Automated Text Analysis in Political Science

Lecture 9: New models

May 16, 2019

dr. Martijn Schoonvelde

School of Politics and International Relations, UCD

Today's class

- Word embeddings
- LTTA
- Flash talks: Manna & Alfredo

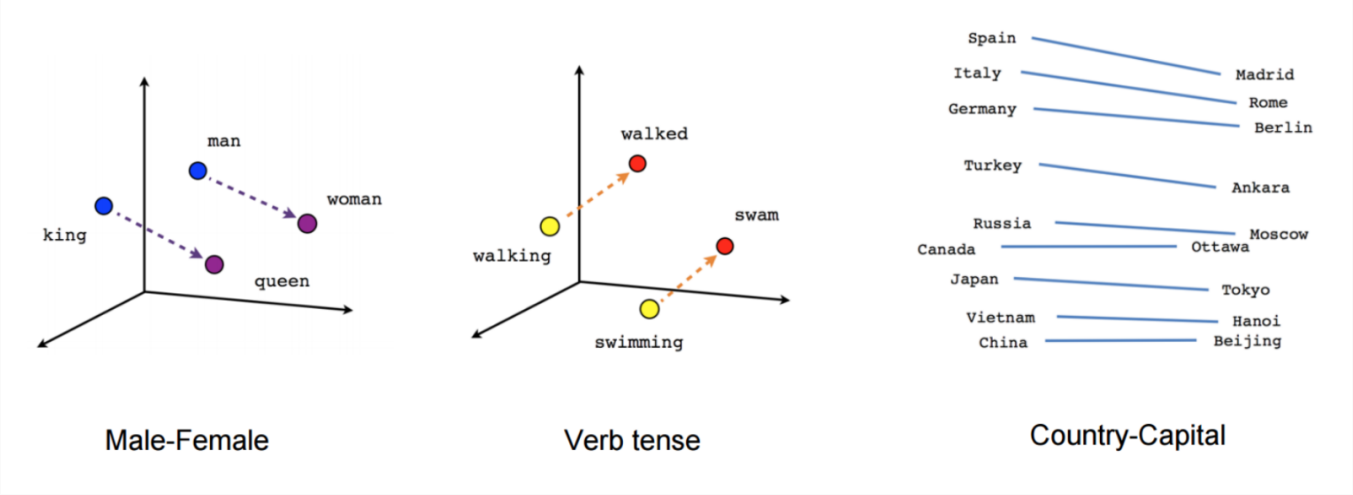
Word Embeddings

- Most applications of text as data in political science: bag of words
  - Context ignored (although there is a little context when using bigrams or trigrams)
  - Usually sparse, lots of zeros
- Word embeddings: different representation of text; words are "embedded" in a semantically space
  - Different algorithms to learn these word embeddings from the local context words appear in: GloVe, word2vec

Word Embeddings

Technique for identifying similarities between words using some type of model to predict the co-occurrence of words within a small chunk of text

- "You shall know a word by the company it keeps" (Firth, 1957)



