CLASSIFICATION OF TWITTER TRENDS USING FEATURE RANKING AND FEATURE SELECTION

Abhishek Shah

Dr. Wenjun Zeng, Thesis Supervisor

ABSTRACT

Twitter scales 500 million tweets per day and has 316 million monthly active users. The majority of tweets are in the form of natural language. Using natural language makes it difficult to understand Twitter's data programmatically. In our research, we attempt to solve this challenge using various machine learning techniques.

This thesis includes a new approach for classifying Twitter trends by adding a layer of feature selection and feature ranking. A variety of feature ranking algorithms, such as TF-IDF and bag-of-words, are used to facilitate the feature selection process. This helps in surfacing the important features, while reducing the feature space and making the classification process more efficient. Four Naïve Bayes text classifiers (one for each class), backed by these sophisticated feature ranking and feature selection techniques, are used to successfully categorize Twitter trends. Using the bag-of-words and TF-IDF rankings, our research provides an average class precision improvement, over the current methodologies, of 33.14% and 28.67% correspondingly.