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.

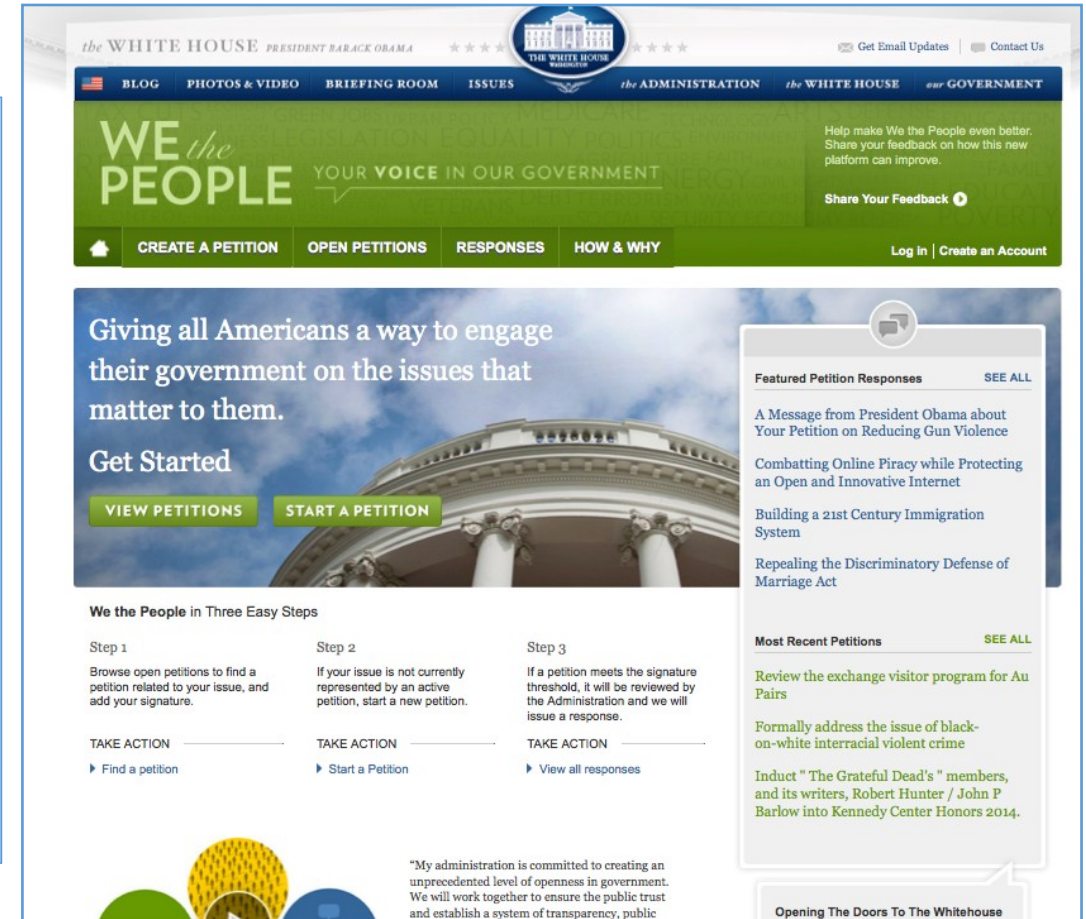80 to 85% of the data in the world is unstructured form, primarily text

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**

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

Extractable information from a text

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

Theory driven

Data driven

Linguistic Characteristics of Persuasive Texts

Semantic Information in E-petition

Extremity, Urgency, Request, Internet Activity, Novelty and Informativeness, Repetition, and Sentiment

Named entities (person, location, organization), and Topics

Methodology

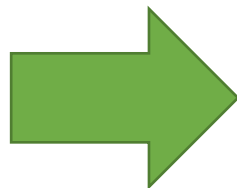- Information Extraction

- Variables

- Regression Analysis

Information Extraction

Unstructured texts

This is a clear violation of federal law. In attempting to undermine our own nation, these 47 senators have committed treason.

Input: 1,671 petition texts

Validation
Validation
Validation
….

