Content+Context=Classification: Examining the Roles of Social Interactions and Linguist Content in Twitter User Classification*

W. M. Campbell
Human Language Technology
MIT Lincoln Laboratory
Lexington, MA 01740
wcampbell@ll.mit.edu

E. Baseman[†]
School of Computer Science
Univ. of Mass. Amherst
Amherst, MA 01003
ebaseman@cs.umass.edu

K. Greenfield
Human Language Technology
MIT Lincoln Laboratory
Lexington, MA 01740
kara.greenfield@ll.mit.edu

Abstract

Twitter users demonstrate many characteristics via their online presence. Connections, community memberships, and communication patterns reveal both idiosyncratic and general properties of users. In addition, the content of tweets can be critical for distinguishing the role and importance of a user. In this work, we explore Twitter user classification using context and content cues. We construct a rich graph structure induced by hashtags and social communications in Twitter. We derive features from this graph structure—centrality, communities, and local flow of information. In addition, we perform detailed content analysis on tweets looking at offensiveness and topics. We then examine user classification and the role of feature types (context, content) and learning methods (propositional, relational) through a series of experiments on annotated data. Our work contrasts with prior approaches in that we use relational learning and alternative, non-specialized feature sets. Our goal is to understand how both content and context are predictive of user characteristics. Experiments demonstrate that the best performance for user classification uses relational learning with varying content and context features.

1 Introduction

In recent years, Twitter has become an extremely prolific social media engine, attracting an extremely diverse user base, ranging from teenagers discussing the latest in pop culture, to businesses looking for free advertising space, to the president of the United States trying to reach a broader audience than traditional media will allow. Twitter is the place to say whatever you want to whoever you want..., as long as it is less than 140 characters. This conciseness constraint forced the Twitter user base to develop innovative ways of maximizing the information content of each letter. As such, the resulting tweets constitute a vast data set of rich textual content. Additionally, these tweets traverse through and define a social network comprised of all kinds of people tweeting to people about people.

In this work, we try to identify who some of these people are by performing user classification. Several prior methods have been proposed. Teng and Chen (2006) performed a similar study on bloggers in order to classify them by interest type. Twitter, however, provides a much richer feature set due to the denser network structure and nearly ubiquitous adoption of user profiles. Romero et al. (2011) defined a social graph structure for Twitter data. While this was the first paper of its kind to explore the benefits of extracting a network structure from textual data, they only considered a limited graph comprising the retweet structure. Rao et al. (2010) and Pennacchiotti and Popescu (2011) expanded the notion of a Twitter graph to more broadly encompass social communication and used this jointly with content features to predict some attributes of Twitter users. Only limited graph analysis was performed and

[†] This work was performed under an internship at MIT Lincoln Laboratory.

[*] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 59–65,
Dublin, Ireland, August 24 2014.

learning was based on the user and not relational methods. In (Wu et al., 2011), a special feature, Twitter lists, was used as a simple approach to user classification.

Our goal is to explore the role of different features and learning methods in user classification. The motivation for this study is multifold. First, prior work has only performed limited analysis on Twitter graph structures and their role in user characterization. Second, the interplay between content, context, and learning method (relational versus propositional) has not been fully explored. Third, user classification using profile information or specific Twitter features may not always be available or accurate. In our study, we find 20% of the users do not have an accessible profile. Additionally, since profile information contains self-reported fields, its accuracy is questionable.

In this paper, our contribution consists of multiple parts. First, we develop a rich network representation of Twitter data that combines the traditionally exploited social network features (retweeting, at-mentions), user tweeting behaviors (hashtag usage), and hashtag co-occurrence without requiring storage of all tweets. We perform network analysis on this graph and show that community structure and centrality in the network are qualitatively interesting and relevant for user classification. Second, we perform tweet level topic clustering and offensiveness detection. We propose a new method for propagating posterior class probabilities to both hashtag and at-mention graph nodes. This provides topic and offensiveness associations for both users and hashtags in the Twitter graph. Finally, in constrast to prior work, we perform relational learning on the Twitter graph and show significant improvements in performance by using information from hashtag-neighbors and user-neighbors of a user node.

2 Corpus Collection

Over the 8 month period from September 2012 to March 2013, we used Twitter's streaming API to collect a $< 1\%$ sample of the entire twitter feed, totaling approximately 686 million tweets, which (Morstatter et al., 2013) showed to be a representative sample of the entire Twitter space. These tweets contain basic information—tweet id, date and time, user id, user location, and tweet text. The tweets are representative of the Twitter population and display a wide variety of user accounts, topics, languages, and locations.

