

EVENT DETECTION IN TWITTER

Jianshu Weng and Bu-Sung Lee, Francis
HP Labs Singapore



AGENDA

- Motivation
- Proposed Method
 - Wavelet analysis
 - Event Detection with Clustering of Wavelet-based Signals (EDCoW)
- Evaluation
 - Experimental study
 - General Election case study
- Conclusion



MOTIVATION

Benefits

Different perspective from the traditional media

Disaster early warning

Challenges

twitter blog

200 million Tweets per day

Thursday, June 30, 2011

- Winner -
"Pointless Babble"

2nd

"Conversational"

3rd

"Pass-Along Value"

MOTIVATION (CONT'D)

- We need a way to detect “events” out of “babble”-flooded twitter content stream.
 - Fast
 - Low storage requirement
 - Can tell how “significant” the event is.
- Existing techniques: usually first-order signals, e.g. TFIDF/DFIDF
 - Redundant information
 - Many features would seem to have similar waveforms when we apply feature-pivot methods



PROPOSED METHOD

- EDCoW: Event Detection with Clustering of Wavelet-based Signals
- Steps:
 - Construct 1st-order signals for individual features based on their DFIDF scores, for a given interval
 - Wavelet analysis on the 1st-order signals to construct 2nd-order signals
 - Filtering
 - Clustering with modularity-based graph cut
 - Measuring the event “significance”

SIGNAL CONSTRUCTION

- First order: DFIDF

$$y_x(t) = \frac{N_x(t)}{N(t)} \times \log\left(\frac{\sum_{i=1}^{T_c} N(i)}{\sum_{i=1}^{T_c} N_x(i)}\right)$$

- $N_x(i)$ is the number of tweets containing feature x at time point i , and $N(i)$ is the total number of tweets at time point i .

SIGNAL CONSTRUCTION (CONT'D)

–Second stage:

- applying wavelet analysis to compress the DFIDF signals to keep only the changes within a number of sample points in the first stage.

- H is the normalized *wavelet entropy* in a sliding window

- Each 2nd-stage sampling point captures how much the change in the wavelet entropy is

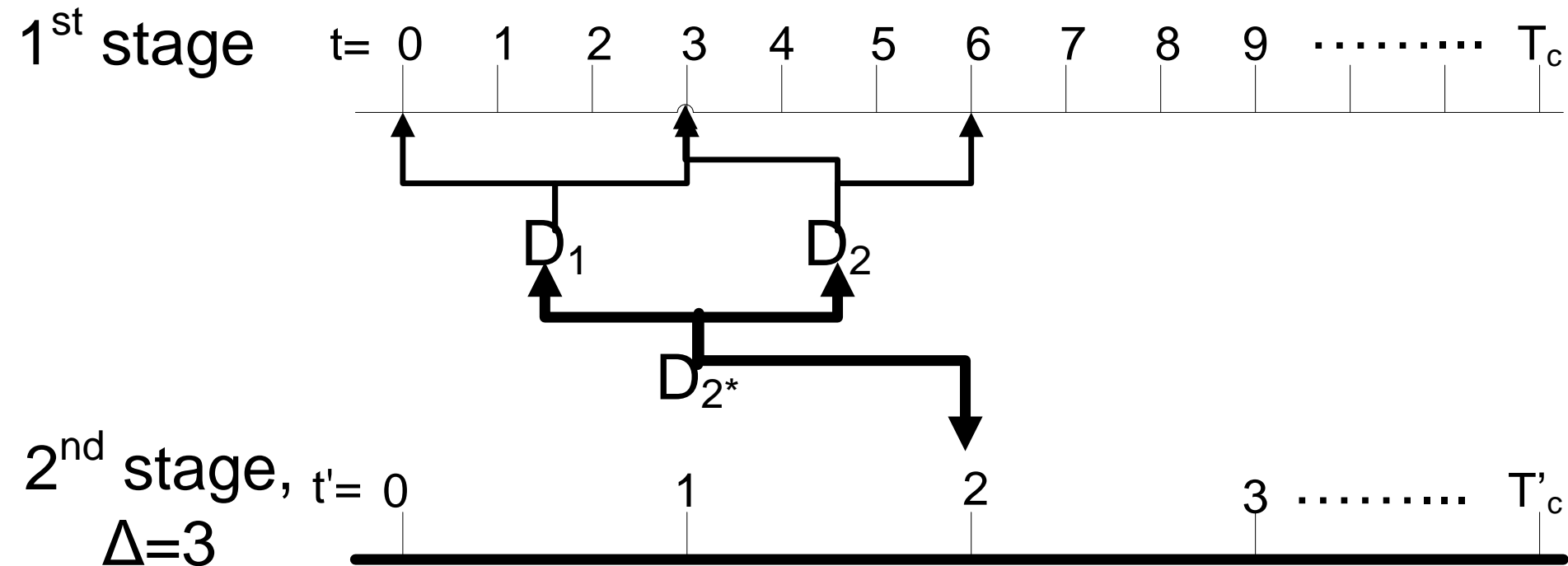
$$s'_w(t') = \begin{cases} \frac{H_{t^*} - H_{t'-1}}{H_{t'-1}} & \text{if } (H_{t^*} > H_{t'-1}); \\ 0 & \text{otherwise} \end{cases}$$

