DATA SCIENCE PIPELINE FOR SOCIAL MEDIA TREND ANALYSIS

Authors: S and B

September 26, 2019

Course Project for Data Driven Research Methods

RESEARCH PROPOSAL: FLEXIBLE PIPELINE TO ANALYZE TRENDS ON SOCIAL MEDIA

Given a main topic, we make a process to analyze data and cluster communities and phrases.

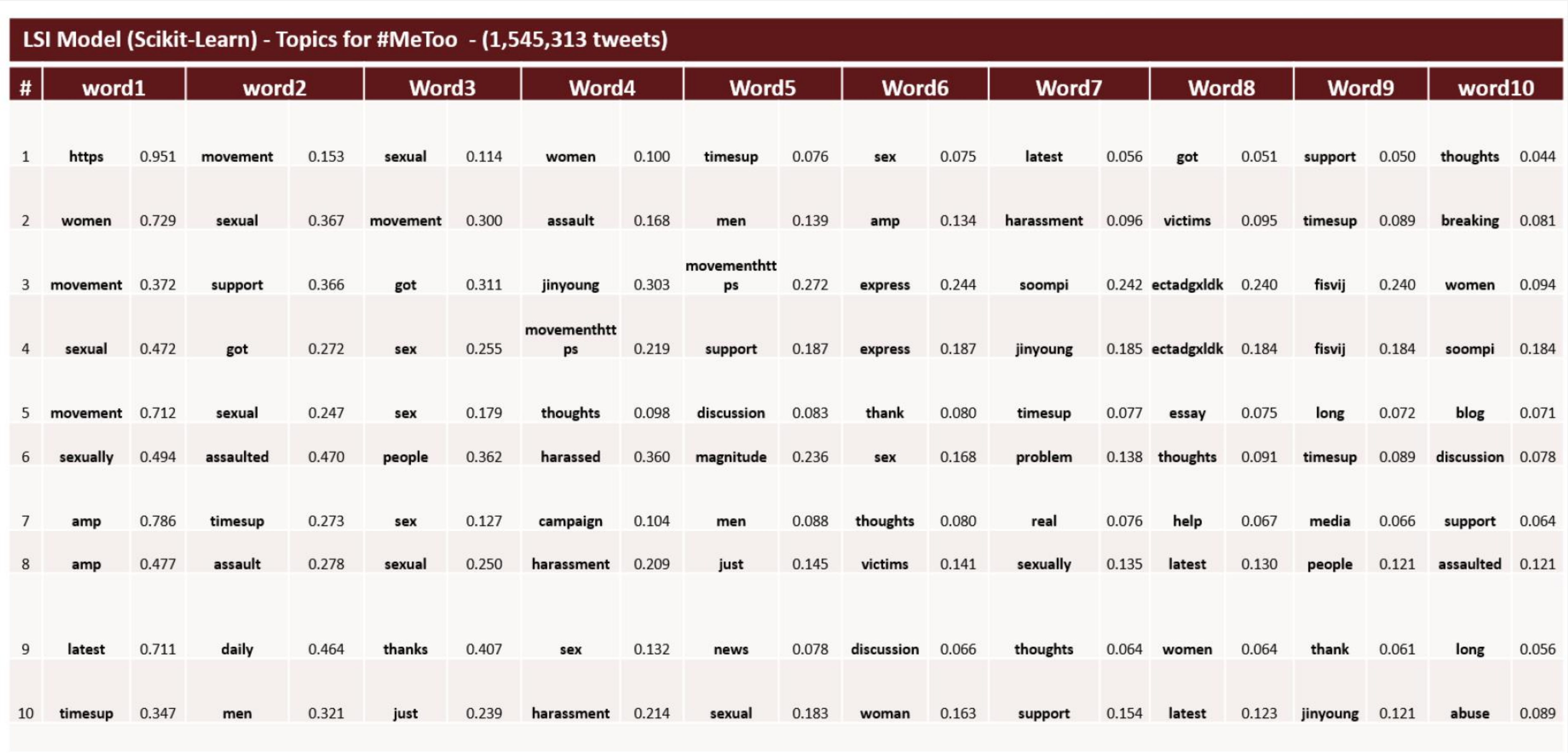
Goal: Find community and phrase clustering to inform further research on data consumption behavior.

CASE STUDY USING DS PIPELINE

What do the human links and events around the #MeToo look like?

PIPELINE TO ANALYZE TWITTER DATA

- Basis work: An automated and flexible pipeline to perform analysis on Twitter data with focus on natural language processing (NLP) research
  - LNM (2019). Trending social topics in Twitter. Unpublished manuscript. Texas State University, San Marcos, Texas.
  - ML models used :
    1. LDA (Latent Dirichlet Allocation)
    2. LSI (Latent Semantic Indexing)
    3. NMF(Non-Negative Matrix Factorization)



LSI Model (Scikit-Learn) - Topics for #MeToo - (1,545,313 tweets)

