

Beyond Binary Labels: Political Ideology Prediction of Twitter Users

Daniel Preoțiu-Pietro

Joint work with Ye Liu (NUS), Daniel J Hopkins (Political Science), Lyle
Ungar (CS)

2 August 2017

Motivation

User attribute prediction from text is successful:

- ▶ **Age** (Rao et al. 2010 ACL)
- ▶ **Gender** (Burger et al. 2011 EMNLP)
- ▶ **Location** (Eisenstein et al. 2010 EMNLP)
- ▶ **Personality** (Schwartz et al. 2013 PLoS One)
- ▶ **Impact** (Lampos et al. 2014 EACL)
- ▶ **Political Orientation** (Volkova et al. 2014 ACL)
- ▶ **Mental Illness** (Coppersmith et al. 2014 ACL)
- ▶ **Occupation** (Preoțiuc-Pietro et al. 2015 ACL)
- ▶ **Income** (Preoțiuc-Pietro et al. 2015 PLoS One)

... and useful in many applications.

Political Ideology & Text

Hypothesis:

Political ideology of a user is disclosed through language use

- ▶ partisan political mentions or issues

[@realDonaldTrump](#) your program last night was top notch! You Sir, are a class act! God bless our Vets [#MakeAmericaGreatAgain](#) [#Trump2016](#)

11:46 PM - 29 Jan 2016



- ▶ cultural differences

Disappointed today. Either I trust God to "have this" or I don't. I truly do, but still disappointed.

7:58 PM - 15 Jun 2015



Political Ideology & Text

Previous CS/NLP research used data sets with user labels identified through:

1. User descriptions



H1 Users are far more likely to be politically engaged

Political Ideology & Text

2. Partisan Hashtags



H2 The prediction problem was so far over-simplified

Political Ideology & Text

3. Lists of Conservative/Liberal users

Right-Leaning Tweets

A public list by [Slate](#)

What top conservatives are saying.

Members **50** Subscribers **415**

[Subscribe](#)





Tweets >

List members >

List subscribers >

More lists by [@Slate](#) · [View all](#)

List members

-  [\[Redacted\]](#) Editor at large [@WeeklyStandard](#) [Follow](#)
-  [\[Redacted\]](#) Writer, political consultant, obsessive but lousy endurance sport junkie. Partner, [@Strat_Media](#), Daily Beast columnist. Author of 7 books. [Follow](#)
-  [\[Redacted\]](#) Editor of the Washington Free Beacon, Contributing editor to The Weekly Standard, author [Follow](#)
-  [\[Redacted\]](#) National Political Reporter [@politico](#). Formerly [@NRO](#) Washington Editor, Fox News producer / gchat: [Follow](#)

H3 Neutral users

Political Ideology & Text

4. Followers of partisan accounts

The image shows the Twitter profile page for President Donald J. Trump. At the top, there is a blue header with the text "PRESIDENT DONALD J. TRUMP" and a red circular profile picture featuring the Republican elephant logo. Below the header, the following statistics are displayed: Tweets (25.8K), Following (1,707), Followers (1.44M), Likes (276), Lists (4), and Moments (8). A red "Follow" button is visible on the right. The main content area shows a grid of suggested accounts, each with a profile picture, a "Follow" button, and a brief bio. One bio reads: "Marketing diva. Mac fanatic. food-obsessed Philly girl getting by in Southern California." The left sidebar shows the account name "GOP" (@GOP), the description "Updates from the Republican National Committee", the location "Washington, DC", the website "gop.com", and the join date "Joined December 2007". There is also a red "Tweet to GOP" button and a link to "1,369 Photos and videos".

H4 Differences in language use exist between moderate and extreme users

- ▶ Political ideology
 - ▶ specific of country and culture
 - ▶ our use case is US politics (similar to **all** previous work)
 - ▶ the major US ideology spectrum is Conservative – Liberal
 - ▶ seven point scale



We collect a new data set:

