Classifying Political Orientation on Twitter: It's Not Easy!

Raviv Cohen and **Derek Ruths**
School of Computer Science
McGill University
raviv.cohen@mail.mcgill.ca, derek.ruths@mcgill.ca

Abstract

Numerous papers have reported great success at inferring the political orientation of Twitter users. This paper has some unfortunate news to deliver: while past work has been sound and often methodologically novel, we have discovered that reported accuracies have been systemically overoptimistic due to the way in which validation datasets have been collected, reporting accuracy levels nearly 30% higher than can be expected in populations of general Twitter users.

Using careful and novel data collection and annotation techniques, we collected three different sets of Twitter users, each characterizing a different degree of political engagement on Twitter — from politicians (highly politically vocal) to "normal" users (those who rarely discuss politics). Applying standard techniques for inferring political orientation, we show that methods which previously reported greater than 90% inference accuracy, actually achieve barely 65% accuracy on normal users. We also show that classifiers cannot be used to classify users outside the narrow range of political orientation on which they were trained.

While a sobering finding, our results quantify and call attention to overlooked problems in the latent attribute inference literature that, no doubt, extend beyond political orientation inference: the way in which datasets are assembled and the transferability of classifiers.

Introduction

Much of the promise of online social media studies, analytics, and commerce depends on knowing various attributes of individual and groups of users. For a variety of reasons, few intrinsic attributes of individuals are explicitly revealed in their user account profiles. As a result, latent attribute inference, the computational discovery of "hidden" attributes, has become a topic of significant interest among social media researchers and to industries built around utilizing and monetizing online social content. Most existing work has focused around the Twitter platform due to the widespread adoption of the service and the tendency of its users to keep their accounts public.

Existing work on latent attribute inference in the Twitter context has made progress on a number of attributes including gender, age, education, political orientation, and even coffee preferences (Zamal, Liu, and Ruths 2012; Conover et al. 2011b; 2011a; Rao and Yarowsky 2010; Pennacchiotti and Popescu 2011; Wong et al. 2013; Liu and Ruths 2013; Golbeck and Hansen 2011; Burger, Henderson, and Zarrella 2011). In general, inference algorithms have achieved accuracy rates in the range of 85%, but have struggled to improve beyond this point. To date, the great success story of this area is political orientation inference for which a number of papers have boasted inference accuracy reaching and even surpassing 90% (Conover et al. 2011b; Zamal, Liu, and Ruths 2012).

By any reasonable measure, the existing work on political orientation is sound and represents a sincere and successful effort to advance the technology of latent attribute inference. Furthermore, a number of the works have yielded notable insights into the nature of political orientation in online environments (Conover et al. 2011b; 2011a). In this paper, we examine the question of whether existing political orientation inference systems actual perform as well as reported on the general Twitter population. Our findings indicate that, without exception, they do not, even when the general population consider is restricted only to those who discuss politics (since inferring the political orientation of a user who never speaks about politics is, certainly, very hard if not impossible).

We consider this an important question and finding for two reasons. Foremost, nearly all applications of latent attribute inference involve its use on large populations of unknown users. As a result, quantifying its performance on the general Twitter population is arguably the best way of evaluating its practical utility. Second, the existing literature on this topic reports its accuracy in inferring political orientation without qualification or caveats (author's note: including our own past work on the topic (Zamal, Liu, and Ruths 2012)). To the reader uninitiated in latent attribute inference, these performance claims can easily be taken to be an assertion about the performance of the system under general conditions. In fact, we suspect that most authors of these works had similar assumptions in mind (author's note: we did!). Regardless of intentions, as we will show, past systems were not evaluated under general conditions and, therefore, the performance reported is not representative of the general use case for the systems.

Far from a harsh critique of existing work, our intention is to establish precisely *how good* political orientation inference systems actually are and, in doing so, set the stage for further progress on the problem. It is also noteworthy that, in the course of this study, we will identify issues and techniques that may be relevant to research on the inference of other attributes, hopefully improving research in these areas as well.

The fundamental issues that we address in this study concern (1) the datasets that were used in prior work to evaluate the political orientation inference systems and (2) the transferability of classifiers trained on one dataset to another dataset (Zamal, Liu, and Ruths 2012; Conover et al. 2011b; 2011a; Golbeck and Hansen 2011; Rao and Yarowsky 2010).

To the first issue, without exception, the datasets in prior work consisted of some mix of Twitter accounts belonging to politicians, to people who registered their account under political groups in Twitter directories (e.g. `www.wefollow.com`), and to people who self-reported their political orientation in their Twitter profiles. It was on *these* datasets that the reported accuracies of 95% were obtained. In this study, we constructed three Twitter user datasets: the first consists of the accounts of US politicians, the second of users with self-reported political orientation, and the third of "modest" users who do not declare their political views, but make sufficient mention of politics in tweets such that their political orientation can be deduced by manual inspection. We consider this third group of politically modest users to the most representative of the general Twitter population. Note that extreme care was taking in collecting these datasets in order to ensure that sources of bias could be controlled for. To our knowledge, our collection is the most exacting in the literature. While improvements certainly can be made, we consider this level of attention to data collection to be an important and often-undervalued aspect of successful latent attribute inference.

