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

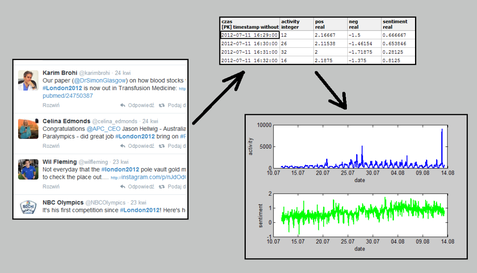
Jan Choloniewski, Julian Sienkiewicz and Janusz Holyst
Faculty of Physics, Warsaw University of Technology

Introduction

- The poster presents results of study on British Twitter activity during Olympic Games 2012 in London. Prediction potential of few classification methods was tested on three benchmark problems. Variables were derived from raw data using an aggregation and a text emotional scoring with a classifier.
- The data and SentiStrength classifier was delivered by partners from University of Wolverhampton in a frame of CyberEMOTIONS project.

Data

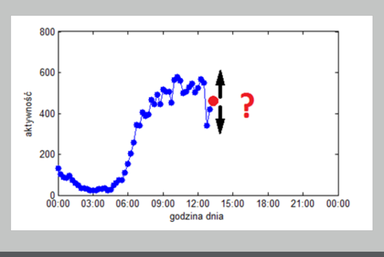
- The data consist of over 1.5 million of tweets (140 character-length text messages published on Twitter website) sent by British users (up to 500 km from London) marked with hashtags related to sports and Olympic Games (e.g. *#100mrun, #london2012...*).

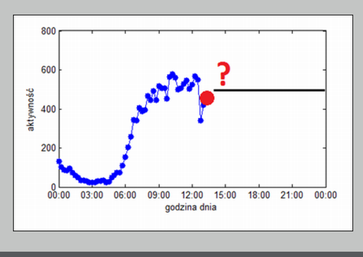


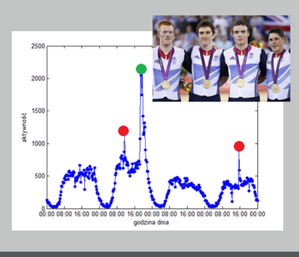
- Each post had its positive and negative emotion scored with SentiStrength
- Data aggregation – whole observation period has been divided into 3134 time windows ($\mathbf{T = 15min}$). A number of posts in a given window would be called activity; a mean value of difference between positive and negative score would be called sentiment.
- Following variables were available for methods: ($\mathbf{t}$ — current time window)
  - ▷ day$(\mathbf{t})$,
  - ▷ activity$(\mathbf{t-2})$,
  - ▷ sentiment$(\mathbf{t-2})$,
  - ▷ 1st and 2nd derivative of activity$(\mathbf{t-2})$,
  - ▷ 1st and 2nd derivative of sentiment$(\mathbf{t-2})$,
  - ▷ mean post length$(\mathbf{t-1})$,
  - ▷ hour$(\mathbf{t})$,
  - ▷ activity$(\mathbf{t-3})$,
  - ▷ sentiment$(\mathbf{t-3})$,
  - ▷ 1st and 2nd derivative of sentiment$(\mathbf{t-1})$,
  - ▷ fraction of unique users$(\mathbf{t-1})$,
  - ▷ activity$(\mathbf{t-1})$,
  - ▷ sentiment$(\mathbf{t-1})$,
  - ▷ 1st and 2nd derivative of activity$(\mathbf{t-1})$,
  - ▷ fraction of reply tweets$(\mathbf{t-1})$;

Benchmark problems

A. Trend forecasting (will activity be higher or lower in the next time window?)
B. Threshold exceed forecasting (will activity exceed given 500 tweets per 15 minutes threshold in the next time window?)
C. Peak classification (is given peak caused by a a medal event?)







Methodology

- For every pair of a method (below) and a set of variables (up to 6 for **A** and **B**, up to 4 for **C**), a classifier has been trained on randomly chosen 80% of observations and then scored on latter 20%. The procedure has been repeated 10 times for each classifier.
- Two main scoring scales were accuracy (fraction of good classifications) and F_1 score (harmonic mean of precision and sensitivity).

Applied classifiers:
- linear discriminant analysis **(LDA)**,
- naïve Bayes **(NB)**,
- quadratic discriminant analysis **(QDA)**,
- regression tree **(REG TREE)**,
- supporting vector machines **(SVM)** with various cores;

SVM cores:
- linear **(SVM LIN)**,
- tangential **(SVM MLP)**,
- polynomial (3rd order) **(SVM POLY3)**,
- quadratic **(SVM QUAD)**,
- radial-based function **(SVM RBF)**;

Results

A. Trend forecasting

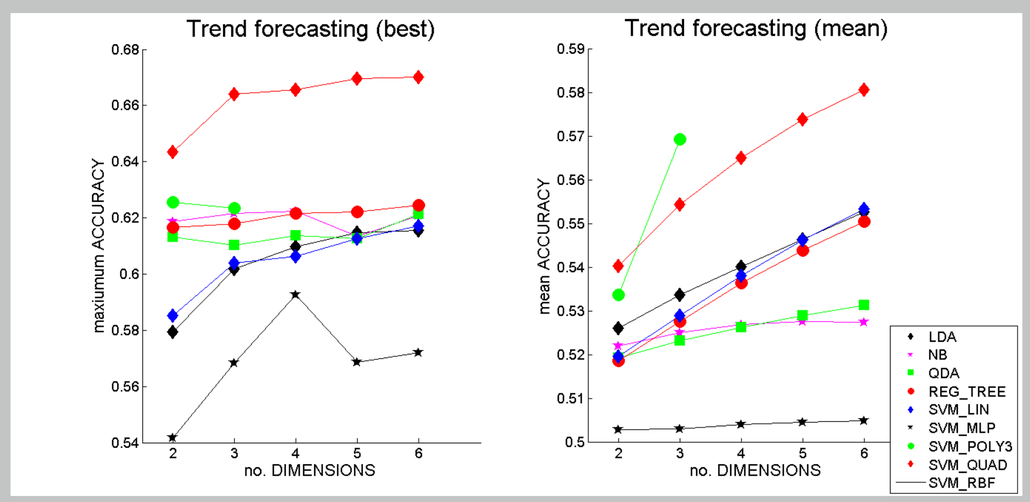


Figure 1: Maximum (left graph) and mean (right) accuracy of classifiers as a function of a number of variables (dimensions); standard deviations were not plotted for the sake of readability, standard deviations are of order of **0.02**