- ▶ 3.938 users (4.8M tweets)
- ▶ public Twitter handle with >100 posts

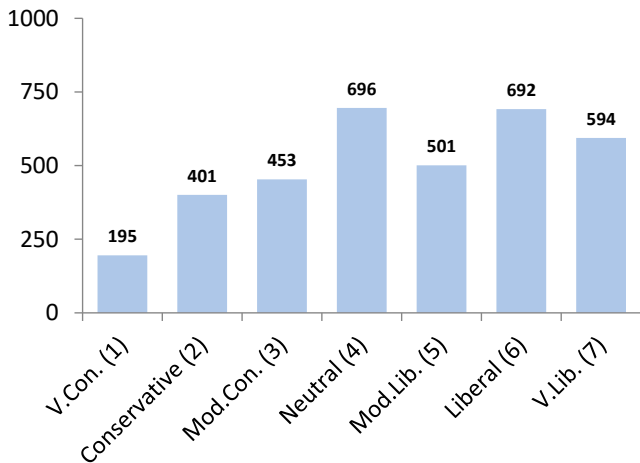
Political ideology is reported through an online survey

- ▶ only way to obtain unbiased ground truth labels (Flekova et al. 2016 ACL, Carpenter et al. 2016 SPPS)
- ▶ additionally reported age, gender and other demographics

- ▶ Data available at `preotiuc.ro`
 - ▶ full data for research purposes
 - ▶ aggregate for replicability
- ▶ Twitter Developer Agreement & Policy VII.A4

“**Twitter Content**, and information derived from Twitter Content, **may not be used** by, or knowingly displayed, distributed, or otherwise made available to any entity **to target**, segment, or profile individuals based on [...] **political affiliation** or beliefs”
- ▶ Study approved by the Internal Review Board (IRB) of the University of Pennsylvania

Class Distribution



For comparison to previous work, we collect a data set:

- ▶ 13,651 users (25.5M tweets)
- ▶ follow liberal/conservative politicians on Twitter

Hypotheses

H1 Previous studies used users far more likely to be politically engaged

H2 The prediction problem was so far over-simplified

H3 Neutral users can be identified

H4 Differences in language use exist between moderate and extreme users

Engagement

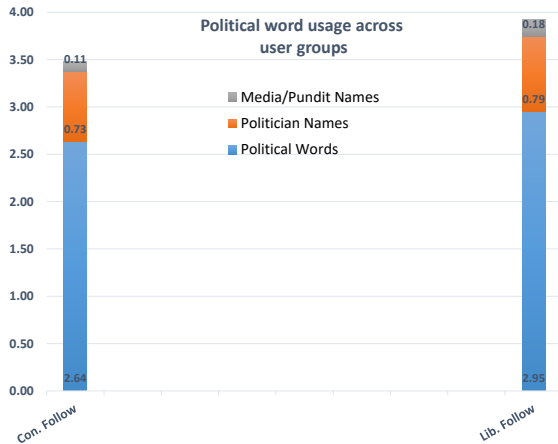
H1 Previous studies used users far more likely to be politically engaged

Manually coded:

- ▶ Political words (234)
- ▶ Political NEs: mentions of politician proper names (39)
- ▶ Media NEs: mentions of political media sources and pundits (20)

Engagement

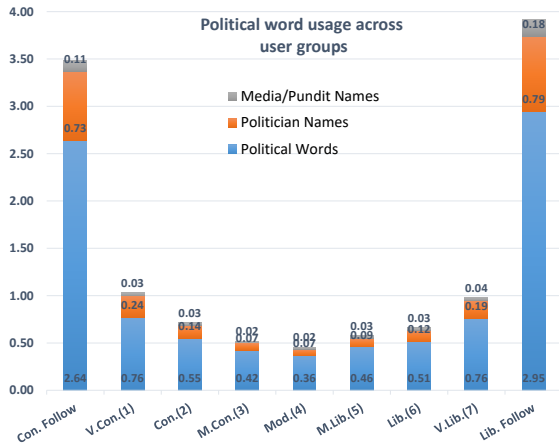
Data set obtained using previous methods



Average percentage of political word usage

Engagement

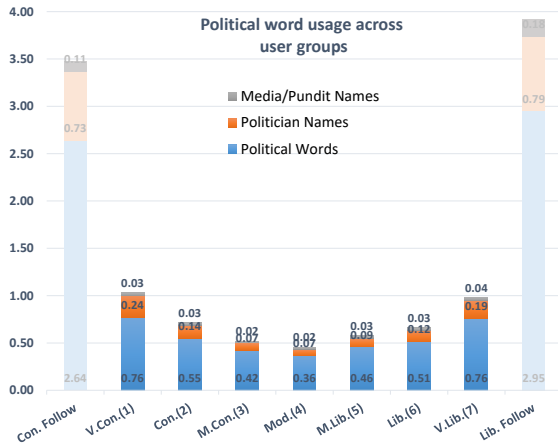
Our data set



Average percentage of political word usage

Engagement

Our data set



Average percentage of political word usage

Engagement

Take aways:

- ▶ 3x more political terms for automatically identified users compared to the highest survey-based scores
- ▶ almost perfectly symmetrical U-shape across all three types of political terms
- ▶ The difference between 1-2/6-7 is larger than 2-3/5-6

Hypotheses

H1 Previous studies used users far more likely to be politically engaged

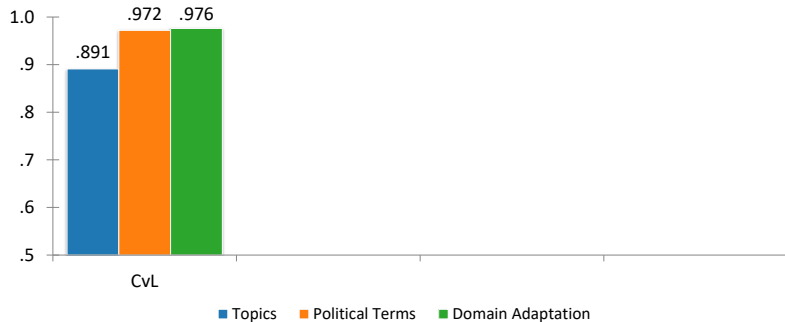
H2 The prediction problem was so far over-simplified

H3 Neutral users can be identified

H4 Differences in language use exist between moderate and extreme users

Over-simplification

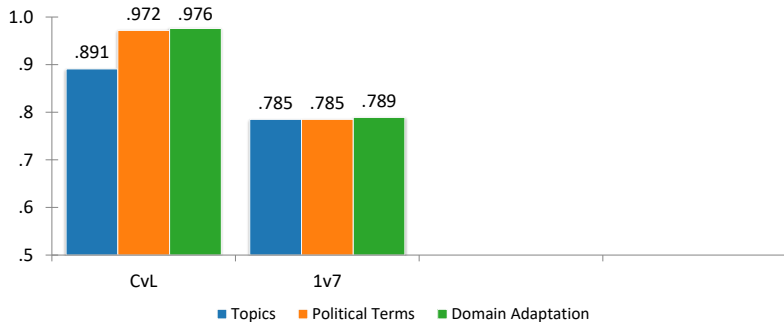
H2 The prediction problem was so far over-simplified



ROC AUC, Logistic Regression, 10-fold cross-validation

Over-simplification

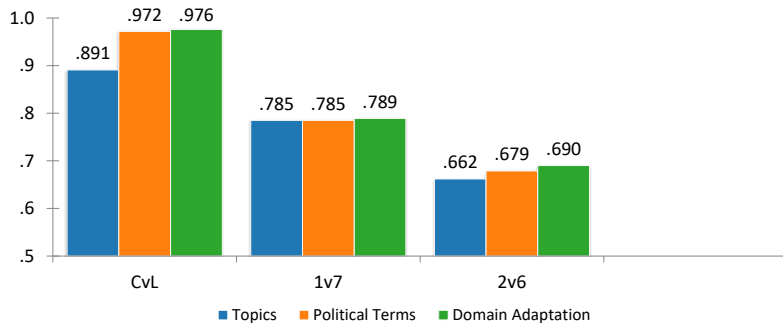
H2 The prediction problem was so far over-simplified



ROC AUC, Logistic Regression, 10-fold cross-validation

Over-simplification

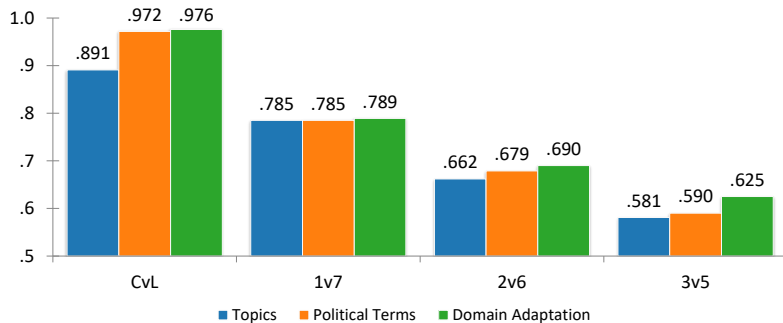
H2 The prediction problem was so far over-simplified