R
Python
Java

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

Structured

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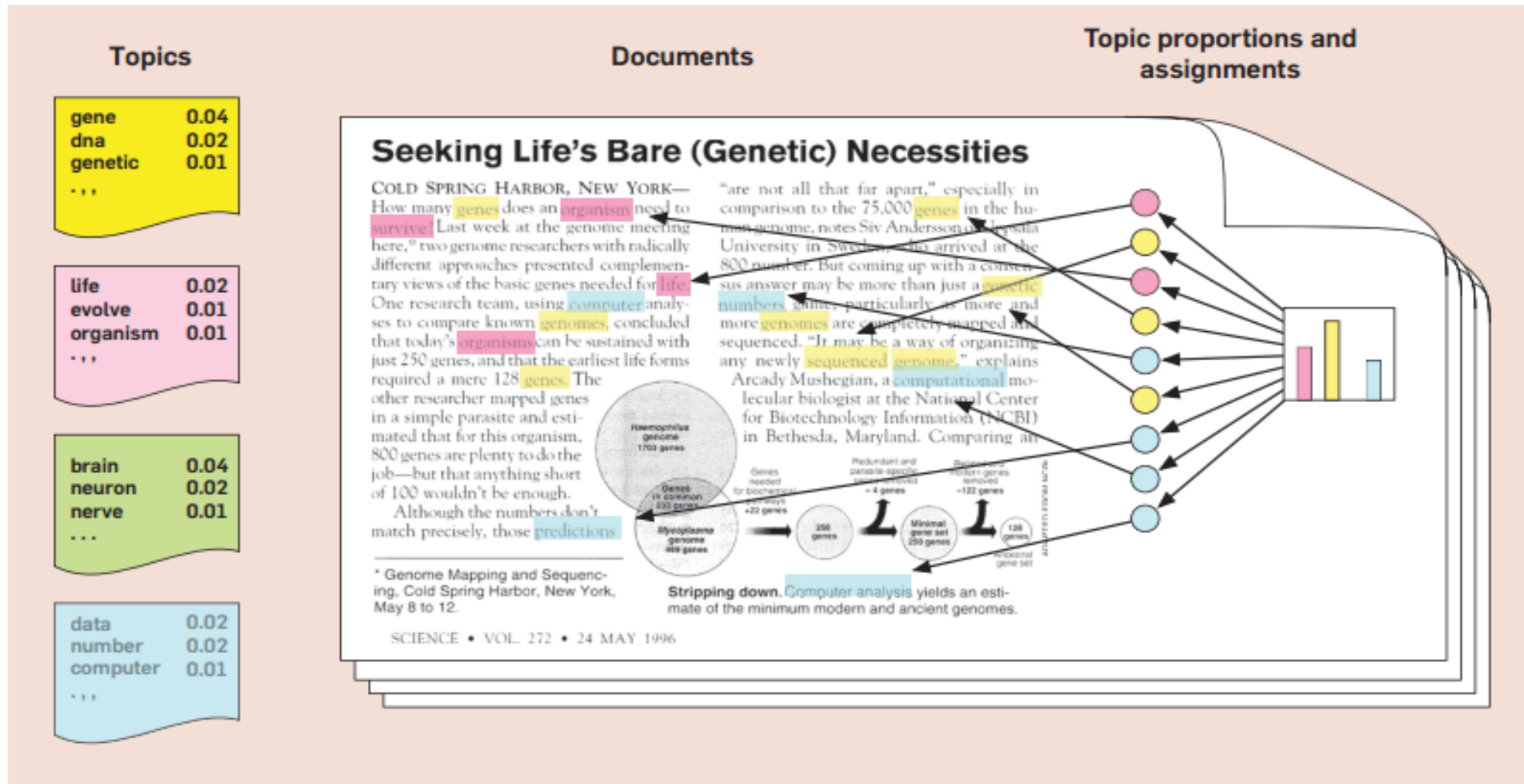
3. Tagging approaches
   - Stanford CoreNLP: Sentiment and named entity recognition tagger

4. Machine learning (MALLET)
   - Topic Modeling

Topic Modeling: unsupervised machine learning to find clusters of words

<Blei (2012), p.78>

Topic Modeling: unsupervised machine learning to find clusters of words

<Topic proportions of a petition>

Topic proportions:

animal 0.49,

people 0.42,

gun: 0.05,

visa: 0.01,

…

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 1: Summary of Variable Development and Evaluation

Variables	Development Strategy	Evaluation Strategy
Extremity	Manual lexicon generation	.
Urgency	Dictionary-based lexicon generation	.
Informativeness	Frequency counting	.
Repetition	Frequency counting	.
Request	Manual lexicon generation	.
Sentiment	Tagging	.
Internet Activity	Manual lexicon generation	.
Named Entity	Tagging	F-measure
Topic	Unsupervised learning	10-fold cross validation