Running the classifier on the datasets, not surprisingly, we find that the accuracy of the inference systems decreases as visible political engagement decreases. What is remarkable is the degree to which the performance decreases - dropping to 65% for the modest user dataset. We also evaluated the capacity for inference classifiers trained on one dataset to be used on other datasets. We found that classifiers based on politicians, while achieving 91% labeling accuracy on other politicians, only achieved 11% accuracy on politically modest users — further underscoring the dramatic and systemic differences between these sets users and the performance that existing classifiers can achieve on them. An analysis of the datasets which considered lexical variation and topic diversity explained the results obtained.

The second issue concerning the transferability of datasets has not been addressed in the latent attribute inference literature at all. On one hand, we recognize that the set of cases on which a classifier is accurate is limited by the training data with which it was constructed. Nonetheless, from a practical perspective, it is instructive to understand *how* accuracy falls off as the dataset changes. The three datasets we constructed presents the opportunity to evaluate exactly this question.

The results are not encouraging. We find that classifiers trained on any of the datasets are highly *non-transferable* to other datasets, despite the fact that they were collected over the same time period and are labeled in consistent ways. While much more investigation must be done on this matter, we consider this a question that should be attacked more frequently and seriously in future work in this area.

Overall, while the core contributions of this paper may be grim, they highlight a number of important open research questions that, when tackled, will propel the field of latent attribute inference forward. To further support research in these areas, we have released the three datasets used in this study, which constitutes the first open set of datasets for political orientation inference.

Political Orientation Classifiers

The goal of our study was to evaluate the effect of different dataset selection on the performance of political orientation classifiers. As a result, we employed existing classifiers for our study. Here we briefly survey how support vector machines (SVMs), boosted decision trees (BDTs), and latent dirichlet allocation-based (LDAs) methods and strategies that have been used in the past.

Latent Dirichlet Allocation. LDAs are topic models that have been employed, to great effect, as text classifiers in a number of areas (Blei, Ng, and Jordan 2003). Despite their strong performance elsewhere, they have been little-used in the domain of latent attribute inference. In the political orientation inference literature, we know of only one study which used LDAs (Conover et al. 2011b). In this work, the output of an LDA was used as one (of many) features that were fed into a larger SVM classifier. As part of the present study we employed a Labeled LDA as a stand-alone classifier and found it to perform as well as SVMs, suggesting that it might be fruitfully used as primary attribute inference system (Ramage et al. 2009). We consider this a promising direction for future work.

Support Vector Machines and Decision Trees. In the literature, support vector machines have enjoyed the most attention as latent attribute classifiers. Where political orientation is concerned, a number of studies have used SVMs, including the study which achieved the best reported performance to date (95%) (Conover et al. 2011b; Rao and Yarowsky 2010; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011). To our knowledge, only one study has used boosted decision trees and reported similar performance to SVMs (Pennacchiotti and Popescu 2011).

SVMs and BDTs share in common a dependence on the decomposition of users into fixed-length feature vectors. A recent study employed an SVM which incorporated a superset of user features from prior work (Zamal, Liu, and Ruths 2012). The features included: k-top words, k-top stems, k-top co-stems, k-top digrams and trigrams, k-top hashtags, k-top mentions, tweet/retweet/hashtag/link/mention frequencies, and out/in-neighborhood size. It is noteworthy that the *k-top X* features (e.g., k-top hashtags) refers to

collecting the k most discriminating items of that type (e.g. hashtags) for each label (e.g., Republicans and Democrats). Thus, k-top words is actually $2k$ features: k words from Republicans and k words from Democrats. For further details on these features, we refer the reader to the original paper (Zamal, Liu, and Ruths 2012).

For the purposes of this study, we primarily employed an SVM classifier based on (Zamal, Liu, and Ruths 2012). While this method did not achieve the best performance to date (93% vs. 95%), it was benchmarked on a less restrictive dataset than (Conover et al. 2011b) which likely contributed to the small different in performance. Furthermore, because it incorporates nearly all features from work that preceded it, we considered it a more fair representation of all existing systems proposed to date. Following prior work, a radial basis function was used as the SVM kernel. The cost and γ parameters were chosen using a grid search technique. The SVM itself was implemented using the library `libSVM` (Chang and Lin 2011).

