Text Mining, Linguistic Analysis, and Computational Methods

Hovig Tchalian
Research Interests

My research focuses on language & innovation...

Or, how new ideas spread, become adopted, and create markets

1. What differentiates GM’s and Tesla’s EV launch strategies?
2. How has corp. governance discourse changed since the 1970s?
3. Why are some public statements trustworthy, others less so?

Examining the fundamental determinants of social consensus
Text Mining & Linguistic Analysis

A simple search provides a great example of language change:

Google Books Ngram Viewer

Graph these comma-separated phrases: revolution,socialism,democracy

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books

Tchalian, Big Data: Separating Hope from Hype
How we say something matters as much as what we say

- New “claims” need justification – Example: Adoption of TQM
  - \( Q \text{ increases Cost (old P)} \rightarrow Q \text{ reduces W : W increases C : Q reduces C (new P)} \)
  - Syllogistic logic (Minor Premise [quality] : Major Premise (waste) : Claim [costs])

Argument Structure and Institutionalization

Green et al, “Suspended in Self-Spun Webs of Significance” (AMJ 2009)
The Number and Type of Each TQM Proposition by Year

- Major premise 1
- Major premise 2
- Minor premise
- Claim


Number of Propositions: 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20
Text Mining & Linguistic Analysis

*Language* is a way of establishing (& perpetuating) *social consensus*

- Language ‘names’ what we see (or, *notice*)
  - Before describing anything, we first need to recognize it

- Language devices make sense of the unfamiliar
  - Amazon *analogies*: browsing, shopping carts, checking out

- The *social conversation* is fundamental to how we think and act
  - Mark Kennedy’s work on ‘social listening’
The common denominator?
The social influence of crowds
So we use radical new tools to analyze crowd-level conversations ("social conversations")
Why?

For example: “Concept car”

Because language use underlies social categories & market labels

Source: Mark Kennedy, Imperial College, London
My Research Project: Electric Vehicle Market Category

• Category emergence and development
• “Extended Nascency”
  • “Next big thing” for the last 100 years
• Influence of valence – i.e., positive / negative associations
  • Social and strategic implications
Empirical Context: EV Technology (Emerged in 1880s)

1990: Impact concept car announced (LA Auto Show)
1996-9: 1,117 EV1 units leased, ~2,000 produced
2002: EV1 discontinued
2003: Tesla Motors founded
2006: Tesla Roadster prototype announced
2008-11: Roadster available for purchase (2,400 sold)
Motivation: Impact of Category Development

• An empirical and theoretical puzzle:
  • Two (*Now Three*) Contrasting Cases:
    Why did the EV1 spectacularly fail to meet expectations while the Roadster stubbornly survived and eventually exceeded expectations, when both vehicles faced similar, seemingly insurmountable adoption challenges?
Data Collection: Comprehensive DB of 80,000+ Texts

Downloaded 110k (using 80k) Factiva texts + metadata – “EV” variants

Comprehensive sources for EV modern formative period

1985-2014

popular sources

press releases

industry magazines
Indicators of Category Development: 1985-2014

EV Article Counts, by Source

- GM Impact announced
- First EV1 prod. model
- First Roadster prod. model
- Tesla Roadster announced

Spike in expert discs. (13-25%)
Analytical Method: Analyzing *Sentence-Level Meaning*

subject + \( \text{verb} + \text{clause(s)} = \text{predicate} \)

**CATEGORY**

*Electric Vehicle*

*Product*

*EV-1 Roadster Leaf*

- is
- is not
- is like
- is unlike
- is AND is not
- is like AND unlike

clusters of consistent associations:

- features
- &
- benefits
Electric Vehicle (EV): Related Second-Order Categories

<table>
<thead>
<tr>
<th>CATEGORY LEVEL</th>
<th>CAR</th>
<th>TRUCK</th>
<th>COMMERCIAL VEHICLE</th>
<th>UTILITY VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Category Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Electric Vehicle (EV): Related Second-Order Categories

<table>
<thead>
<tr>
<th>CATEGORY LEVEL</th>
<th>CAR</th>
<th>TRUCK</th>
<th>COMMERCIAL VEHICLE</th>
<th>UTILITY VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Category Level</td>
<td>Sedan</td>
<td>Luxury Car</td>
<td>Sports Car</td>
<td>Coupe</td>
</tr>
<tr>
<td>Feature Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Electric Vehicle (EV): Related Second-Order Categories