In addition to collecting a corpus of tweets, we used Twitter's REST API to collect the user profiles for all users who were either an author of or at-mentioned in at least one of the previously collected tweets. User profiles contain both user generated and automatically generated content. The user generated content includes information such as webpage link, location, time zone, screen name, language, and a textual description of the account. Since these are self-reported attributes, their veracity is often noisy or over-generalized—e.g., location reported as "all around the world." The automatically generated user account content consists of account information and statistics such as number of followers, date of account creation, number of tweets (statuses_count), verified status, and favorite count.

We augmented our tweet + profile corpus by labeling a subset of user profiles with their user type. Annotation was performed by multiple annotators. We considered the following partitioning of user profiles into user classes. **individual:famous** : Famous person: writer, actor, former politician, etc. **individual:generic** : Generic (everyday) user tweeting. **individual:other** : An individual that doesn't fall into the categories above. **organization** : Business, non-profit, government organization. **fake** : Fictional characters, celebrity impersonations, deceased individuals, etc. **info** : Information source–news, trivia, jokes, quotes. **bots** : Produces automated posts via Twitter. **missing** : User doesn't exist (deleted page or misspelled user id). **dontcare** : Spam, offensive services. **other** : Not in one of the above categories. We selected this schema as an initial exploration of interesting categories and many alternate schema are possible.

3 Twitter Graph Construction

In order to construct an analytics platform, we applied a mapping that converts a corpus of tweets and corresponding user profiles into a Twitter graph. There are several methods of representing Twitter data as a graph which capture diferent levels of data richness at inversely proportional computational expense (both in processing and storage requirements). In this work, we consider a graph with multityped nodes in one-to-one correspondence with the union of the set of user profiles (e.g., `@blueman`) and the set of

Table 1: Example communities obtained from Infomap community detection with the Twitter Graph

Communities and High-Pagerank Members	Highest Pagerank Node
#me, #cute, #instagood, #beautiful, #fashion	#love
#500aday, #tfb, #instantfollowback, #teamautofollow, #followback	#teamfollowback
#breakoutartist, #musicfans, #onedirection, #popartist, #celebrityjudge	#peopleschoice
#android, #androidgames, #ipad, #ipadgames, #iphone	#gamesinsight
@harry_styles, @real_liam_payne, @louis_tomlinson, @zaynmalik, @onedirection	@niallofficial
#football, #49ers, #packers, #ravens, #redskins	#nfl
#p2, #teaparty, #tlot, #gop, #obama	#tcot
#believecoustic, #believe, #believetour, #kiss, @alfredoflores	@justinbieber

hashtags (e.g., #yankees) that occur in one or more of the collected tweets. We chose not to include individual tweets as nodes in the graph in order to maintain tractability.

In addition to the two classes of nodes, we considered five classes of edges. There are three classes of edges that connect two user profile nodes. The first is a directed, weighted edge corresponding to the number of times one user at-mentions another user (e.g., @blueman writes a tweet containing @greenman). The second type of user to user edge is a directed, weighted edge corresponding to the number of times one user retweets another user (e.g., @blueman writes a tweet containing RT @greenman). Unlike the at-mentions and retweets edges, the third type of user to user edge doesn't map to communication between users; rather this edge classification refers to an undirected, weighted count of the number of times at-mentions of two users co-occur in the same tweet (e.g., @redman writes a tweet containing @greenman and @blueman). Similarly to the user to user co-occurence edge classification, there is an undirected, weighted edge classification corresponding to the co-occurrence of two hashtags. The final class of edges are weighted, directed edges corresponding to the number of times a given user tweets a particular hashtag.

4 Graph Features

We extracted network features from the Twitter graph based on community detection and centrality (Pagerank). We partitioned the node set into commuties by leveraging the infomap approach (Rosvall and Bergstrom, 2008). We use Pagerank to calculate the centrality of each node and we define community centrality as the sum of the Pageranks for all nodes in the community. The community "Pagerank" allows us to rank communities by centrality. After computing both Pagerank and communities, we added these node features to the original (directed) graph.

Optimizing the communities in the Twitter graph yields communities with both user and hashtag nodes. Nodes with high-pagerank in a community serve as a community summarizations. We show a summary of some of the largest Pagerank sum communities in Table 1. Cursory analysis reveals that qualitatively the communities are very interpretable—user Justin Bieber is associated with the Believe tour, the hashtag #gameinsight is associated with different platforms and game types, followbacks are grouped together, and the #love community has the highest community Pagerank.