We also evaluated the accuracy of a Labeled-LDA-based classifier, following work that shows how the LDA can be used on labeled data (Ramage et al. 2009; Ramage, Dumais, and Liebling 2010). Note that little work in latent attribute inference has used LDA-based classifiers alone. We applied it here primarily as a way of characterizing the lexical diversity of different datasets. We used a Labeled-LDA implementation available in Scala as part of the Stanford Modeling Toolbox[1]. In evaluating the accuracy the LLDA, we found it to be extremely similar to the SVM and, due to space limitations, we do not report results for this classifier except as a means of characterizing the difference in topics present across the three datasets we considered. It is worth noting that the similar accuracy achieved by both the LDA and the SVM methods allayed a concern we had about the SVM's use of only top-discriminating features. The LDA, by construction, employs all words present in the corpus. Thus, if additional accuracy could be gained by incorporating less discriminating words, presumably this would have manifested as significant improvement in the LDA performance. This was not observed under any of the conditions considered in this study.

Construction of Testing Datasets

As mentioned earlier, the goal of this study was to evaluate the extent to which the dataset selection criteria influenced the performance of political orientation classifiers. In particular, our concern was to determine the performance of classifiers on "ordinary" Twitter users.

Our definition of "ordinary" primarily concerned the extent to which users employed political language in tweets. Our intuition was that very few Twitter users generate politically overt tweets. This intuition was proven out by the fraction of randomly sampled users who had even a single tweet with political content. Given this bias towards little, if any, political commentary in Twitter, benchmarks based on politically verbose users would not properly gauge the performance of classifiers on the general population. Of course, the discovery that the political orientation of ordinary users is harder to discern than that of political figures is hardly news. *How much harder* it is, however, is quite important: this is the difference between the problem still being rather easy and the problem of political orientation inference suddenly being largely unsolved. As we have already indicated, our findings suggest the latter to a profound degree.

Table 1: Basic statistics on the different datasets used. Total size of the Figures dataset was limited by the number of federal level politicians; size of the Modest dataset was limited by the number of users that satisfied our stringent conditions - these were culled from a dataset of 10,000 random individuals.

Dataset	Republicans	Democrats	Total
Figures	203	194	397
Active	860	977	1837
Modest	105	157	262
Conover	107	89	196

To conduct this study, we built three datasets which acted as proxies for populations with different degrees of overt political orientation. Each dataset consisted of a set of Twitter users whose political orientation was known with high confidence. The basic statistics for these datasets are shown in Table 1.

Political Figures Dataset (PFD). This dataset was intended to act as a baseline for the study. In many ways it also served to proxy for datasets that were used in other papers since, with the exception of the Conover 2011 dataset described below, we were unable to obtain or recreate datasets described and used in other papers (Conover et al. 2011b). In prior work, the primary way for labeled political orientation datasets to be built was by compiling a list of self-declared Republicans and Democrats. In surveying such lists, we observed that a (super)majority of the Twitter users were, in fact, politicians. To mimic this design, we populated this dataset entirely with state governors, federal-level senators, and federal-level representatives.

We created this dataset by scraping the official websites of the US governors[2], senators[3], and representatives[4]. This provided us a complete list of their individual websites and political orientation. From the individual websites, we obtained their Twitter username and used the Twitter search API to obtain their latest 1000 tweets.

Once this dataset was constructed, we derived two aggregate statistics from it in order to generate the Politically Modest Dataset (described later): the *politically discriminative hashtag set*, H_Δ, and the *politically neutral hashtag set*, H_N. These are mutually exclusive hashtag sets that are a subset of all abundant hashtags, H_A, in the PFD. By *abundant* hashtags, we refer to all those that are used at least

[1] `http://nlp.stanford.edu/software/tmt/tmt-0.4`

[2] `http://www.nga.org/cms/governors/bios`

[3] `http://www.senate.gov`

[4] `http://www.house.gov/representatives`

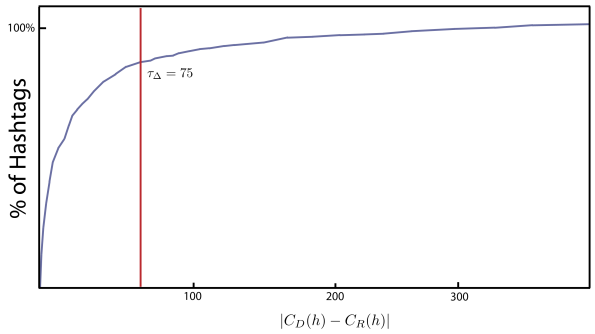

Figure 1: The distribution of hashtags according to their discriminatory value. τ_Δ was selected at the inflection point of the curve.