B. Threshold exceen forecasting

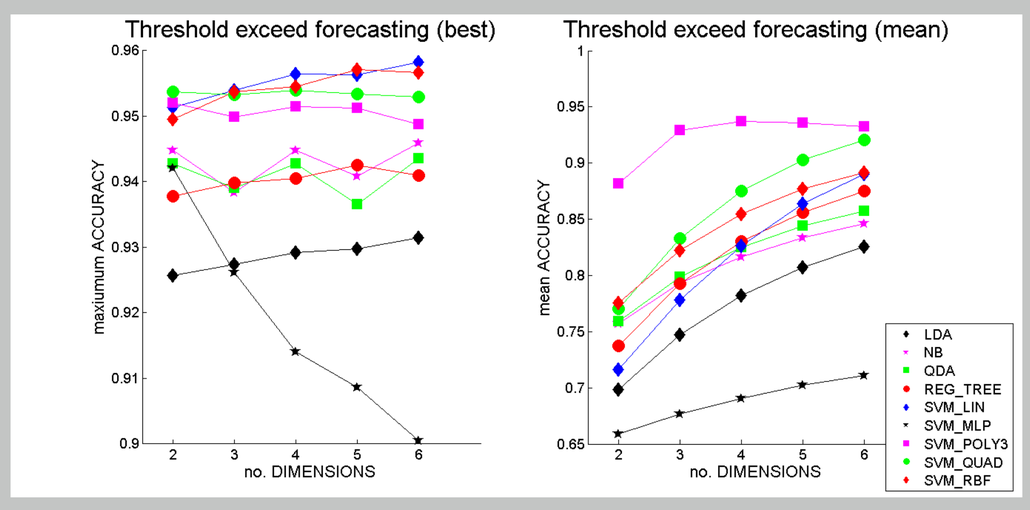


Figure 2: Maximum (left graph) and mean (right) accuracy of classifiers as a function of a number of variables (dimensions); standard deviations were not plotted for the sake of readability, standard deviations are of order of **0.02**

C. Peak classification

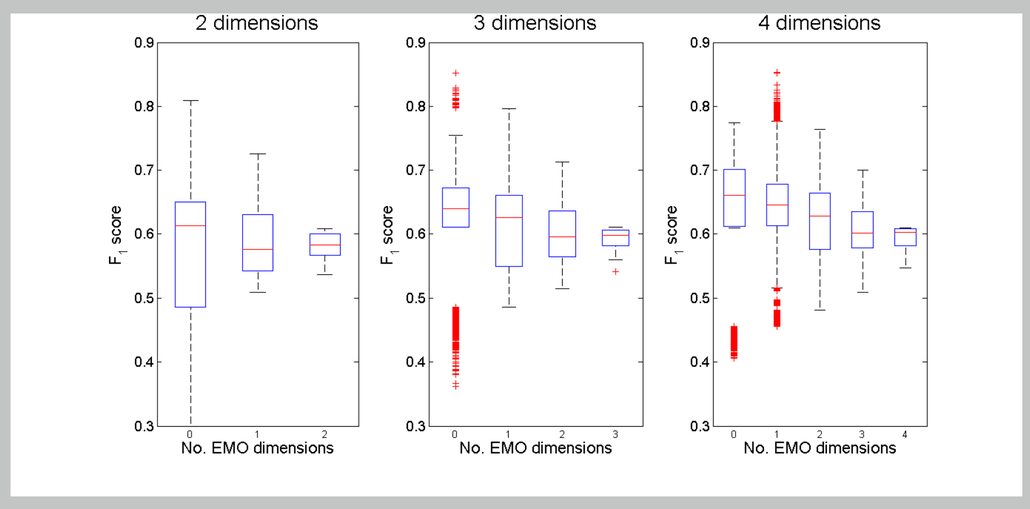


Figure 3: Box plot of F_1 score for different numbers of variables used by classifier as a function of number of emotion-related variables;

Conclusions

- In most cases, the more dimensional classifier is the mean accuracy is higher.
- In some cases, adding dimensions results in lowering maximum accuracy (e.g. tangential cored SVMs).
- Best classifiers for each problem:
  - A. **SVM** with radial function core (accuracy $= \mathbf{0.67 \pm 0.02}$)
  - B. **SVM** with radial function core (accuracy $= \mathbf{0.96 \pm 0.01}$), linear (accuracy $= \mathbf{0.96 \pm 0.01}$) or quadratic (accuracy $= \mathbf{0.95 \pm 0.01}$)
  - C. **Naïve Bayes** (accuracy $= \mathbf{0.97 \pm 0.02}$) and **regression tree** (accuracy $= \mathbf{0.96 \pm 0.03}$) – only these classifiers used EMO variables.

Contact Information

- choloniewski@if.pw.edu.pl
- julas@if.pw.edu.pl
- jholyst@if.pw.edu.pl