Text Analysis: New Methods

Shahryar Minhas

July 19, 2018

Overview

Fredén

Summary

- Conduct an exploratory analysis on the relationship between threat, negative emotions and voting in the 2016 Brexit referendum using a "mortality salience manipulation experiment"
- Focus is on explaining how latent semantic analysis (LSA) can be used to understand voter choice in the 2016 British EU Referendum
- Key benefit study shows with regards to LSA is that it can uncover "relationships that were hidden in the standard analysis of numeric data"

Mortality salience manipulation experiment

To test the relationships between threats and the preference for Brexit or not, the authors conducted a mortality salience manipulation experiment in which:

- One group was randomly assigned to a scenario in which they would describe their thoughts related to what would happen to their body when they die
- the other group had to describe their associations related to a dentist visit

"The death threat scenario was supposed to evoke existential threat, whereas the dentist condition was expected to provoke milder reactions"

- **More detail was necessary here:**
  - What are the expected effects of being given one of these questions on a vote for Brexit and why?
  - Also should there have been a third group that was not asked to either give their thoughts about the death scenario or the dentist visit?

Initial findings

Authors report that initial look at the data indicated little difference between two groups.

- "neither the share for leave or remain differed between control and threat conditions"
- "Another expectation was that the impact of threat would vary depending on personal characteristics ... However, the relationships were rather weak, and the reactions from the death threat treatment were difficult to distinguish from dentist treatment"
- **What analyses were conducted to test for differences?**

Estimating underlying structure

- The dataset contained approximately 400 individual responses to the open-ended questions, and to examine whether there was underlying structure in these responses the authors turn to LSA.
- LSA involves decomposing the term-document matrix via SVD.
- **A few notes:**
  - Use of LSA is really interesting and it has connections to how some in political science have studied networks via the latent factor model
  - Should give a nod to the Simon & Xenos (2004) Political Analysis piece that built a dimension reduction framework similar to LSA
  - How was k chosen in the SVD?
  - More detail would have been useful in how exactly a respondent level measure of threat was obtained from the LSA.

Jerzak, King, & Strezhnev

Updating readme

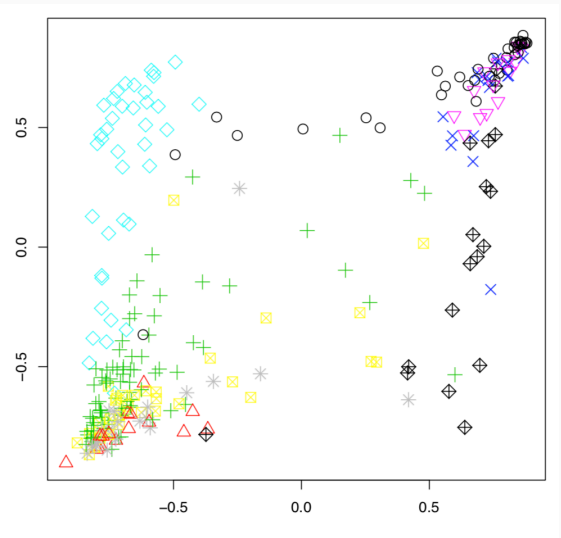
- This paper presents an update to the readme approach developed by Hopkins & King (2010).
- Goal of readme versus other text analytic approaches is to estimate percent of documents in each category using a supervised approach rather than classifying the documents directly.
- First half of the paper is devoted to reintroducing readme in a way that emphasizes two key drawbacks:
  - Feature space created in the readme process can be inefficient and sparse
  - Semantic changes in language over time

Innovation 1

- When converting text to data, the readme procedure followed relatively standard procedures. As noted by Denny & Spirling (2018), preprocessing decisions can have important effects.
- readme2 provide a very interesting shift away from the typical document-term matrix to one where each term is projected onto a low-dimensional vector space represetation.
- In this revised format, each term in a document feature meatrix is now represented by a low-dimensional continuous vector estimated using a novel dimension reduction method designed for direct estimation of category proportions.

Innovation 1 cont'd

- **A couple of questions:**
  - To generate the continuous vector representations the authors optimize over category (CD) and feature distinctiveness (FD): $\Gamma^* = \arg \max_{\Gamma \in \mathbb{R}^{W \times W'}} \lambda \times CD(\Gamma) + (1 - \lambda) * FD(\Gamma)$
  - λ regulates the weighting scheme for CD and FD. In the applications is λ set to 0.5? Are there circumstances in which we would want to weight differently?
  - Also it was unclear to me how W should be set and what it was set at for the applications.



Innovation 2

- The authors further alter the feature space to take into account semantic changes over time.
- To do this, they prune the set of observations in the labeled set that have covariate profiles that are very distinct from those in the unlabeled set
- **One question:**
  - It seems that this solution hinges on making sure the labeled set is represenative. Is there any further general guidance to give here or will advice just vary from case to case?

Evaluation Exercises

- Results show great support for the benefits of readme2
- Last question I have is does the novel method the authors have produced for constructing a feature space result in improvement for any classifier? For example, does the random forest classifier with the revised feature space perform better than the random forest classifier which utilizes a feature space that is constructed in the typical way?

Windsor, Cai, & Cupit

Cognitive Framework of Fake News

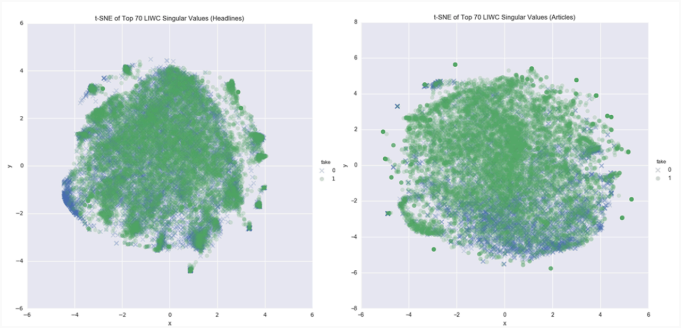
- Authors conduct a very interesting study of what syntactic patterns differentiate fake and real news
- To explore this difference they utilize a variety of methods such as LIWC, truncated SVD, and topic modeling

Clarification questions

- Syntax & Political Language
  - In the Syntax and Political Language section, a number of hypotheses are stated with regards to the level of syntactic complexity, narrativity, etc. in fake vs real news. It was not clear in the paper, how the texts were coded for these criteria.
- Getting Sentimental about Fake & Real News
  - The hypotheses here relate to the level of deceptive/honest language between fake and real news. How is the LIWC procedure applied to document headlines providing you with measures of honesty and deception?
  - Why the use of truncated SVD, the LIWC generates 93 measures and after performing the SVD you stil retain 70 features? Why not just apply the t-SNE on the features generated by the LIWC?

Clarification questions cont'd

- Getting Sentimental about Fake & Real News
  - Additionally, it's not clear to me what differences emerge in the application of the t-SNE algorithm



- Topics in Fake and Real News
  - More details should be provided on the steps you took to run the topic model. For example, what type of text processing was done and what was the gram size?
  - From the paper it's unclear to me what the topic model is buying you in terms of your interest in differentiating between fake and real news.

Next steps?

- A lot of interesting work is done in this project!
- Through this project you have been able to better hone in on features that might be important in differentiating between fake and real news. **But** what I was hoping to see towards the end of this paper was you using that knowledge to develop a classifier for predicting fake vs real news.
- One last question ... given that the creators of fake news may not be unsophisticated as research like this emerges (e.g., Rubin et al. 2016; Horne & Adali 2017; Potthast et al. 2017), do you imagine them shifting tactics?

Thanks!