The Role of Emotional Variables in the Classification and Prediction of Collective Social Dynamics

J. Chołoniewski[a,*], J. Sienkiewicz[a], J. Hołyst[a,b] and M. Thelwall[c]

[a]Physics in Economy and Social Sciences Research Group, Faculty of Physics, Warsaw University of Technology, Koszykowa 75, PL–00662 Warszawa, Poland

[b]ITMO University, 19, Kronverkskiy av., 197101 Saint Petersburg, Russia

[c]Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom

We demonstrate the power of data mining techniques for the analysis of collective social dynamics within British Tweets during the Olympic Games 2012. The classification accuracy of online activities related to the successes of British athletes significantly improved when emotional components of tweets were taken into account, but employing emotional variables for activity prediction decreased the classifiers' quality. The approach could be easily adopted for any prediction or classification study with a set of problem-specific variables.

1. Introduction

Knowledge of human collective behaviour is a powerful tool in the hands of those who know how to use it. Advances in economics and social sciences can be insightful but quantitative descriptions are often absent. The reason might be the high complexity of many social systems, their highly nonlinear dynamics and internal relationships. Physicists have encountered these kinds of problems many times and have provided valuable conclusions — if one knows only a little about the underlying microscopic system dynamics (such as a phase space, short- and long-distance interactions), then one can focus on the macroscopic outputs of the system using statistical methods [1–7].

Following mass adoption of the internet, it became clear that gathering massive amounts of data about a people's behavior (on-line stock markets, social media) — nearly impossible before — had become easy. It is possible to track the activities of an entity but describing a person would be a great challenge for the scientists of the future. Nevertheless, internet data are extremely useful to study collective human behaviors. This approach has already been tested within some recent studies [8–10]. Moreover, it provides quantitative methods for social dynamics and emotion analyses [11–14]. The last cited model describes possible proceedings of online discussions taking into account the emotional context of posts — e.g., if the first post in a discussion has strong positive emotions, what tends to happen to the rest of the discussion?

These new possibilities always create new challenges — including how to effectively manage the necessary big data.

———

*corresponding author; e-mail: choloniewski@if.pw.edu.pl

1.1. Data mining in science

Astronomy and astrophysics had to deal with a data abundance since an invention of an automatic optical telescope and digital photograpy. An amount of highly magnified high resolution pictures of the sky heavily outnumbered a size of the labor force.

Traditional methods to analyse big data are too time-consuming — even an army of astronomers would need to study such big datasets for years. This problem led to the development of statistical data exploration methods [15]. For example, extracting the features of an object from a photo and applying a classification algorithm trained on empirical data enables the detection of various kinds of stars and galaxies. Data mining techniques have also been employed in large CERN projects — e.g., in particle classifications conducted during the Higgs boson search experiments [16], but also in many other physical sciences — for example predicting certain kinds of crystals [17] or graphen-like structures [18] in solid-state physics.

Of course, data mining applications are not limited to physics and science in general (at the moment, these are not even their main areas). There are also commercial (e.g., basket analyses [19], customer behaviour research [20]), medical (e.g., automated diagnostics [21]) and economics (e.g., credit scoring [22]) uses. The development of data mining techniques is inevitable and will greatly impact on our everyday life.

1.2. Science in data mining

Data mining perhaps began in the 1700s when Bayes presented his famous theorem about a conditional probability to the Royal Society but the real advances happened in the 20th century — cluster analysis [23], genetic algorithms [24], classification trees [25], support vector machines [26] etc. Nowadays, these are also part of the new discipline of big data science. While most findings should be considered as being contributions from mathematics and computer science, data mining has bidirectional interactions with physics.

Physics-based approaches have led to new interesting methods, such as a clustering algorithm based on quantum mechanics [27], support vector machine studies based on statistical physics [28] or a cluster analysis method inspired by theoretical spin glasses research [29].

In the face of the great benefits and potentials of data mining, it seems that developing new methods and improving existing ones could be a great help for the next generation of physicists.

1.3. Main focus

The main aims of the paper are to check (i) if one can predict online social dynamics using data mining techniques, (ii) if one can detect special events taking place in the real world from observations of online discussions, (iii) if detecting emotional content can contribute to a better performance for some social data mining techniques.

Twitter is a social website that reacts to some kinds of external impacts (i.e., medal winning during the 2012 Olympic Games in London). In this paper, its activities are predicted and its dynamics are classified. The relationships between the number of dimensions available, the methods applied, emotional data usage and the accuracy of trained classifiers are also analyzed.

Note that previous statistical Twitter analyses have also yielded interesting results, such as stock market predictions [30] or extreme social-media events predictions [31].

2. Data

2.1. Original dataset

The original dataset was gathered by the Statistical Cybermetrics Research Group from University of Wolverhampton and shared as part of the EU funded CyberEmotions project.

From 2012-07-11 to 2012-08-13 (2012 Olympic Games in London and about two weeks before), all tweets written in English and sent within a radius of 50 km around London were gathered if they contained a sports-related hashtag (e.g., *#olympics*, *#100mrun*). The information about each tweet consists of its unique ID, text (up to 160 characters, including hashtags), date and time sent, author's user name, if a tweet was a response to another tweet — ID and author's user name of that tweet, positive sentiment strength (scored by the SentiStrength classifier [32]; from 1 to 5, where 1 — no positive sentiment and 5 — very strong positive sentiment), negative emotion strength (scored by the SentiStrength classifier; from –1 to –5, where –1 — no negative sentiment and –5 — very strong negative sentiment).

Tweets with no sentiment (i.e., a positive sentiment score of 1 and a negative sentiment score of –1) were discarded from the analysis.

2.2. Aggregated and derived data

The original dataset was aggregated into 3134 non-overlapping 15-minute time windows. In each window, features were calculated as presented in Table I.

TABLE I

Features in 15-minute windows used in the classification aggregated and derived from original dataset; features marked with star (*) are considered *sentiment* dimensions in the study.

Abbre-viation	Description	How it was calculated?
ACT	Activity	A number of tweets sent in a given 15-minute window
SENT*	Sentiment	A mean difference between a positive and a negative emotions score
PERC UNIQ	Percent of unique users	A number of unique users (tweet authors) divided by the number of tweets
PERC REP	Percent of reply tweets	A number of tweet that are replies divided by the number of tweets
MEAN PL	Post length	A mean tweet number of characters
dACT	First derivative of activity	A difference between ACT in the time windows and ACT in the previous time window
dSENT*	First derivative of sentiment	A difference between SENT in the time windows and SENT in the previous time window
d2ACT	Second derivative of activity	A difference between dACT in the time windows and dACT in the previous time window
d2SENT*	Second derivative of sentiment	A difference between dSENT in the time windows and dACT in the previous time window
SENT PEAK*	Sentiment peak	*description in 2.2.*

These variables were used as predictors for benchmark problems.

Peaks of activity and sentiment in the Twitter time series were detected using the algorithm described in [33] and the presence of a sentiment peak was marked as a feature of each time window (sentiment peak — SENT PEAK*). The presence of an activity peak was later used as a benchmark for problem 3.

2.3. Medal data

For some analyses, dates and times of British athletes winning a medal were recorded. Medal colors and sports events were also gathered but not used in the study. The complete list can be found in Appendix A.

3. Benchmark problems

Classifiers (i.e., data mining methods trained on a set of observations of given dimensionality) were provided information about the features in previous time windows and were tested on three two-class classification problems. A summary of class populations for each problem is in Table II.

TABLE II

Class counts for each benchmark problem.

Problem	Observations	Class 1 count	Class 2 count
Trend prediction	3140	1614 (51.4%)	1516 (48.6%)
Threshold exceed	3140	1768 (56.3%)	1362 (43.7%)
Peak classification	516	482 (93.4%)	34 (7.6%)

The combination of features listed in Table I was considered in each time window as an observation vector.

Below we describe the tasks that the classifiers had to solve.

3.1. Trend prediction

Each time window was considered as an observation. The classifier had to determine if the activity component in the next time window would be lesser (class 1) or greater (class 2) than the activity in the present time window, knowing the present and the past feature vector.

3.2. Threshold forecasting

The procedure was as in 3.1., but the classifier had to determine if the activity in a given 15-minute window would be lesser (class 1) or greater (class 2) than 500 tweets.

3.3. Peak classification

Each time window where an activity peak was detected was considered as an observation. The classifier had to determine if the peak co-occurred with an event in which a British athlete(s) had won a medal.

4. Methodology

9 data mining methods were chosen for testing along with 17 features for the first and the second problem and 15 for the third problem. The accuracy for each pair "method-set of features" (sets of features constructed from up to 6 elements) was tested on the benchmark problems.

All 2-, 3-, 4-, 5- and 6-element permutations of features were employed to train each method on a randomly chosen 80% of the observations and tested on the remaining 20% of the observations. The procedure was repeated 20 times for each pair.

4.1. Data mining methods

Applied methods:

- linear discriminant analysis (LDA),

- naive Bayes (NB),

- quadratic discriminant analysis (QDA),

- regression tree (REG TREE),

- support 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).

4.2. Variables sets

For each observation in the set, the following features were selected (t means "in the time window where the peak occurred", $t-1$ means "in the previous time window" and so on) for first two problems (abbreviations described in Table I): ACT($t-1$), ACT($t-2$), ACT($t-3$), SENT($t-1$), SENT($t-2$), SENT($t-3$), dACT($t-1$), dACT($t-2$), dSENT($t-1$), dSENT($t-2$), d2ACT($t-1$), d2ACT($t-2$), d2SENT($t-1$), d2SENT($t-2$), PROC REP($t-1$), PROC UNIQ($t-1$), MEAN PL($t-1$). For the third problem: ACT(t), ACT($t-1$), ACT($t-2$), SENT(t), SENT($t-1$), SENT($t-2$), SENT PEAK(t), SENT PEAK($t-1$), SENT PEAK($t-2$), PERC UNIQ(t), PERC UNIQ($t-1$), PERC UNIQ($t-2$), MEAN PL(t), MEAN PL($t-1$), MEAN PL($t-2$).

5. Results

For each benchmark problem, the relationships between the number of dimensions and classifier accuracy were analysed as well as the impact of applying a specific number of emotional dimensions.

5.1. Trend prediction

In Fig. 1 the best (left) and the mean (right) accuracy of trend predicting (see subsection 3.1.) classifiers with a given dimensionality is presented. The best results are achieved for support vector machines with a radial-basis function as their core. These classifiers achieved an accuracy of 66%. It is interesting that adding dimensions does not necessarily enhance the quality of the best classifiers. The mean accuracy of a classifier tends to increase with every additional dimension, however.

Figure 2 shows the relationship between the accuracy of the trend predicting classifiers and the number of emotional dimensions for all methods, while in Fig. 3 — for support vector machines with radial basis core (the most accurate classifiers for the problem). It is clear that in this case using emotional dimensions does not enhance classification. Emotional features seem to carry no valuable information for this kind of exercises — comparing accuracies for 2 dimensional classifiers with 0 emotional features (2–0 emo) with 3–1 emo, 4–2 emo, 5–3 emo and 6–4 emo gives nearly the same result.
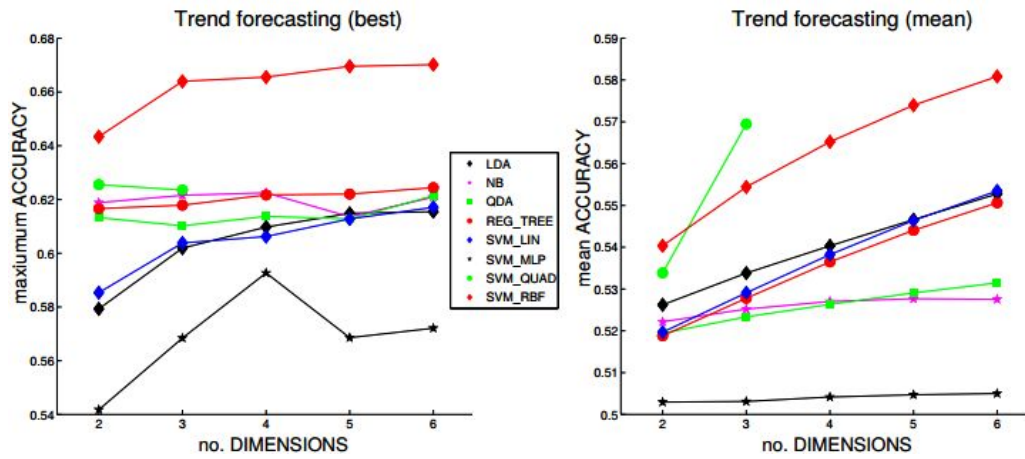
Fig. 1. Maximum (left graph) and mean (right) accuracy of classifiers as a function of a number of variables (dimensions); standard deviations were not plotted for the sake of readability, they are of the order of 0.02.
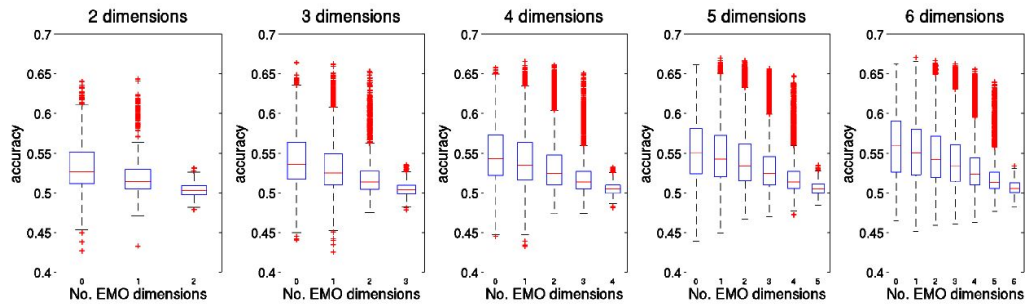

Fig. 2. Box plot of accuracy of trend forecasting for different numbers of variables used by various classifiers as a function of number of emotion-related variables.
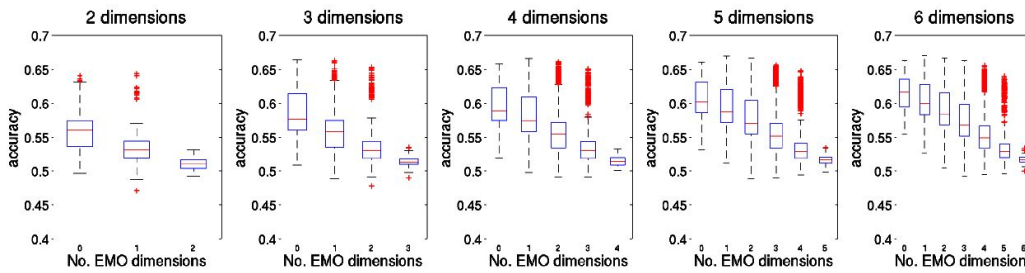

Fig. 3. Box plot of accuracy of trend forecasting for different number of variables used by SVM with a radial basis core as a function of number of emotion-related variables.
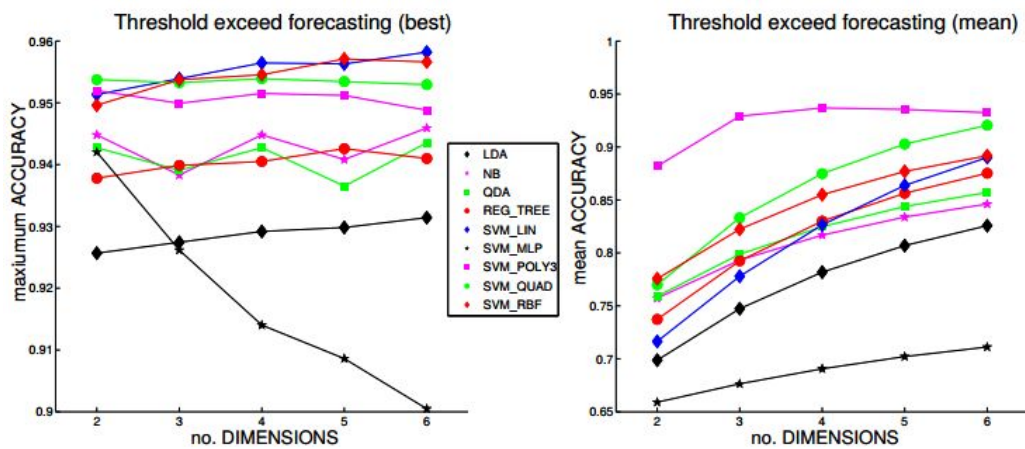

Fig. 4. Maximum (left graph) and mean (right) accuracy of classifiers as a function of a number of variables (dimensions); standard deviations were not plotted for the sake of readability, they are of the order of 0.02.
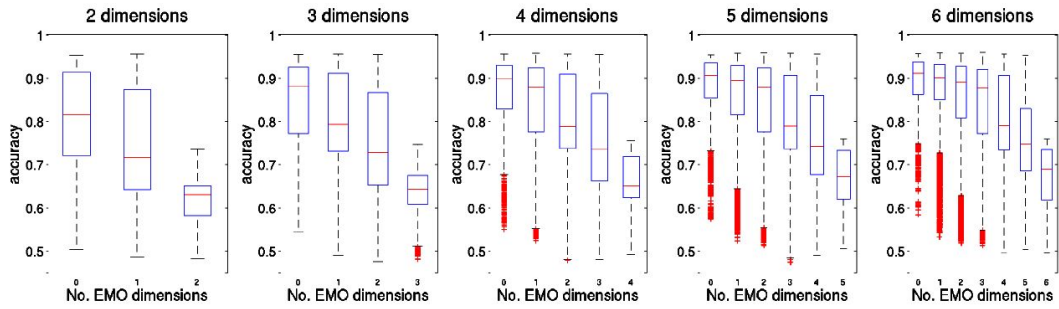
Fig. 5.   Box plot of accuracy of threshold exceed forecasting for different numbers of variables used by various classifiers as a function of number of emotion-related variables.
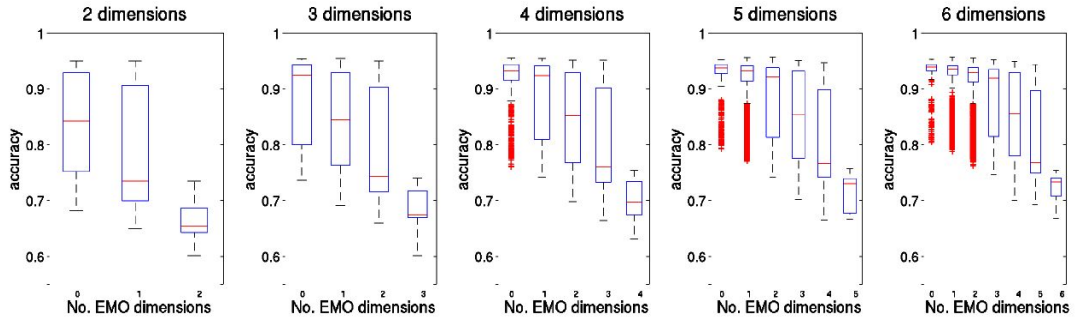

Fig. 6.   Box plot of accuracy of threshold exceed forecasting for different number of variables used by SVM with a radial basis core as a function of number of emotion-related variables.
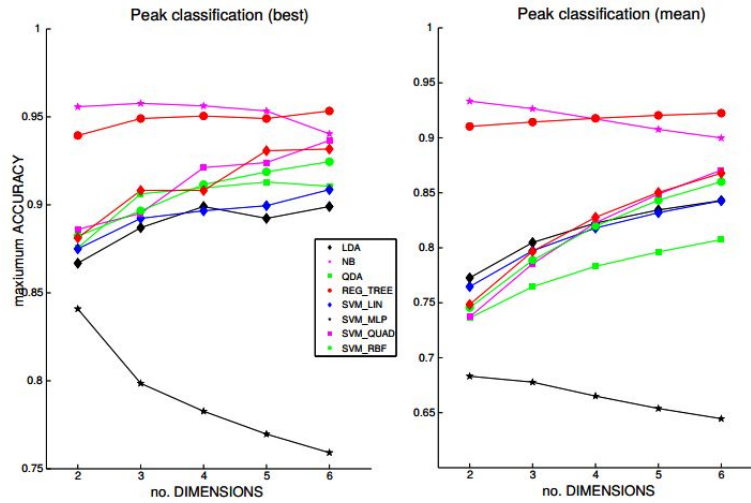

Fig. 7.   Maximum (left graph) and mean (right) accuracy of classifiers as a function of a number of variables (dimensions); standard deviations were not plotted for the sake of readability, they are of the order of 0.02.

5.2. Threshold exceed forecasting

In Fig. 4 the best (left) and the mean (right) accuracy of threshold forecasting (see subsection 3.2.) classifiers with a given dimensionality is presented. The best results are achieved for support vector machines with a radial-basis core function. These classifiers achieved an accuracy of 96%. Also in this case adding dimensions does not necessarily increase the quality of a classifier; for SVM with a tangential core adding new dimensions decreases the

accuracy of the best classifier. Again, the mean accuracy of the classifier tends to increase with every additional dimension, however.

Figure 5 shows the relationship between the accuracy and the number of emotional dimensions for all methods, while in Fig. 6 — for support vector machines with a radial-basis core (the most accurate classifier for the problem). Also in this case, using emotional variables provided no classifier enhancing information.

5.3. Peak classification

The third problem is fundamentally different from the previous ones. In Fig. 7 the best (left) and the mean (right) accuracy of the peak type classifiers (see subsection 3.3.) with a given dimensionality is presented. The results are a little surprising — the best accuracy (about 96%) was achieved for the naive Bayes classifier with 3 features. Similar results were achieved by naive Bayesian classifiers with a different number of dimensions and a regression tree. The mean accuracy of the classifier tends to increase with every additional dimension for most of the methods, but not for the best and the worst method.