5 Content Analysis

Content analysis uses natural language processing to extract additional structured features from unstructured tweets. We discussed simple content analysis based on parsing tweets in order to identify communication (at-mentions and retweets) and content (hashtags) in a previous section. In this section, we cover more advanced techniques—topic modeling and offensiveness detection.

Before covering the details of content analysis, we describe the general framework for incorporating content features into our classification framework. We assume that content analysis produces a vector of posterior probabilities,

$$\mathbf{c}_j = \left[p(\omega_i|\text{tweet}_j)\right] \tag{1}$$

where ω_i is the indicator for the class label i. In general, including all tweets as nodes in a graph for classification leads to a graph of prohibitive size. Instead, we propagate the information contained in tweets to corresponding user and hashtag nodes and compute the expected posterior probability; i.e., we average all of the vectors propagated to a certain node.

The rules for propagating $\mathbf{c}_j$ for at-mentions are as follows. If user `@blueman` tweets with no recipient or multiple recipients, then $\mathbf{c}_j$ is propagated to the `@blueman` node. If user `@blueman` retweets from user `@greenman`, then $\mathbf{c}_j$ is propagated to both `@blueman` and `@greenman`. Similarly for hashtags nodes, we propagate the $\mathbf{c}_j$ vector to all hashtags used in tweet_j.

Averaging the content vectors that were propagated to user and hashtag nodes defines representative content vectors for those nodes. A drawback of this aggregation strategy is that the average estimator can have a variable variance proportional to the number of vectors propagated to a node, which can introduce noise into the classification process.

5.1 Topic Modeling

Our topic modeling is based upon probabilistic latent semantic analysis (PLSA). PLSA models the joint probability between documents and words by introducing a latent variable z representing possible topics in a document. We trained the PLSA model with an EM algorithm.

An analysis of the topics produced by PLSA provides meaningful interpretation of many of the high scoring topics. For instance, topics such as money, sleep habits, the ubiquitous Justin Bieber, birthdays and Valentine's day, and love are easily seen. Expressions of happiness via emoticons are also a topic. The topics in general represent broad categories (love, football) and specific events (video awards, Valentine's day). In addition, topics that represent common linguistic phenomena—African American vernacular English (AAVE), various expletives, and teen lifestyle (class, teacher, parents, ...). We remark that the topics are related to but not the same as the community detection applied to the Twitter graph. Qualitatively, the Twitter graph communities appear to be better defined and more easily interpreted than their PLSA counterparts.

5.2 Offensiveness

Another significant attribute of a tweet is linguistic register—the variation due to the social setting. In an attempt to capture some of the phenomena that occurs due to formality, familiarity, etc., we trained an offensiveness detector. The goal of building this detector is that variations in offensiveness might distinguish user type (e.g., politician versus generic user).

Offensiveness is a broad term and could be defined in many different ways. We define offensiveness using a pragmatic two-stage approach. First, we obtained a set of offensive tweets by issuing queries via Lucene of offensive terms. To obtain a set of putative non-offensive tweets, we took a random sample of the remaining tweets from a large pool, assuming that in this case offensiveness has a lower prior. We then trained a classifier with the offensive and non-offensive data sets. The resulting output of the detector yields a consistent definition of offensiveness. Additionally, training a detector rather than using just a dictionary approach captures some of the additional co-occurring terms and allows learning appropriate weightings of terms.

We split the annotated data into distinct train and test sets for performance measurement and calibration of the detector. Approximately 14000 tweets were in both sets. We trained an SVM by using normalized word-count vectors and a linear kernel. The SVM regularization parameter was tuned for optimal performance to $c = 0.1$. The equal error rate is 12.4% for the detector. The detector appears to produce consistent results. Given that tweets are short messages, this performance level appears reasonable.

We needed two additional steps in order to apply the detector to all English tweets. First, we converted the output of the SVM to a posterior probability using the standard approach in Platt (Platt, 2000) and optimized by using a conjugate gradient method. Second, there were numerous cases in the data where unseen vocabulary in the training set resulted in a zero-vector for $\mathbf{v}_i$. In these cases, we used an imputed posterior of offensiveness by taking the average value across all nodes.

6 User Classification

Our goal in this work is to identify the saliency of network and content features as well as relational versus propositional learning methods for classifying Twitter users. We cast the user classication problem as a detection problem; i.e., for each label (verified, generic, etc.), we build a 1-versus-rest detector to predict that attribute of the user.

Since our representation of the Twitter data involves user-user, user-hashtag, and hashtag-hashtag relationships, we apply a relational learning approach to user classification. Relational learning techniques leverage the structure of the neighborhood around a user of interest as well as the attributes of that user and its neighbors. In this work, we construct queries consisting of the user of interest, and the nodes adjacent to and edges incident to that user. We use relational probability trees (RPTs) (Neville et al., 2003) to leverage these subgraphs for classification. An RPT is a decision tree which automatically calculates and considers aggregate features within the subgraphs.