- Why 2nd-stage signal?

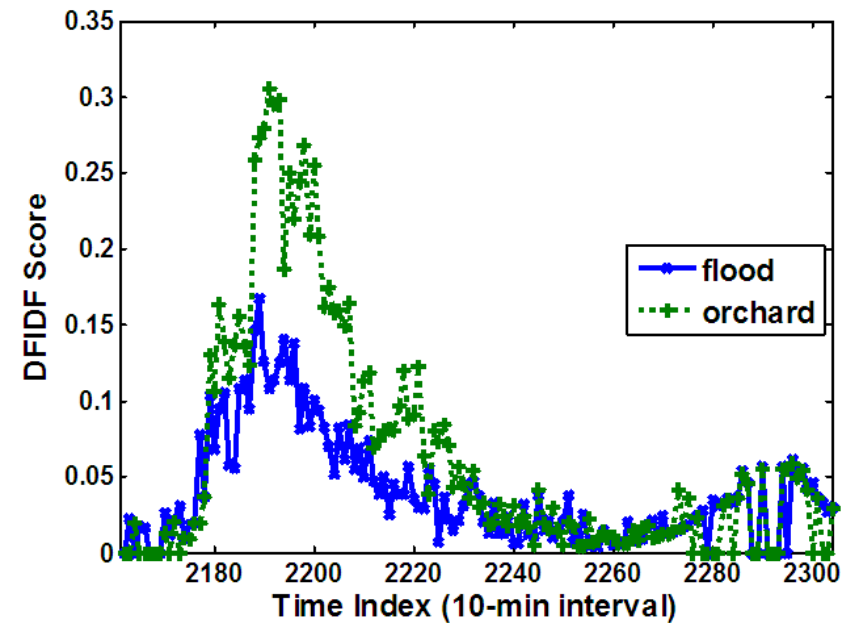
- Make the change in usage pattern more salient, easier to measure the similarity between features

- Compression: only non-trivial changes are stored

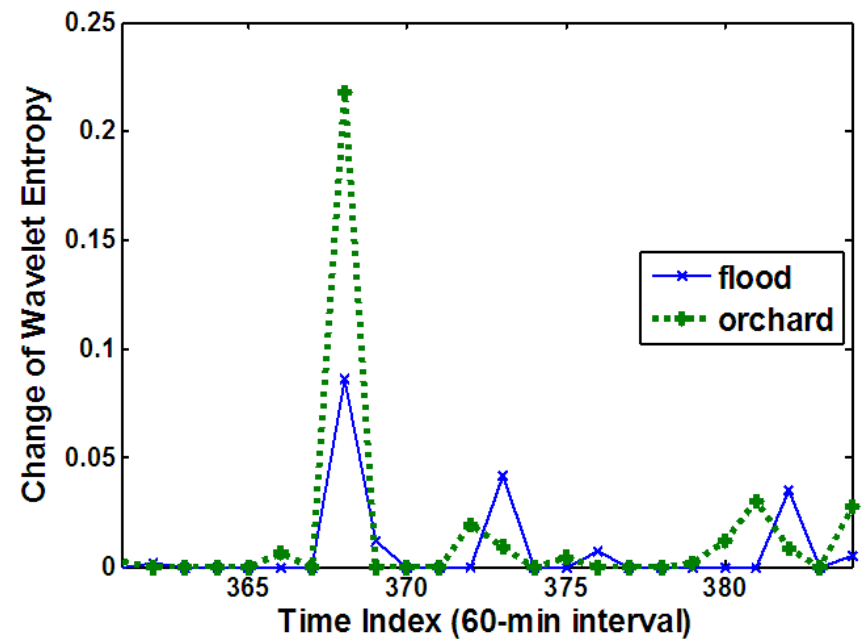
SIGNAL CONSTRUCTION (CONT'D)



