Inferring Audience Partisanship for YouTube Videos

Ingmar Weber
Qatar Computing Research Institute
ingmarweber@acm.org

Venkata Rama Kiran Garimella
Qatar Computing Research Institute
vgarimella@qf.org.qa

Erik Borra
University of Amsterdam
e.k.borra@uva.nl

ABSTRACT

Political campaigning and the corresponding advertisement money are increasingly moving online. Some analysts claim that the U.S. elections were partly won through a smart use of (i) targeted advertising and (ii) social media. But what type of information do politicized users consume online? And, the other way around, for a given content, e.g. a YouTube video, is it possible to predict its political audience? To address this latter question, we present a large scale study of anonymous YouTube video consumption of politicized users, where political orientation is derived from visits to "beacon pages", namely, political partisan blogs. Though our techniques are relevant for targeted political advertising, we believe that our findings are also of a wider interest.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

YouTube; political polarization; audience prediction; partisan blogs

1. INTRODUCTION

In the run-up to the U.S. presidential elections 2012, political campaigning happened increasingly through social media and social networks. Such campaigning goes beyond maintaining a Facebook page or a Twitter account; it tries to target individual users who are likely to be responsive to the political message. For a given type of apolitical content, say, a YouTube video of a cute kitten, is it possible to predict the political leaning of its audience?

We use a large data set of anonymous browsing behavior to analyze "the political YouTube user". We infer a user's likely political orientation through visited political blogs. Using such a set of labeled users, we study the problem of predicting which audience, in terms of political orientation, watches a given YouTube video. We do this not with political campaign videos in mind, but rather to unearth relationships with, say, the music genre of a music video. To the best of our knowledge, this is the first study of online browsing and video consumption from the angle of political orientation.

2. RELATED WORK

The importance of targeting political online ads to "the right users" has been recognized by politicians, in particular in the United States [6, 12, 5, 13] and parties have undergone painstaking efforts to collect huge amounts of data "by hand" [8], gathering details about individual voters. In our work, we take an algorithmic approach to data collection. The problem of classifying YouTube videos using tags according to political leaning of the viewership is related to political text classification in general. This problem has been studied before in the context of party programs and labeled text sets [9, 14], news articles [4] and hyper-text documents [3] and users [11, 15]. Though certain elements such as using tokens as features are shared with these works, the data set and overall setting are completely different.

3. DATA SET

The anonymous browsing data was collected through a toolbar of a large internet company. For users who give their explicit consent, this toolbar logs all of the browser's page views, including redirect page views caused when, e.g., clicking a search result on a search engine. For secure HTTPs connections, no dynamic URL parameters and only the static URL prefix is stored. The user-based sample we collected contained all logged page views for 13 months (March 2011 - April 2012) for millions of distinct users. For simplicity, we always equate an anonymous toolbar ID with an individual user. We removed users with less than 1,000 or more than 2,000,000 page views as the former are likely to use another "main" computer and the latter are likely internet cafes, virus infected computers or other abnormal cases. We also removed users who, based on their IP address were located outside the United States of America. For all page views to `youtube.com/watch?v=` we extracted the corresponding video id. Using YouTube's API we obtained the video's title, tags and category.

As the political leaning of blogs read online largely aligns with the leaning of the user reading them [10], we obtained the handcrafted lists of blogs annotated with a political leaning from [1] and the Wonkosphere Blog Directory[1]. Re-directs were corrected and abandoned blogs removed. In the end 1,099 blogs (644 right, 387 left, 68 center) remained. We then used these blogs to label users as "L" if (i) all of their political page views[2] were on left leaning blogs and (ii) these views occurred on at least two distinct sites. The labeling as "R" was analogous. We labeled a user as "C" if no more than 70% of his political page views on political blogs were on sites of either a left or right leaning. All other users were considered unlabeled. The final user counts were 18.3k, 1.1k, 4.0k, for L, C, and R respectively. Users without a ground truth label were dropped from further analysis.

[1] `http://wonkosphere.com/directory.htm`

[2] We trained a token-based URL classifier to tell political from apolitical blog pages as `huffingtonpost.com` and other sites also host non-political, celebrity related content.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

