Author Identification on Twitter (Milestone Report)

Antonio Castro
antonio.alfredo.castro@gmail.com

Brian Lindauer
brian@shendauer.com

I. Introduction

As of June 2012, Twitter has 500M users, 140M of whom are in the United States. These users, especially those outside of the United States, may assume that they have a certain level of anonymity among this sea of tweets. Our project investigates whether the identity of an anonymous Twitter user can, in fact, be uncovered using only linguistic stylometry. Authorship recognition is a very well-studied domain, but the scale is almost always limited to no more than a few hundred authors. Narayanan, et al. [2] study authorship recognition at Internet scale, by looking at the characteristics of different classifiers when applied to a corpus of approximately 100k blogs. Starting with the set of features, classifiers, and normalization methods that yielded the best results over blog data in that paper we adapt them to Twitter data and measure the results.

II. Data Collection

Our goal was to obtain a training set across at least 1000 users and have at least 1000 tweets per user to enable our feature set to have enough data due to the sparsity of our planned data features. Our first data collection task was to identify an appropriate subset of the Twitter users and Twitter users in which we have prior knowledge that they are authored by the same user.

Table I describes the various sources we attempted gathering and how we ended up using the data. We initially utilized a web site called discovertext.com to follow the desired accounts and begin collecting data. However, we discovered that the site did not capture all of the data we would like to use in our feature set and the rate in which new tweets were produced did not provide enough history to give us the desired data size per user. A key data element we wanted to capture that was not available was the text of the original tweet in the case of a retweet. As a result we utilized the Twitter API to gather the last 1000 tweets of each of the users identified in Table I. Primarily due to Twitter rate limit limitations the data gathering took a few days to complete.

We identified a collection of Twitter account pairs, where one author is responsible for both accounts. Our primary methods for identifying those accounts were a request to employees of Dell with official Dell Twitter accounts, and using Google to search for phrases indicating multiple Twitter accounts. We also obtained some meta-information about Google Plus profiles from the authors of [4] and used that to target a crawl of Google Plus profile data. This produced an additional set of account pairs, but those cannot be used until we manually filter them to eliminate non-English feeds, feeds containing only links, etc.

Due to the small number of valid account pairs in which we have prior knowledge of the account being authored by the same user, we are currently opting to split the data set and use our generalization error calculations to simulate

Source	**Desired Usage**	**Actual Usage**
Klout	We attempted to gather the top thousand klout users as a method of gathering users with rich content.	Not used. Only small lists of top klout rankings are published and we were not able to obtain a list longer than 10-20.
Twitter Firehose	Obtain tons of tweets across a massive number of users.	Not used. While the data was broad, it did not provide us with a long enough history of any individual user to populate our sparse set of features.
Twitaholic	Provides a top 1000 most followed list of Twitter users.	Used as a starting point for a list of users to gather data from.
Dell Solicitation	Obtain a list of employees who maintain two seperate Twitter accounts.	Used to populate a set of known Twitter accounts that are authored by the same person.
Web Search	Search for phrases such as also follow me at on Twitter profile pages.	Used to populate a set of known Twitter accounts that are authored by the same person.
Google Plus Profile Scrape	Discover users that report having multiple Twitter feeds to be followed.	Not yet used as manual filtering of unusable feeds is still required.

Table I
Data Sources

separate users. This is not ideal, but it allows us to begin to hone the algorithm as we continue to collect a more comprehensive list of pairs.

Additionally, as we collected data we have noticed a handful of issues we will need to address, as they might introduce confounds into the experiment. Examples of these include foreign languages and tweets automatically generated by applications such as foursquare.

The results in this milestone report are based on a collection of 830 Twitter stream containing around 750k total tweets. These include 22 pairs of accounts that are known to be maintained by the same author. We plan to continue adding to this data collection between now and the final project report.

III. Features

Our initial selection of features was inspired by Narayanan [2], Writeprints [1], and Ireland [3]. It is, in fact, nearly a subset of the Narayanan features. These features aim to focus on the style of the tweet rather than its content. So, for example, we specifically look at function/stop words rather than ignoring them and focusing on words with large TF.IDF scores.

Category	Description	Count
Length	words/characters per post	2
Word shape	frequency of words in uppercase, lowercase, capitalized, camelcase, and other capitalization schemes	5
Word length	histogram of word lengths from 1-20	20
Character frequencies	frequency of letters a-z (ignoring case), digits, and many ASCII symbols	68
Unicode	frequency of non-ASCII characters	1
Function/stop words	frequency of words like “the”, “of”, and “then”	293

Table II
Current features, mostly from [2]

We are currently working with 389 features. There are a number of additional feature categories we plan to investigate over the next few weeks, listed in Table III.