τ_a times, establishing a minimum popularity a hashtag must achieve before it is considered. The discriminative hashtag set consists of all abundant hashtags used at least τ_Δ more times by one political orientation than another — thus every hashtag in the set is discriminating of one political orientation or the other. The neutral hashtag set consists of all remaining abundant hashtags (those that are not preferentially used by one orientation or another). Formally:

$$H_A = \{h \in H_{PFD} : C_D(h) + C_R(h) > \tau_a\},$$
$$H_\Delta = \{h \in H_A : |C_D(h) - C_R(h)| \geq \tau_\Delta\},$$
$$H_N = \{h \in H_A : |C_D(h) - C_R(h)| \leq \tau_\Delta\}.$$

where H_{PFD} is the set of all hashtags used by tweets in the PFD and $C_R(h)$ ($C_D(h)$) is the number of times the hashtag h is used in Republican-labeled (Democrat-labeled) user tweets. For the remainder of the study, we used $\tau_a = 50$, though the choice does not seem to matter as long as it is not very large, $\tau_a > 500$, or very small, $\tau_a < 10$. We chose τ_Δ by evaluating the fraction of hashtags with a discriminating value ($|C_D(h) - C_R(h)|$) less than or equal to some τ_Δ, shown in Figure 1. The curve levels off around $\tau_\Delta \approx 75$, indicating that the small number of hashtags beyond this value have similarly high discriminatory value. We used this value for the remainder of the study. We considered other values in the vicinity of the inflection point without observing any notable differences in results.

Politically Active Dataset (PAD). Despite the fact that datasets in literature were largely composed of politicians, many also included self-identified Republicans and Democrats. Insofar as dataset collection has been concerned, self-identification is typically done in one of two ways. A user may indicate a political orientation in her Twitter profile (e.g., "Republican and loving it!" or "Born a Democrat, will die a Democrat.") or be flagged as belonging to one party or the other in a Twitter listing such as the WeFollow service[5]. While Twitter listings generally do not reveal their list construction methods, we observed that in nearly all cases Republicans and Democrats in these lists featured their orientation prominently in their profile. As a consequence, we used explicit mentions in a user's profile of a political orientation to construct this second dataset. This dataset represents a milder set (still) politically vocal Twitter users. A notable qualitative distinction between the political figures dataset and this dataset is that nearly all political figure tweets are political in nature whereas political topics are no longer the dominating topic in tweets generated by politically active users.

This dataset was created by first walking the profiles appearing in tweets read off the Twitter gardenhose. To exclude non-US individuals (to avoid including non-US citizens), only tweets written in English, geotagged with a location inside the continental United States were considered. Furthermore, only profiles explicitly indicating their location to be a US city or state were considered. Subject to these stringent criteria, any profile which mentioned "Republican", "Democrat", "Conservative", or "Liberal" were flagged as candidates for inclusion in the dataset. Users already appearing in the Political Figures dataset were removed to eliminate any overlap (and any political figures in this dataset). In a second pass, the remaining accounts were manually inspected and only those indicating a clear association with one of the two major US parties or political orientations were kept. The second pass was required to avoid the possibility of a user invoking a party name without being actually associated with it (e.g., "Can't stand liberals!"). For each user account that remained, the Twitter API was used to collect their 1000 most recent tweets. This became the Politically Active dataset.

Politically Modest Dataset (PMD). The goal of this dataset was to collect a set of "typical" Twitter users who expressed *some* political view. Note that the expression of some political view was important to the purpose of this study as it allowed us to label users. Admittedly, many, if not most, Twitter users never express *any* political views. While it might still be possible to recover the political orientation of such users, this would be done through distinctly non-political features that correlated with political orientation (e.g., perhaps Democrats talk about sports more than Republicans). Obtaining high quality political orientation labels for such users given current resources and technology is virtually impossible. As a result, in this study we opted for the next best set of users — those who talk about politics, but do not explicitly label themselves in any way. We implemented this by restricting our pool of candidate users to those geographically restricted to the US (per the methods described for the active users) who do not make any mention of political parties or politically discriminative hashtags in their profiles. Using the same method as described for the PAD, we easily identified 10,000 users who satisfied these criteria. As done before, the latest 1000 tweets were retrieved for each user.

At this point, however, we faced a serious problem: the nature of our selection criteria made these users exceedingly difficult to label. In fact, we anticipated that most of the sampled users could not even be labeled, based on the assumption that most of the Twitter population never

[5] http://www.wefollow.com

says anything political. Left with no explicit signal for political orientation that a regular expression could process, any attempt at labeling would involve the nuances of human judgement (determining non-explicit political orientation from language in individual tweets). One approach frequently employed to doing such work at scale is to use a crowd sourcing platform such as Amazon Mechanical Turk[6]: in this case, having several AMT workers look at each sampled user and independently propose a label as either *Republican*, *Democrat*, or *Unknown* (if too little signal is present in their tweets) (Schnoebelen and Kuperman 2010; Buhrmester, Kwang, and Gosling 2011). Those users for which multiple AMT workers applied the same label would be given the corresponding label. This exercise would be theoretically possible except that asking an AMT worker to code a single user would involve her parsing through that user's 1000 tweets to find a piece of political signal: reading 1000 tweets is beyond the scope of a reasonable AMT job.