7 Experiments

7.1 Experimental Setup

From a 1% sample of raw tweets, we created a Twitter graph combining content and network features. We included network structure in the graph by letting edges indicate retweets, communication and cooccurance of users in tweets, or co-occurrence of hashtags within tweets. In addition, we annotated each edge with counts for each interaction type. We added additional content features for topic and offensiveness as well as network features for cluster and cluster pagerank for each user and hashtag. Topic vectors were assigned using a PLSA approach, and offensiveness was determined using an SVM. Clustering and pagerank were achieved using the infomap approach. In addition, we hand-annotated approximately 1300 users (individual:famous, individual:generic, individual:other, organization, info source, bot, etc.) by examining their profiles and tweets. The resulting graph had 252K nodes (189K users and 64K hashtags), and 1.16M edges.

We performed experiments using multiple feature subsets (content, network and content+network) and different learning methods (propositional, relational). The content features we generated for users and hashtags are topic and offensiveness. We include the top three most likely topics for each user and hashtag in our feature set for these experiments. Our network features for users and hashtags are cluster and cluster pagerank. We cast the user classification process as a detection problem. For each user label (verified, generic, etc.), we create a detector using an RPT that uses a one-versus-rest labeling.

To predict whether or not a user has a verified account, we learn decision trees with Proximity[*] with varying features included in the analysis. There were a total of 84k users with a verified label in our graph from downloaded Twitter profiles. We subsampled 10% of this data set to give reasonable run times for Proximity.

For the remaining hand-annotated classes, we used all available labeled data. We learned decision trees using Proximity with a maximum tree depth of 3. For some user types, there were not enough labeled users to make reliable models and predictions; we excluded user types with probability less than or equal to 1% (fake, other, spam). We also found prediction of organization to be unreliable. Therefore, we focused our experiments on high-prior hand-annotated classes—generic, info, and famous.

For both relational and propositional (user-only) learning, we divided the data into 5 random splits with 80% of the data in training and 20% of the data in test. For each split, we measured the area under the curve (AUC) from the trained detector on the heldout test set. Results are reported in terms of the mean and standard deviation (SD) of the AUC across all splits.

Relational learning was somewhat complicated by the varying structure of user neighborhoods. Ideally, we would like our graph queries to return subgraphs that consist of a central labeled user, and all users and hashtags that are immediate neighbors of this labeled user. However, the RPTs need to be able to calculate the same aggregate features for each subgraph. A problem arises because some labeled user nodes only have neighbors that are users, while other user nodes have only hashtag neighbors, and still

[*] Software and documentation is available at http://kdl.cs.umass.edu/proximity

more nodes have both user and hashtag neighbors. There is also an unusual case where some users do not have any neighbors. We find that user nodes from these four neighborhood structure cases (no neighbors, only user neighbors, only hashtag neighbors, and both user and hashtag neighbors) cannot all be mixed together in the same training and test sets. This diffculty occurs because aggregate attributes that are well-defined on the users with only hashtag neighbors, such as average offensiveness of neighboring hashtags, are undefined for nodes with only user neighbors. We handle this by running separate sets of experiments for each of these four neighborhood structure cases.

7.2 User Classification Experiments

7.2.1 Verified User Results

Table 2 shows average AUC results for propositional and relational prediction of account verification. Note that an AUC of 0.5 indicates chance performance. From the table, we see that we can perform reasonable user classification of verified users using our *extracted* features.

From the table, we can reach multiple conclusions. First, propositional learning performs similarly with either network-only or content-only features. Second, content and network features together provide a significant boost in performance. This observation has been noted in other domains such as Enron e-mail (Coppersmith and Priebe, 2012). Third, the introduction of relational learning gives substantial performance improvements over propositional learning. For instance, the performance of content features is substantially increased in the relational case with “users only.” A fourth observation is that “not all neighbors are created equal.” In all cases, the use of information about user neighbors is substantially more important than hashtag neighbors. We found that most neighborhoods contained users; the distribution was none (11%), user only (43%), hashtag only (9%), and both (37%).

7.2.2 Hand-Annotated Results

Additional experiments on the three labels famous, generic, and info were also performed and results are shown in Table 2. Note that the small number of labels is most likely impacting two aspects of performance. First, since we have a smaller training set size, best absolute AUC is typically lower than the verified case. Also, in some cases performance is worse than chance showing that it is difficult to generalize well from a small training set. In general, though, similar trends in AUC performance are similar to the verified case.