SIGNAL CONSTRUCTION (CONT'D)



1st-order



2nd-order

FILTERING OF TRIVIAL FEATURES

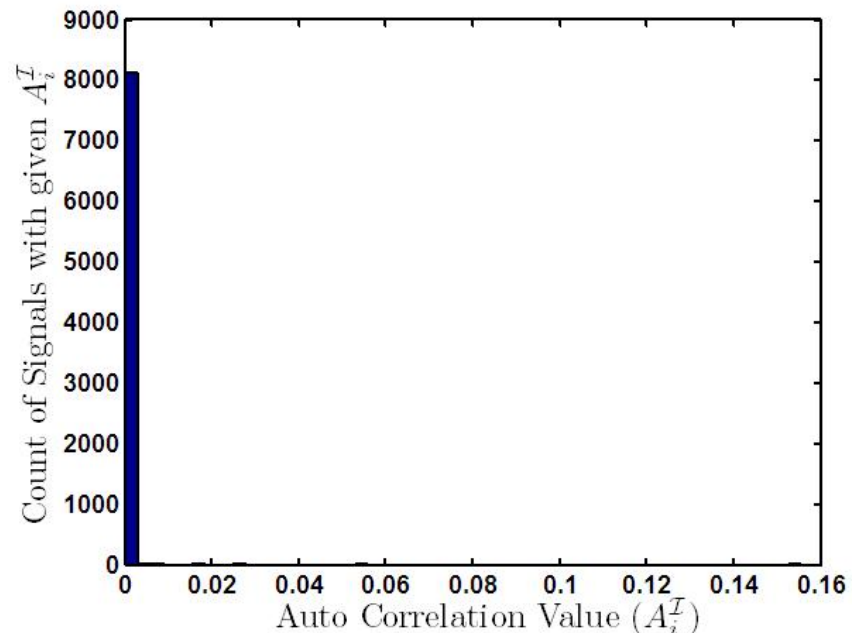
- Measure the *auto-correlation* of the signal corresponding to a feature

$$(f \star g)(t) = \sum f * (\tau)g(t + \tau)$$

- Filtering with *median absolute deviation* (MAD)

- $MAD(\mathcal{S}^{\mathcal{I}}) = \text{median}(|A_i^{\mathcal{I}} - \text{median}(A_i^{\mathcal{I}})|)$

- 5% features remain



MEASURE THE SIMILARITY BETWEEN FEATURES

- Similarity measured as the *cross-correlation* between signals

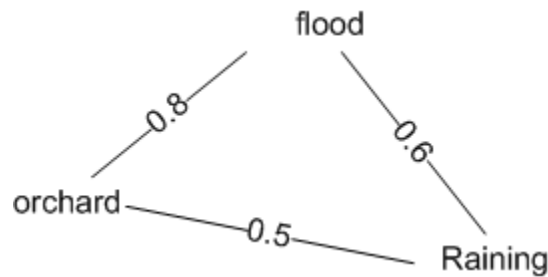
$$(f \star g)(t) = \sum f * (\tau) g(t + \tau)$$

- Ignore the similarity between features if it is too weak
 - Using Median Absolute Deviation
- $\mathcal{O}(n^2)$ complexity, but still tractable with small portion of remaining features



CLUSTERING FEATURES TO DETECT EVENTS

- Organize the cross-correlation between features as a matrix \mathcal{M}
 - Adjacency matrix of a graph \mathcal{G} .



\mathcal{G}

0	0.8	0.5
0.8	0	0.6
0.5	0.6	0

\mathcal{M}

CLUSTERING FEATURES TO DETECT EVENTS (CONT'D)

- Event detection can be formulated as a graph partitioning problem.
- Modularity can be applied to measure the quality of the partitioning
 - the sum of weights of all the edges that fall within sub-graphs (after partitioning) subtracted by the expected edge weight sum if edges were placed at random

$$Q = \frac{1}{2m} \sum_{ij} \left(w_{ij} - \frac{d_i \cdot d_j}{2m} \right) \delta_{c_i, c_j}$$

- The goal is to partition the graph so that Q value is maximized.



MEASURE THE SIGNIFICANCE OF THE EVENT

- $\epsilon = (\sum w_{ij}^c) \times \frac{e^{1.5n}}{(2n)!}$, $n = |V^c|$
- The first part $\sum w_{ij}^c$ sums up all the cross correlation values between signals (i.e. features) associated with an event
- The second part $\frac{e^{1.5n}}{(2n)!}$ discounts the significance if the event is associated with too many features
 - it is not reasonable for an event to be associated with too many words



EVALUATION

–Experimental

–Data

- Obtain the top 1000 Singapore-based *Twitter* users with the most followers from <http://twitaholic.com/>. U
- For each user in U , include her Singapore-based followers and friends within 2 hops. U^*
- For each user in U^* , collect the tweets published in June 2010.

–Remove stop words, apply stemming



Day	Event	€ value	Event Description
1-3	No event detected		
4	1. democrat, naoto	0.417	Ruling Democratic Party of Japan elected Naoto Kan as chief.
	2. ss501, suju	0.414	Korean popular bands Super Junior's and SS501's performance on mubank.
	3. music, mubank	0.401	Related to Event 2, mubank is a popular KBS entertainment show.
	4. shindong, youngsaeng	0.365	Related to Event 2, Shindong and Youngsaeng are member of the two bands.
	5. junior, eunhyuk	0.124	Related to Event 2, Eunhyuk is a member of super junior.
5	6. robben, break	0.404	No clear corresponding real-life even
6	No event detected		
7	7. kobe, kristen	0.417	Two events: Kristen Stewart won some MTV awards, and Kobe Bryant in a NBA match.
	8. #iphone4, ios4, iphone	0.416	iPhone 4 released during WWDC 2010
8	9. reformat, hamilton	0.391	No clear corresponding real-life event
	10. avocado, commence, ongoing	0.124	No clear corresponding real-life event
9	11. #failwhale, twitter	0.360	A number of users complained they could not use twitter due to over-capacity. A logo with whale is usually used to denote over-capacity.
10	12. vuvuzela, soccer	0.387	People started to talk about world cup.
11	13. #svk, #svn	0.418	#svk and #svn represent Team Slovakia and Slovenia in World Cup 2010.
12	14. #kor, greec, #gre	0.102	A match between South Korea and Greece in World Cup 2010.
13	15. whale, twitter	0.417	Similar as Event 10.
14	16. lippi, italy	0.326	Italy football team coach Marcello Lippi made some comments after a match in World Cup 2010.
15	17. drogba, ivory	0.417	Football player Drogba from Ivory Coast is given special permission to play in World Cup 2010.
	18. #prk, #bra, north	0.114	A match between North Korea and Brazil in World Cup 2010.
16	19. orchard, flood	0.357	Flood in Orchard Road.
17	20. greec, #gre, nigeria	0.122	A match between Greece and Nigeria in World Cup 2010.
18	21. #srb, podolski	0.403	A match between Germany and Serbia in World Cup 2010. Podolski is a member of Team Germany in World Cup 2010.
19-30	No event detected		



Table 1: All Events Detected by EDCoW in June 2010

EVALUATION (CONT'D)








- Practical case study
- Singapore General Election 2011
 - 7 parties
 - Most Singaporeans vote for the first time
 - Social media is allowed as a campaign platform for the first time



