-</td>
</tr>
<tr>
<td>Person</td>
<td>-</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>-</td>
<td>Positive</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Named entities</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>religion &amp; gay</td>
<td>Positive</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>secession</td>
<td>Positive</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>gun</td>
<td>Positive</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>china</td>
<td>-</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>children</td>
<td>-</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>student &amp; visa</td>
<td>-</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>white genocide</td>
<td>-</td>
<td>Negative</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Topics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Social events and petition activities?</td>
<td>&lt;Example: A Petition Title&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Democracy crisis in Malaysia: foreign workers were employed for fraud voting in Malaysian General Election”

223,913 signatures
Discussion

• Petition popularity and social events
• Many person names and criminal investigation topics: successful
• Global participation (Repetition, Location)
  • 44% of successful petitions from foreign countries
• Extremity
  • Small set of data (laboratory experiments) vs. big data
Conclusion

• Investigate feasibility of using text as data for model building
  • Linguistic and semantic feature selection for model building
• Uncover latent patterns of e-petition texts associated with petition popularity
• Discoveries from big data analyses are sometimes contradictory to findings based on small set of sample experiments
• It is important to establish valid processes for understanding online political participation when using text as data and computational tools for analysis
Main references


Thank you!!

Comments/Questions?
E-mail: lonihagen@usf.edu
Validation: Data-driven variables

• NER: F-measure
  • Person (0.926), organization (0.718), location (0.870)

• Topics: 10-fold cross validation
  • Selecting topic variables: Human coding and cross validation (average mean squared error)
  • Validation of topic variables: use of test dataset
Variable selection: 18 topics => 15 topics

• Regression:
  • Dependent variable: logarithm of signature counts
  • Independent variable: 18 topic variables

• 10 fold cross-validation:
  • average mean squared error
  • Removing one of the 18 topics from the variables

• Selected 15 topics
<table>
<thead>
<tr>
<th>Topic Variables</th>
<th>Regression on Training Set Coefficient (SE)</th>
<th>Regression on Test Set Coefficient (SE)</th>
<th>Topic Variables</th>
<th>Regression on Training Set Coefficient (SE)</th>
<th>Regression on Test Set Coefficient (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>veteran</td>
<td>-0.14 (0.33)</td>
<td>-1.08** (0.39)</td>
<td>student_visa</td>
<td>-1.01** (0.32)</td>
<td>-1.65*** (0.43)</td>
</tr>
<tr>
<td>religion_gay</td>
<td>0.55 (0.38)</td>
<td>1.9*** (0.49)</td>
<td>military</td>
<td>-1.03** (0.35)</td>
<td>-0.3542</td>
</tr>
<tr>
<td>children</td>
<td>-1.34*** (0.39)</td>
<td>-2.20*** (0.55)</td>
<td>national park</td>
<td>-1.42*** (0.41)</td>
<td>-1.82** (0.78)</td>
</tr>
<tr>
<td>investigation</td>
<td>-0.86** (0.37)</td>
<td>-0.50 (0.50)</td>
<td>white_genocide</td>
<td>-2.33*** (0.24)</td>
<td>-2.39*** (0.26)</td>
</tr>
<tr>
<td>marijuana</td>
<td>0.91** (0.38)</td>
<td>0.61 (0.60)</td>
<td>gun</td>
<td>0.61 (0.40)</td>
<td>1.38** (0.54)</td>
</tr>
<tr>
<td>sentence</td>
<td>-0.60 (0.36)</td>
<td>-1.07** (0.55)</td>
<td>Intercept</td>
<td>7.83*** (0.06)</td>
<td>7.87*** (0.07)</td>
</tr>
<tr>
<td>cancer_research</td>
<td>-1.06*** (0.33)</td>
<td>-0.90 (0.49)</td>
<td>N</td>
<td>1,671</td>
<td>1,671</td>
</tr>
<tr>
<td>secession</td>
<td>1.22*** (0.27)</td>
<td>1.78*** (0.29)</td>
<td>F</td>
<td>14.73***</td>
<td>16.77***</td>
</tr>
<tr>
<td>china</td>
<td>-2.02*** (0.33)</td>
<td>-2.23*** (0.39)</td>
<td>R²</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>awareness</td>
<td>-1.36*** (0.40)</td>
<td>-1.71** (0.78)</td>
<td>Adjusted R²</td>
<td>0.11</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Note: * p ≤ .1, ** p ≤ .05, *** p ≤ .001