Detecting Sentiment Change in Twitter Streaming Data

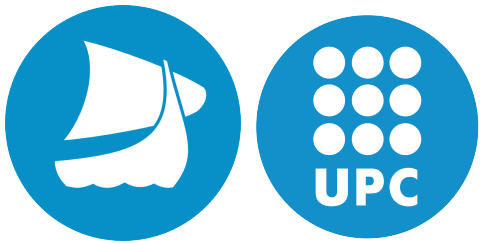
Albert Bifet, Geoff Holmes, Bernhard Pfahringer and Ricard Gavaldà

University of Waikato
Hamilton, New Zealand

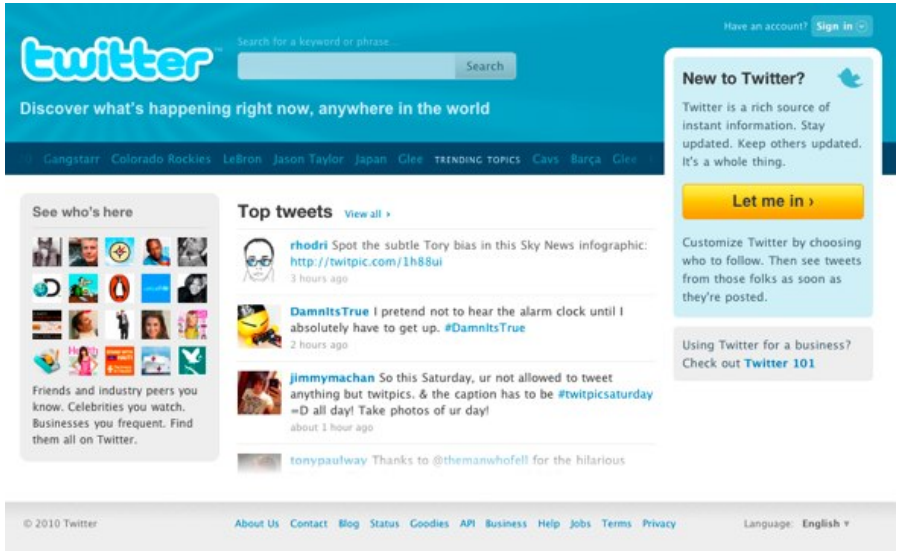
Laboratory for Relational Algorithmics, Complexity and Learning **LARCA**
UPC-Barcelona Tech, Catalonia

Castro, 19 October 2011
WAPA 2011

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato



Detecting Sentiment Change in Twitter Streaming Data



twitter

moa

Twitter is a micro-blogging service built to discover what is happening **now** anywhere

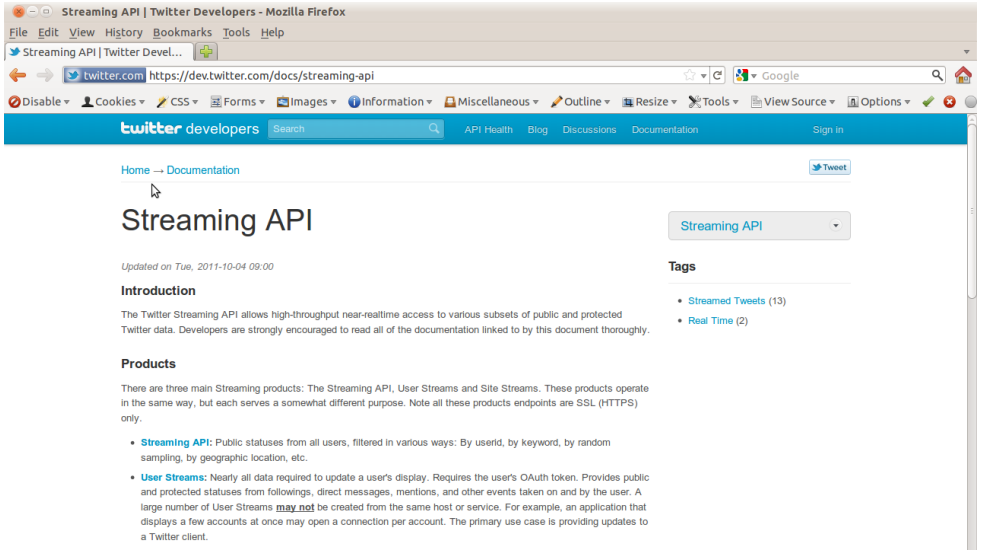
Detecting Sentiment Change in Twitter Streaming Data

twitter

moa

Problem: How to analyze Twitter data on real time

Detecting Sentiment Change in Twitter Streaming Data

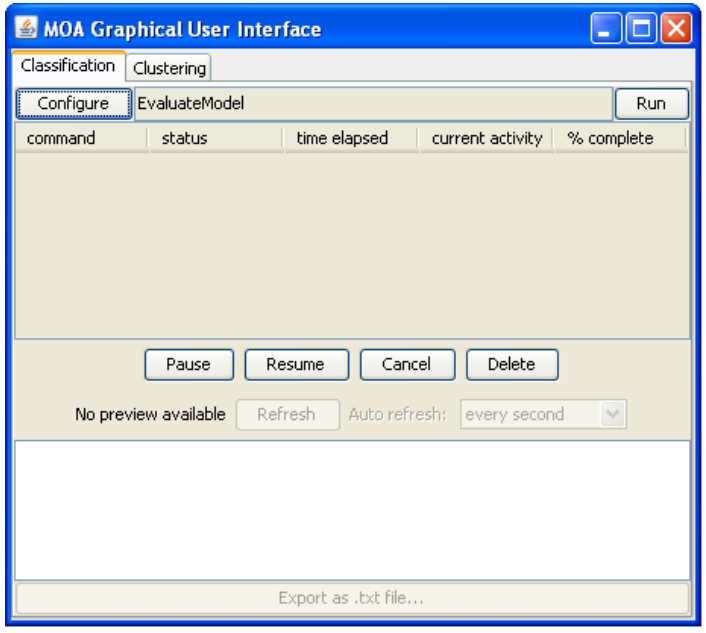


twitter

moa

Twitter Streaming API: API for accessing Twitter in real-time

Detecting Sentiment Change in Twitter Streaming Data



twitter

moa

MOA is an open source project for data stream mining, for analyzing big data on real time

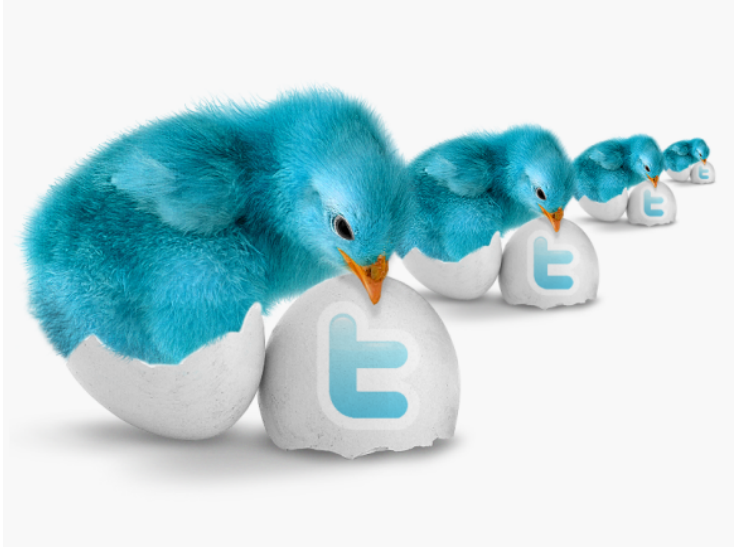
Detecting Sentiment Change in Twitter Streaming Data

twitter

moa

Sentiment analysis: analyze tweets with positive :) or negative :(tweets

Detecting Sentiment Change in Twitter Streaming Data



Problem: We need to convert tweet texts in sparse vector of features on real-time