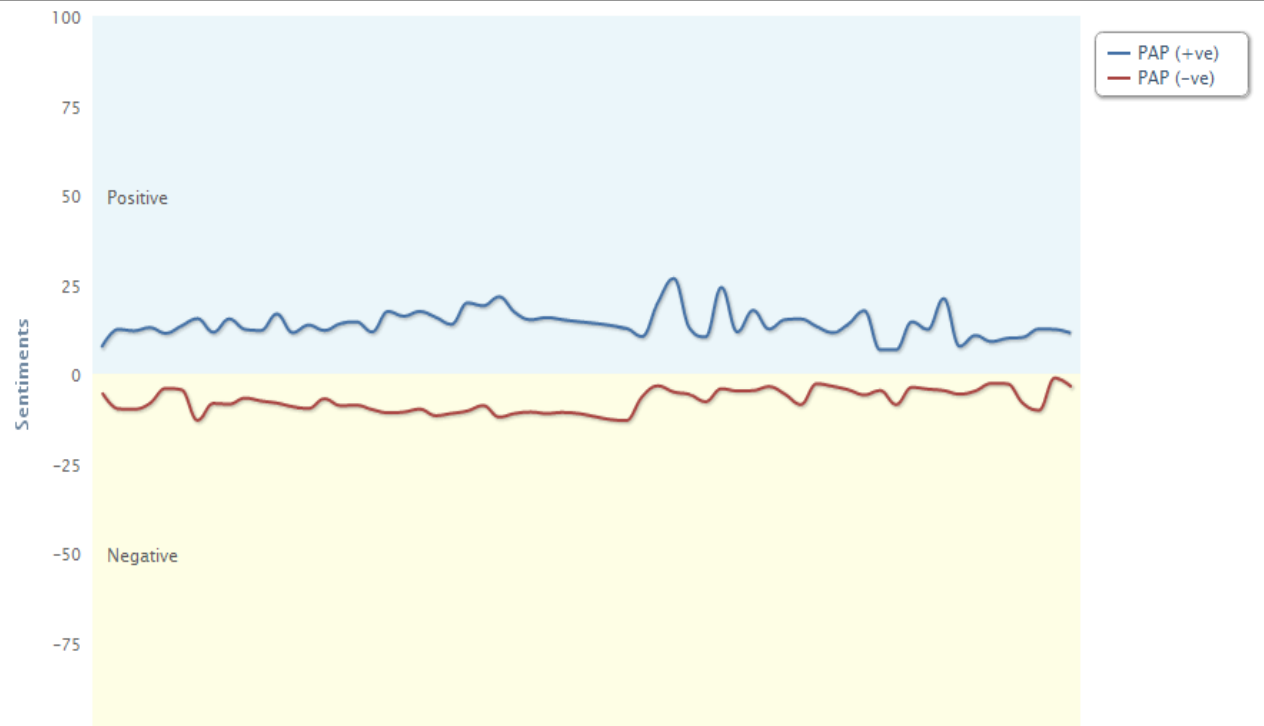
Political Party Sentiments

Sentiments on:  

Show All Parties

	National Solidarity Party (NSP)	16.3%
	People's Action Party (PAP)	55.3%
	Reform Party (RP)	1.7%
	Singapore Democratic Alliance (SDA)	2.0%
	Singapore Democratic Party (SDP)	10.7%
	Singapore People's Party (SPP)	8.2%
	Worker's Party (WP)	5.7%

Tag Cloud



Hot Topics #2

« Back to the main page

Sentiments on:



Political party

People's Action Party

Date

2011-05-08

☒ Negative ☒ Neutral ☒ Positive

Rate +ve = -ve

Sun May 08 2011 @ 02:58:40

RT @eisen: RT @jennyteo: I hope PAP finally realized that the loss of George Yeo as FM is the fault of no one but their GRC system. #sge ...

Rate +ve = -ve

Sun May 08 2011 @ 02:58:42

Finally, Sitoh Yih Pin has become the mayor of Potong Pasir after 27 years of check-in. #SgElections

Rate +ve = -ve

Sun May 08 2011 @ 02:59:48

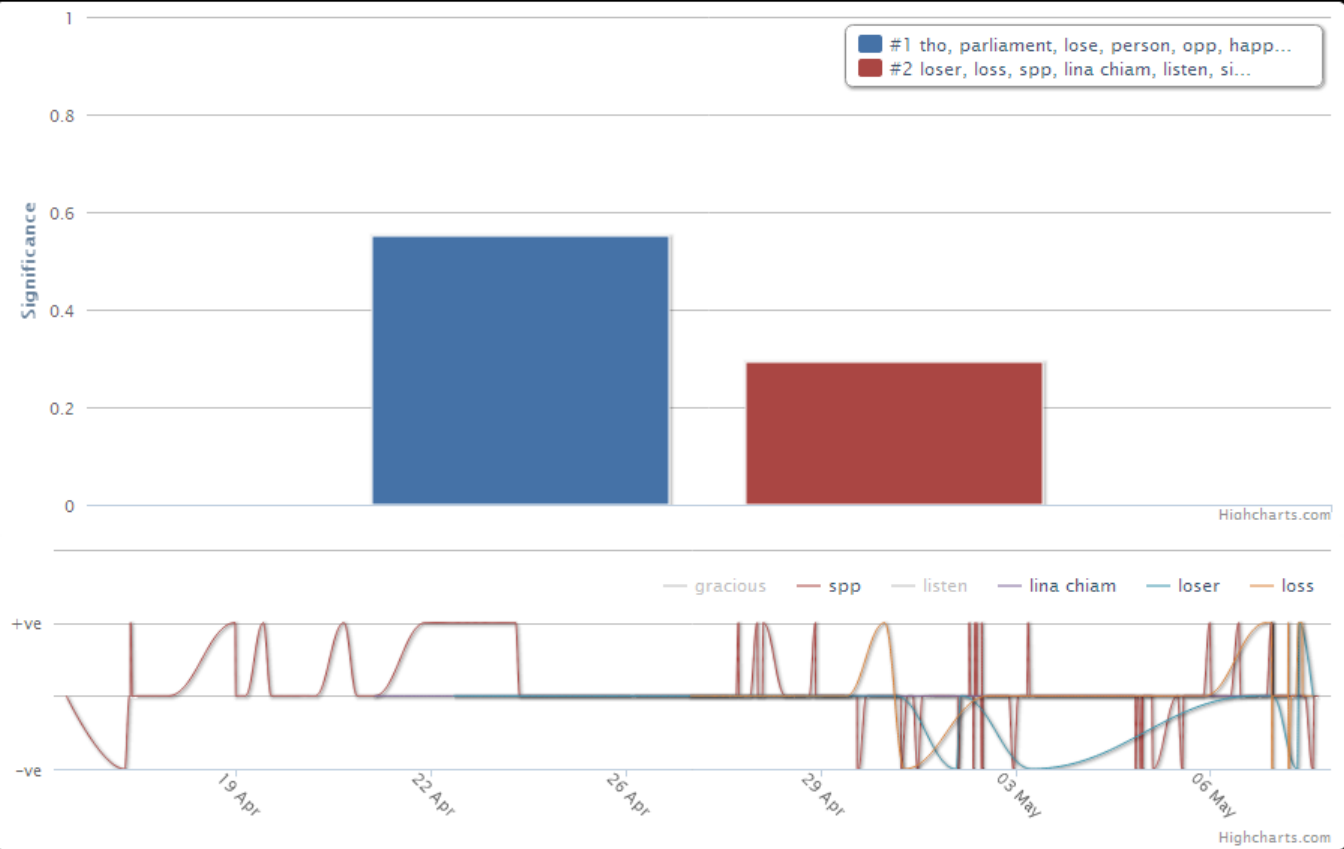
@flyrene well.. Everyones watching the PAP v closely for nxt 5 yrs. Next elections.. Would be worth waiting for. Poor lina chiam. Sighhhh.

Rate +ve = -ve

Sun May 08 2011 @ 02:59:49

RT @fakeMOE: Potong Pasir comes back to big boss PAP. Let them upgrade already, 5 years later go back to SPP. Smart tactic residents. #S ...

Rate +ve = -ve



SINGAPORE GENERAL ELECTION 2011

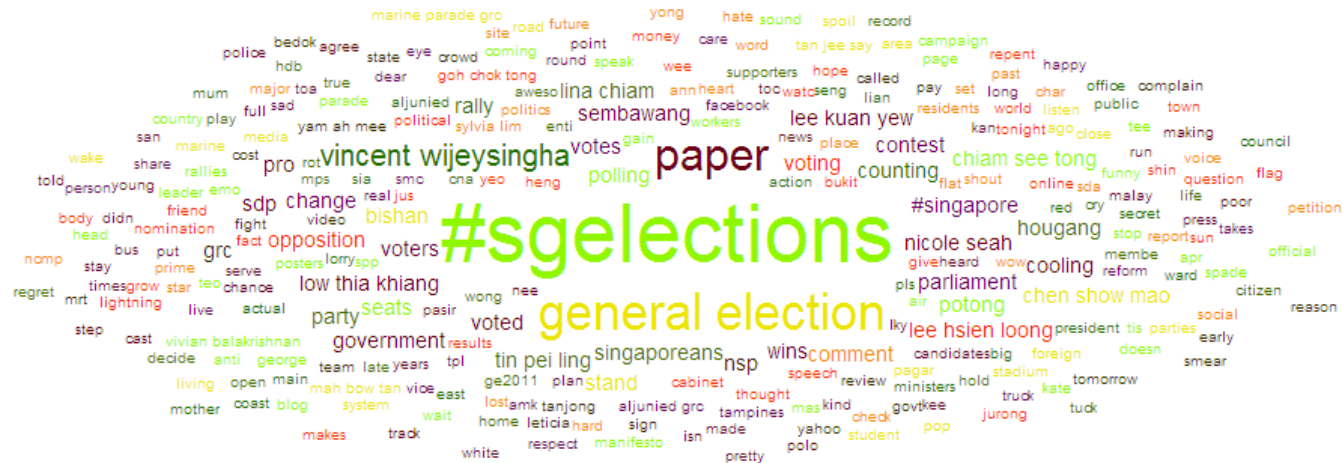


[Search](#) [About Us](#)