4. PREDICTING AUDIENCE COMPOSITION

Among brand advertisers, online video advertising is one of the fastest growing segments, together with mobile and social media advertising. In the political domain the corresponding targeting problem translates to: given a video, can we predict the political leaning of its online audience? Here, we limit our scope to YouTube, the biggest online video platform. For each video watched directly on www.youtube.com/watch, rather than re-branded channels, we counted how many labeled users watched it. Note that embedded video views via `barackobama.com/videos/` or `mittromney.com/videos/` are *not* contained in this set. For videos with at least 50 distinct users we computed the fraction of L, C and R users. There were 10,659 such videos. Macro-averaged across all videos, political or not, the fractions were 73.6%L, 17.1%C and 9.3%R, differing considerably from the overall user distribution. The videos with highest L, C and R fractions were `bit.ly/AexUVS`, `bit.ly/nH2UI` and `bit.ly/xdzHtX` respectively, of which only the last one is actually political. With advertising opportunities in mind, we were more interested in apolitical videos, e.g. political trends concerning music videos. Finding such correlations is similar to the work described in [2] and the targeting of reruns "TV Land" for political advertising [6]. Thus, we removed videos with popular political tags containing "obama", "republican", "politic" and a dozen others. We split the set of apolitical videos into 9,972 training videos and 500 test videos. For the training set we trained three separate linear SVM regression models to predict each of the L, C and R fractions given only its category as well as its title tokens and tags, ignoring tokens appearing in less than 10 videos. To better understand the prediction performance, we also evaluated a simple baseline model which always predicts the average value for each of L, C and R.

		baseline		SVM	
Category	*n*	abs	rmse	abs	rmse
All	500	5.72	7.67	5.20	7.14
Music	252	5.81	7.46	5.10	6.75
Entertainm.	73	5.08	6.63	5.07	6.50
Comedy	30	4.21	5.46	4.56	5.61
People	22	6.84	8.73	6.41	8.83
Film	30	6.51	9.85	7.04	10.67
Education	11	5.74	7.32	4.26	5.69
Animals	11	3.39	4.22	3.90	4.86
News	10	6.76	8.88	7.04	8.58
Shows	10	2.91	3.70	2.93	3.76

(a) Performance, both absolute error and RMSE, for the linear SVM regression models for the L fraction for the biggest categories. All values are for the same, category-agnostic classifier with the category as one of its features.

top "L"	top "C"	top "R"
tit:beyonce	tag:country	tit:mp4
tag:r&b	tag:rock	tag:marine
tag:soul	tag:nashville	tag:land
tag:whitney	tag:car	tag:marines
tag:viral	cat:autos	tit:commercial
tit:temptations	tag:jason	tag:navy
tit:houston	tag:ray	tag:every
tit:spongebob	tit:live	tag:were
tit:mating	tit:road	tag:israel
tag:boyz	tit:mom	tit:haggard

(b) The 10 strongest positive features from each of the three linear SVM models. R&B music appears indicative of a left leaning, while interest in military videos indicates a right leaning.

Table 1a summarizes the performance of our L model, with the other two models performing comparably. Overall the reductions in root mean squared error (RMSE), compared to the baseline, are L: 7.67 → 7.14, C: 5.28 → 5.07, R: 5.89 → 5.55, all fairly small. This bad news for the classifier is, arguably, good news for society as it indicates that general YouTube videos are not polarized enough to be easily told apart. However, there are peculiar differences across video categories, with the "Education" category having a relative improvement of 22% in RMSE over the baseline for L, with 13% for C and 8% for R. Apart from the overall regression performance, we analyzed which features were indicative of a given leaning. Table 1b shows the 15 strongest positive features for each of the linear SVM models. A "tag:" prefix indicates a tag, "tit:" a title and "cat:" a category token.

We conclude by noting that videos matching the string "kitten" follow the overall L-C-R background distribution for videos. Reassuringly, kittens appear to be apolitical and universally adored. However, adding features may make such seemingly innocuous information reveal sensitive personal attributes too [7].

5. REFERENCES

[1] Y. Benkler and A. Shaw. A tale of two blogospheres: Discursive practices on the left and right. *American Behavioral Scientist*, 56(4):459–487, 2012.

[2] N. L. Center and Z. International. The zogby/lear center survey on politics and entertainment, 2008.

[3] M. Efron. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. In *CIKM*, pages 390–398, 2004.

[4] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. König. Blews: Using blogs to provide context for news articles. In *ICWSM*, pages 60–67, 2008.

[5] K. Hart. The 2012 tech primary. *Politico*, January 2012.

[6] S. Issenberg. The definitive story of how president obama mined voter data to win a second term. *MIT Technology Review*, 2013.

[7] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 2013.

[8] D. Kreiss. Yes we can (profile you) – a brief primer on campaigns and political data. *Stanford Law Review*, February 2012.

[9] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *APSR*, 97(02):311–331, 2003.

[10] E. Lawrence, J. Sides, and H. Farrell. Self-Segregation or deliberation? blog readership, participation, and polarization in american politics. *POP*, 8(01):141–157, 2010.

[11] R. Malouf and T. Mullen. Taking sides: user classification for informal online political discourse. *Internet Research*, 18:177–190, 2008.

[12] J. W. Peters. As viewing habits change, political ads switch screens. *N.Y. Times*, April 2012.

[13] E. Schultheis. Romney counters obama's barnard speech with targeted web ads. *Politico*, May 2012.

[14] J. Slapin and S. Proksch. A scaling model for estimating time-series party positions from texts. *AJPS*, 52(3):705–722, 2008.

[15] D. X. Zhou, P. Resnick, and Q. Mei. Classifying the political leaning of news articles and users from user votes. In *ICWSM*, 2011.