Robust and Effective Feature Selection for Opinion Spam Detection

Rakesh Verma[1] *Ekene Sibeudu*[2]
(1) Computer Science Dept., University of Houston, TX 77204
(2) Computer Science Dept., University of Maryland Baltimore, MD
rmverma@cs.uh.edu, bac1087@gmail.com

Abstract
Social networks such as FaceBook and Twitter are prone to the problem of postings by automatic programs called Bots. In this paper we consider the problem of automatic detection of posting author: human or bot. In previous work, researchers had introduced a third category called Cyborg, which denotes either a human assisted by a bot or a bot assisted by a human. For simplicity, we collapse the cyborg with the human category here. Using robust techniques, we present a classifier that is more than 86% accurate on this task using Twitter Data.

1 Introduction

Recently there has been a dramatic increase in the number of social networking websites such as FaceBook and Twitter. In this paper we focus on Twitter, which is online social network and micro-blogging tool released in 2006. The growing population and open nature of Twitter has made it a target for exploitation by automated programs knows as bots.

We study the problem of automatically identifying whether a tweet was authored by a human or a bot, and build a classifier that achieves over 86% accuracy using a small set of robust features for classification.

The organization of the rest of this paper is as follows. Section 2 presents the most relevant related work. Section 3 describes the data collection process. Section 4 provides the data parsing procedure and Section 5 the data analysis and results. Section 6 concludes the paper.

2 Related Work

For brevity, we refer the reader to [2] for a more thorough discussion.

Chu et al. [1] classified Twitter users as either Human, Bot, or Cyborg. Their classification system was based on four components, (1) Entropy, which looked at the timing of the messages; (2) Machine Learning, uses the content of the tweet to detect spam; (3) Account properties, including url ratio, device makeup, and verified accounts among others; and a (4) Decision maker which using the previous three components for classification.

Zhang and Paxson's work [3] focuses on detecting automated activity on Twitter. They believe that if an account is highly automated it will exhibit timing patterns that non automated users do not exhibit. They concluded the presence of automation if tweet times are either not uniform enough or too uniform.

Song et al. looks at the sender receiver relationship to determine spam. Their work focuses on looking at the distance and connectivity of user pairs. They postulate that almost all messages that come from a user whose distance is more than four are spam.

Our work uses metrics based on the features of tweets. They are the url ratio, re-tweets, hashtags, and mentions. These are features of Twitter that users can use in their tweets. We found that bots

have a tendency to have a high number of these features present in their tweets. Our work has some similarity, but important differences from these works. A similarity is the common use of a URL ratio metric. Where these works focus on the overall picture including the use of timing we are focused more on individual tweets. We believe that timing is not a robust feature, since bots could use randomization in the future to escape detection.

3 Data Collection

We used Twitter4j - the Java library for the Twitter API. Using the Twitter REST API, which is rate-limited to 350 calls per hour, we were able to collect 20 tweets for 15 users per hour. We switched to the Streaming API to collect a random sample of all public statuses. Using the Streaming API allowed us to collect a significantly larger amount of tweets, but they were chosen at random. This process, though increasing the number of overall users, decreased the number of tweets per user. To circumvent this issue we collected for long periods of time to increase our chances of having more tweets per user.

4 Data Parsing

We developed a parser to parse through files containing millions of tweets. The parser grouped users into different folders based on the first two characters of the username. In these folders, the parser creates a file for each user containing all of their collected tweets. Once these files were created for each user another parser we created was used to extract the metrics we used for our classification system.

5 Data Analysis

Ground Truth Creation. One of the biggest challenges is to construct a significant amount of gold standard test data since this task must be done manually. We selected a random subset of 100 users from our tweet data to classify manually as either Bot or Human.

Out of the 100, 20 did not have enough tweets for classification. We used the classification tool Weka[1] to build two different kinds of classifiers: naive bayes and naive bayes updatable. We used stratified 10-fold cross-validation.

Results. We were able to correctly classify over 86% percent of the ground truth data using the naive bayes updatable classifier. Hence, we believe that our metrics shows promise and with more ground truth data we can get better results.

6 Conclusions

We presented metrics that are more robust and close to being as effective as earlier work. For the future, we plan to construct more ground truth data and explore more detailed tweet analysis.

References

[1] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *ACSAC*, pages 21–30, 2010.

[2] J. Song, S. Lee, and J. Kim. Spam filtering in twitter using sender-receiver relationship. In R. Sommer, D. Balzarotti, and G. Maier, editors, *RAID*, volume 6961 of *Lecture Notes in Computer Science*, pages 301–317. Springer, 2011.

[3] C. M. Zhang and V. Paxson. Detecting and analyzing automated activity on twitter. In N. Spring and G. F. Riley, editors, *PAM*, volume 6579 of *Lecture Notes in Computer Science*, pages 102–111. Springer, 2011.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`