In order to resolve this issue, we elected to subsample each user's tweet history — effectively reducing each user to 10 politically-relevant tweets that they had generated. An AMT worker would then read the 10 tweets for a user and, based on these tweets, decide on the most appropriate label for that user. To identify these 10 political tweets, we made use of the politically neutral hashtag set extracted from the PFD as described above. For each user, a set of (at most) 10 tweets containing at least one politically neutral hashtag were obtained. Any user who had no tweets that intersected with the neutral hashtag set were discarded and not included in the AMT labeling exercise. In this way, we obtained politically-relevant representations for "normal" Twitter users. To run these AMT tasks, each task consisted of coding 25 users, for which an AMT worker received \$0.10. In total $1,500$ Twitter users were assigned a label by 5 AMT workers and only Twitter users that received 3 or more labels in agreement on the party were placed in the actual Politically Modest Dataset. The final PMD contained 327 Twitter users, which constitutes approximately 3% of the original $10,000$ random, anglophone, US-based users we started with. This, incidentally, gives a rough sense for the fraction of such users that engage in some form of politically oriented discussion on Twitter.

Why neutral hashtags? Before moving on, it is instructive to explain why we could not use the politically discriminative hashtag set (or the complete hashtag set) from the PFD. Fundamentally, using anything but the neutral hashtags would favor including an artificially enriched population of politically polarized, easy-to-classify users in the PMD - key properties we were trying to avoid.

To see this, consider the implications of using politically discriminative hashtags. All these hashtags, by construction, have a strong statistical association with one party or another. By keeping only users who used those hashtags, the PMD would consist of a population of users whose political hashtag usage would echo the PFD. Thus, at the very least, the PFD classifier would be guaranteed to do well on the resulting PMD.

[6]http://mturk.amazon.com

However, there is an even more serious issue: since there is a strong correlation of these hashtags with political orientation in the PFD, it is plausible that there is a strong correlation of these hashtags with the political labels that the PMD users would be assigned (Conover et al. 2011a). The effect of selecting users that use discriminating hashtags is that nearly every Democrat-labeled PMD user would have tweets that contain a Democrat-discriminating hashtag and nearly every Republican-labeled PMD user would have tweets that contain a Republican-discriminating hashtag. The effect would be a PMD dataset in which a user's tweets literally contain one or more keywords identifying their class. This would be tantamount to picking users who are self-declaring their affiliation in their tweets - a stronger signal of political orientation than is even present in the Politically Active Dataset. Note that using all the political hashtags in the dataset would somewhat soften this effect - but since the dataset still contains strongly discriminative hashtags, the effect could persist with severity depending on the relative abundance of politically discriminative hashtags in the dataset.

By using only the most neutral political hashtags in the PFD, we avoid using hashtags that strongly correlate with either party, thereby giving the resulting classifier no advantage through dataset construction. It is important to appreciate that we are not *guaranteeing* that these users are hard to classify — we are simply ignoring any particularly good politically discriminative signal during the dataset construction phase.

Conover 2011 Dataset (C2D). At the outset of the project, we had hoped to incorporate datasets from other studies in order to evaluate the extent to which our constructed datasets compared to the datasets used in prior work. Due to various restrictions in Twitter and university data sharing policies, we were unable to obtain any datasets, save one. The authors of (Conover et al. 2011b) were able to share their dataset in a restricted form: each user was represented as a bag of words with no tweet- or user-level features (e.g., tweet frequencies, user degree). The absence of these features certainly impacted classifier accuracy. Nonetheless, we included the dataset in the study and report results for it. Overall classifier performance on the C2D loosely tracks that of the PFD, which is consistent with the way in which it was constructed.

Other data collection details. With the exception of the Conover2011 dataset, all data was collected using a combination of the Twitter REST API and access to the Twitter firehose. All tweets and user profile data were gathered over the time period, May 11-15, 2012. This period falls significantly before the final run up to the 2012 US Presidential Election. Most importantly, during the collection period there were no major incidents that might have introduced substantial variation in tweet content collected over this timeframe. Furthermore, while some top political hashtags in all three sets involved some referring directly to election activities, none referred to particular election events.

Table 2: Top hashtags in the Political Figures dataset compared to their new ranking in the other datasets. Taking ranking as a rough proxy for the importance of political discussion in user tweeting behaviors, differences in the user groups become immediately clear.

Hashtag	Figures Ranking	Active Ranking	Modest Ranking
#obama	1	147	568
#tcot	2	312	26
#gop	3	448	482
#jobs	4	35	502
#obamacare	5	128	4113
#budget	6	67	2848
#medicare	7	415	4113
#healthcare	8	440	1436
#debt	9	510	3370
#jobsact	10	613	2458

To underscore, at the outset, the differences present within the three datasets, Table 2 shows the rankings of the top hashtags present in the PFD across the other two hashtags. Taking rank simply as a rough proxy for the relative importance of political discourse in these different datasets, we can see how the different sets of users engage differently with political topics.

