Diffusion of Lexical Change in Social Media

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, Eric P. Xing

PloS ONE, vol. 9, no. 11, 2014
https://doi.org/10.1371/journal.pone.0113114

Presented by Bhanuka Mahanama
November 21, 2022

What is the Full Form?

- Lol
- Brb
- Btw
- DM
- HBD
- TBH
- G2G
- XD
- RT
- FAQ
- AFK

- Where did you find out?
- What is the origin of these terms?

Lexical Change Diffusion

- Process by which change in the meaning or use of a word is spread
  - AFK: Developed from chat rooms in the 1990s
- Importance
  - Identify influential groups
  - Groups evolve together
  - Hidden structures that shape the society
- Challenges
  - Building a robust model
  - Capturing channels of communication

https://slang.net/

Methodology

- Data collection
  - Sample Tweets using Twitter API
  - Data pre-processing
- Step 1: Modeling lexical dynamics
  - Word frequency across time
  - Word frequency based on other MSAs
- Step 2: Constructing a network for diffusion
  - Use lexical model to influential regions
  - Networks of influential flow
- Step 3: Demographic and geographic correlation analysis
  - Relationship to the constructed network
  - Can we predict MSA network using demographic features?

Dataset

- 107M tweets
  - Between 2009 - 2012
  - 165 weeks
  - 20.7M unique users
  - GPS coordinates of tweets
  - 200 largest Metropolitan Statistical Areas (MSAs)
- Demographic data
  - 2010 United States census data

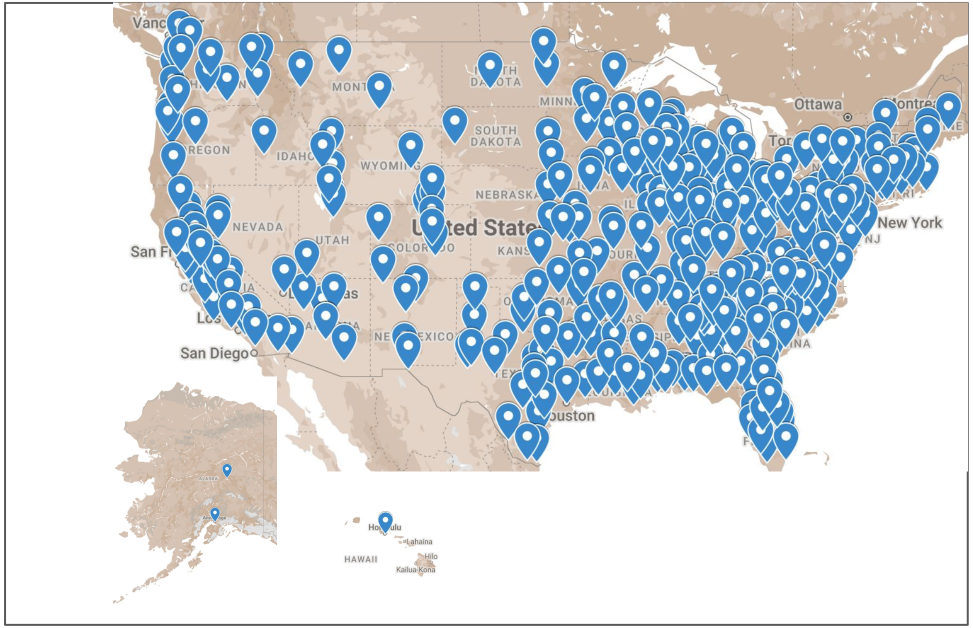

Illustration: Metropolitan Statistical Areas (MSAs) in the dataset (supplementary material)

Dataset Pre-Processing

- Remove marketing oriented accounts
- 100k most frequent terms
- 4854 terms with highest significant frequency change
- Manually refine to filter
  - Foreign worlds (e.g. bendiciones, y)
  - Hashtags (e.g. #nyc, #fb)
  - Names
- 2603 English words

smashin	some1
duin	evryone
doinn	evry1
doiin	every1
doin	evrybdy
doinq	everyonee
eatin	evrybody
grilling	oomf
cookn	oomfs
eattin	meeka
bakin	no1

