Finding Social Media Trolls: Dynamic Keyword Selection Methods for Rapidly-Evolving Online Debates

Anqi Liu[1], Maya Srikanth[1], Nicholas Adams-Cohen[2], R. Michael Alvarez[3], and Anima Anandkumar[1]
[1]Department of Computing and Mathematical Sciences, California Institute of Technology
[2] Immigration Policy Lab, Stanford University
[3]Division of Humanities and Social Sciences, California Institute of Technology

Alvarez thanks the John Randolph Haynes and Dora Haynes Foundation for supporting his research in this area. Prof. Anandkumar is supported by Bren endowed Chair, faculty awards from Microsoft, Google, and Adobe, DARPA PAI and LwLL grants. Anqi Liu is a PIMCO postdoctoral fellow at Caltech.

Prevention of online harassment requires rapid detection of harassing and offensive social media posts, and current data collection methods rely on **outdated** keywords.
Can we propose a **dynamic, interpretable, and efficient** method of tracking fast-changing topics in a social media debate to aid the detection of online abuse?

Data Sources

Twitter #MeToo: social media movement against sexual harassment.

Reddit RedPill: toxic, misogynistic subreddit that has now been quarantined.

Wikipedia corpus

Keyword Selection Techniques

GloVE: Global Vectors for Word Representation

(Pennington, Socher, and Manning, 2014)

Cosine Similarity Metric

- Indicates linguistic and semantic similarity between words
- Can be used to rank *most similar* keywords to use as the "next set" of query parameters for data collection
- Sheds light on the differences in discussion across various social media platforms

K-Means Clustering

- Clustering on word vectors is useful in topic detection

Ongoing and future explorations:

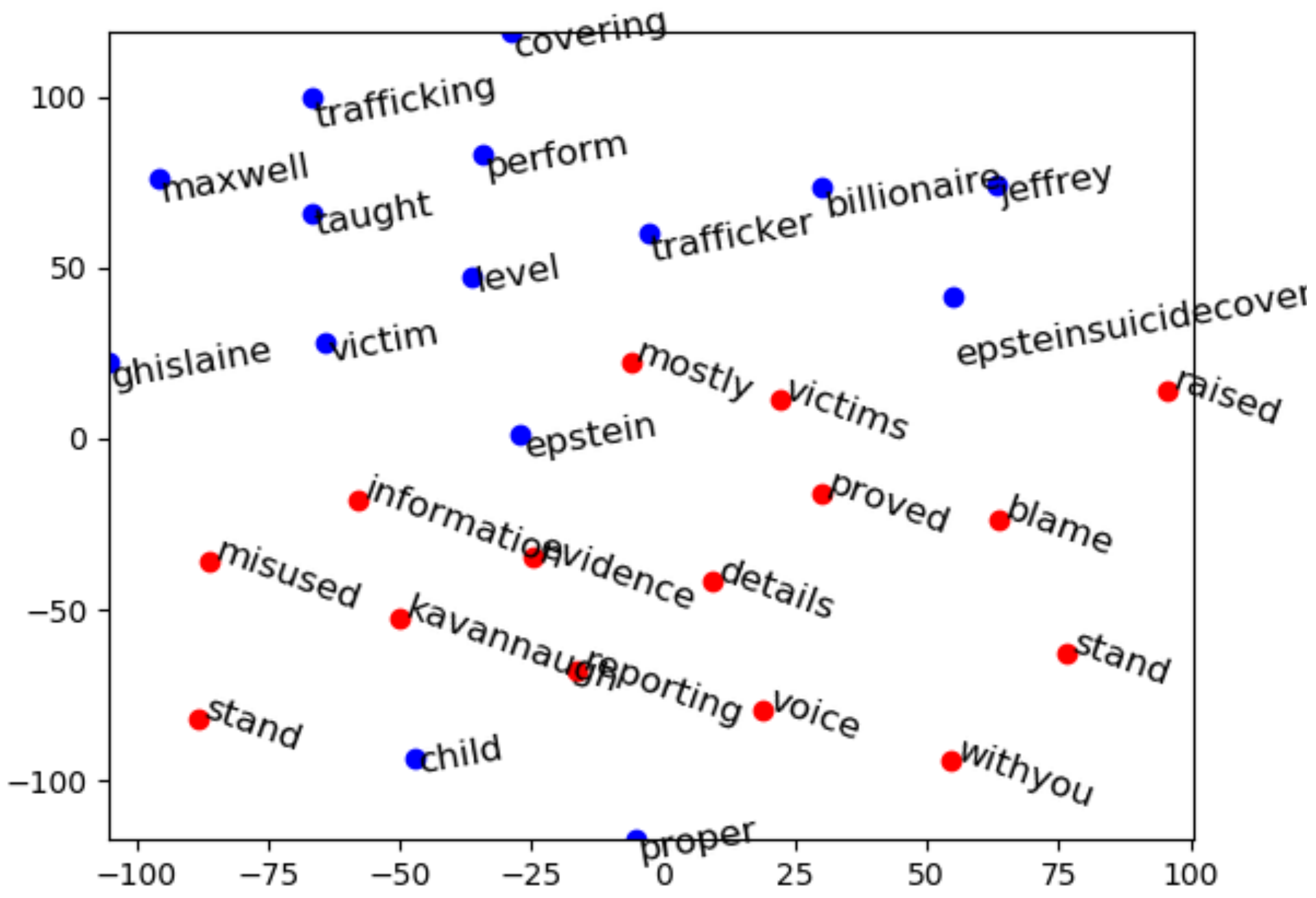
- Devise a method to **rank** the value of new keywords?
- Propose metric to determine how keyword selection procedure is performing?
- Analyze a social media network and detect trolling or abusive interactions ?

