Development of an Authorship Identification Algorithm for Twitter Using Stylometric Techniques

Zou, Cherry

I developed software that implements semi-supervised learning to dramatically improve accuracy when stylometrically attributing an unidentified tweet to the correct author from a set of known Twitter authors. Existing stylometric techniques generally do not perform well on short texts. Software written in Python streamed, preliminarily processed, and stored 1000 tweets each from up to 30 prolific authors on Twitter. Traditional and flexible bigrams, as well as their frequencies of occurrence, were extracted from both the authors' known tweets and the unknown tweet, forming each author's profile. These bigrams were then used as tokens for a Naïve Bayes classifier which returned the probability of each author having written the unknown tweet. The first, second, and third most likely authors were determined by the classifier and written as output. After repeating this process multiple times, the percent accuracy of identifying the correct author was calculated. A program was completed that would, to a significant degree of accuracy, identify the author of an unknown tweet. Furthermore, it was found that excluding retweets, using a combination of flexible and traditional bigrams, and other techniques produced the most effective algorithm for stylometrically identifying the author of a tweet. With 10 authors, the algorithm correctly identified the author of the tweet with 73 percent accuracy on the first guess and with 87 percent accuracy within the top three guesses, showcasing the potential of stylometric techniques in application to extremely short messages. Moreover, this algorithm has significant potential in investigating anonymous cyber-crimes committed over social media.

Awards Won:

National Security Agency Research Directorate : Award of $5,000 for outstanding project in the systems software category.