Examples of selected words
(supplementary material)

Sample Diffusions

- Ion
  - = I don’t
  - Common in Southeast
- -_-
  - = Annoyance
  - Nationwide spread
- Ctfu
  - = cracking the f* up
  - Midwest to mid-Atlantic

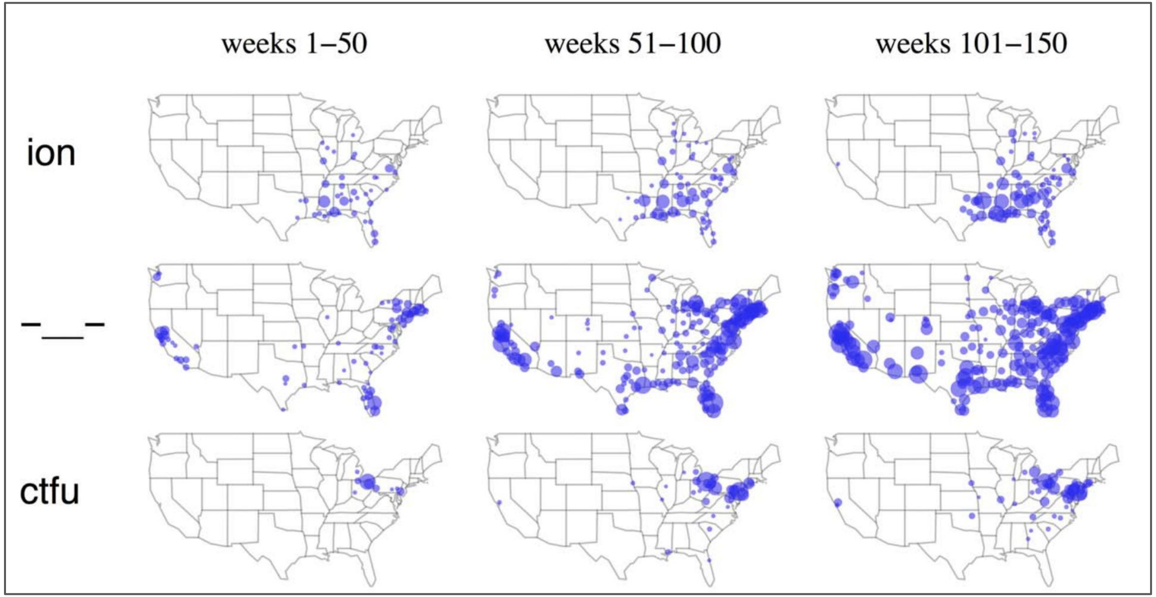

Figure 1: Change in frequency of ion, -_-, ctfu. Circle size proportional to word probability

Step 1: Modeling Lexical Dynamics

- Simplest approach: Autoregression
  - Output depends on previous values
- Directly operate on word count

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

Autoregressive model of order p

Challenges

- Challenge 1: Dissimilar users and tweets in MSAs
  - NYC = 4x San Francisco-Oakland, CA (10th)
  - NYC = 20x Oklahoma City, OK (50th)
- Challenge 2: Varying sampling rate in dataset
  - 5-15% sampling rate in 2010 and earlier
  - 10% sampling rate onwards

Solution 1: Data Normalization

- Convert word count to probabilities
  - Word = w
  - Region (MSA) = r
  - Time = t

Probabilities

- Not invariant to frequency change
- Different MSAs, different engagement levels
- Word count = 0 for many MSAs

\# of individuals used word ***w***

Probability of tweeting ***w***

$$p_{w,r,t} = \frac{c_{w,r,t}}{s_{w,r,t}}$$

\# of individuals tweeted

Solution 2: Latent Vector Autoregression

- Latent variables for word activation
  - Derived with logistic transformation of probability
  - Latent spatio-temporal activation
  - Underlying activation of word in region
- Global activation
  - Word becomes popular everywhere at once
- Regional activation
  - Word becomes popular in region

$$p_{w,r,t} = Logistic(\eta w, r, t)$$

$$\overbrace{c_{w,r,t}}^{\text{\# of individuals used word } \boldsymbol{w}} \sim Binomial(s_{r,t}, Logistic(\overbrace{\eta_{w,r,t}}^{\text{Latent spatio-temporal}} + \overbrace{\nu_{w,t}}^{\text{Global activation}} + \overbrace{\mu_{r,t}}^{\text{Regional activation}}))$$