Detecting Sentiment Change in Twitter Streaming Data

twitter

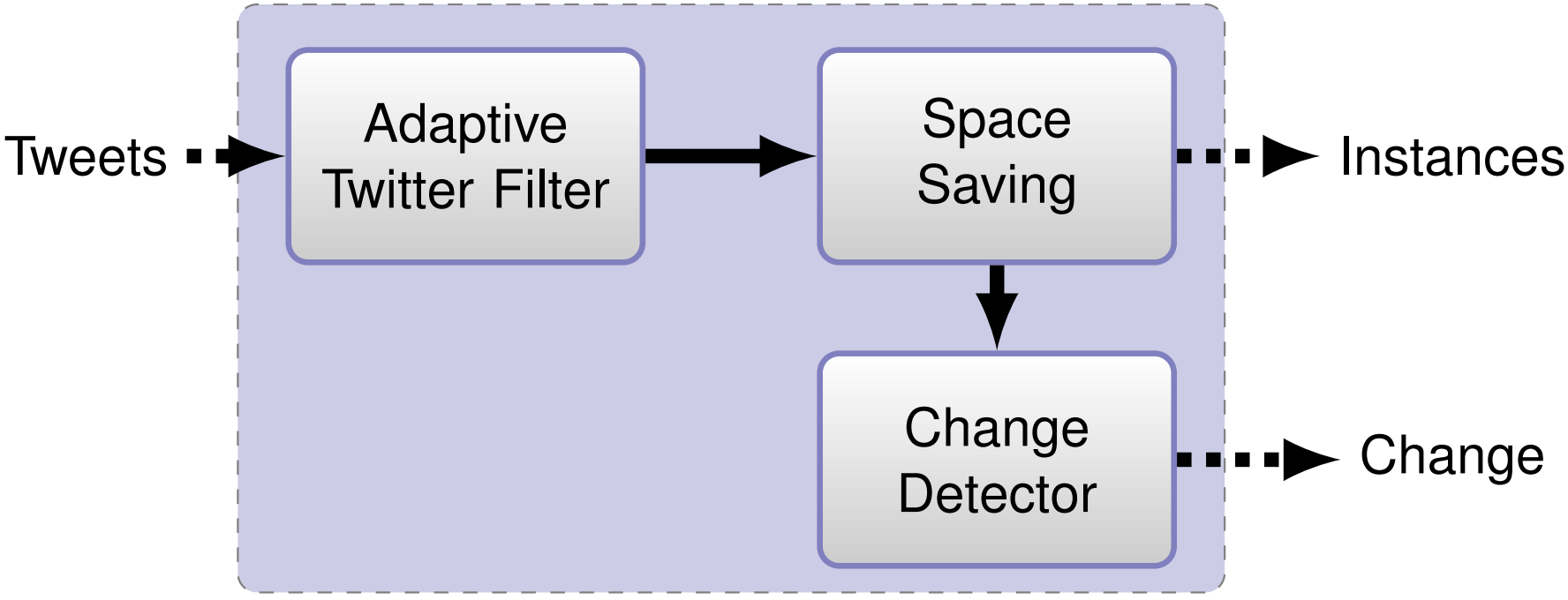
moa

Real-time means (i) change adaption
(ii) fast: can not store tweets on memory

MOA-TweetReader

Solution: MOA-TweetReader, a package to connect MOA with Twitter

Detecting Sentiment Change in Twitter Streaming Data



MOA-TweetReader consists in (i) Adaptive Twitter filter (ii)Frequent item miner (iii) Change Detector

Detecting Sentiment Change in Twitter Streaming Data

$$f_{i,j} = \frac{\text{freq}_{i,j}}{\sum_\ell \text{freq}_{\ell,j}} \quad \text{(number of times a word appears in the document)}$$

$$idf_i = \log \frac{N}{n_i} \quad \text{(inverse frequency of the word in the corpus)}$$

$$w_{i,q} = f_{i,j} \cdot idf_i$$

MOA-TweetReader Adaptive Twitter filter: online tf-idf

Detecting Sentiment Change in Twitter Streaming Data

SPACE SAVING (METWALLY ET AL.)

```
1   T ← ∅
2   for every term i
3       do if i ∈ T
4           then freq[i] ← freq[i] +1
5           else if |T| < k
6                   then ▷ Add a new item
7                        T ← T ∪ {i}
8                        freq[i] ← 1
9                   else ▷ Replace the item with lower freq.
10                       j ← arg min_{j∈T} freq[j]
11                       T ← T ∪ {i} \ {j}
12                       freq[j] ← freq[j] +1
```