Code available at:
https://github.com/mayasrikanth/TwitterStudiesCode.git

Preliminary Results

Topic Detection: K-Means Clustering on GloVE

Figure 1. t-SNE 2D visualization of 2 clusters following K-means clustering on #MeToo data (K = 100). Blue dots are "Epstein" cluster, red dots are "Kavanaugh" cluster.



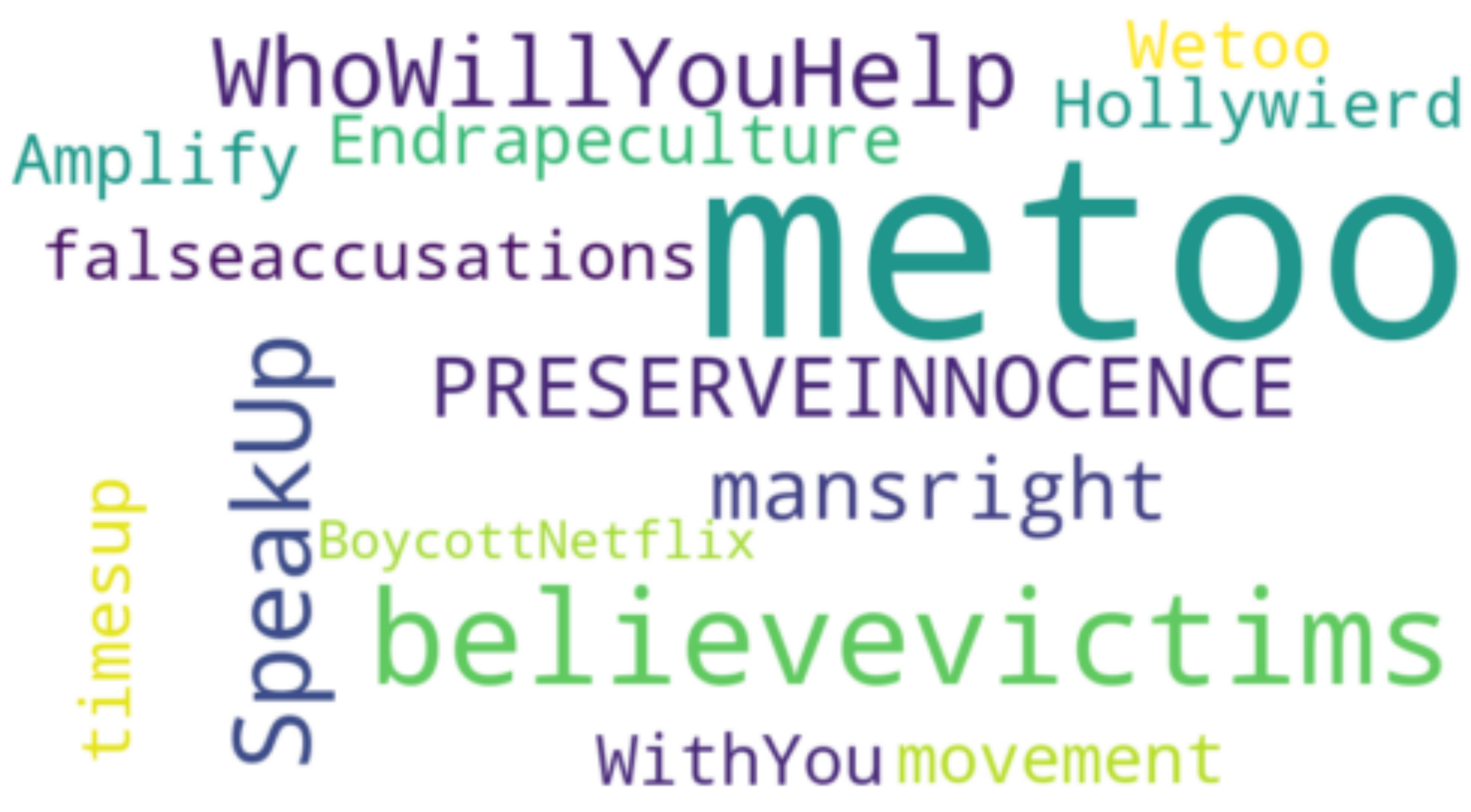
Domain Shift:

Domains	"Female"	"Male"
Wikipedia	adult young woman teenage girl individual age child older	female adult woman girls individual older young age child
#MeToo	companies founded startups desire oppressor employee victims capitalist	venture leader female committed dominated junior referred day
RedPill	sexual negative sexuality intercourse self physical dialogue respective	alpha female plight attraction beta equivalent sexual value emotional plight

Figure 2. Topmost similar words to "female" and "male" across various domains.

New Keywords: Most similar words to "MeToo"

Figure 3. Word cloud containing words with closest cosine distance to "MeToo". Larger font corresponds with more proximity to MeToo.


WhoWillYouHelp
Wetoo
Hollywierd
Amplify
Endrapeculture
falseaccusations
metoo
SpeakUp
PRESERVEINNOCENCE
timesup
mansright
BoycottNetflix
believevictims
WithYou
movement