CASOS Summer Institute 2016

# Extracting Identities from Tweets

**Kenneth Joseph**
kjoseph@cs.cmu.edu

**Prof. Kathleen M. Carley**
kathleen.carley@cs.cmu.edu

## Overview

In this work we developed a method to extract *identities*, words and phrases we use to label other people, from tweets.

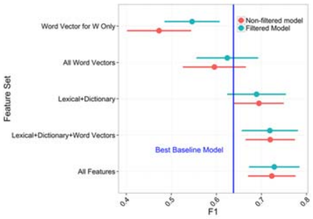
An example is below; the method tags each word in a tweet as being either **I**niside an identity label or **O**utside an identity label. It would extract the identities "teacher", "student" and "police officer" from the tweet below

| The | teacher | told the | student | to draw a | police officer |
|---|---|---|---|---|---|
| O | **I** | O O | **I** | O O O | **I** I |

## Evaluation

We evaluated our model against a baseline model that used a large dictionary to search for identities in tweets

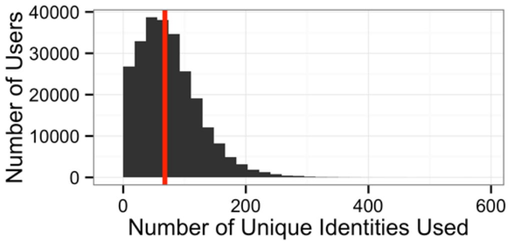
**Our model improves by 33% over the baseline mode.**



## Case Study

We ran the model on over 750M tweets from 250K geotagged Twitter users who were actively involved in discussion of the Eric Garner and Michael Brown tragedies

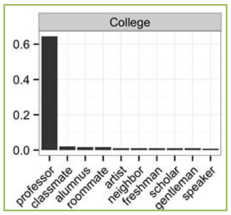
We then performed a case study on this data to show what the identities expressed in the data could tell us about the users within it and narrative they expressed.

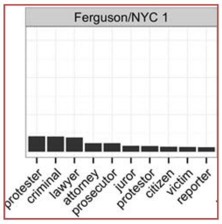


We found that the median user expressed **68(!)** unique identity labels in their tweets

## Clustering Identities

In our case study, we used latent Dirichlet allocation (LDA) to cluster identities based on how often they were expressed by users. In other words, users were our "documents" and identities were our "words"

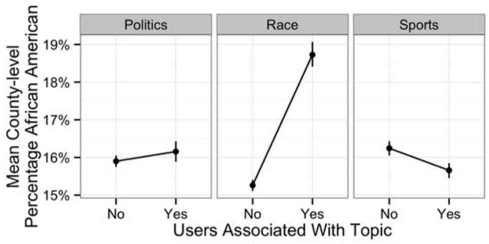




These two example topics (y-axis is weight in posterior) show that results **confirm** and **extend** existing sociological notions of how identities cluster into "institutions".

## Mapping to "Offline" Contexts

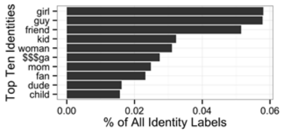
We took three identity clusters (politics, race and sports and compared whether or not users ever expressed identities in these clusters to the mean county-level percentage of African Americans in the county they tweeted from most often



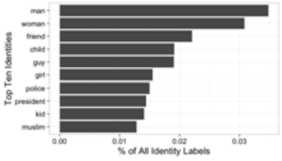
Results suggested that identity usage mapped to offline context

## Other datasets

We also completed a case study on data relevant to the Arab Spring. Results were consistent but different identities were important in the two datasets



**Garner/Brown top identities**



**Arab Spring top identities**

**This work is from the Computational Analysis of Social and Organizational Systems (CASOS) group of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). Support was provided, in part, by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning and a MINERVA N000141310835 on state stability. The views and proposal contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. government.**

**www.casos.cs.cmu.edu**