#	word1		word2		Word3		Word4		Word5		Word6		Word7		Word8		Word9		word10	
1	https	0.951	movement	0.153	sexual	0.114	women	0.100	timesup	0.076	sex	0.075	latest	0.056	got	0.051	support	0.050	thoughts	0.044
2	women	0.729	sexual	0.367	movement	0.300	assault	0.168	men	0.139	amp	0.134	harassment	0.096	victims	0.095	timesup	0.089	breaking	0.081
3	movement	0.372	support	0.366	got	0.311	jinyoung	0.303	movementhttps	0.272	express	0.244	soompi	0.242	ectadgxldk	0.240	fisvij	0.240	women	0.094
4	sexual	0.472	got	0.272	sex	0.255	movementhttps	0.219	support	0.187	express	0.187	jinyoung	0.185	ectadgxldk	0.184	fisvij	0.184	soompi	0.184
5	movement	0.712	sexual	0.247	sex	0.179	thoughts	0.098	discussion	0.083	thank	0.080	timesup	0.077	essay	0.075	long	0.072	blog	0.071
6	sexually	0.494	assaulted	0.470	people	0.362	harassed	0.360	magnitude	0.236	sex	0.168	problem	0.138	thoughts	0.091	timesup	0.089	discussion	0.078
7	amp	0.786	timesup	0.273	sex	0.127	campaign	0.104	men	0.088	thoughts	0.080	real	0.076	help	0.067	media	0.066	support	0.064
8	amp	0.477	assault	0.278	sexual	0.250	harassment	0.209	just	0.145	victims	0.141	sexually	0.135	latest	0.130	people	0.121	assaulted	0.121
9	latest	0.711	daily	0.464	thanks	0.407	sex	0.132	news	0.078	discussion	0.066	thoughts	0.064	women	0.064	thank	0.061	long	0.056
10	timesup	0.347	men	0.321	just	0.239	harassment	0.214	sexual	0.183	woman	0.163	support	0.154	latest	0.123	jinyoung	0.121	abuse	0.089

Partial results from original work – Unlabeled classification of phrases

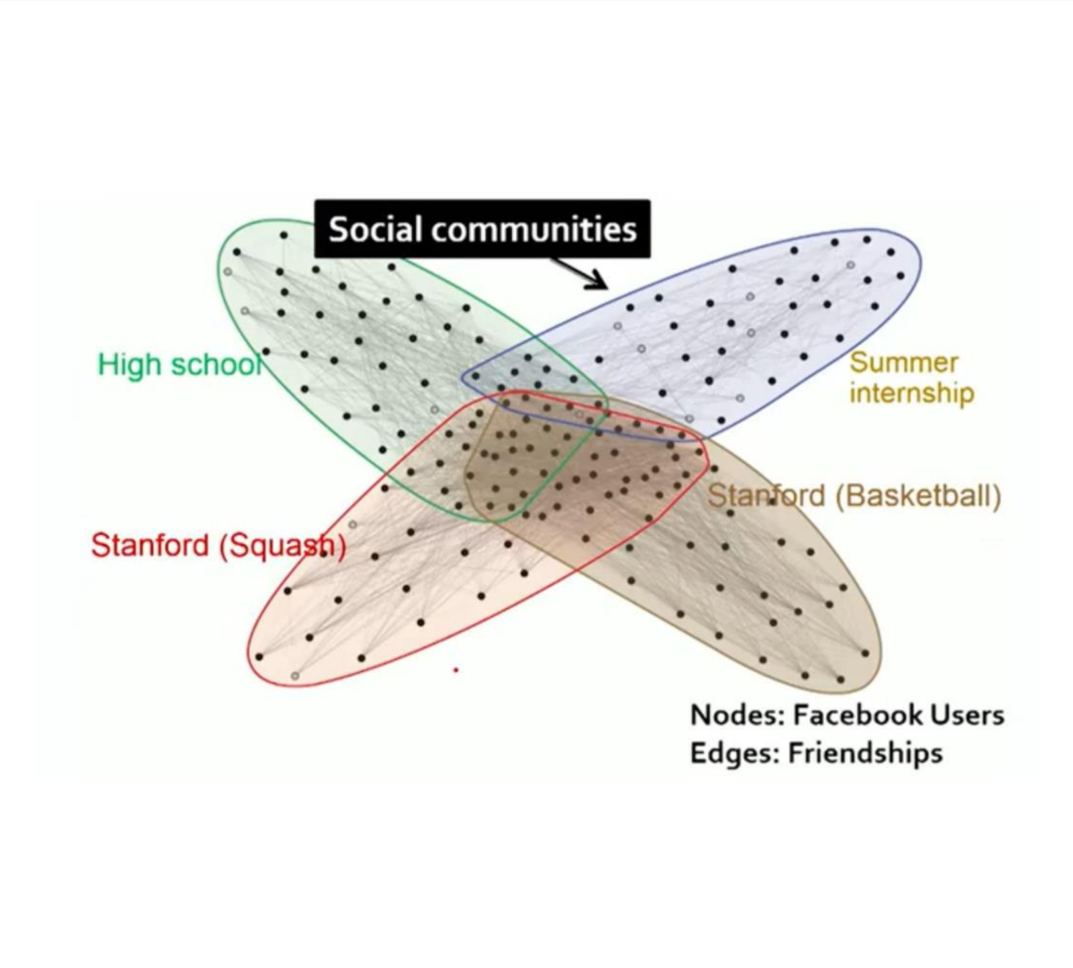
PIPELINE TO ANALYZE TWITTER DATA

- **We propose an improvement on original pipeline to include data cleaning, network analysis, and a method for community detection**

1. Integrate domain experts' feedback on relevant phrases to create n-grams to include in model
2. Integrate bot detection to refine classification
3. Graph data and use Louvain method to discover baseline communities
4. Post-processing module: From top 5% communities (by number of members) pull top 5% users (split by "followers" and "following")
    - Provide some descriptive analysis for those groups

FRAMEWORK & TOOLS

- Dataset
  - 3.1 million tweet documents provided by lead sociologists in the Family and Consumer Sciences department at TXST
  - https://archive.org/details/twitterstream (binary json .tar files) (usage tdb)
- Python: Scikit-learn, Gensim, https://github.com/IUNetSci/botometer-python, natural language tool kit, …
- Data storage: MongoDB
- Models: Latent Dirichlet Allocation, Non-negative Matrix Factorization, Latent Semantic Indexing, and **Louvain modeling method**



Example graph clustering of social network communities

THANK YOU