Stylometry for Online Forums

Kurt Barry
Katherine Luna

CS 229 : Machine Learning
Stanford University
Autumn Quarter 2012

Abstract

We apply stylometric techniques to determine the authors of posts in online forums. After selecting a small number (15) of features by information gain, we used the Naive Bayes and Support Vector Machine algorithms to perform multiclass classification. The SVM achieves the best performance, being 53% accurate on the six most frequent posters in the data set and 33% accurate on the top 46 (using ten-fold cross validation). Accuracy is limited by short posts and the number and types of features used, but proof-of-concept is achieved.

1 Introduction

One of the world wide web's most integral features is the possibility, in principal, of complete anonymity. Though today's web is not anonymous by default, tools like Tor [1], encryption, and careful browsing habits can render users difficult to trace and hamper the association of multiple online identities with a single behind-the-scenes individual. While this may be a good thing for political dissidents living under oppressive regimes, in other circumstances it would be desirable to remove anonymity to whatever extent possible, e.g. in the case of law enforcement monitoring the communications of criminals or terrorists over cyber-channels. In particular, the discussion forums of such malicious actors are of great interest. Interesting questions include: who are these people (could anonymous posters in forums be linked to Facebook accounts)? On what other sites do they post? Are any of them using multiple usernames (i.e. how many distinct individuals exist in an organization)?

This project applies machine learning techniques to perform stylometric analysis on online forum posts. Stylometry, the study of linguistic style, was used long before the internet existed to resolve historical authorship disputes. In the information age, it has been studied as a means to identify bloggers. It is essentially the only tool one has to attempt to reveal or track anonymous entities in web forums. This environment presents challenges, however. The very short length of many forum posts yields little stylometric information. Many users are only active on a forum for a brief time, making few posts before their activity ceases. It is an open question as to what extent stylometry can succeed in such an environment.

2 Data

The data were obtained by scraping posts from an online forum (name omitted to respect privacy) via a custom-written PHP script. The collected corpus contains 331 unique usernames and 16,052 posts. A small core of long-term participants accounts for a majority of these posts; the 31 most frequent posters are responsible for 73% of all posts. The six most frequent posters account for 41% of all posts. Some basic sanitization was performed, primarily removing quotations to prevent confusing text written by different authors.

3 Methods

There are essentially two components to stylometry: feature selection and classification algorithm. Most features were selected from previous literature [2, 3], though a few, like time-of-posting and use of emoticons, were unique to this project. Information gain was used to select the most useful features. This quantity is defined, for a feature distributed among posts according to a probability distribution X, as: $IG(Y|X) = H(Y) - H(Y|X)$. Y is the probability distribution for a post to be written by a each author, $H(Y)$ is the entropy[1] of that probability distribution; $H(Y|X)$ is the conditional entropy of the author distribution given the distribution of the feature in question. Since all the features were continously-valued, they had to be discretized before information gain could be calculated. This was done via a heuristic algorithm based on the notion that bins should on average have a width proportional to the average standard deviation of that feature among all users, but should also contain roughly equal numbers of examples. We used the top fifteen features as ranked by information gain. Using more features than this number involved diminishing returns, likey linked in part to the high frequency of short, ambiguously authored posts. In some cases adding features even worsened performance. Keeping the list short has the added benefit of computational efficiency. A very serious implementation would use hundreds of features, but this was unnecessary for a proof-of-principle exercise. Brief descriptions of the selected features are found in Table 1.

Feature	**Description**
Punctuation Fraction	Fraction of characters that are punctuation
Time	Time of posting
Apostrophes Per Word	(# of apostrophes)/(# of words)
Uppercase Fraction	Fraction of letters that are uppercase
Characters Per Word	(# of characters)/(# of words)
Letter Fraction	Fraction of characters that are letters
Complexity	(# of unique words)/(# of words)
Number of Characters	Number of characters in the post
Number of Words	Number of words in the post
Whitespace Fraction	Fraction of characters that are whitespace
Digit Fraction	Fraction of characters that are decimal digits
Bigraphs lc, co, me, and we	Frequency of these two-character strings

Table 1: **The features used in the first implementation of the algorithm.**

[1]Entropy is defined for a probability distribution P as $H(P) = -\sum_i p_i \log p_i$.

The top features can be seen in Figure 3.1. The same set of features was found to be best for all users with > 50 posts and > 500 posts, though the ordering changed slightly; for example, time-of-posting is more important when users with fewer posts are included, possibly because the more frequent posters have broader distributions of posting times. It is interesting that some useful features are either not available or less powerful in other stylometric applications. For example, time-of-posting makes little sense in the context of books or newspaper articles (although it could conceivably be applicable in the blogosphere). Others, like the fraction of characters that are punctuation, the average number of apostrophes per word, and the fraction of characters that are upper case, are likely so useful in forums precisely because there is little incentive for individuals to follow the established rules of English orthography beyond making the effort necessary to be understood. This is an important aspect of stylometry in forums that differs from other application regimes.