Credit: `https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html`

Word Representations

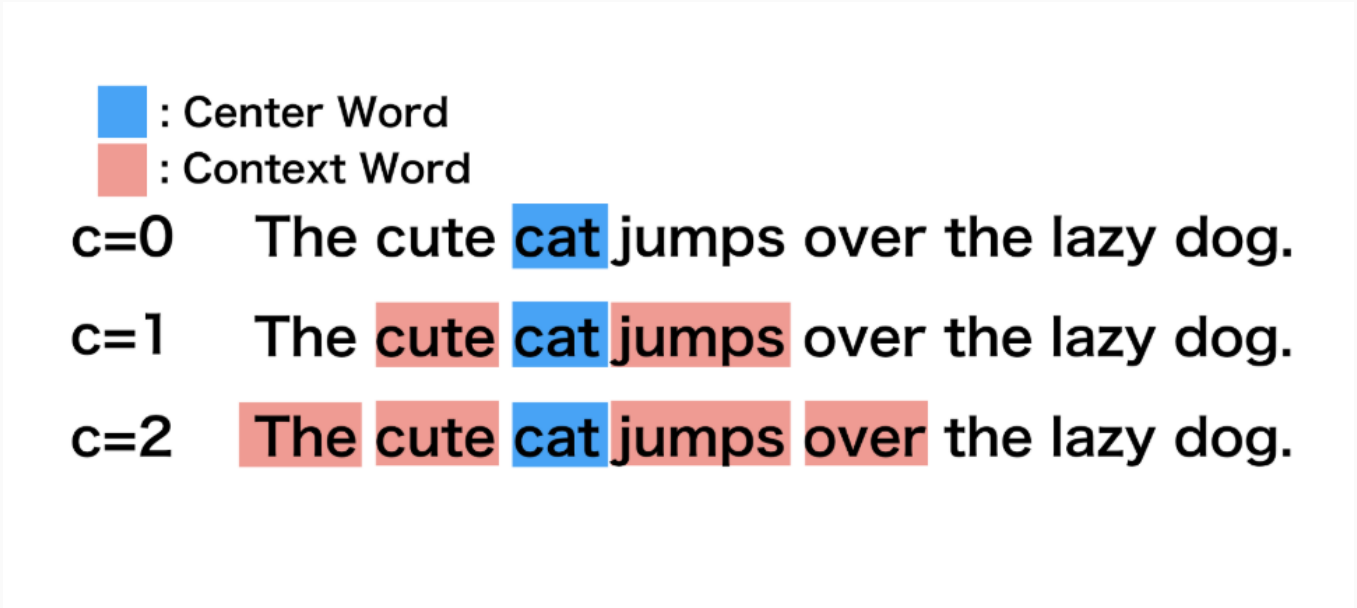
Bag of Words	Word embeddings
One-hot encoding	Vector in a semantic space
DxN	*NxV*
No context	Estimated from context
Meaning exogenous	Meaning learned
Input to a model	Output from a model

D = number of documents

N = number of words

V = number of embedding dimensions

Context Window



Credit: `https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html`

Training word embeddings