<table>
<thead>
<tr>
<th>CATEGORY LEVEL</th>
<th>CAR</th>
<th>TRUCK</th>
<th>COMMERCIAL VEHICLE</th>
<th>UTILITY VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Category Level</td>
<td>Pickup Truck</td>
<td>Van</td>
<td>Minivan</td>
<td></td>
</tr>
</tbody>
</table>

Feature Level
Electric Vehicle (EV): Related Second-Order Categories

<table>
<thead>
<tr>
<th>CATEGORY LEVEL</th>
<th>CAR</th>
<th>TRUCK</th>
<th>COMMERCIAL VEHICLE</th>
<th>UTILITY VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Category Level</td>
<td></td>
<td></td>
<td>Long-Haul</td>
<td>Big rig</td>
</tr>
<tr>
<td>Feature Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Electric Vehicle (EV): Related Second-Order Categories

<table>
<thead>
<tr>
<th>CATEGORY LEVEL</th>
<th>CAR</th>
<th>TRUCK</th>
<th>COMMERCIAL VEHICLE</th>
<th>UTILITY VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Category Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature Level</td>
<td>Tractor</td>
<td>Lawn Mower</td>
<td>Golf cart</td>
<td></td>
</tr>
</tbody>
</table>
GM: Cognitive Categorization

CATEGORICAL LEVEL

CATEGORY LEVEL

Sub-Category Level

Facet Level

CAR

Sedan

Coupe

EV

COMMERCIAL VEHICLE

UTILITY VEHICLE

~ Sports car
Unlike 1 or 2

Conventional categorization
Prominent member of ‘failed’ set
Electric Vehicle (EV): Related Second-Order Categories

**Strong Associations**

**Weak Associations**

<table>
<thead>
<tr>
<th>CATEGORY LEVEL</th>
<th>CAR</th>
<th>TRUCK</th>
<th>COMMERCIAL VEHICLE</th>
<th>UTILITY VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Category Level</td>
<td>Sports Car</td>
<td>Van</td>
<td>Electric</td>
<td>Golf Cart</td>
</tr>
</tbody>
</table>

**Feature Level**

{associational linkages: direct, indirect, ambiguous}
Text Analysis – Some Basic Tools & Approaches

- Simple word frequencies
  - Google Ngrams: [https://books.google.com/ngrams](https://books.google.com/ngrams)
  - Wordle / Word Clouds: [http://www.wordle.net/create](http://www.wordle.net/create)
Text Analysis – Some Basic Tools & Approaches

- **Simple word frequencies**
  - Google Ngrams: [https://books.google.com/ngrams](https://books.google.com/ngrams)
  - Wordle / Word Clouds: [http://www.wordle.net/create](http://www.wordle.net/create)

- **Motif Investing & idea clusters**
  - [http://video.cnbc.com/gallery/?video=3000285380](http://video.cnbc.com/gallery/?video=3000285380)
  - [https://www.motifinvesting.com/motifs#catalog=our&checked=featured](https://www.motifinvesting.com/motifs#catalog=our&checked=featured)
Text Analysis – Some Basic Tools & Approaches

- Simple word frequencies
  - Google Ngrams: https://books.google.com/ngrams
  - Wordle / Word Clouds: http://www.wordle.net/create

- Motif Investing & idea clusters
  - http://video.cnbc.com/gallery/?video=3000285380
  - https://www.motifinvesting.com/motifs#catalog=our&checked=featured

- Vector-Space Models
Vector Space Models: A Brief Overview

Documents → Vector-space representation

We study the complexity of influencing elections through bribery. How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election’s winner? We study this problem for election systems as varied as scoring …

<table>
<thead>
<tr>
<th>Term</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
</tr>
</thead>
<tbody>
<tr>
<td>complexity</td>
<td>2</td>
<td></td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>algorithm</td>
<td>3</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>entropy</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>traffic</td>
<td></td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>network</td>
<td></td>
<td></td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Vector Space Models: A Canonical Example
Text Analysis – Some Basic Tools & Approaches

- **Simple word frequencies**
  - Google Ngrams: [https://books.google.com/ngrams](https://books.google.com/ngrams)
  - Wordle / Word Clouds: [http://www.wordle.net/create](http://www.wordle.net/create)