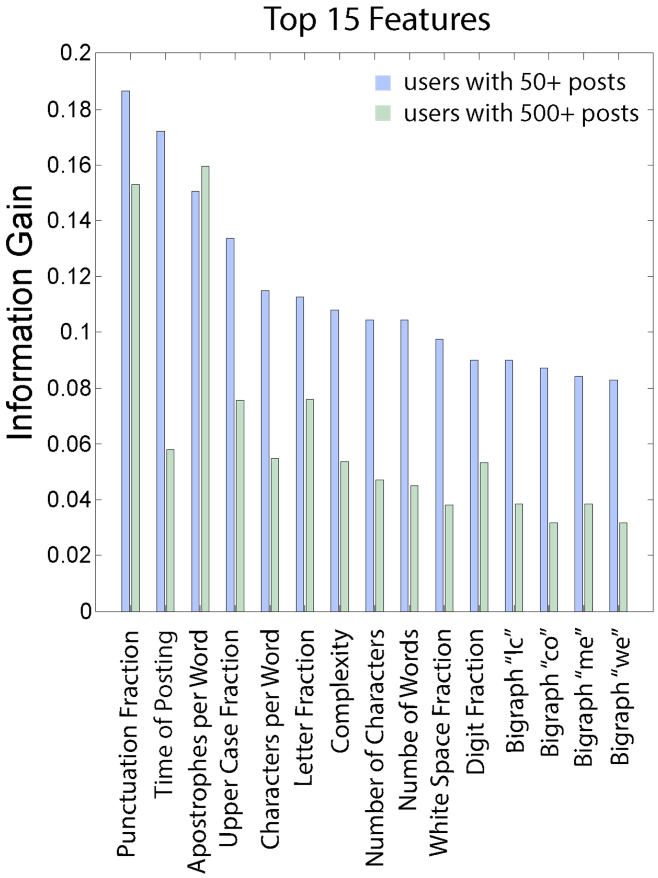

Figure 3.1: **The top features, ranked by information gain.**

Two algorithms were used. The first was a multinomial Naive Bayes classifier with Laplace smoothing. It was implemented in Python and used the same discretizations used to calculate the information gain. The second was a support vector machine (SVM). Rather than implement our own, we used the SVC support vector machine in the open source `scikit-learn` [4] Python module (which itself uses libsvm [5]). Multi-class classification is handled by constructing a binary classifier for each pair of users; the decision process is then treated as an "election" in which each classifier "votes" for a particular user [5]. Since the various features all had different value ranges, we rescaled the extracted features to have mean 0 and standard deviation 1 before running the SVM. A number of different kernels were experimented with: linear, exponential ($e^{-\gamma<x,x'>^2}$), and polynomial kernels of various degree. The best

	Algorithm Accuracy		
	Naive Bayes	SVM	Random Chance
users with > 500 posts	40.5%	53.8%	16.7%
users with > 50 posts	20.4%	32.9%	2.2%

Table 2: **Algorithm performance on the top six and top forty-six most frequent posters, compared to random chance.**

performing kernel was found to be quadratic in the inner product of the feature vectors, specifically $K(x, x') = (0.5 < x, x' > +9)^2$. To give some quantitative sense of this kernel's superiority, its accuracy is roughly six percentage points better than a linear kernel. Using higher-degree polynomials does not yield improvements in accuracy.

4 Results

The performance of these algorithms is summarized in Table 2. Two (non-disjoint) subsets of users were tested: those with more than 500 posts (six individuals) and those with more than 50 posts (forty-six individuals). Ten-fold cross validation was used to assess the accuracies. The selected folds were the same for both algorithms. In all cases random chance is outperformed by a significant amount. Though the accuracy is reduced for the larger set of users, it is actually better compared to random than for fewer users. A major factor limiting performance is very short posts, which contain little stylistic information and are thus difficult to attribute. As an example of this, training and testing on only posts with more than ten words for those users with greater than 500 posts improves the accuracies to 44.6 and 57.1 percent, for Naive Bayes and the SVM, respectively. To give some indication of the challenges involved in separating the data with the chosen features, Figure 4.1 shows an example SVM classification using three posters and two features. The substantial overlap of the data highlights the need to use many features and nonlinear decision boundaries.

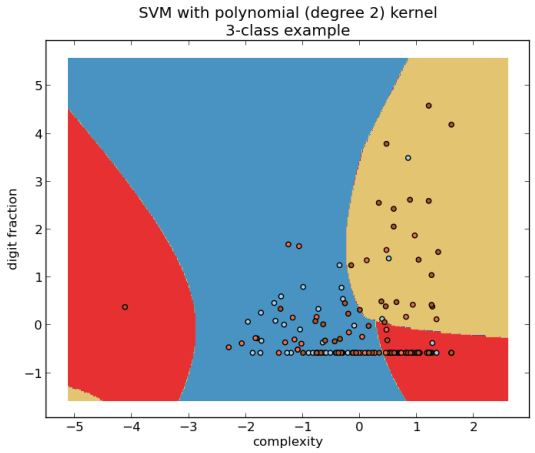

Figure 4.1: **An example SVM classification with limited data, using only two features.**

5 Conclusions

This project represents proof-of-concept for stylometry in online forums where data is often noisy and less abundant than in other written environments. Improvements could certainly be made. It is difficult to imagine significant progress on attributing extremely short posts (e.g., "Welcome to the forum!"), but the use of more extensive feature sets may improve accuracy for longer posts. Examples of promising features include "function word" frequencies (see [3]) and recognition of particular phrases or grammatical constructs preferred by posters (which would require some level of natural language processing to extract). The objective of long-term interest, which was not demonstrated here, remains cross-domain attribution, so that actors in an anonymous context can be linked with known actors in a non-anonymous context by posting style.

References

[1] Tor Project: Anonymity Online `https://www.torproject.org/`

[2] Andrew W.E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, Rachel Greenstadt "Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization" The 12th Privacy Enhancing Technologies Symposium. 2012.

[3] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, Dawn Song: On the Feasibility of Internet-Scale Author Identification. IEEE Symposium on Security and Privacy 2012: 300-314

[4] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

[5] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`