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

Lexicon generation

Dictionary-based
Tagging
Frequency counting

1. **Extremity**
2. **Request**
3. **Internet Activity**
4. **Urgency**
5. **Sentiment**
6. **Informativeness**
7. **Repetition**

Semantic Variables

1. **Named entities (Stanford CoreNLP)**
   - Person
   - Location
   - Organization
2. **15 Topics (MALLET)**

Regression Analysis

Hierarchical Ordinary Least Squares Regression

	Model 1	Model 2	Model 3	Model 4
Dependent variable	LogSigCount	LogSigCount	LogSigCount	LogSigCount
Control variable block	Control Variable Block	Control Variable Block	Control Variable Block	Control Variable Block
Independent variable blocks		Linguistic Style Variable Block	Linguistic Style Variable Block	Linguistic Style Variable Block
			NER Variable Block	NER Variable Block
				Topic Variable Block

Results

Hierarchical Ordinary Least Squares Regression

	Model 1	Model 2	Model 3	Model 4
Dependent variable	LogSigCount	LogSigCount	LogSigCount	LogSigCount
Control variable block	Control Variable Block	Control Variable Block	Control Variable Block	Control Variable Block
Independent variable blocks		Linguistic Style Variable Block	Linguistic Style Variable Block	Linguistic Style Variable Block
			NER Variable Block	NER Variable Block
				Topic Variable Block
N	1,671	1,671	1,671	1,671
R^2 Change		0.01***	0.004**	0.08***
Adjusted R^2	0.24	0.25	0.26	0.32

Note: *$p \leq .1$, ** $p \leq .05$, *** $p \leq .001$

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 epetition setting, in contrast to previous studies.
  - Attributable to the discrepancies in the study settings (Lab; well-written).
  - **<u>The analysis of large volumes of texts led us to findings seemingly contradictory to previous findings</u>** 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

10/15/2011   1/13/2013   1/3/2015

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

Model B:
100,000 signatures required for White House response

Result: Post Hoc Analysis

		Model A	Model B
Linguistic style	Extremity	Negative	-
	Repetition	-	**Negative**
Named entities	Person	-	Negative
	Location	-	Positive
Topics	religion & gay	Positive	Positive
	secession	Positive	Negative
	gun	Positive	-
	china	-	Negative
	children	-	Negative
	student & visa	-	Negative
	white genocide	-	Negative

<Example: A Petition Title>

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

223,913 signatures

Social events and petition activities?

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
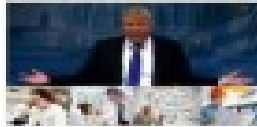 - Small set of data (laboratory experiments) vs. big data

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

- Craig, T. Y., & Blankenship, K. L. (2011). Language and persuasion: Linguistic extremity influences message processing and behavioral intentions. *Journal of Language and Social Psychology, 30*(3), 290–310. http://doi.org/10.1177/0261927X11407167

- Davenport, T. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press.

- Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363–370). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1219840.1219885

- McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. Retrieved from http://mallet.cs.umass.edu/

- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (Vol. 1631, p. 1642). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.383.1327&rep=rep1&type=pdf

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

Topic Variables	Regression on Training Set Coefficient (SE)	Regression on Test Set Coefficient (SE)	Topic Variables	Regression on Training Set Coefficient (SE)	Regression on Test Set Coefficient (SE)
veteran	-0.14 (0.33)	-1.08** (0.39)	student_visa	-1.01** (0.32)	-1.65*** (0.43)
religion_gay	0.55 (0.38)	1.9*** (0.49)	military	-1.03** (0.35)	-0.3542
children	-1.34*** (0.39)	-2.20*** (0.55)	national park	-1.42*** (0.41)	-1.82** (0.78)
investigation	-0.86** (0.37)	-0.50 (0.50)	white_genocide	-2.33*** (0.24)	-2.39*** (0.26)
marijuana	0.91** (0.38)	0.61 (0.60)	gun	0.61 (0.40)	1.38** (0.54)
sentence	-0.60 (0.36)	-1.07** (0.55)	Intercept	7.83*** (0.06)	7.87*** (0.07)
cancer_research	-1.06*** (0.33)	-0.90 (0.49)	N	1,671	1,671
secession	1.22*** (0.27)	1.78*** (0.29)	F	14.73***	16.77***
china	-2.02*** (0.33)	-2.23*** (0.39)	R^2	0.12	0.13
awareness	-1.36*** (0.40)	-1.71** (0.78)	Adjusted R^2	0.11	0.12

Note: * $p \leq .1$, ** $p \leq .05$, *** $p \leq .001$