Figure 8 shows the relationship between the accuracy and the number of emotional dimensions for all methods, while in Fig. 9 — for naive Bayes classifiers (the most accurate classifiers for the problem). In this case, using emotional dimensions clearly increases the accuracy of the classifiers.
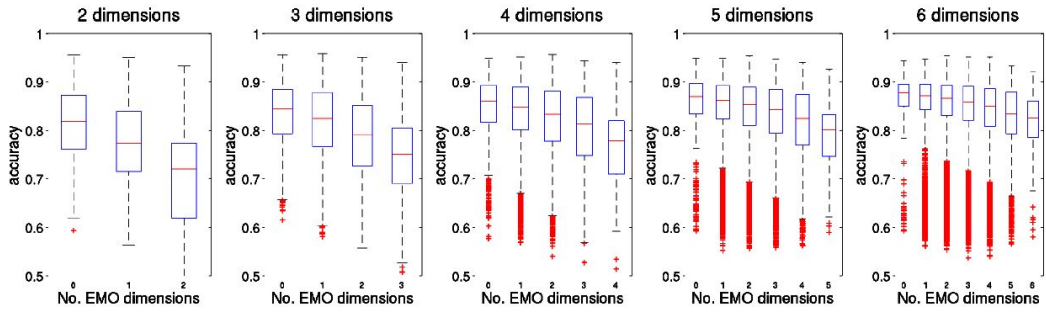

Fig. 8.   Box plot of accuracy of peak classification for different numbers of variables used by various classifiers as a function of number of emotion-related variables.
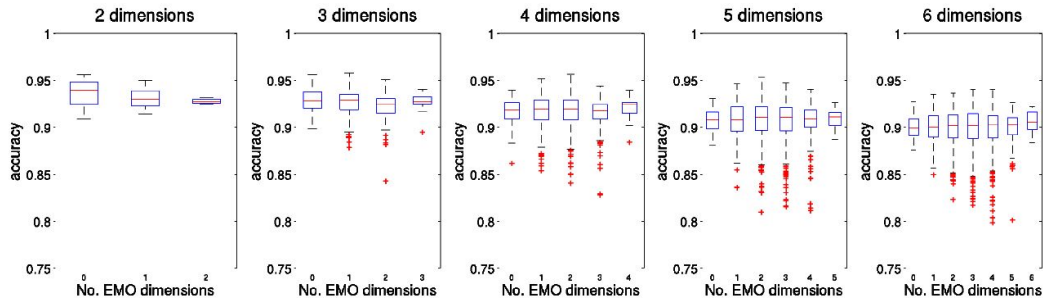

Fig. 9.   Box plot of accuracy of peak classification for different number of variables used by SVM with a radial basis core as a function of number of emotion-related variables.

6. Conclusions

This research has been conducted on Twitter activity and sentiment data. We investigated the role of emotional dimensions in classification but our approach could be easily adapted to any classification (or prediction) study of certain systems. To achieve this, change the dataset to another aggregated macroscopic collection of data from a system with unknown internal microscopic dynamics and emotional dimensions to problem-specific variables. The insights provided may confirm the usefulness of the variables and identify the best suited methods and variable sets for a specific problem.

The results show the significance of the emotional content of Twitter posts in the peak classification study related to special events (i.e., medal winning), but employing emotional dimensions in predicting activity brings no additional insights.

A set of popular data mining techniques' performance was analyzed for activity prediction and peak classification. The results are interesting because of evidence about the importance of emotions in social media and a wide spectrum of applications (e.g., predicting time series, classifying the origins of peaks).

The relationship between the accuracies of a set of basic data mining methods and number of features as well as between their accuracy and the number of emotional features was also analyzed. The benchmark problems consisted of (A) trend prediction, (B) forecasting of threshold crossing and (C) peak classification.

Depending on the problem, the best classifiers had an accuracy of (A) 66%, (B) 96% and (C) 96%.

Adding a dimension to the benchmark problems increased the mean classifier accuracy for most cases but the best classifier for a given method is not necessarily

the one with the largest number of dimensions. In a few cases, adding dimensions decreases the best accuracy and the mean classifier accuracy.

Using an emotional dimension has no impact for (A) and (B), while clearly enhances accuracy for (C).