MOA-TWEETREADER Frequent item miner :
SPACE SAVING

Detecting Sentiment Change in Twitter Streaming Data

SPACE SAVING (METWALLY ET AL.)

```
1   T ← ∅
2   for every term i
3       do if i ∈ T
4           then freq[i] ← freq[i] +1
5           else if |T| < k
6                   then ▷ Add a new item
7                        T ← T ∪ {i}
8                        freq[i] ← 1
9                   else ▷ Replace the item with lower freq.
10                       j ← arg min_{j∈T} freq[j]
11                       T ← T ∪ {i} \ {j}
12                       freq[j] ← freq[j] +1
```

SPACE SAVING is the frequent item algorithm for streams with best performance results

Detecting Sentiment Change in Twitter Streaming Data

SPACE SAVING EXPONENTIALLY DECAYED (CORMODE ET AL)

```
1  T ← ∅
2  for every term i with timestamp t_i
3       do if i ∈ T
4            then freq[i] ← freq[i] + exp(λ t_i)
5            else if |T| < k
6                 then ▷ Add a new item
7                      T ← T ∪ {i}
8                      freq[i] ← 1
9                 else ▷ Replace the item with lower freq.
10                     j ← arg min_{j∈T} freq[j]
11                     T ← T ∪ {i} \ {j}
12                     freq[j] ← freq[j] + exp(λ t_i)
```

Improvement to SPACE SAVING: space saving with exponential decay, or using ADWIN

Detecting Sentiment Change in Twitter Streaming Data

ADWIN: ADAPTIVE WINDOWING ALGORITHM

1 Initialize Window W
2 **for** each $t > 0$
3 **do** $W \leftarrow W \cup \{x_t\}$ (i.e., add x_t to the head of W)
4 **repeat** Drop elements from the tail of W
5 **until** $|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}| < \varepsilon_c$ holds
6 for every split of W into $W = W_0 \cdot W_1$
7 output $\hat{\mu}_W$

Improvement to SPACE SAVING: space saving with exponential decay, or using ADWIN

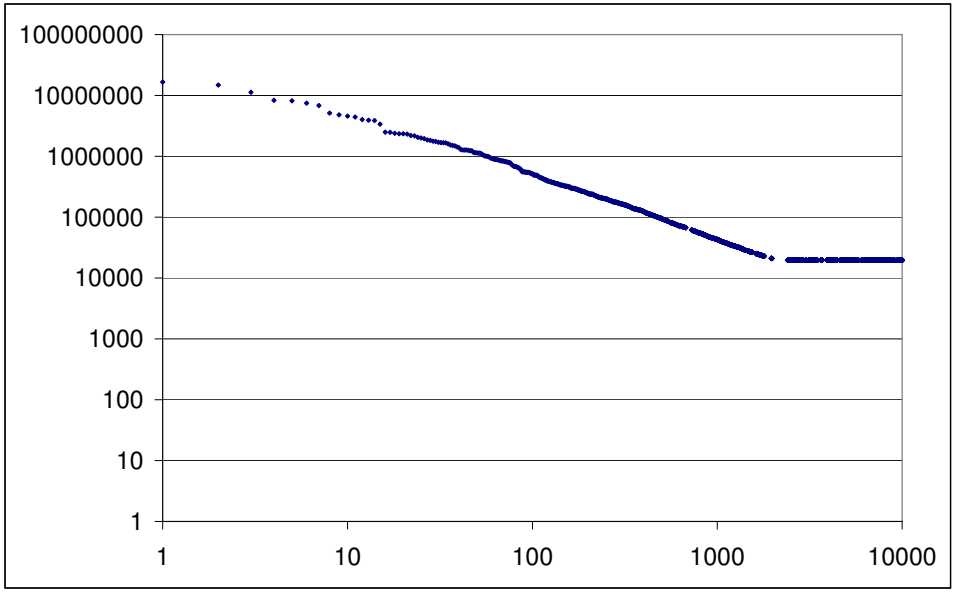
Detecting Sentiment Change in Twitter Streaming Data