For both the famous and generic labels with propositional learning, we found that network-only features work substantially better than content-only features. For info labels, content-only features are slightly better demonstrating that info is a unique case.

Further examination of relational results shows similar trends to the verified user case. Relational methods are superior to propositional methods. In addition, using both network and content features is helpful or at least not detrimental to performance. The case of the info user class is interesting from the viewpoint of neighbors; we see that the gap in performance between hashtag-only neighbors and user-only neighbors is less than other user classes. In terms of features appearing in the RPTs, we found that all features were valuable in user classification. Aggregate features were common as decision points in the relational case.

8 Conclusions

Twitter contains a rich set of social network, content, and individual user cues that give insight into user characteristics. In this paper, we explored features that captured these characteristics via topic clustering, offensiveness detection, and network analytics. Additionally, we examined various classification strategies which used both propositional and relational methods. Our different approaches demonstrated that the different feature types are complementary and all indicative of user classification. For example, finding “interesting” Twitter accounts (the opposite of generic users) can be accomplished with content and network features. This process emphasizes the fact that user classification can be accomplished with many strategies. Possible future work includes performing alternative classification studies, analyzing the effects of different sample sizes, extending to other languages, and predicting additional user classes.

Table 2: Average and standard deviation AUC for detection of verified and annotated accounts. Propositional results are indicated by a “-” in neighborhood structure.

Feature	Neighborhood	Verified	Famous	Generic	Info
Sets	Structure	AUC Avg (SD)	AUC Avg (SD)	AUC Avg (SD)	AUC Avg (SD)
Content	-	0.6201 (0.0219)	0.5549 (0.0215)	0.5725 (0.0390)	0.7716 (0.0436)
Network	-	0.6916 (0.0339)	0.7685 (0.0529)	0.7110 (0.0468)	0.6438 (0.0791)
Content+Network	-	0.7443 (0.0349)	0.7901 (0.0136)	0.7301 (0.0290)	0.7002 (0.1682)
Content	No Neighbors	0.5935 (0.0306)	0.5970 (0.1583)	0.5842 (0.1535)	0.4006 (0.1777)
Content	Users Only	0.8945 (0.0248)	0.8172 (0.0375)	0.7558 (0.0560)	0.6813 (0.1321)
Content	Hashtags Only	0.6501 (0.0416)	0.4939 (0.1548)	0.5557 (0.1518)	0.6153 (0.1151)
Content	Users and Hashtags	0.9213 (0.0202)	0.8420 (0.0564)	0.7790 (0.0589)	0.5583 (0.0954)
Network	No Neighbors	0.5123 (0.1292)	0.6227 (0.1707)	0.4085 (0.0505)	0.4040 (0.1224)
Network	Users Only	0.8760 (0.0310)	0.8383 (0.0262)	0.7708 (0.0527)	0.6433 (0.1336)
Network	Hashtags Only	0.4821 (0.0657)	0.4502 (0.1418)	0.5995 (0.0576)	0.5778 (0.1241)
Network	Users and Hashtags	0.9091 (0.0257)	0.8437 (0.0157)	**0.8014** (0.0559)	**0.8038** (0.0760)
Content+Network	No Neighbors	0.5939 (0.0801)	0.5654 (0.1141)	0.6572 (0.1299)	0.3917 (0.1487)
Content+Network	Users Only	0.8823 (0.0317)	0.8019 (0.0520)	0.7566 (0.0474)	0.5815 (0.1708)
Content+Network	Hashtags Only	0.6699 (0.0525)	0.4510 (0.1312)	0.5543 (0.0954)	0.4260 (0.2057)
Content+Network	Users and Hashtags	**0.9325** (0.0087)	**0.8431** (0.0535)	0.7831 (0.0192)	0.8021 (0.0608)

References

Glen A Coppersmith and Carey E Priebe. 2012. Vertex nomination via content and context. *arXiv preprint arXiv:1201.4118*.

Fred Morstatter, Jurgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. *Proceedings of ICWSM*.

Jennifer Neville, David D. Jensen, Lisa Friedland, and Michael Hay. 2003. Learning relational probability trees. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, Washington, DC, August. ACM Press, New York, NY. Poster session: Research track.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *ICWSM*.

John C. Platt. 2000. Probabilities for SV machines. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. The MIT Press.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. 2011. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*, pages 18–33. Springer.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*.

Chun-Yuan Teng and Hsin-Hsi Chen. 2006. Detection of bloggers’ interests: using textual, temporal, and interactive features. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 366–369. IEEE Computer Society.

Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM.