

Master's Thesis

USING MICROBLOGS FOR CROWDSOURCING AND PUBLIC OPINION MINING

Cuneyt Gurcan Akcora
Ubiquitous Computing Laboratory
Department of Computer Science and Engineering
University at Buffalo
August 6, 2010



Overview

- **Background of Microblogging tools, and Twitter**
- A crowdsourcing approach: RainRadar Project
- Public Opinion Mining: Upinion Project

Microblogging Tools

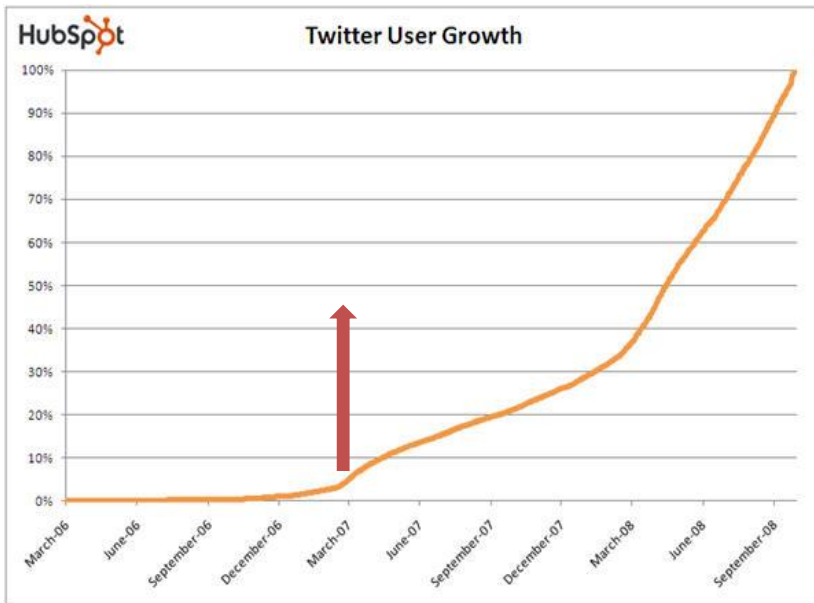
- Expression of the self in 140 characters.

- Facebook status update was the pioneer.



- Twitter has been evolving in its creator Jack Dorsey's mind since 2000, but started in 2005.

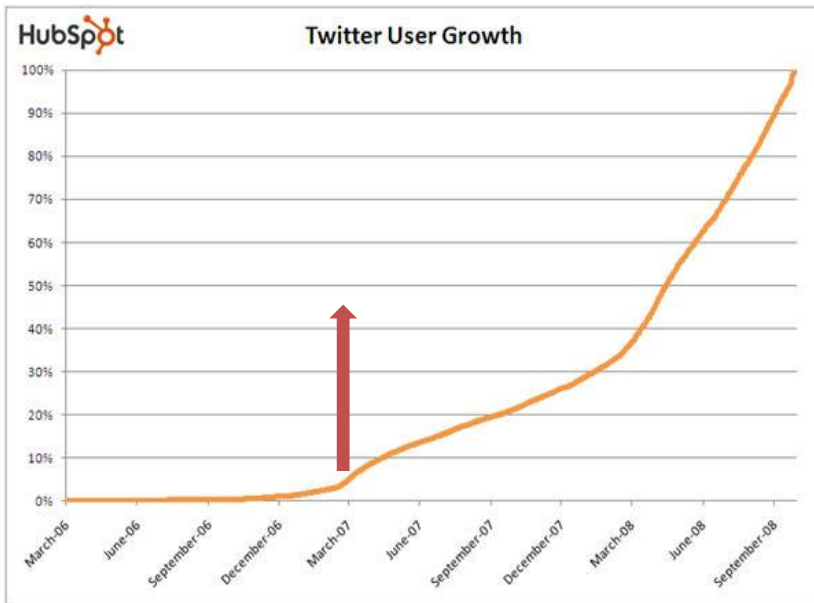
Twitter Growth



- Twitter took off after 2007 South by Southwest festival.



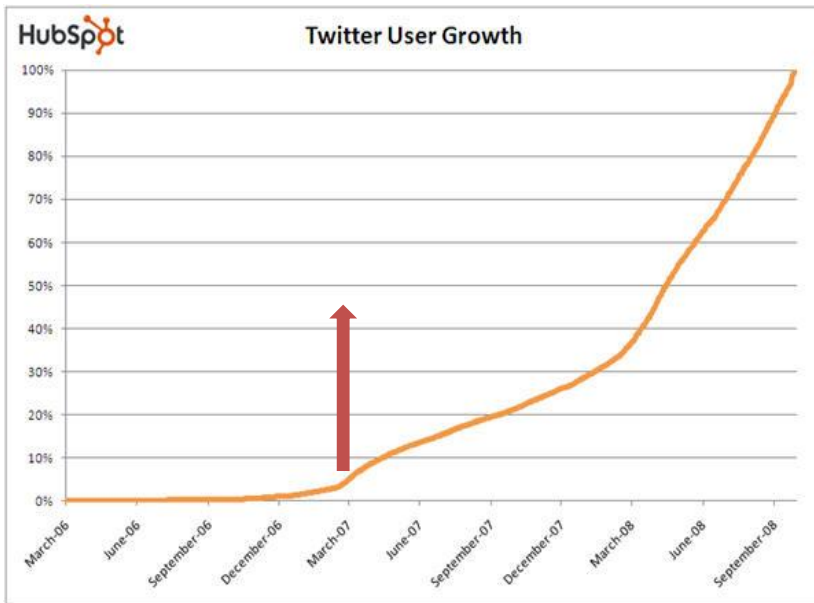
Twitter Growth



- Twitter took off after 2007 South by Southwest festival.
- SXSW is not a conference, it is an art festival. Twitter was able to reach many influential bloggers in the festival, and they helped spread Twitter's fame.

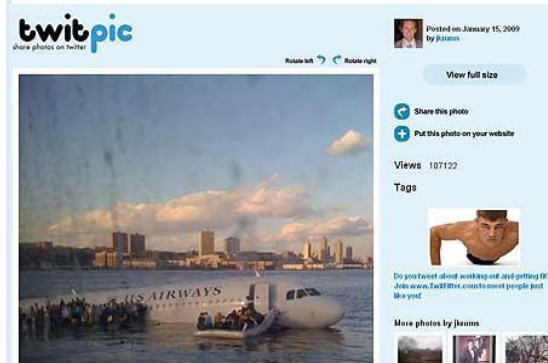


Twitter Growth



- Twitter took off after 2007 South by Southwest festival.

- Breaking news stories fueled Twitter growth.



- Emergency landing of a US Airways flight into the Hudson River

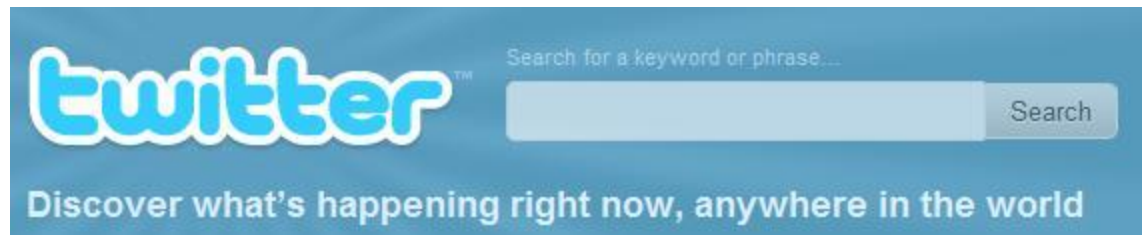


- Mumbai Terror attacks



A Vast Information Source

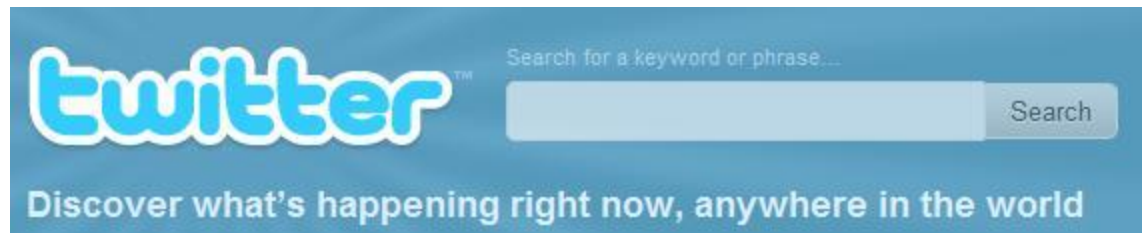
- Users do not need coherent ideas, or expertise. Anyone can post anything.



- Twitter posts are largely public, but you can also post privately.

A Vast Information Source

- Users do not need coherent ideas, or expertise. Anyone can post anything.