Classifiers on the Datasets

To evaluate the performance of the SVM and LLDA classifiers on each dataset, we ran 10-fold cross validation and report average accuracy achieved in the individual folds. The results are shown in Table 3. Note that while accuracy generally is an insufficient statistic to characterize a classifier's performance, here we are considering a binary classification problem. In this special case, accuracy ($\frac{TP_{Dem}+TP_{Rep}}{Total_{Dem}+Total_{Rep}}$) correctly reports the overall performance of the classifier. What we lose is detailed knowledge about how much members of each label contributed more to the overall error. In the present study, we do not consider the classifier performance at this level and, therefore, omit these details.

Due to space limitations, we only report the SVM classifier performance here. The LLDA performance was effectively the same in values and trends to the SVM values reported.

Table 3: The average of ten 10-fold cross-validation SVM iterations. The test was performed on each one of our datasets respectively.

Dataset	SVM Accuracy
Figures	91%
Active	84%
Modest	68%
Conover	87%

Table 4: Percentage of tweets that used highly discriminating hashtags vs. those that did not. Highly discriminating hashtags were those that had a discriminating value greater than 75% of all hashtags in the dataset.

Dataset	High Disc.	Low Disc.
Figures	44%	56%
Active	32%	68%
Modest	24%	76%

In the results given in Table 3, two trends are apparent. First, we find that the Conover dataset falls somewhere between the Political Figures and Politically Active Datasets. Recall that we were only able to obtain a heavily preprocessed version of the dataset which made a number of user-level features impossible to compute. In prior work these features have been identified as significant contributors to overall classifier accuracy. As a result, we suspect that the classification accuracy would have been substantially higher on the original dataset - bringing it in line with the Political Figures Dataset, which it is most similar to. Regardless of exactly how classifiers would fair on the original, its reported accuracy is substantially higher than both PAD and PMD, indicating that our approximation of past datasets with highly politically active and politician accounts was fair.

A second, and more central, observation is that the accuracy implications of different dataset selection policies are evident: politically modest users are dramatically more difficult to classify than political figures. Surprisingly, politically active users also are markedly more difficult to classify. At the outset, we had expected these users would be more like the political figures, assuming that a willingness to self-identify with a particular political orientation would translate into tweets that clearly convey that political orientation. Evidently this is true for many active users, but is notably less ubiquitous than is the case for politicians.

In order to better understand the nature of this loss in accuracy, we evaluated how lexical diversity, a major source of noise to both SVM and LLDA classifiers, varied across the datasets. Fundamentally, by lexical diversity, we refer to the extent to which users in a particular dataset simply employ a larger, less uniform vocabulary. A more diverse vocabulary could mean that individuals are talking about the same political topics, but using a broader lexicon, or that individuals are actually discussing a wider range of topics. In either case, a less controlled vocabulary can make a class difficult to summarize or difficult to discern from another class (when the vocabularies overlap).

We evaluated this diversity in two different ways. First, we looked at the percent of tweets that used highly discriminating hashtags in each dataset, shown in Table 4. The null model of interest in this case is one in which Democrats and Republicans have different lexicons but do not favor any particular word out of that lexicon. This is equivalent to users from both groups draw a hashtag from the set of available Democrat/Republican discriminating hashtags with uniform

Table 5: Jaccard Similarity between the topics used by Republicans and Democrats.

Dataset	Jaccard Similarity
Figures	37%
Active	48%
Modest	61%

probability. In this case, no hashtag would be statistically favored over another. As a result, the hashtags that have a discriminating value greater than or equal to $X\%$ of all hashtags would occur in no more than $(1 - X)\%$ of tweets — if a particular hashtag were favored (more discriminating), then it would be chosen more often than the other hashtags in the hashtag set and appear in more tweets. This is *exactly* what we find is true for the Politically Modest users — the top 25% of discriminating hashtags occur in almost exactly 25% of the tweets. Politically active users exhibit a high degree of affinity for particular discriminating hashtags and Political figures deviating the most from the null model. In other words, politically modest users show no collective preference for a particular political (or non-political since a discriminating hashtag must not be, itself, political in nature) hashtag set. As a result, in the modest dataset neither Democrat or Republican groups of users are distinguishing themselves by focusing around a particular set of hashtags. This gives the classifier little within-label consistency to hold onto.

Another way of quantifying lexical diversity is by considering the cross-label lexical similarities within a dataset using the topic models constructed by the LLDA classifier. Briefly, the LLDA classifier builds a set of topics for the documents it encounters and attempts to associate some probability distribution over those topics with each label (in this case Republicans and Democrats). Individual topics are probability distributions over the complete vocabulary encountered in the document set (e.g., a sports-like topic would assign high probabilities to "baseball" and "player", low probabilities to "Vatican" and "economy" since these are rarely associated with sports).