Tag Cloud

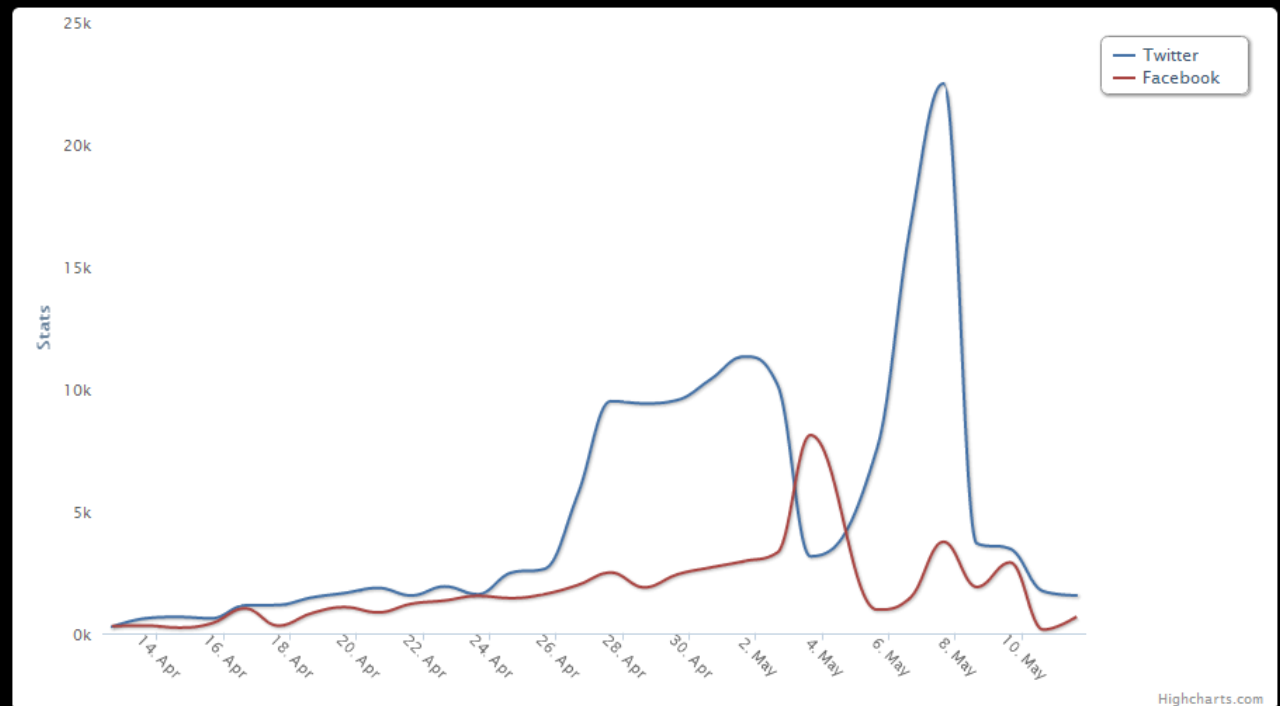
[« Back to the main page](#)

Sentiments on:



Volume Monitoring

[« Back to the main page](#)



Online popularity alone won't get you elected

Being a hot topic in cyberspace will not determine whether you win or lose an election.
By JASMINE OSADA

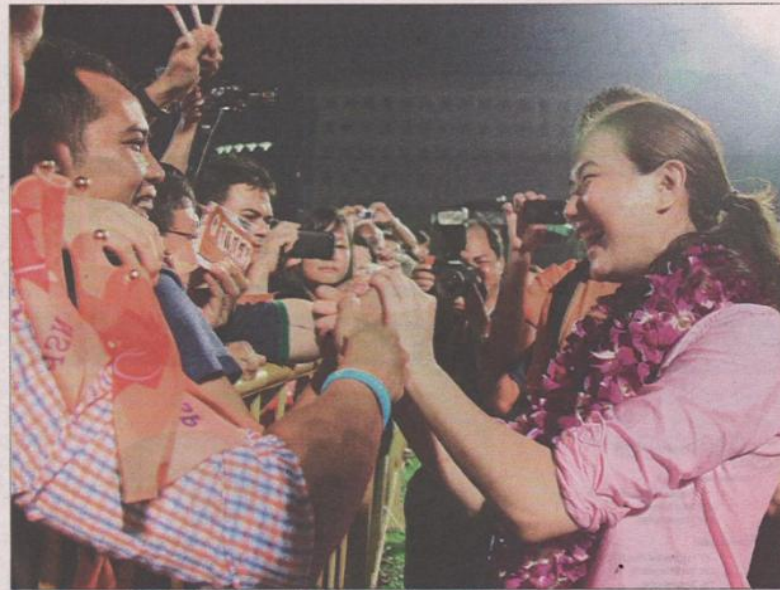
The General Election 2011, which saw candidates and voters engage each other directly on Facebook and Twitter pages, was a watershed for social media.