SPACE SAVING ADWIN

```
1   T ← ∅
2   for every term i with timestamp t_i
3       do if i ∈ T
4           then  Insert 1 into ADWIN[i] and 0 to other ADWINs
5           else if |T| < k
6                   then ▷ Add a new item
7                        T ← T ∪ {i}
8                        Init ADWIN[i]
9                        Insert 1 into ADWIN[i] and 0 to other ADWINs
10                  else ▷ Replace the item with lower freq.
11                       j ← arg min_{j∈T} freq[j]
12                       T ← T ∪ {i} \ {j}
13                       Insert 1 into ADWIN[j] and 0 to other ADWINs
```

Improvement to SPACE SAVING: space saving with exponential decay, or using ADWIN

Detecting Sentiment Change in Twitter Streaming Data

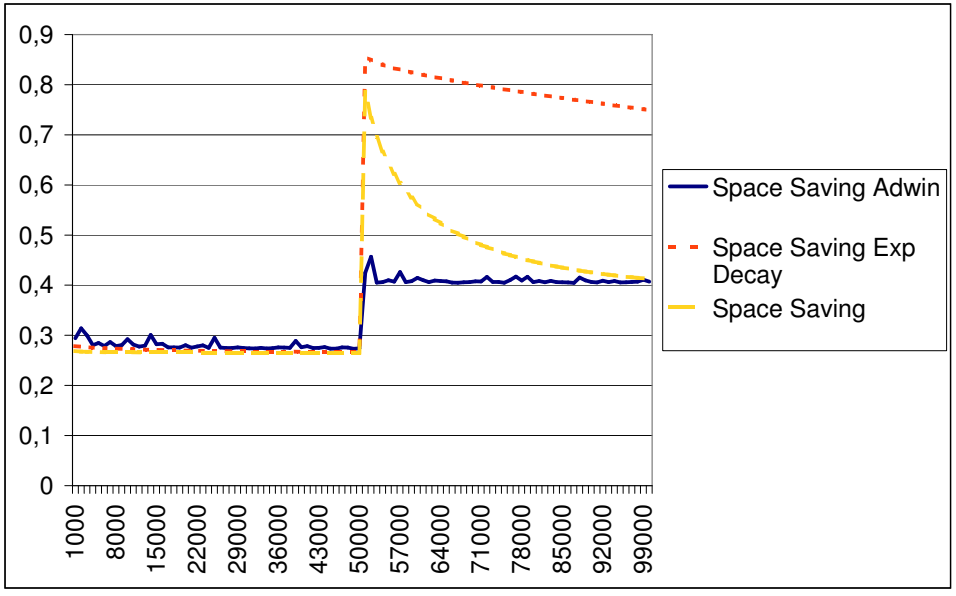


twitter

moa

Experiments: Frequency and ranking of twitter data follows a Zipf distribution

Detecting Sentiment Change in Twitter Streaming Data



twitter

moa

SPACE SAVING ADWIN is able to adapt automatically

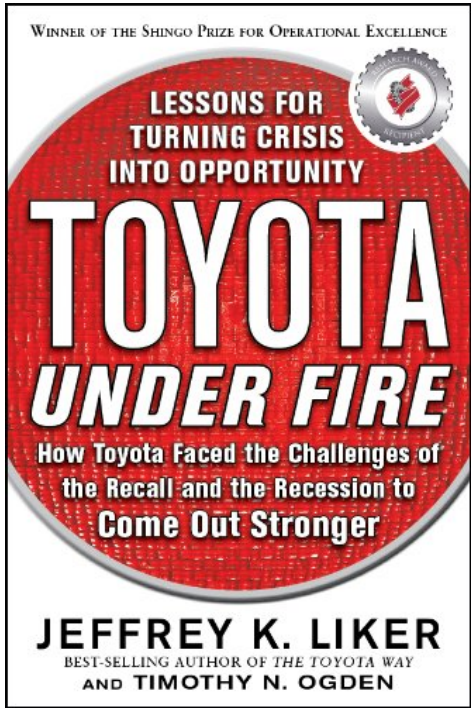
Detecting Sentiment Change in Twitter Streaming Data

twitter

moa

Toyota crisis: during end of 2009 and beginning of 2010 Toyota had problems with accelerator pedals and had to recall millions of cars