Modelling Spatial Diffusion

Model diffusion using latent spatio-temporal variable as first order linear dynamical system with Gaussian noise

Latent spatio-temporal of region ***r*** at ***t***

Autoregressive coefficient between ***r*** and ***r'***

Latent spatio-temporal of region ***r'*** at ***t -1***

Autoregressive variance

$$\eta_{w,r,t} = N\left(\sum_{r'} a_{r'r}\, \eta_{n,r',t-1},\ \ \sigma^2_{w,r}\right)$$

Step 2: Constructing Diffusion Network

- Autoregressive dynamics matrix
  - Use autoregressive coefficients
- Generate ordered set of coefficients
  - Computed over all samples
  - Coefficient significantly greater than zero
- Use coefficients to form edges of network
  - Different thresholding => multiple networks
  - Model differentiates directionality => Bidirectional

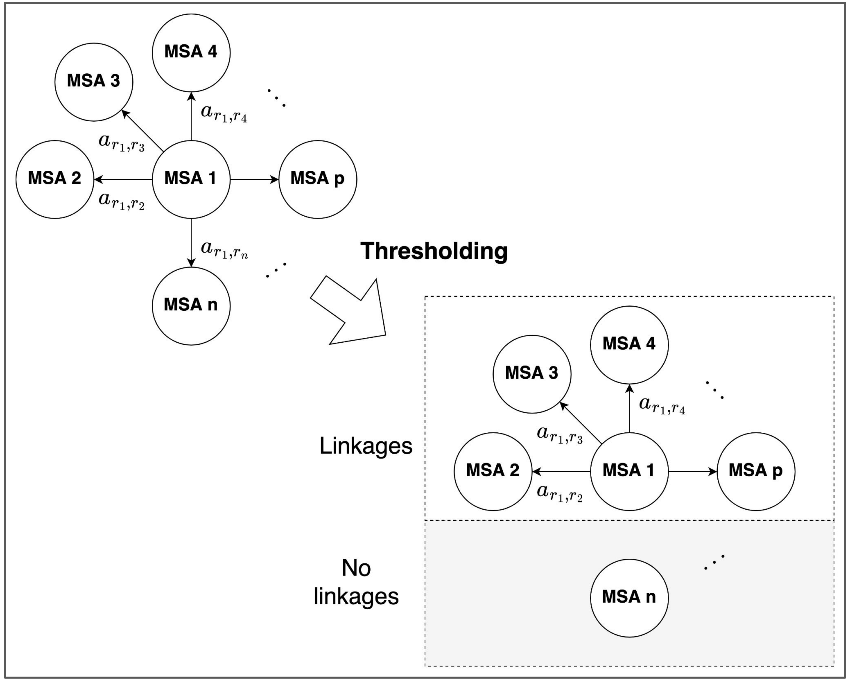

Illustration: Thresholding to form an induced network

An Example Induced Network

- Dense connections
  - Northeast
  - Midwest
  - West Coast
- Few cross-country connections