In Table 5, we show the average percent overlap in the topics associated with Republicans and Democrats. Under ideal classification conditions, labels would be associated with topics distinctive to that category. Here we see that for modest users, there is extensive overlap between the topics that characterize each label. Thus, not only is the within-label vocabulary diffuse (as evidenced by the hashtag analysis above), but the across-label vocabulary is highly similar (giving rise to similar categories being used to characterize each label). This combination of lexical features make the Politically Modest dataset exceedingly challenging for lexicon-based classifier.

Cross-Dataset Classification

Beyond the raw accuracy of political orientation classifiers, a question of significant practical and theoretical interest is the extent to which a classifier trained on one dataset can be used to classify another dataset. Such a property would be highly desirable in the political orientation case: as we have seen, political figures are easy to find and easy to label whereas ordinary users are quite difficult to label properly. Thus, building a classifier on political figures is a much easier endeavor. Can such a classifier be used to obtain meaningful labels on arbitrary users?

Table 6: Performance results of training our SVM on one dataset and inferring on another, italicized are the averaged 10-fold cross-validation results

Dataset	Figures	Active	Modest
Figures	*91%*	72%	66%
Active	62%	*84%*	69%
Modest	54%	57%	*68%*

Table 7: Jaccard Similarity between the features across datasets.

	Figures	Active	Modest
Figures	100%	18%	9%
Active	18%	100%	23%
Modest	9%	23%	100%

From the outset, the degree of classifier transferability across datasets is not clear. While, certainly, the PFD, PAD, and PMD are different, if the features that correlate with Democrat/Republican labels in the PFD also correlate in the PAD and PMD, then the classifier might maintain a moderate degree of accuracy. To evaluate this, we built a classifier for each dataset using all labeled users. This classifier was then used to classify the political orientation of users in the other datasets.

The results of this experiment, shown in Table 6 reveal that the classifiers not only lose accuracy, but perform profoundly worse than even a random classifier, which would achieve $\sim 50\%$ accuracy on each dataset. Interestingly, the PFD-based classifier performs just as badly on PAD users as it does on PMD users. The same phenomenon is observed for the PAD and PMD classifiers as well.

This trend also emerges by considering the overlap in features employed by the classifiers for the different datasets, shown in Figure 7. Here we find that the chasm between classifiers is somewhat exacerbated. On one hand, this is somewhat to be expected since the classifiers are only able to incorporate the k-top items of each feature of interest (hashtags, words, etc...). However, by comparing the k-top features, we obtain a stronger sense of the extent to which these datasets differ: no two datasets share more than 20% of their most discriminative features in common.

Taken together, these results strongly suggests that, while politician and politically active users are substantially easier to classify than politically modest users, their actual useful, politically-discriminating features are utterly different. This is both a remarkable and concerning result. For two collections of users who share the same designations (Republican/Democrat) and similar classify-ability to be so dif-

ferent in features suggests substantial behavioral differences (Twitter-based behavior, in this case). On one level, this is not surprising given that one group consists of politicians. Perhaps more remarkable is the gap between politically active and modest users. The behavioral differences suggest that, not only are politically active users more politically vocal on Twitter, but *what* they say politically is also quite different. To our knowledge, this has not been documented in the literature. Better understanding the nature of these differences will be an exciting direction for future work.

These cross-dataset results have severe implications for the immediate utility of political orientation classifiers: they simply will not transfer across datasets. This has two practical ramifications. First, model building remains hard: we cannot assume that we can build models on easy-to-obtain datasets and then lift these up and apply them to harder-to-label, but conceptually similar datasets. Second, one must be very attentive to when one can use a particular classifier. As the accuracy degradation between the active and modest users revealed, even seemingly ordinary users who are perceived to simply exhibit different degrees of a behavior may, actually, manifest different behaviors that a classifier cannot accommodate.

Moving Forward

The overall implication of this work is that classification of political orientation, and potentially many other latent attributes, in Twitter is a hard problem — harder than portrayed in the current literature — in several key ways.

Good, labeled datasets are hard to build. As demonstrated by the complex process involved in building the politically modest dataset, it seems that, for some latent attributes, assembling unbiased datasets with high-confidence labels is a non-trivial endeavor. Furthermore, beyond being complicated and intensive, there are subtle pitfalls that must be avoided, such as the choice of using neutral vs. all political hashtags in the construction of the PMD. This calls for renewed attention on the part of researchers and reviewers to the assumptions and biases implicit in the construction of datasets and assignment of labels. Furthermore, in this paper we have demonstrated one technique that might be used to construct other datasets, both for political orientation and for other attributes. However, more broadly, there is a need for principled approaches to building labelled datasets, particularly where latent attribute inference is concerned.

In the interest of eliminating the significant time and effort involved in building such datasets, there is also the desperate need for existing datasets to be shared. In the case of this study, we have released all three datasets we used in an anonymized form. We encourage other researchers to do the same.