Category	Description
Parse tree	features relating to the English syntax of the tweet
Function word categories	frequency of words indicating certain states of mind
Bigrams	frequency of character bigrams
Twitter conventions	frequency of Twitter-specific abbreviations such as RT, HT, and MT
Retweets	features relating to the users handling of retweets

Table III
Prospective Features

IV. Classifiers

Because Narayanan, et al. reported their best results with a combination of nearest neighbors (NN) and regularized least squares classification (RLSC), we decided to first implement these classifiers and run them against the data. We soon discovered that RLSC is quite complicated and decided to begin with just NN to get a working end-to-end example with which we could begin refining our classifier. During the remainder of the quarter, we do plan to return to RLSC to compare performance of NN and RLSC separately and in combination.

Focusing on NN, we implemented the variation described by Narayanan along with the normalization procedure from that same paper. Rather than keeping all data points in memory, which would be very expensive considering the number of Twitter users, we compute one centroid for each Twitter account. To do this, we read in all extracted tweet features from all users, then normalize by both column and row. First, each column value is normalized by the mean of the non-zero values in that column. Then, each row value is divided by the norm of that row.

At prediction time, we read the extracted features of each tweet in the test stream. For each of those tweets, we measure the Euclidean distance to each of the centroids computed in training. Then we take the sum of the distances of all the tweets to each of the centroids and rank the centroids by their average nearness to a tweet in the test stream. We ask whether the account by the same author appears in the top N% of the ranking, including whether it is ranked first. Because at the time of this writing, we only have 22 labeled pairs of Twitter accounts, we also measured generalization error using 70/30 cross validation on tweets from the same account. In computing both training error and cross validation generalization error, we counted a prediction as correct only if the correct account appeared first in the ranked list. Our computed training error was 0.61%. Our computed cross validation generalization error was only 2.5%. This generalization accuracy is surprising considering that we have yet to implement any grammar-based or Twitter-specific features. Our feature set is composed largely of one-grams and function words.

Though our sample size is small, we are able to use our 22 labeled account pairs to get a measure of accuracy when the algorithm is applied to our intended use case. Given one of the accounts in the pair, the classifier ranks the other account first 18% of the time – an error rate of 82%. But the correct author is ranked at least 2nd 27% of the time. And they appear in the top 10% of the ranking nearly 60%

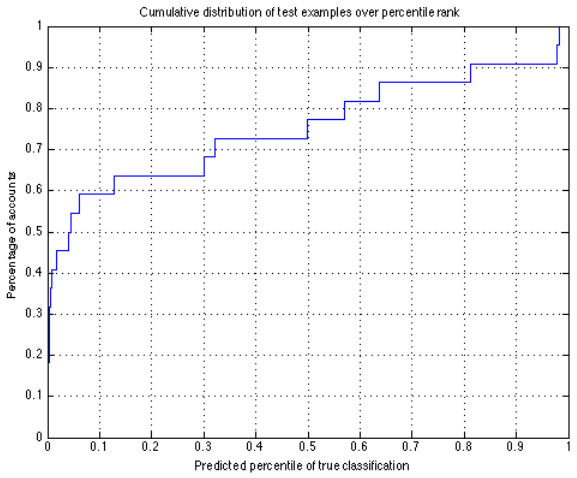

Figure 1. Cumulative distribution of percentile rank over test examples

of the time. Figure 1 shows the cumulative distribution over these rankings.

V. Continuing Work

We believe that by improving our featureset and adding RLSC, we will be able to further improve our classification accuracy. There are a number of tasks we hope to work on in order to both increase our accuracy and to increase our understanding of our results.

- Adding the regularized least squares classifier.
- Adding additional features.
- Studying which features are most predictive, and possibly removing those that are not.
- Studying the effect of aggregating tweets for prediction rather than evaluating each tweet's distance individually.
- Studying the effect of the number of available tweets in a feed on prediction accuracy.
- Continuing to collect data from additional arbitrary streams to better understand how the algorithm might perform at larger scales.
- Add more labeled Twitter account pairs to help smooth the effect of odd accounts (both positive and negative) in our test data.

VI. Conclusion

We are encouraged by our initial result. The algorithm exhibits excellent accuracy in cross validation, and even does fairly well in the general use case where the test stream comes from a different Twitter account by the same author. We expected less accuracy from Twitter than from the blog or email data used in previous work becuase the amount of data is so much smaller. So the work is promising.

References

[1] Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." ACM Transactions on Information Systems 26.2 (2008): 7.

[2] Narayanan, Arvind, et al. "On the feasibility of internet-scale author identification." Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012.

[3] Ireland, M.E., & Pennebaker, J.W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. Journal of Personality and Social Psychology, 99.

[4] Perito, Daniele, et al. "How unique and traceable are usernames?." Privacy Enhancing Technologies. Springer Berlin/Heidelberg, 2011.