ROC AUC, Logistic Regression, 10-fold cross-validation

Over-simplification

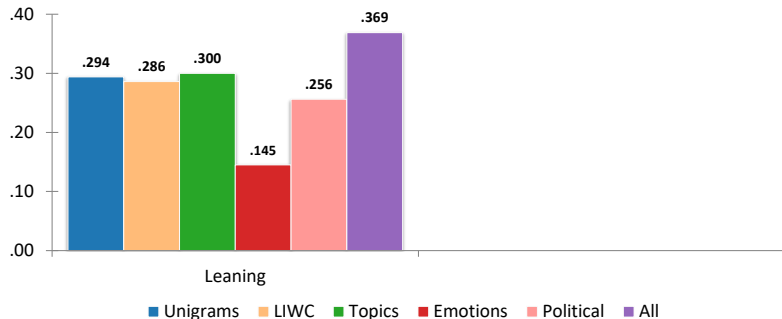
H2 The prediction problem was so far over-simplified



ROC AUC, Logistic Regression, 10 fold-cross validation

Over-simplification

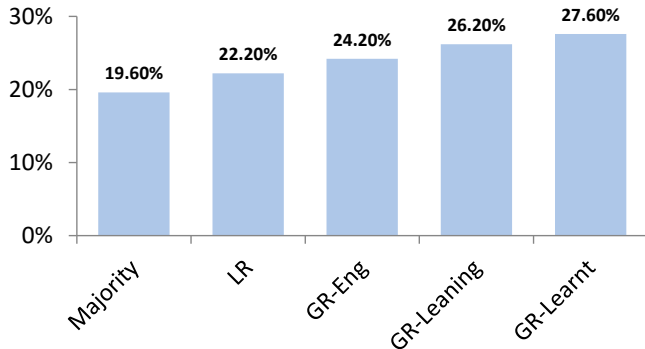
Predicting continuous political leaning (1 – 7)



Pearson R between predictions and true labels, Linear Regression, 10-fold cross-validation

Over-simplification

Seven-class classification



Accuracy, 10-fold cross-validation

GR – Logistic regression with Group Lasso regularisation

Hypotheses

H1 Previous studies used users far more likely to be politically engaged

H2 The prediction problem was so far over-simplified

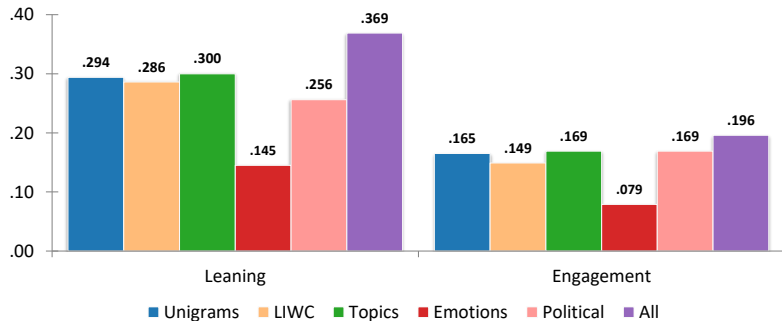
H3 Neutral users can be identified

H4 Differences in language use exist between moderate and extreme users

Political Engagement

H3a There is a separate dimension of political engagement

Combine the classes into a scale: 4 – 3&5 – 2&6 – 1&7



Pearson R between predictions and true labels, Linear Regression, 10 fold-cross validation

Hypotheses

H1 Previous studies used users far more likely to be politically engaged

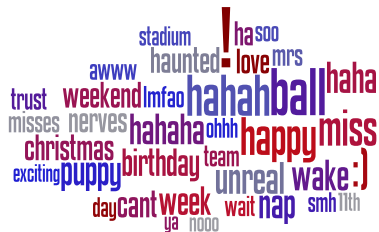
H2 The prediction problem was so far over-simplified

H3 Neutral users can be identified

H4 Differences in language use exist between moderate and extreme users

Moderate Users

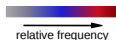
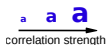
H4 Differences between moderate and extreme users



Words associated with moderate liberals (5 and 6).



Words associated with extreme liberals (7).



Correlations are age and gender controlled

Take Aways

- ▶ User-level trait acquisition methodologies can generate non-representative samples
- ▶ Political ideology:
 - ▶ Goes beyond binary classes
 - ▶ The problem was to date over-simplified
 - ▶ New data set available for research
 - ▶ New model to identify political leaning and engagement

Questions?

www.preotiuc.ro

wwbp.org