- **Motif Investing & Idea clusters**
  - [http://video.cnbc.com/gallery/?video=3000285380](http://video.cnbc.com/gallery/?video=3000285380)
  - [https://www.motifinvesting.com/motifs#catalog=our&checked=featured](https://www.motifinvesting.com/motifs#catalog=our&checked=featured)

- **Vector-Space Models: LDA / Topic Modeling**
  - **Latency**: Word vectors $\rightarrow$ topic vectors $\rightarrow$ documents
  - **Unlabeled** approach – domain experts or automated process
  - **Labels**: Recognizability (frequencies) $\leftarrow \rightarrow$ discerningness (PMI)
Table 1. Unsupervised Eight-Topic Model. Table 1 displays the 40 highest-ranked words for each topic. Words were “stemmed” in the model (e.g., propose, proposes, and proposal are treated as the same word, propos, for analysis) but have been rewritten as full words here for clarity when applicable. $\alpha$ for this model was set to .01.

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep</td>
<td>work</td>
<td>shift</td>
<td>utility</td>
<td>propose</td>
<td>propose</td>
<td>electronic</td>
<td>propose</td>
</tr>
<tr>
<td>work</td>
<td>work</td>
<td>regulate</td>
<td>cost</td>
<td>company</td>
<td>construction</td>
<td>day</td>
<td>day</td>
</tr>
<tr>
<td>shift</td>
<td>operate</td>
<td>operate</td>
<td>eobr</td>
<td>carrier</td>
<td>propose</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>fatigue</td>
<td>propose</td>
<td>study</td>
<td>require</td>
<td>carrier</td>
<td>construction</td>
<td>address</td>
<td>address</td>
</tr>
<tr>
<td>day</td>
<td>propose</td>
<td>safety</td>
<td>system</td>
<td>eobr</td>
<td>industry</td>
<td>make</td>
<td>make</td>
</tr>
<tr>
<td>perform</td>
<td>vehicle</td>
<td>period</td>
<td>data</td>
<td>day</td>
<td>duty</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>schedule</td>
<td>study</td>
<td>vehicle</td>
<td>motor</td>
<td>attach</td>
<td>safety</td>
<td>company</td>
<td>company</td>
</tr>
<tr>
<td>night</td>
<td>operate</td>
<td>white</td>
<td>crash</td>
<td>work</td>
<td>period</td>
<td>rest</td>
<td>rest</td>
</tr>
<tr>
<td>study</td>
<td>effect</td>
<td>data</td>
<td>accident</td>
<td>industry</td>
<td>day</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>operate</td>
<td>circadian</td>
<td>crash</td>
<td>carrier</td>
<td>propose</td>
<td>attach</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>effect</td>
<td>period</td>
<td>rest</td>
<td>rest</td>
<td>propose</td>
<td>work</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>period</td>
<td>duty</td>
<td>motor</td>
<td>propose</td>
<td>propose</td>
<td>work</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>de</td>
<td>device</td>
<td>motor</td>
<td>propose</td>
<td>propose</td>
<td>work</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>circadian</td>
<td>cost</td>
<td>electronic</td>
<td>propose</td>
<td>propose</td>
<td>work</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>report</td>
<td>safety</td>
<td>safety</td>
<td>propose</td>
<td>propose</td>
<td>transport</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>research</td>
<td>status</td>
<td>increase</td>
<td>propose</td>
<td>propose</td>
<td>concrete</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>highway</td>
<td>log</td>
<td>addition</td>
<td>propose</td>
<td>propose</td>
<td>limit</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>day</td>
<td>electronic</td>
<td>industry</td>
<td>propose</td>
<td>propose</td>
<td>delivery</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>data</td>
<td>device</td>
<td>duty</td>
<td>propose</td>
<td>propose</td>
<td>maximum</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>operate</td>
<td>crash</td>
<td>fatigue</td>
<td>propose</td>
<td>propose</td>
<td>company</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>state</td>
<td>carrier</td>
<td>agency</td>
<td>propose</td>
<td>propose</td>
<td>product</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>period</td>
<td>company</td>
<td>company</td>
<td>propose</td>
<td>propose</td>
<td>washington</td>
<td>work</td>
<td>work</td>
</tr>
<tr>
<td>worker</td>
<td>benefit</td>
<td>benefit</td>
<td>propose</td>
<td>propose</td>
<td>clerk</td>
<td>road</td>
<td>road</td>
</tr>
<tr>
<td>safety</td>
<td>impact</td>
<td>impact</td>
<td>propose</td>
<td>propose</td>
<td>owner</td>
<td>road</td>
<td>road</td>
</tr>
<tr>
<td>fhwa</td>
<td>type</td>
<td>type</td>
<td>propose</td>
<td>propose</td>
<td>regulate</td>
<td>road</td>
<td>road</td>
</tr>
</tbody>
</table>