There are various algorithms to train word embeddings vectors:

- Word co-occurrence matrix and SVD
- Word2Vec
- GloVe

Important to keep in mind: researcher determines the size of the context window, the length of the word embeddings vector and whether to use pre-trained word embeddings or not (see Spirling & Rodriguez, 2019)

Some applications of word embeddings for social science

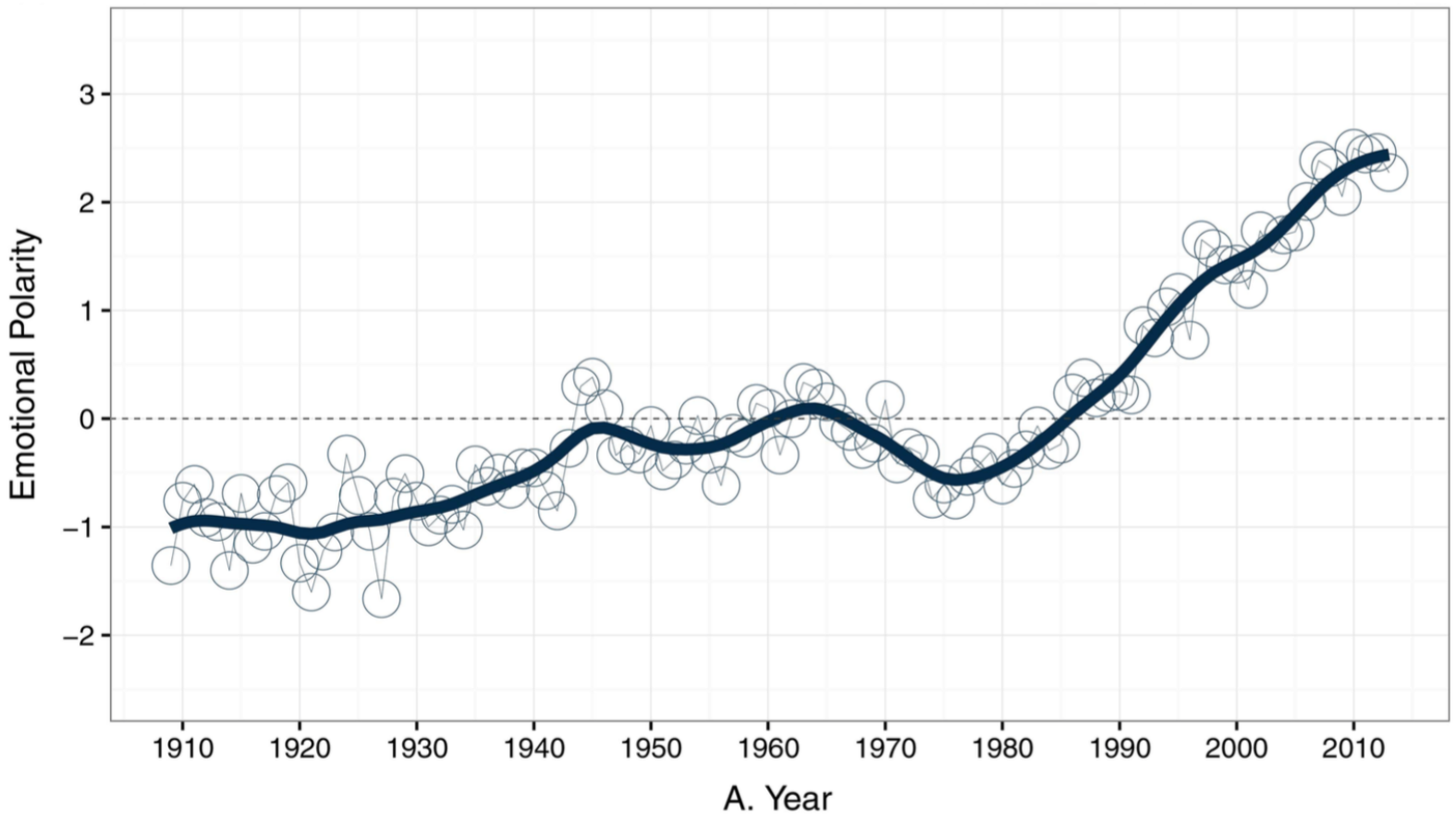
- Recommender systems
  - If I like movie A, I will also like movie B
- Improve dictionaries
- Understanding semantic shifts

Word Embeddings in Political Science

- Sentiment analysis
  - Rheault et al (2016) use a word embeddings algorithm to develop a "domain specific sentiment dictionary"
  - British House of Commons speeches between 1909 and 2013
  - After preprocessing, total of 108,506 unique tokens
    - Use word embeddings algorithm (Glove) to train their model
    - Then locate 200 positive and negative 'seed' words in this space
    - With these words located, they can locate other words nearby, leading to a total of 4200 words denoting positive and negative sentiment

Overall sentiment in the HoC

PLOS ONE | Measuring Emotion in Parliamentary Debates



Government and opposition sentiment in the HoC

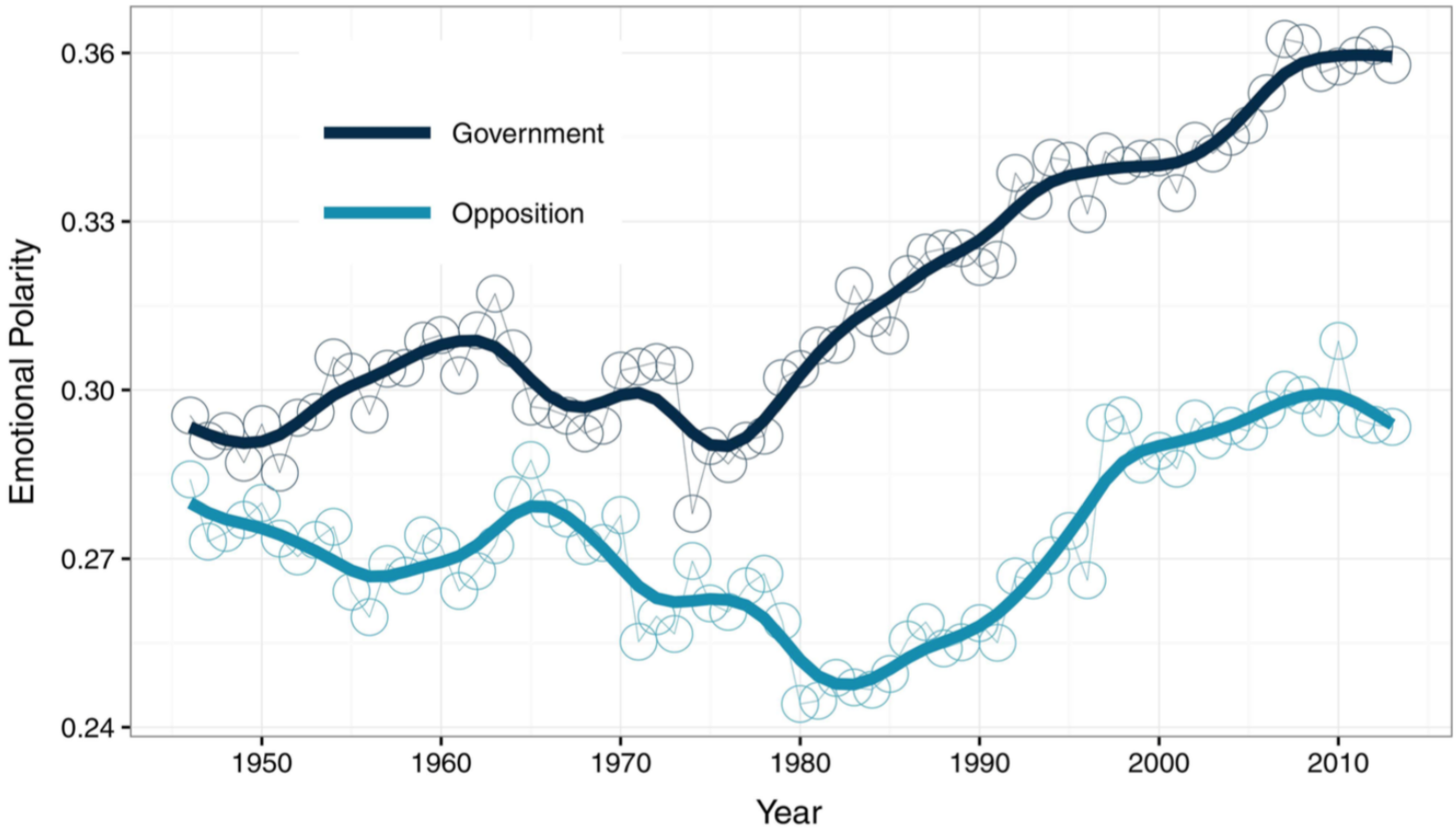


Fig 2. Emotional Polarity of Government and Opposition in Britain, 1946-2013.

doi:10.1371/journal.pone.0168843.g002

Word Embeddings

- Lots of cool possibilities: For example, how does the semantic meaning of words change over time (e.g., liberal and conservative)?
- Do parties shift in *how* they use particular words? For example, does debate vocabulary change over time?
    - See, e.g., work by Milan van Lange and Ralf Futselaar on War debates in Dutch parliament `https://github.com/MilanvanL/debating_evil`

Overview

For a set of validations of word embeddings models in political science, see Spirling & Rodriguez (2019)

