Jan Chołoniewski, Julian Sienkiewicz, Janusz Hołyst

Data mining-driven prediction of Twitter activity during Olympic Games 2012

The poster presents results of data mining-driven analysis of correlated emotional data as well as analyses of classification accuracy and possibility of data mining methods applications to forecast changes in activity and emotional data series derived from Twitter during sport events. The data was delivered by CyberEMOTIONS project partners from University of Wolverhampton. It consists of comments added by Twitter users at most 500 km from London, just before and during Olympics 2012 in London, and marked with hashtags related to sports. Each tweet had its emotional content scored, for both positive and negative emotion, using SentiStrength classifier. Also exact timings of events when British athletes won a medal were available.

The sets of observations and assigned classes were used for classifiers training and testing. A naive Bayes classifier, linear and quadratic discriminant analyses, a regression tree and supporting vector machines were applied. The principal component analysis was also performed and classifications using new components were conducted. A performance of classifiers was analyzed with application to three different classification problems and comparison between scoring parameters. Each time all the possible combinations of ,,method-set of variables" were tested with restraint of maximum of 4 variables in a set.

The first problem solved by classifiers was to forecast next step's change (increase or decrease) in the activity series.The second task was to forecast if the activity in the next step would exceed the threshold of 500 posts per 15 minutes. The last of the studied problems was to assess if a given activity peak was caused by significant sports event (i.e., a British athlete scoring a medal).