Considering the relation of emotional to other dimensions in the context of our conclusions, it is important to note that using the dimensions enhances a classification quality for the last benchmark problem only and decreases it for another two problems. It seems that emotional data is strictly related to medal events but provides little information about activity changes. Probably, there are some kinds of problems that can be described with problem-specific dimensions which provide a significant improvement in particular kinds of analyses and decrease in others. The approach presented in the paper would be helpful to identify the problems and the related dimensions.

The results confirm the usefulness of classifiers, give insights about applying additional features and dimensions and show that the methods analyzed can be applied to quantitative descriptions of social media phenomena.

Acknowledgments

References

[1] J. Kwapien, S. Drozdz, *Phys. Rep.* **515**, 115 (2012).

[2] K. Malarz, K. Kulakowski, *Acta Phys. Pol. A* **117**, 695 (2010).

[3] P. Gawronski, K. Malarz, M.J. Krawczyk, J. Malinowski, A. Kupczak, W. Sikora, K. Kulakowski, J. Was, J.W. Kantelhardt, *Acta Phys. Pol. A* **123**, 522 (2013).

[4] M. Wilinski, A. Sienkiewicz, T. Gubiec, R. Kutner, Z.R. Struzik, *Physica A* **392**, 5963 (2013).

[5] K. Urbanowicz, J.A. Hołyst, *Physica A* **344**, 284 (2004).

[6] A. Sienkiewicz, T. Gubiec, R. Kutner, Z.R. Struzik, *Acta Phys. Pol. A* **123**, 615 (2013).

[7] P. Oswiecimka, S. Drozdz, J. Kwapien, A.Z. Gorski, *Acta Phys. Pol. A* **123**, 597 (2013).

[8] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, V. Loreto, *PNAS* **106**, 10511 (2009).

[9] G. Bello, H. Menendez, S. Okazaki, D. Camacho, *Computational Collective Intelligence: Technologies and Applications*, Eds: C. Badica, N.T. Nguyen, M. Brezovan, Springer-Verlag, Berlin, Germany 2013, vol. 8083, p. 622.

[10] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, J.A. Hołyst, *PLoS ONE* **6**, e22207 (2011).

[11] P. Sobkowicz, M. Kaschesky, G. Bouchard, *Government Information Quarterly* **29**, 470 (2012).

[12] P. Sobkowicz, A. Sobkowicz, *SSCR* **30**, 448 (2012).

[13] P. Pohorecki, J. Sienkiewicz, M. Mitrovic, G. Paltoglou, J.A. Hołyst, *Acta Phys. Pol. A* **123**, 604 (2013).

[14] A. Chmiel, P. Sobkowicz, J. Sienkiewicz, G. Paltoglou, K. Buckley, M. Thelwall, J. A. Hołyst, *Physica A* **390**, 2936 (2011).

[15] U.M. Fayyad, N. Weir, S. Djorgovski, *Proceedings of the Second International Conference on Information and Knowledge Management*, ACM, New York, USA 1993, p. 527.

[16] D. Rousseau, *IEEE Software* **29**, 11 (2012).

[17] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, *Nat. Mater.* **5**, 641 (2006).

[18] S. Lebègue, T. Björkman, M. Klintenberg, R.M. Nieminen, O. Eriksson, *Phys. Rev. X* **3**, 031002 (2013).

[19] Y.L. Chen, K. Tang, R.J. Shen, Y.H. Hu, *Decision Support Systems* **40**, 339 (2005).

[20] E.W.T. Ngai, L. Xiu, D.C.K. Chau, *Expert Systems with Applications* **36**, 2592 (2009).

[21] N. Lavrac, E.T. Keravnou, B. Zupan, *Intelligent data analysis in medicine and pharmacology*, Kluwer Academic Publishers, Norwell, MA, USA 1997, vol. 414, p. 1.

[22] C.L. Huang, M.C. Chen, C.J. Wang, *Expert Systems with Applications* **33**, 847 (2007).

[23] F. Murtagh, M.J. Kurtz, arXiv:1209.0125, 2012.

[24] J.J. Grefenstette, *IEEE Transactions on Systems, Man and Cybernetics* **16**, 122 (1986).

[25] D.L. Verbyla, *Canad. J. For. Res.* **17**, 1150 (1987).

[26] C. Cortes, V. Vapnik, *Mach. Learn.* **20**, 273 (1995).