Levy and Franklin, “Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry” (Social Science Computer Review, 2014)
Topic Modeling / LDA

What are topic models?

- Most popular = LDA (Latent Dirichlet Allocation)
  - Bayesian / Inferential method – *backing into* a generative model
  - Assumption: authors sample from a set of discourse-specific topics

- Bag-of-words approach
  - Words are only “visible” (!= *latent*) feature
  - Treated as “random” variable – independent of sequence, linguistic context

- Mixed-membership assumption
  - Words can appear in >1 topic (approximates meaning / nuance)
  - Each *document* is a (vector-based) probability / likelihood distribution over *topics*
  - Each *topic* is a (vector-based) probability / likelihood distribution over *words*

David Blei, one of the originators, has expanded to L-LDA, hLDA, ...
Advantages

- Helps “interpret” texts
- Allows longitudinal analysis of category development
- Can handle large corpora

Drawbacks

- Difficult to interpret (w rigor)
- Challenging to apply method appropriately
- Can be demanding to implement

Topic Modeling Tradeoffs (Hannigan et al. 2018)

Developed in IE (Info Extraction) and IR (Info Retrieval) – has to be adapted for theoretical research
Our Analytical Methods: Variants of Topic Modeling

- These and similar text-analytic techniques made it possible to explore how the strategic efforts of GM and Tesla influenced / were influenced by the problematic EV category.
Analytical Details: *EV Two-Part Empirical Model*

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms

- percent
- year
- govt
- i
- million
- police
- people
- bush
- united
- new
- president
- government
- two
- years
- first
- billion
- state
- south
- court
- test
- today
- west
- officials
- party
- company
- home
- federal
- dukakis
- house
- military
Another Alternative *Supervised* LDA Method

<table>
<thead>
<tr>
<th>hLDA</th>
<th>L-LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>- <em>Supervised</em>, hierarchical (rank-ordered) topic generation</td>
<td>- <em>Supervised</em> (pre-labeled) topic generation</td>
</tr>
<tr>
<td>- Fewer parameters to choose</td>
<td>- Constrained to topics of interest</td>
</tr>
<tr>
<td>- Potentially more rigorous (Jordan)</td>
<td>- Provides framework for apples-to-apples comparison</td>
</tr>
</tbody>
</table>


L-LDA Example: Cross-Disciplinary Dissertations

Key Question: how well do cross-disciplinary dissertations (e.g., computer science and computational linguistics) fit their labels? (– And secondarily, how close are corresponding departments?)

PROCESS

1. “Learn” topics based on department designations

2. Use departments as tags for L-LDA (i.e., departments = topics)

3. Ignore labels & rerun algorithm → compare results

Chuang et al, Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis (CHI’12, May 5–10, 2012)
L-LDA Output: Cross-Disciplinary Dissertations

Figure 4. The Thesis View shows individual dissertations as small circles placed between the focus department and the next most similar department. Reading the original text of the dissertation enables experts to evaluate observed dept-dept similarities, and confirm the placement of three computational linguistics Ph.D.s that graduated in 2005.
Getting Started: Implementations

1. User-friendly / GUI tools – e.g., Topic Modeling Tool (TMT)
   ✓ G Code Archive: https://code.google.com/archive/p/topic-modeling-tool/

2. Mallet (Java) + Hierarchie for hLDA (caveat: Mallet hLDA in beta)
   ✓ Mallet for Windows: http://mallet.cs.umass.edu/

3. R and/or Python for “conventional” LDA and some variants
   ✓ Both have learning curve but have become standard in the field
Using Text Analysis and Computational Methods

• Explore on your own, get a feel for output – start with GUI

• Partner with a technical expert – esp. Mallet implementation

• Experiment w R / Python – 6-mo.+ learning curve but worth it
Thank You