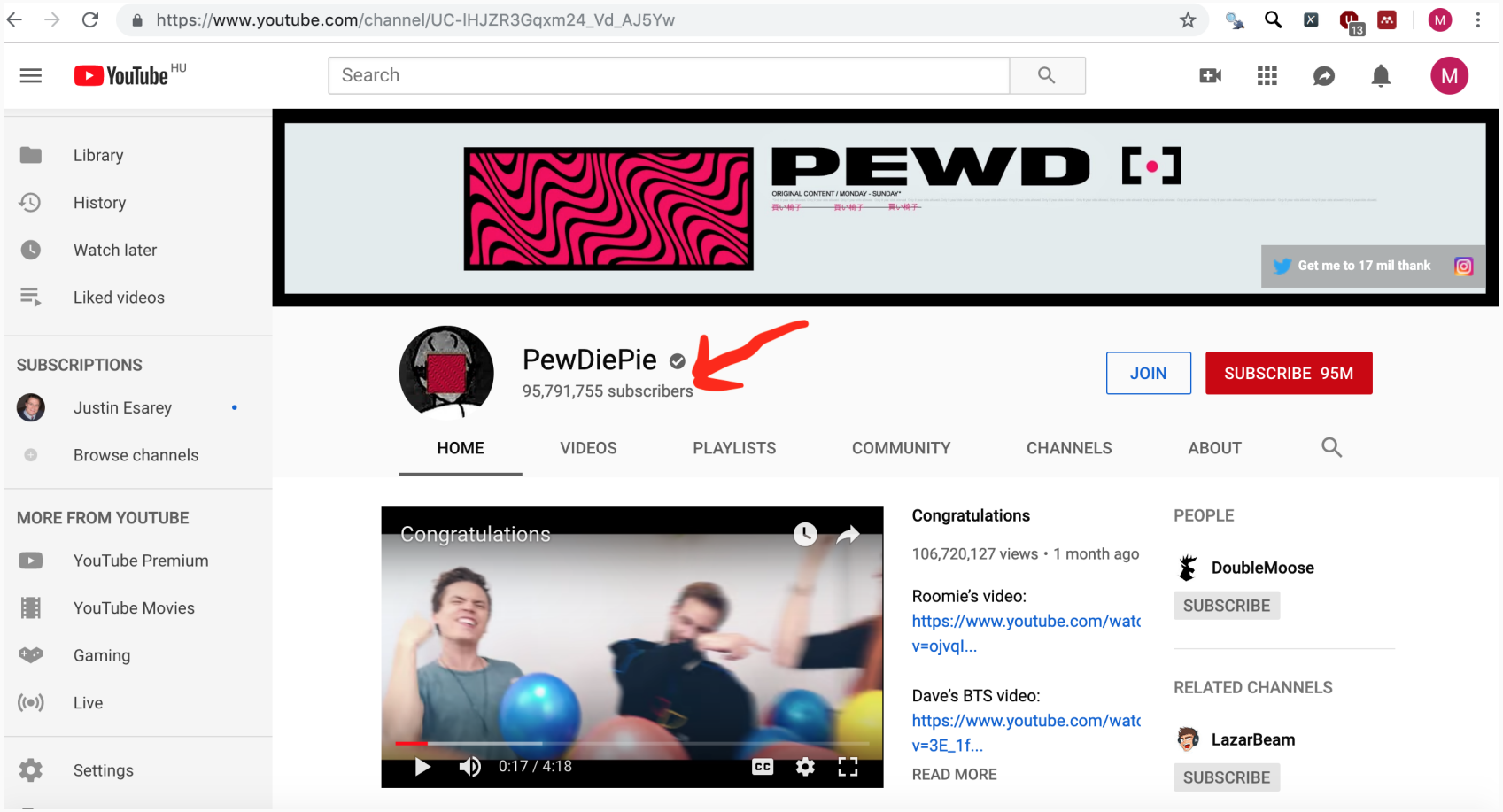
Word Embeddings
What works, what doesn't, and how to tell the difference for applied research*

Arthur Spirling[†] Pedro L. Rodriguez[‡]

