Classification of twitter data from the 2012 Emilia-Romagna earthquake by machine learning: comparison of k-nearest neighbors, kernel support vector machine, and string kernel methods

Franc JM, Ingrassia PL, Boniolo E, Carenzo L, Della Corte F.

INTRODUCTION: During large-scale disasters, social media may give insight into the societal implications, concerns, and sentiments of the affected area. Twitter is a commonly used social media and may represent a valuable source of information. However, as tweets are generally formed of unstructured text, they can be difficult and time consuming to analyze. The present study compares the ability of several machine learning algorithms to classify tweets from the 2012 Emilia-Romagna earthquake into meaningful categories. Machine learning was performed using three algorithms: k-nearest neighbors (KNN), kernel support vector machine using a term-document matrix (KSVM), and string kernel support vector machine (SKSVM). Accuracy of machine learning classification was compared to that of a group of three skilled reviewers. The null hypothesis of no difference between the three machine learning methods in accuracy of classification was tested against the alternative hypothesis that significant difference exists.

METHODS: A listing of Italian language tweets for the first 24 hours following the Emilia-Romagna earthquake was obtained from the commercial vendor GNIP using the keywords “Terremoto” and “Emilia”. Two trained reviewers independently classified each tweet to one of five groups: A) seeking information, B) giving general information, C) opinion, sentiments, or emotions, D)media or organizational information, and E) links to other information. When the two reviewers did not agree, a third reviewer chose between the two discordant assignments to determine the final assignment. For the purpose of machine learning, the set of tweets were randomly divided into a training set (10%) and a test set (90%). Machine learning was performed using R: A Language and Environment for Statistical Computing. The text mining package ‘tm’ was used to form a term-document-matrix from the unstructured tweet data. The R function knn() from the package ‘class’ was used for the KNN analysis of the term-document-matrix. For the KSVM analysis the function ksvm() from the package ‘kernlab’ was used to analyze the term-dot-matrix using the ‘LaPlace’ kernel. Finally, for the SKSVM analysis the string kernel ‘stringdot’ from the package ksvm() was used to analyze the unstructured tweet data directly. Differences between the cross-agreement accuracy of the three groups was tested using a three sample proportion test with $p<0.05$ considered significant.

RESULTS: The data consisted of 14190 tweets. Cross-agreement between the initial two reviewers was 82%; in 18% (2603 cases) the two reviewers did not agree and a third reviewer was needed to assign the final score. Final number of tweets per group was A) 144, B) 6642, C) 3129, D) 1120, and E)

3155. Cross-agreement between the machine learning assignment and the final reviewers assignment for the three machine learning algorithms was: 62% for KNN, 78% for KSVM, and 81% for SKSVM. SKSVM was statistically significantly more accurate ($p<0.0001$).

DISCUSSION: Of the three methods of machine learning, string kernel support vector methods were more accurate in replicating the assignments given by the three human raters. String kernel methods appeared to be nearly as accurate as the human raters for classification of tweets. Manual analysis of such large amounts of information would be difficult in real-time during large-scale disasters. An automated machine-learning based system would enable investigators to monitor twitter feeds and select useful information from the large amount of noise.