UTAH STATE UNIVERSITY, COMPUTER SCIENCE DEPARTMENT

CS 7930 Social Media Mining
Homework 2

Gopal Menon

March 26, 2016

1. **Task 1**
   These are the five features I decided to use in order to distinguish between legitimate users and spammer, and the reason why:

   a) **Number of Followings:** I reasoned that a legitimate user would follow the users he was interested in, while a spam user would follow as many people as possible. I have seen from personal experience, that my Twitter account is followed by some people who do not seem to be legitimate users.

   b) **Number of Followers:** I thought that a legitimate user would have some followers, while a spam user would not have many. It was possible that spam users would follow each other, but I thought I would include this feature all the same.

   c) **Number of Tweets:** I assumed that a spam user would have a large number of tweets, while a legitimate user would only have a reasonable number.

   d) **Number of Url Tweets:** Since spam users try to sell services or products, I decided to have a url count in tweets as a feature in order to identify spammers.

   e) **Change in Following:** Following churn would be a good indicator of a spammer and so I decided to use this as a feature.

   In order to extract the features from the training and testing data sets, I used Microsoft Excel to import the data as tab separated text. In the case of the file with the following count, I used both tabs and commas as separators while doing the import. I used the Excel Search function to look for a url within the tweet text. Then I used a pivot table to summarize the data so that I got the total count of tweets containing a url for a user. For finding the following churn, I used the absolute value of the difference between the first and last following counts. Once all data files were imported into Excel and the url containing tweets were summarized, I pasted the data side by side onto a single worksheet. I used an Excel formula to identify places where the Twitter User Id did not match on the same row as some data was missing. For these cases, I aligned the data by shifting it down, till the Excel formula showed a match.

Once the data were available, I saved the Excel worksheets for training and testing data as tab separated text files that I could import using Python. The extracted data contained the columns listed above and in addition, also contained columns for length of screen name, length of user profile description, number of tweets in the tweets file per user and the class label.

2. **Task 2**
   I used the following classifiers for identifying the spammers:

   a) AdaBoost Classifier

   b) Decision Tree Classifier

   c) Logistic Regression Classifier

   The classifiers report spam as the positive, and so a false positive rate corresponds to the legitimate users predicted to be spam users. The classifiers print the confusion matrix, F1 score, Precision and Recall. The False Positive Rate (FPR) and False Negative Rate (FNR) were computed using a calculator on the numbers in the confusion matrix.

Classifier		
AdaBoost	Decsion Tree	Logistic Regression
F1 - 0.95 FPR - 0.028 FNR - 0.076	F1 - 0.90 FPR - 0.110 FNR - 0.089	F1 - 0.93 FPR - 0.104 FNR -0.029

3. **Task 3**
   Here are some ways in which the quality of the classifiers could be improved:

   a) **Better Following churn measure:** I had used the absolute value of the difference between the first and last following count as a feature. This feature could be improved to better represent following churn so that ups and downs in the following count are captured. If the absolute value of the difference between subsequent following counts are added up, we would get a better picture of the churn and the feature would be a better spam predictor as spammers have a higher value of following churn.

   b) **Identify spam Urls in tweets:** Since spammers try to influence other users, the urls present in their tweets would be related to advertising, trolling or anti-social behavior. If we could classify urls based on the above and other spam related types, we could use the count of such urls as a feature and not just the url count. This would help in better identifying spamers.

   c) **Categorize Tweet Sentiment:** If we categorize tweet text sentiment as related to advertising, trolling or anti-social behavior, we could use the count of such tweets as a feature to better identify spammers.