- Twitter posts are largely public, but you can also post privately.
- Many mobile and web applications
- Millions of users, billions of posts: A vast information source



RainRadar Project

- Microblogging tools, and Twitter background
- **A crowdsourcing approach: RainRadar Project**
- Public Opinion Mining: Upinion Project

➤ *Murat Demirbas, Murat Ali Bayir, Cuneyt Gurcan Akcora, Yavuz Selim Yilmaz, Hakan Ferhatosmanoglu,*
"Crowd-Sourced Sensing and Collaboration Using Twitter", WOWMOM 2010, Montreal, Canada, acceptance rate=21%.

Motivation

- With Twitter's open user base, we focused on designing location based collaborative crowd-sourced sensing systems for solving real life problems with wisdom-of-crowd effect.



Motivation

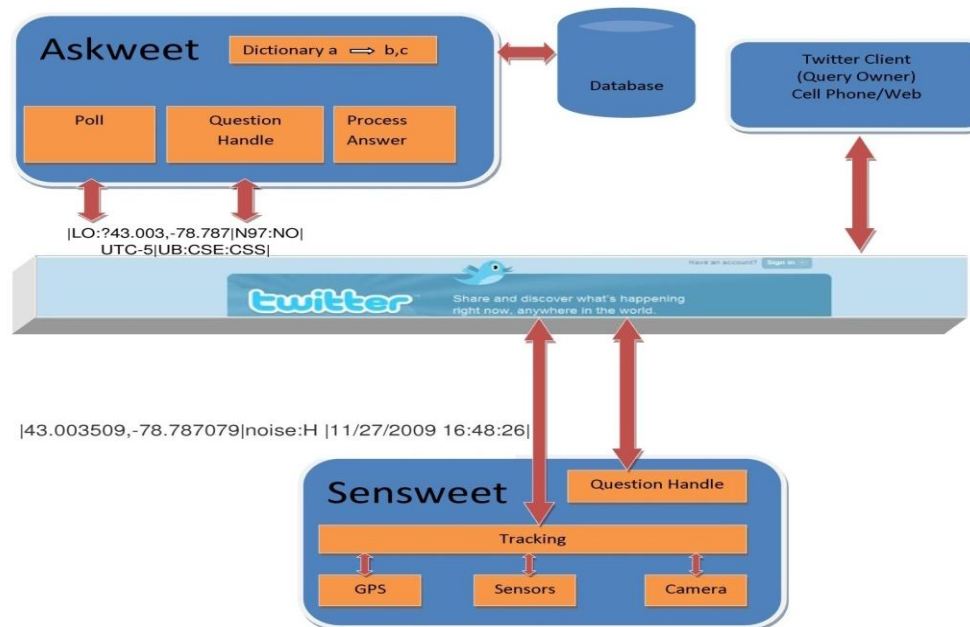
- With Twitter's open user base, we focused on designing location based collaborative crowd-sourced sensing systems for solving real life problems with wisdom-of-crowd effect.



- We propose that Twitter can provide an “open” publish-subscribe infrastructure for sensors and smartphones and pave the way for ubiquitous crowd-sourced sensing and collaboration applications.

Architecture

- **Askweet:** A server side application that pushes questions to Sensweets (based on their locations) and collects answers for Query owners (Twitter Clients).



- **Sensweet:** A smart-phone client or Sensor gateway device to publish sensed data including location information to Twitter.

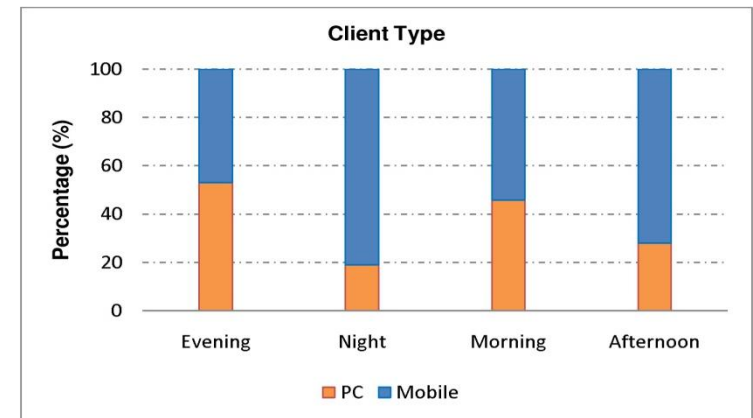
Application

- We created very fine granularity weather maps of cities in USA and Canada.