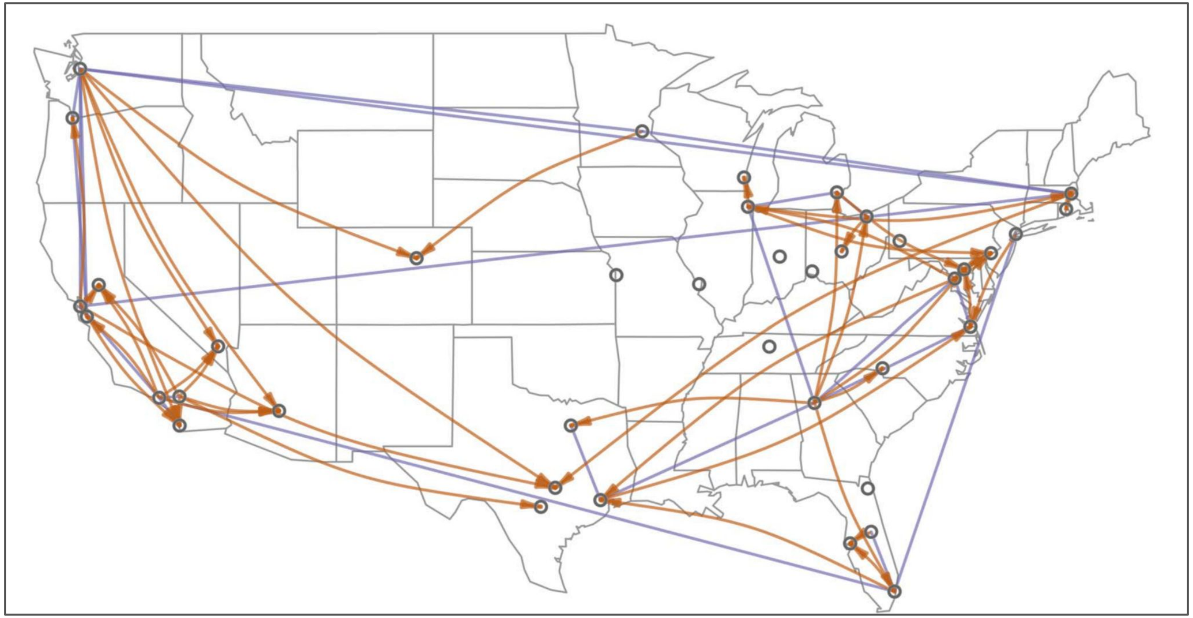

Figure 4: An induced network showing linkages among 40 most populous MSAs. Blue: Bidirectional, Orange: Unidirectional

Step 3: Geographic and Demographic Correlation

- Consider linkage between cities
  - Set of unlinked cities
  - Set of linked cities
- Compare geographic and demographic feature distance
  - Geographic: Distance
  - Demographic: Urbanized %, Income, Age, Renter %, Racial composition
- Symmetry
  - Symmetric: Absolute distance
  - Asymmetric: Raw distance

Geographic and Symmetric Demographic Features

Lower distance or difference

Higher distance or difference

	linked mean	linked s.e.	nonlinked mean	nonlinked s.e.
geography				
distance (km)	919	36.5	1940	28.6
symmetric				
abs diff % urbanized	9.09	0.246	13.2	0.215
abs diff log median income	0.163	0.00421	0.224	0.00356
abs diff median age	2.79	0.0790	3.54	0.0763
abs diff % renter	4.72	0.132	5.38	0.103
abs diff % af. am	6.19	0.175	14.7	0.232
abs diff % hispanic	10.1	0.375	20.2	0.530

Table 3: Differences between linked/non-linked cities

Predicting Links Between MSAs

- Use features to predict linkage
  - Geography
  - Symmetric demography
  - Asymmetric demography
  - Population
    - Raw log difference
- Using a logistic regression
  - Cross-validation accuracy

The network can be reconstructed using demographic and geographic features.

	Mean Accuracy	Std. Error
Geography + Symmetric + Asymmetric	**74.37**	**0.08**
Geography + Symmetric	74.09	0.07
Geography + Asymmetric	73.13	0.08
Geography + Population	67.33	0.08
Geography	66.48	0.09

Table 4: Average accuracy in predicting linkages

Conclusion

- Social media can uncover language evolution
  - Reveal hidden structures
  - Transmission in demographically similar areas
  - Language is homophilous
    - Demography and geography
  - Homophily at macro level
- Homophily between communities is an important factor driving the observable diffusion of lexical change
- Latent autoregressive model
  - Varying sampling rate
  - Different engagement levels
  - Dissimilarly populated MSAs
- Diffusion network
  - Relates to geographic and demographic features
- Challenges
  - Diffusion within MSAs
  - Only uses word frequencies
  - Doesn’t capture structural changes
    - I don’t know => ion