Predicting users' political support from their Reddit comment history

Aaron Acosta [ateam91], Silviana Ciurea-Ilcus [smci], and Michal Wegrzynski [michalw]

Introduction

How large is the conservative-liberal divide on the Internet and does it extend beyond politics? In this project we explore **the Reddit usage patterns** of users who are either Donald Trump, Hillary Clinton, or Bernie Sanders supporters. **We attempt to predict their political inclination based on their activity in the non-political parts of the website**.

Dataset

Reddit is a website consisting of forums named 'subreddits', organized around various topics. We assume that users whose comments are rated highly in subreddits supporting a particular political candidate are supporters of that candidate.

Dataset: 100,000 comments made in politically-neutral subreddits by the users associated with the top 10,000 comments on each of *r/The_Donald, r/HillaryClinton*, and *r/SandersForPresident*, for a total of 300,000 comments

Processing: We stemmed the words using Porter Stemmer 2, for a 0.3-1.7% increase in accuracy. We did not remove stop words because previous studies have found a correlation between political leanings and the use of stop words like 'I', 'am' vs. 'we', 'are' etc.

Training/testing examples: each selected user's comment history from non-political subreddits, labelled as a Sanders, Trump, or Clinton supporter, based on them being active users with highly upvoted comments in the candidate-specific subreddits.

Feature selection

Previous research found correlations between political leaning and emotions expressed online, use of swear words, first-person singular vs plural words, attitude towards new information, choice of hobbies, taste in art etc. We thus selected as features:

a. Term frequency-inverse document frequency of each 1-gram (stemmed), 2-, 3- 4-gram

b. Positive, negative, and neutral sentiment score/user (%)

c. Positive, negative, and neutral sentiment score per word /user (%), to capture the context in which each word is used

d. Number of posts in each subreddit/ user

Results and analysis

Our five top performing classifiers were Logistic Regression, Linear SVM, Random Forest (with 100 estimators and √n features), Voting Classifier (with the three aforementioned classifiers with weights 1), and Multinomial Naive Bayes. We obtained these results by using stemmed 1-grams, which performed better than 2-, 3-, or 4-grams.

	Log Reg		**Linear SVM**		**Random Forest**		**Voting Classifier**		**Naive Bayes without priors**	
num. features	train acc.	test acc.	train acc.	test acc.	train acc.	test acc.	train acc.	test acc.	train acc.	test acc.
a.	0.760	0.480	0.719	0.471	0.948	0.482	0.816	0.491	0.847	0.421
a+b	0.871	0.482	0.869	0.473	0.955	0.487	0.890	0.494	0.911	0.501
a+b+c	0.913	0.490	0.943	0.478	0.993	0.501	0.945	0.546	0.931	0.554
a+b+c+d	0.989	0.513	0.974	0.490	0.995	0.512	0.951	0.556	0.943	0.567

- The top 3 subreddits by number of comments posted were the same for the three groups. Despite some overlap in the top 20 subreddits, **we found that users that supported different candidates tended to be active in different nonpolitical forums**. This supports the correlation between areas of interest and political leanings proposed by previous studies.
- The performance for each candidate varied by algorithm. Most classifiers tended to have more true and false positives for Sanders. Distinguishing between Clinton and Trump/Sanders was easier than between Trump and Sanders. Hillary Clinton supporters proved to be hard to detect, possibly because they seem to be less polarized than either Sanders or Trump supporters.
- The average neutral and positive sentiment per user for supporters of either candidate was similar, however Trump supporters' comments were **on average 10% more negative** than other users'.

	Clinton supporters	**Sanders supporters**	**Trump Supporters**
SecretSubreddit	**- (0)**	**- (0)**	**+ (2522)**
apple	**+ (1170)**	**+ (1039)**	**- (418)**
Science	**+ (1612)**	**- (349)**	**- (310)**
programming	**+ (815)**	**- (198)**	**- (50)**
TexasRangers	**- (249)**	**+ (755)**	**- (14)**
MMA	**- (226)**	**- (252)**	**+ (916)**
CringeAnarchy	**- (48)**	**- (116)**	**+ (1337)**
KotakuInAction	**- (49)**	**- (385)**	**+ (1633)**
Overwatch	**- (373)**	**- (650)**	**+ (750)**
leagueoflegends	**- (668)**	**+ (884)**	**- (489)**
Showerthoughts	**- (410)**	**+ (807)**	**- (483)**
BigBrother	**+ (856)**	**- (289)**	**- (162)**
WTF	**+ (932)**	**- (498)**	**- (526)**

- We obtained an average of **only 3-7% higher accuracy** when classifying the same users' comments from **political subreddits,** excluding candidate-specific ones.
- **In conclusion**, **the Clinton-Sanders-Trump divide in non-political content is strong,** and surprisingly only a few percentage points lower than the one in political content.

References

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences*, *37*(3), 297–307.

Sylwester, K., & Purver, M. (2015). Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLoS ONE*, *10*(9), e0137422.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.