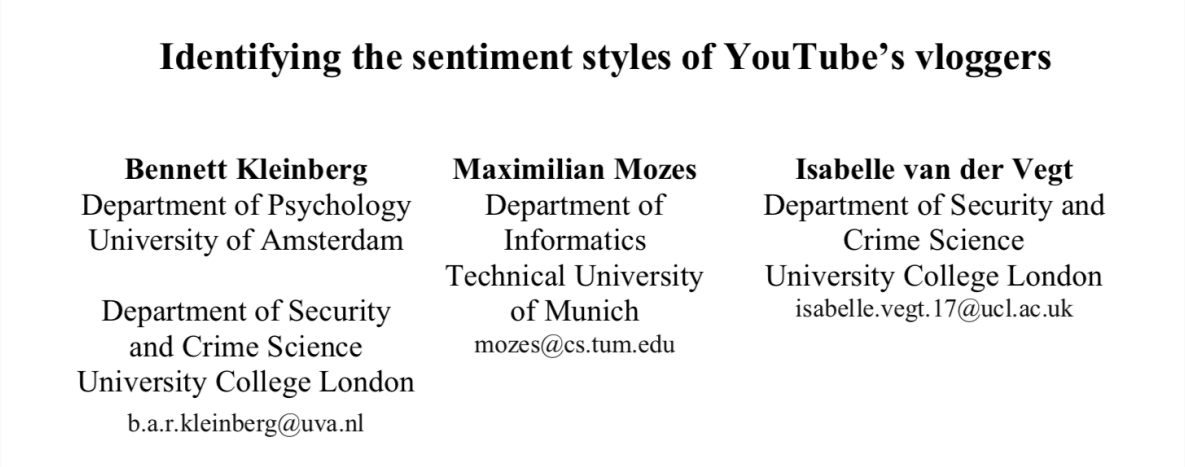
Abstract

We consider the properties and performance of word embeddings techniques in the context of political science research. In particular, we explore key parameter choices—including context window length, embedding vector dimensions and the use of pre-trained vs locally fit variants—in terms of effects on the efficiency and quality of inferences possible with these models. Reassuringly, with caveats, we show that results are robust to such choices for political corpora of various sizes and in various languages. Beyond reporting extensive technical findings, we provide a novel crowd-sourced "Turing test"-style method for examining the relative performance of any two models that produce substantive, text-based outputs. Encouragingly, we show that popular, easily available pre-trained embeddings perform at a level close to—or surpassing—both human coders and more complicated locally-fit models. For completeness, we provide best practice advice for cases where local fitting is required.

Who is this?


https://www.youtube.com/channel/UC-IHJZR3Gqxm24_Vd_AJ5Yw
YouTube HU
Search
Library
History
Watch later
Liked videos
SUBSCRIPTIONS
Justin Esarey
Browse channels
MORE FROM YOUTUBE
YouTube Premium
YouTube Movies
Gaming
Live
Settings
PEWD
ORIGINAL CONTENT / MONDAY - SUNDAY*
Get me to 17 mil thank
PewDiePie
95,791,755 subscribers
JOIN
SUBSCRIBE 95M
HOME
VIDEOS
PLAYLISTS
COMMUNITY
CHANNELS
ABOUT
Congratulations
0:17 / 4:18
Congratulations
106,720,127 views • 1 month ago
Roomie's video:
https://www.youtube.com/watc
v=ojvql...
Dave's BTS video:
https://www.youtube.com/watc
v=3E_1f...
READ MORE
PEOPLE
DoubleMoose
SUBSCRIBE
RELATED CHANNELS
LazarBeam
SUBSCRIBE

LTTA: Linguistic Temporal Trajectory Analysis



Identifying the sentiment styles of YouTube's vloggers

Bennett Kleinberg
Department of Psychology
University of Amsterdam

Department of Security and Crime Science
University College London
b.a.r.kleinberg@uva.nl

Maximilian Mozes
Department of Informatics
Technical University of Munich
mozes@cs.tum.edu

Isabelle van der Vegt
Department of Security and Crime Science
University College London
isabelle.vegt.17@ucl.ac.uk

- Method to inspect how linguistic markers like sentiment shift over time in a text

