Detecting Fake Tweet Activity based on NLP and Sentiment Analysis

Final Project Report

TCSS-554 : Information Retrieval

J. McFadden, E. Han, K. Chandrasekaran,
S. Ananthapadmanabha, K. R. Kalyanam

1 Problem Description

Our proposal is to analyze Twitter data to detect fake/bot accounts and fake retweet activity by any account. Specifically, our first goal is to be able to determine if a given account is a fake/bot account. If time permits we would like to extend this by quantifying the probability of a given account being fake as a percentage. This will be followed, and complimented, by our second goal, detecting fake retweet activity. Should time permit, we will again seek to quantify the probability that a given retweet is fake, in other words automated, activity. This problem is important due to the large quantity of false and misleading information circulating on social media; and solving even part of it would help to improve the dialogue that all free and open societies must have.

2 Proposed Solution

Our solution aims to provide an efficient method of determining if a given post is generated by an actual human or an AI bot. Since even sophisticated AI bots have difficulty expressing emotions, we are using this weakness to help hunt for them. We hypothesize that the manner in which bots express emotions in text should be markedly different from the manner in which humans express emotions. Therefore our model seeks to quantify the emotions expressed in the text of a given tweet so that various machine-learning algorithms can be trained to detect the difference between human and bot generated Tweets. If time permits, we would like to extend this model so that it

provides a probability that a given tweet is real or fake.

We quantify the emotions of a given tweet using the NRC and Afinn natural language processing and sentiment analysis frameworks. The NRC framework consists of ten emotions with each word in the English language being associated with one of those ten emotions. This was used to quantify emotions in a given tweet by determining the frequency with which emotions occurred in each tweet. The other framework used, Afinn, provided an overall score for the emotion of each tweet based on the words in the tweet. Each word in a given tweet was scored according to the Afinn framework, then these scores were all added for that tweet to determine its Afinn score. A negative Afinn score would represent a negative tweet, while a positive Afinn score would represent a positive tweet.

This model is then used to produce what we are calling the *Tweet-Emotion Matrix*. This matrix has thirteen columns. The first column contains the ID numbers of the tweets. The next ten columns contain the scores for the ten emotions in the NRC framework. The twelfth column contains the score from the Afinn framework; and the thirteenth column contains a binary classifier which classifies tweets as their *real* (1) or *fake* (0). Finally, the rows of this matrix represent individual tweets, while the elements of the matrix represent the strength of a given NRC emotion, the Afinn score, or the real/fake classifier for the tweet represented by the row. Ultimately, this *Tweet-Emotion Matrix* will be run through various machine learning algorithms.

Running the data from the *Tweet-Emotion Matrix* through machine-learning algorithms serves two purposes. The first purpose is to determine if there is any statistically significant correlation between the emotional data in the matrix and the classification of a tweet as real or fake, as well as quantify any statistical significance based on both the t-test and a ten-fold cross validation for accuracy.