Detecting Sentiment Change in Twitter Streaming Data



twitter

moa

Recommended reading “Toyota under fire”

Detecting Sentiment Change in Twitter Streaming Data

There was a gap between the time that our U.S. colleagues realised that this was an urgent situation and the time that we realised here in Japan that there was as urgent situation going on in the U.S. It took ***three months*** *for us to recognise that this had turned into a crisis. In Japan, unfortunately, until the middle of January we did not think that this was really a crisis.*

Akio Toyoda

Recommended reading “Toyota under fire”

Detecting Sentiment Change in Twitter Streaming Data

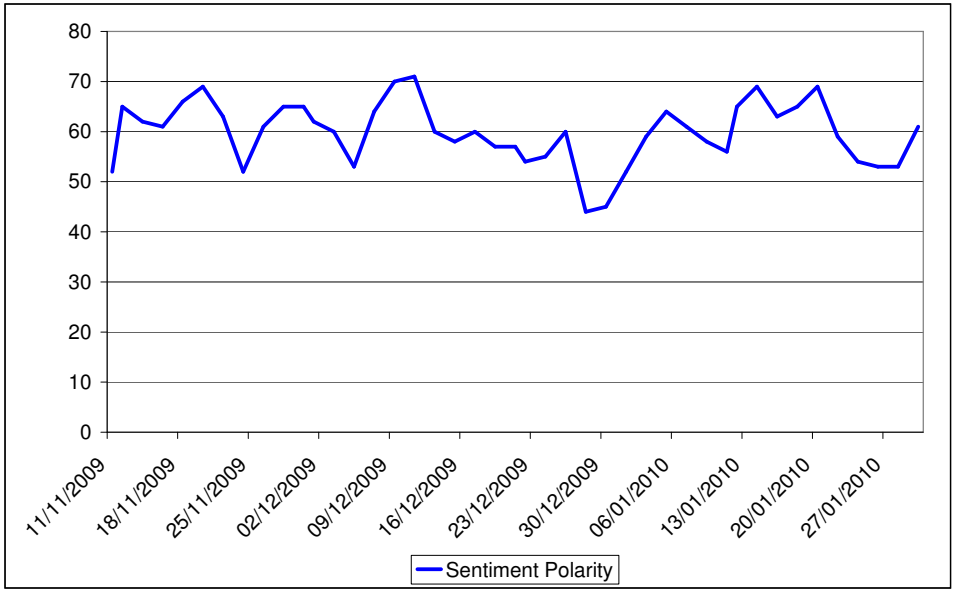
Term	Before	After	Diff
gas	0.122	0.484	0.363
pedals	0.129	0.438	0.309
wonder	0.017	0.214	0.198
problem	0.163	0.357	0.194
good	0.016	0.205	0.190
recalling	0.012	0.106	0.095
gm	0.011	0.089	0.077
#heard_on_the_street	0.040	0.113	0.072
social	0.031	0.099	0.068
sticking	0.070	0.125	0.055
fix	0.026	0.076	0.050
popularity	0.016	0.037	0.021
love	0.017	0.024	0.008

twitter

moa

Following twitter data sentiment, and changes in MOA-TweetReader it is possible to know faster when problem starts

Detecting Sentiment Change in Twitter Streaming Data



twitter

moa

A tool like MOA-TweetReader would have helped Toyota to understand the crisis sooner and to respond more appropiately

Detecting Sentiment Change in Twitter Streaming Data

twitter

moa

CONCLUSIONS. Our goal: how to do real time analysis of twitter data. Our proposal: MOA-TweetReader