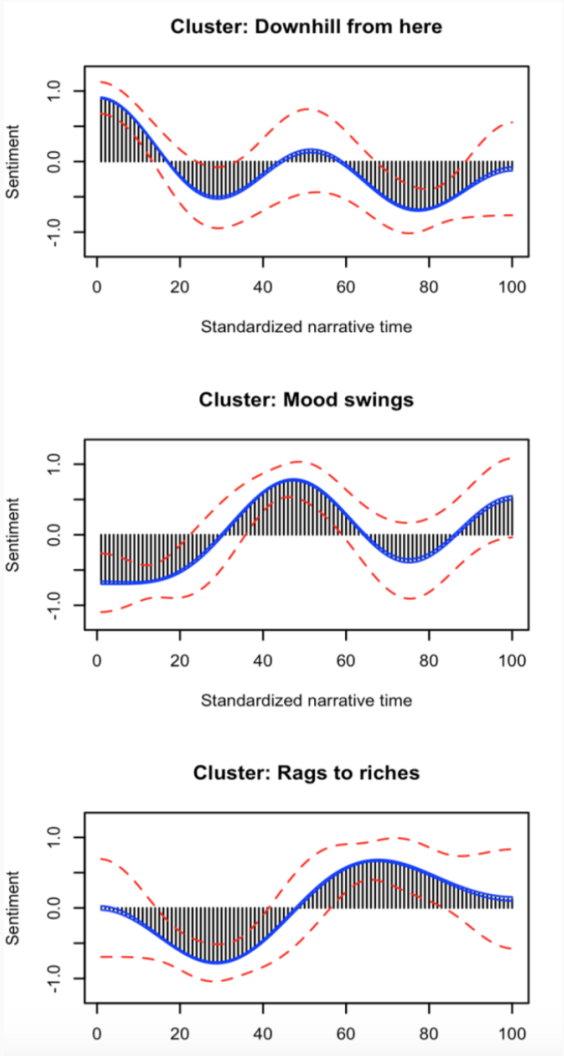
LTTA

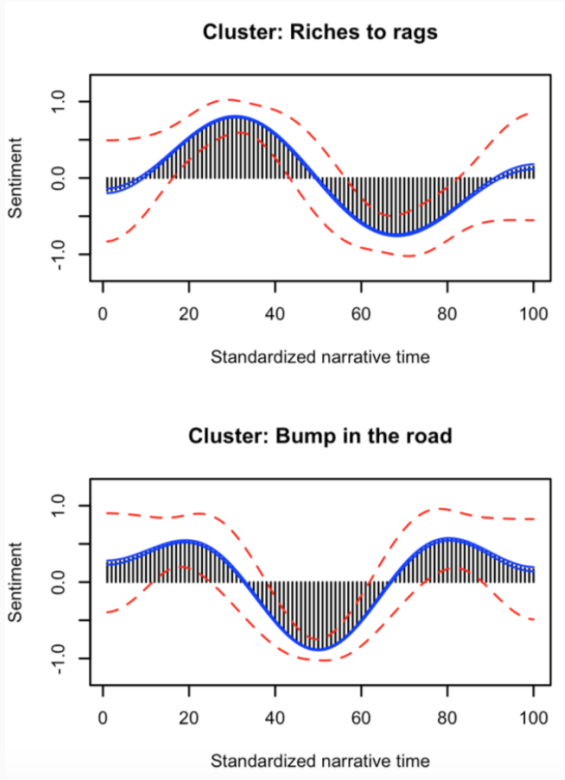
Corpus: selection of vlogs from the most popular vloggers

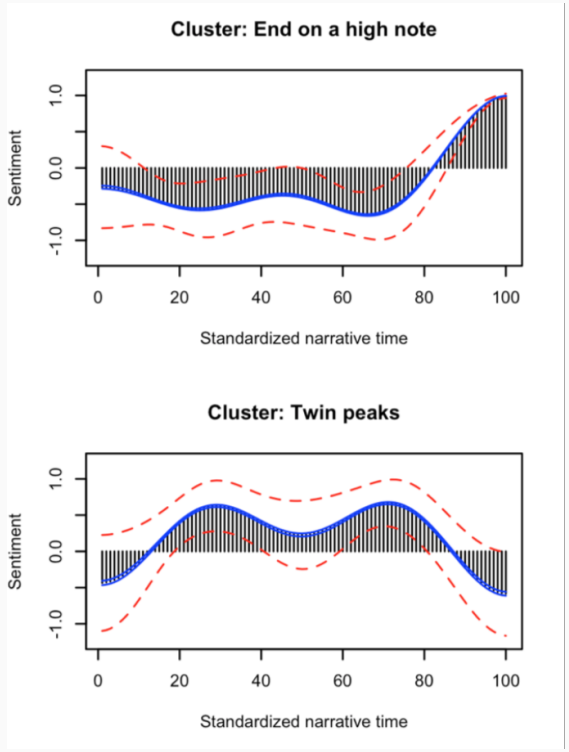
Obtain transcripts from all vlogs produced

- Method: sentiment detection in these using a sliding window – standardize within fixed time periods

LTTA


Cluster: Downhill from here
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100
Standardized narrative time
Cluster: Mood swings
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100
Standardized narrative time
Cluster: Rags to riches
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100


Cluster: Riches to rags
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100
Standardized narrative time
Cluster: Bump in the road
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100
Standardized narrative time


Cluster: End on a high note
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100
Standardized narrative time
Cluster: Twin peaks
Sentiment
1.0
0.0
-1.0
0
20
40
60
80
100
Standardized narrative time

Cluster	Family	Female	Male
Downhill from here	2.23	1.26	-2.88*
Mood swings	-2.31	1.96	1.25
Rags to riches	2.13	-1.95	-1.08
Riches to rags	-2.05	4.88*	-0.56
Bump in the road	1.69	-1.12	-1.08
End on a high note	-5.16*	-6.03*	8.32*
Twin peaks	3.83*	2.25	-4.99*

Table 3. Standardized residuals for the cluster-by-gender association.

Conclusions

LTTA could be a very neat way to study political rhetoric

Lots of cool new developments / tools outside of political science, under the realm of computational social science