Existing methods fail on "ordinary" users. While prior work (including one of the author's own past papers) has suggested that the political orientation inference problem can be solved with high accuracy, these findings do not apply to more normal, less politically vocal Twitter users. Perhaps the single greatest root cause for this is illustrated in Figure 2 which shows the normalized differentiating values for the 1000 most differentiating hashtags in each of the three datasets we considered. Politically modest users, "normal" users, simply lack strongly differentiating language between political orientation classes (in this case, Republicans and Democrats). This suggests that in order to identify political leanings (and other attributes that encounter similar issues), it will be necessary to leverage more subtle behavioral signals, such as following behavior, the behavior of neighbors, and the greater network context in which an individual is situated (e.g., (Zamal, Liu, and Ruths 2012)).

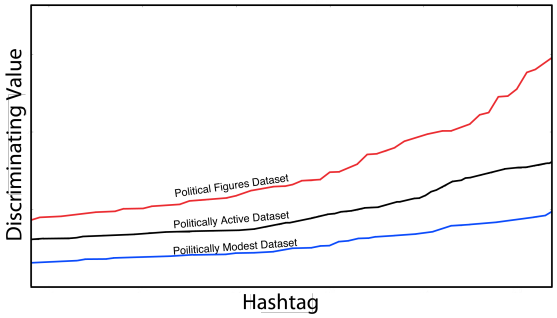

Figure 2: The discriminating values for the top 516 discriminating hashtags in each dataset, in increasing order. The plot shows that politically modest users utterly lack discriminating hashtags. The discriminating values are normalized by the largest discriminating value in all the datasets.

Because gaining insight into the behavior of arbitrary and ordinary users is so central to the goals of research into and commercialization of online social environments, recognizing and addressing the lack of support for such users is crucial to the forward progress of latent attribute inference and the delivery of tools which will serve the needs of social scientists, companies, and other organizations.

Classifiers do not transfer across types of users. On some level, the fact that applying a classifier to a dataset it was not designed for hurts accuracy is unsurprising. However, our results quantify, for the first time, just how severe the effects of transferring a classifier across seemingly related datasets can be. We suspect that this issue is not unique to political orientation. An important question for all researchers working in latent attribute inference is the extent to which their methods and classifiers generalize to different populations — populations separated by behavior (as was the case in this study), but also separated by other features such as geography and even time. A number of natural, interrelated research topics emerge out of this result as well: how can we build classifiers that do transfer across datasets? How can we know when a classifier will transfer? Answers to these and related questions will significantly advance the utility and theoretical foundations of latent attribute inference.

Conclusion

In this work we have shown how the ways in which political orientation-based datasets have been built have lead to significant overestimation of the accuracy of political orientation classifiers as applied to populations of normal Twitter users. We showed a painstaking way in which a labeled, unbiased Twitter dataset could be built for the political orientation problem. Using this dataset, we quantified the extent and nature of the accuracy overestimation. We emphasize our belief that prior work on this topic has consistently shown good faith in making sound forward progress on the problem of political orientation inference. The results we have presented here are intended to offer new perspectives on an established problem that, hopefully, will invigorate further productive research on the topic and related areas.

Acknowledgements

The authors thank Twitter for providing elevated Gardenhose access and Conover and co-authors for sharing an anonymized version of their political orientation dataset. This manuscript benefited from the feedback from anonymous reviewers. This work was supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight #435-2012-1802).

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1):3–5.

Burger, J.; Henderson, J.; and Zarrella, G. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chang, C., and Lin, C. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Conover, M.; Goncalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011a. Political polarization on twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

Conover, M.; Goncalves, B.; Ratkiewicz, J.; and Flammini, A. 2011b. Predicting the political alignment of twitter users. In *Proceedings of the International Conference on Social Computing*.

Golbeck, J., and Hansen, D. 2011. Computing political preference among twitter followers. In *Proceedings of Conference on Human factors in computing systems*.

Liu, W., and Ruths, D. 2013. What's in a name? using first names as features for gender inference in twitter. In *Symposium on Analyzing Microtext*.

Pennacchiotti, M., and Popescu, A. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the International Conference on Weblogs and Social Media*.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *The Conference on Empirical Methods in Natural Language Processing*.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *International Conference on Weblogs and Social Media*.

Rao, D., and Yarowsky, D. 2010. Detecting latent user properties in social media. In *Proceedings of the NIPS Workshop on Machine Learning for Social Networks*.

Schnoebelen, T., and Kuperman, V. 2010. Using amazon mechanical turk for linguistic research. *Psihologija* 43(4):441–464.

Wong, F.; Tan, C. W.; Sen, S.; and Chiang, M. 2013. Media, pundits and the u.s. presidential election: Quantifying political leanings from tweets. In *Proceedings of the International Conference on Weblogs and Social Media*.

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *The International Conference on Weblogs and Social Media*.

Erratum: Table 6 values were incorrect in the first online version of this paper. They have been corrected. The print version contains the corrected values.