Data from **HP Labs Singapore** the computer maker's research arm – showed that social media chatter increased 30 times between April 13, when the online activity was first tracked, and Nomination Day on April 27.

Due to the relaxation of online campaigning rules and the ease in which information can be shared through social platforms, the online realm had created overnight celebrities out of some candidates while subjecting others to scrutiny and criticism.

Ms Tin Pei Ling, from the People's Action Party (PAP) who won the contest for Marine Parade GRC, bore the brunt of online criticisms. At least nine Facebook groups were created against the newly elected Member of Parliament. The largest, called "I Do Not Want Tin Pei Ling In Parliament" had some 45,000 members as of Monday.

On the other hand, candidates like the National Solidarity Party's (NSP) Ms Nicole Seah, shot to fame with the help of social media. The 24-year-old candidate, who was part of the NSP team which contested in Marine Parade GRC, was Singapore's most "liked" political figure by press time. She had 96,500 fans on the



PHOTOS: REUTERS, TERENCE TAN

Ms Nicole Seah of the National Solidarity Party is the most well-liked local political figure online but her popularity in cyberspace did not translate into enough votes to win her a seat in Parliament.

social networking site. Coming in second was Minister Mentor Lee Kuan Yew's Facebook page with 85,610 fans.

However, being well-liked online does not necessarily mean that one will score a victory at the polls. Social networking posts analysed by **HP Labs Singapore's** sentiment analysis software indicated that the NSP had garnered the most number of positive posts of all the seven political parties. Yet, the party failed to win a single contest at the polls. The ruling party PAP, which won 81 of 87 seats, came in at third place in HP's list after the Workers' Party.

HP's sentiment analysis software makes use of data from both

Facebook and Twitter to determine whether a certain topic is discussed positively or negatively online.

A large percentage of the social networking posts were made during the party rallies, as well as netizens blogging and uploading photos of

the rallies to their Facebook pages. Telcos here have reported a marked increase in the use of mobile data during the two-week campaign at rally sites, where networks experienced congestion issues from too many users trying to access data services at the same time.

Mr Ivan Lim, M1's deputy director of corporate communications and investor relations, told Digital Life that the telco saw network traffic increase two to four times at the more popular rallies, especially at those held by the Workers' Party.

Ms Cassie Fong, StarHub's corporate communications manager, said the telco saw a traffic increase of up to 15 per cent at certain rally sites as well.

Both telcos said there had been some congestion issues, such as users having to make more than one attempt to establish a voice call or experiencing slow data speeds. However, they noted that services promptly returned to normal once the crowd had thinned out.

A SingTel spokesman, who also acknowledged a significant increase in mobile broadband traffic, said traffic peaked between 7pm and 11pm during election rallies.

osadaj@sph.com.sg

WHAT NETIZENS WERE INTERESTED IN

Political parties ranked based on the amount of positive posts they received in social networking sites:

1. National Solidarity Party
2. Workers' Party
3. People's Action Party
4. Singapore Democratic Party
5. Singapore Democratic Alliance
6. Singapore People's Party
7. Reform Party

(Source: **HP Labs Singapore**, Sentiment analysis software)



SOME TWEETS

- PAP is changing! @PAPSingapore: Join PM Lee's live chat at the PAP FB page TONIGHT at 8pm.
<http://on.fb.me/kg4qiS> #sgelections #sgelection
- say Yes to Tin Pei Ling & get a free Kate Spade wallet
- Potong Pasir comes back to big boss PAP. Let them upgrade already, 5 years later go back to SPP. Smart tactic residents.



CONCLUSION

- Text analysis on Twitter content reveals discussion about “real-life” events
- Spaces for improvement:
 - Combine discussion-based similarity (signal cross correlation) with other factors
 - N-Gram
 - semantic similarity
 - Further exploit the social relationship among users
 - Larger scale study



Q&A

