

AMA Team server

Machine learning: models & algorithms

"Data and Knowledge Processing at Large Scale" Axis

Joint research team between CNRS, Grenoble INP, UJF, and UPMF

Dataset — Buzz in social media

FILES

- Classification task :
 - [browse](#)
 - [download compressed version](#)
- Regression task : browse
 - [browse](#)
 - [download compressed version](#)

OVERVIEW

This dataset contains two different social networks: [Twitter](#), a micro-blogging platform with exponential growth and extremely fast dynamics, and [Tom's Hardware](#), a worldwide forum network focusing on new technology with more conservative dynamics but distinctive features. Tom's Hardware (TH) and Twitter (TW) have very distinctive properties:

- Both German and French contributions are found on (TH), while English contributions are also available in the sample of (TH) we have collected
- In (TW) there is no direct audience estimator, and we use the `nad` feature as the target feature, while it is given in (TH) by the number of displays, that measures the number of times a content is displayed to visitors
- (TW) shows a higher reactivity of exchanges than (TH) 80% of re-tweets take place in the day following the initial tweet, whereas in (TH) the replies to a thread are mostly produced during the subsequent week;
- The (TW) community is broader than the (TH) one, with more than 500 million visitors per month against nearly 41 millions in (TH)

In this study, we focus on 6671 topics, such as: over-cloaking; grafikarten; disque dur; android; etc. related to the technology domain. Here is a summary per language.

	NB. Users			Nb. Discussions			Nb. Examples			Nb. Topics
	FR	EN	DE	FR	EN	DE	FR	EN	DE	
(TH)	$72 \cdot 10^3$	0	$8 \cdot 10^3$	$50 \cdot 10^4$	0	$1 \cdot 10^4$	4879	0	3026	4957
(TW)	$24 \cdot 10^6$	$30 \cdot 10^6$	$10 \cdot 10^6$	$23 \cdot 10^7$	$28 \cdot 10^7$	$46 \cdot 10^6$	76292	35317	29089	6671

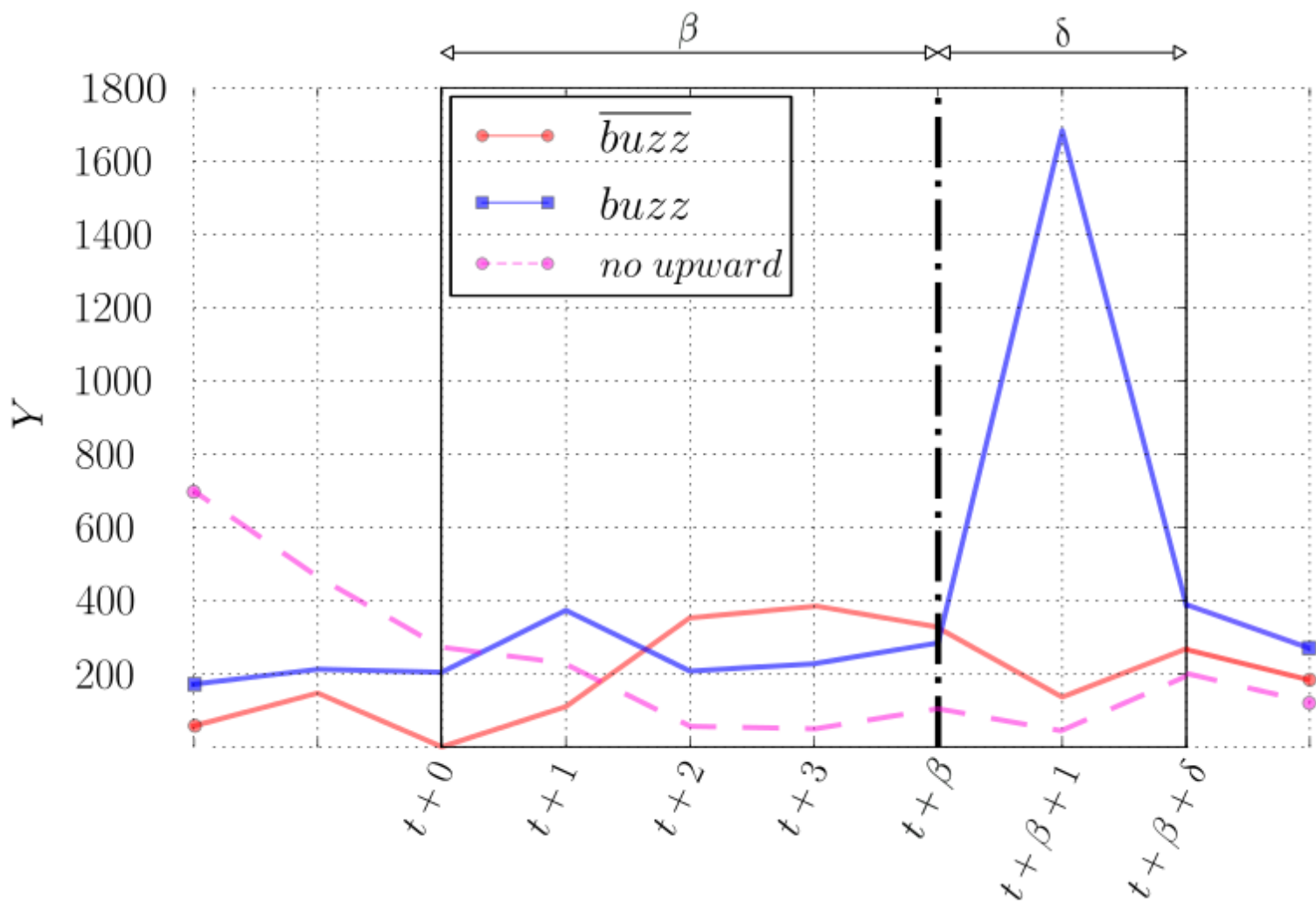
DATA FORMAT

This dataset is published using the [UCI guidelines](#). Hence examples are stored using a standard comma separated value (CSV) format. You will find an example per line, and one feature per column. Each dataset is provided with an additional instructions file, as suggested by the UCI.

CLASSIFICATION TASK

In the classification task you will be provided with time-windows showing an upward trend. The objective of this task is to determine whether or not these time-windows are followed by buzz events. In this task:

- Each example matches an upward window. Such an example is a multivariate time-series ranging from t to $t+\beta$.
- The labeling (ie. buzz; non-buzz) of an example, as well as the upward detection, are performed considering an univariate time-series. This time series (Y , the target feature, presented bellow) is meant to reflect the popularity of a topic.
- There is two ways to label examples: Absolute labeling and Relative labeling. The first one labels as buzz any example which is "sufficiently" popular between $t+\beta+1$ to $t+\beta+\delta$ while the second one is based on the increment of popularity level before and after β
- For both of these labeling methods, the threshold value σ varies in order to qualify buzz of distinct magnitude. Concretely $\sigma = 500$ implies that an example is labeled as a buzz if:
 - (Absolute labeling) the sum of Y 's values ranging from $t+\beta+1$ to $t+\beta+\delta$ is greater than 500
 - (Relative labeling) the difference between (a) the Y 's mean value between $t+\beta+1$ and $t+\beta+\delta$ and (b) the Y 's mean value between t to $t+\beta$ is greater than 500



REGRESSION TASK

As in the classification task you will be provided only with upward-windows. The value to be predicted will be the value of the time-series used to determine the popularity of a topic (Y , as presented before)