[27] D. Horn, A. Gottlieb, *Phys. Rev. Lett.* **88**, 018702 (2001).

[28] M. Opper, R. Urbanczik, *Phys. Rev. Lett.* **86**, 4410 (2001).

[29] A. Decelle, Ph.D. Thesis, Université Paris-Sud, Paris, 2011.

[30] J. Bollen, H. Mao, X. Zeng, *J. Comput. Sci.* **2**, 1 (2011).

[31] J.M. Miotto, E.G. Altmann, *PLoS ONE* **9**, e111506 (2014).

[32] M. Thelwall, K. Buckley, G. Paltoglou, *J. Assoc. Inf. Sci. Technol.* **63**, 163 (2012).

[33] Billauer's homepage, http://www.billauer.co.il/peakdet.html.

Appendix A — Events with British medalists

No.	Date	Time	Sport	Event	Medal
1	29 July	18:45	Cycling	Women's road race	Silver
2	29 July	23:45	Swimming	Women's 400 m freestyle	Bronze
3	30 July	22:00	Gymnastics	Men's artistic team all-around	Bronze
4	31 July	16:00	Equestrian	Team eventing	Silver
5	1 August	15:00	Rowing	Women's coxless pair	Gold
6	1 August	15:30	Rowing	Men's eight	Bronze
7	1 August	19:15	Cycling	Men's time trial	Bronze
8	1 August	19:15	Cycling	Men's time trial	Gold
9	2 August	18:45	Canoeing	Men's slalom C-2	Gold
10	2 August	18:45	Shooting	Men's double trap	Gold
11	2 August	18:45	Canoeing	Men's slalom C-2	Silver
12	2 August	19:00	Judo	Women's 78 kg	Silver
13	2 August	21:15	Cycling	Men's team sprint	Gold
14	3 August	15:15	Rowing	Men's single sculls	Bronze
15	3 August	15:15	Rowing	Women's double sculls	Gold
16	3 August	15:45	Rowing	Men's coxless pair	Bronze
17	3 August	21:45	Cycling	Men's team pursuit	Gold
18	3 August	22:15	Cycling	Women's keirin	Gold
19	4 August	14:45	Rowing	Men's coxless four	Gold
20	4 August	15:00	Rowing	Women's lightweight double sculls	Gold
21	4 August	21:15	Cycling	Women's team pursuit	Gold
22	4 August	00:00	Athletics	Women's heptathlon	Gold
23	4 August	00:30	Athletics	Men's long jump	Gold
24	4 August	01:00	Athletics	Men's 10,000 m	Gold
25	5 August	17:45	Sailing	Finn class	Gold
26	5 August	19:15	Gymnastics	Men's pommel horse	Bronze
27	5 August	19:15	Gymnastics	Men's pommel horse	Silver
28	5 August	19:45	Tennis	Men's singles	Gold
29	5 August	20:30	Tennis	Mixed doubles	Silver
30	5 August	22:00	Cycling	Men's omnium	Bronze
31	5 August	00:15	Athletics	Women's 400 m	Silver
32	6 August	17:45	Gymnastics	Women's uneven bars	Bronze
33	6 August	20:00	Equestrian	Team jumping	Gold
34	6 August	21:15	Cycling	Men's sprint	Gold
35	7 August	16:15	Sailing	Men's sailboard	Silver
36	7 August	17:15	Triathlon	Men's triathlon	Bronze
37	7 August	17:15	Triathlon	Men's triathlon	Gold
38	7 August	19:15	Equestrian	Team dressage	Gold
39	7 August	20:15	Cycling	Women's omnium	Gold
40	7 August	21:15	Cycling	Men's keirin	Gold
41	7 August	21:30	Cycling	Women's sprint	Silver
42	7 August	00:00	Athletics	Men's high jump	Bronze
43	9 August	19:00	Equestrian	Individual dressage	Bronze
44	9 August	19:00	Equestrian	Individual dressage	Gold
45	9 August	19:45	Boxing	Women's flyweight	Gold
46	9 August	01:30	Taekwondo	Women's 57 kg	Gold
47	10 August	16:45	Sailing	Women's 470 class	Silver
48	11 August	12:45	Canoeing	Men's K-1 200 m	Gold
49	11 August	23:00	Athletics	Men's 5,000 m	Gold
50	11 August	00:15	Boxing	Men's bantamweight	Gold
51	11 August	01:15	Diving	Men's 10 m platform	Bronze
52	12 August	18:45	Boxing	Men's super heavyweight	Gold