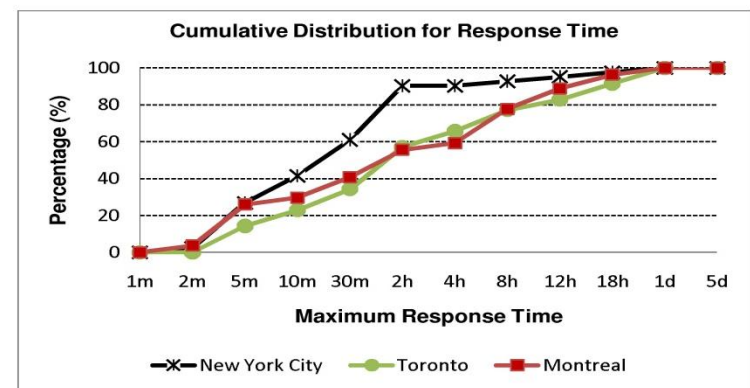
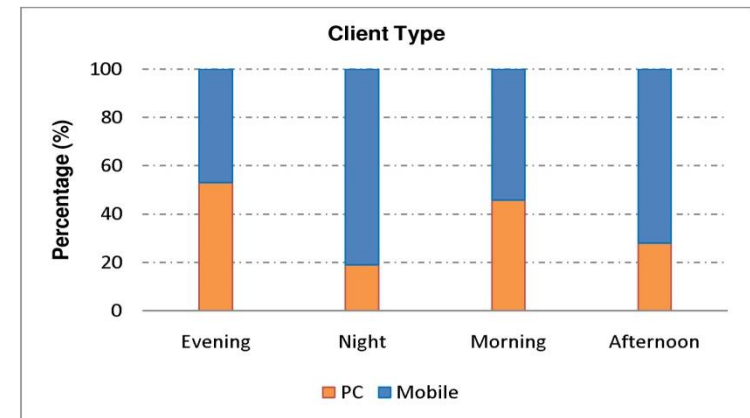
Application

- We created very fine granularity weather maps of cities in USA and Canada.



Application

- We created very fine granularity weather maps of cities in USA and Canada.



Upinion Project

- Microblogging tools, and Twitter background
- A crowdsourcing approach: RainRadar Project
- **Public Opinion Mining: Upinion Project**

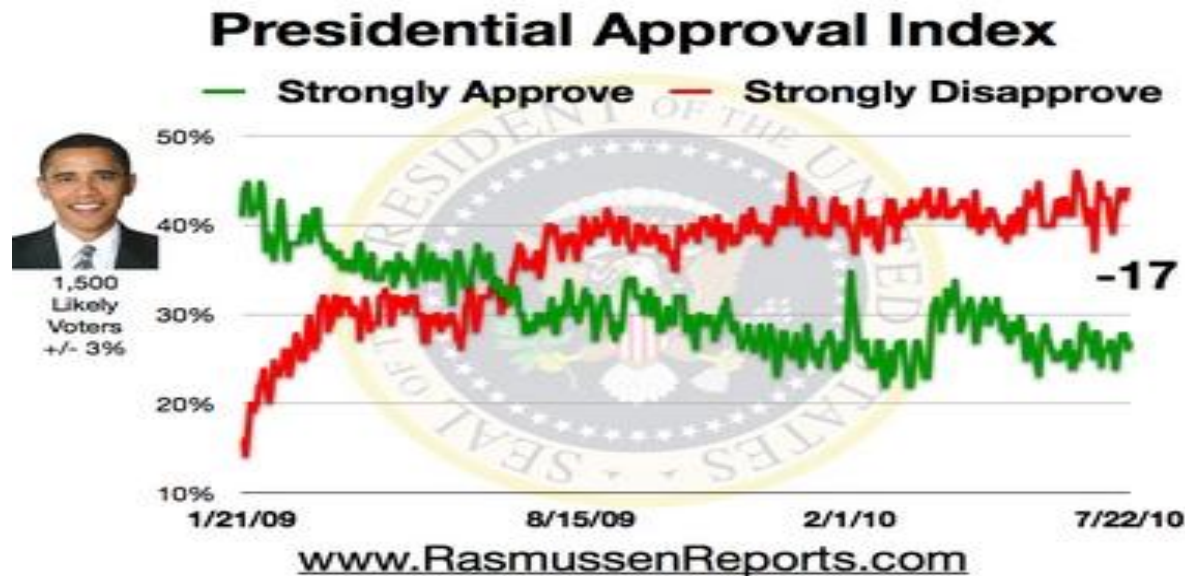
➤ *Cuneyt Gurcan Akcora, Murat Ali Bayir, Murat Demirbas, Hakan Ferhatosmanoglu, "Identifying Breakpoints in Public Opinion", SIGKDD, Social Media Analytics 2010, Washington D.C, USA, also accepted to talks, talk acceptance rate=20.6%.*

Motivation

- Polling enables us to understand what public thinks about a certain topic.

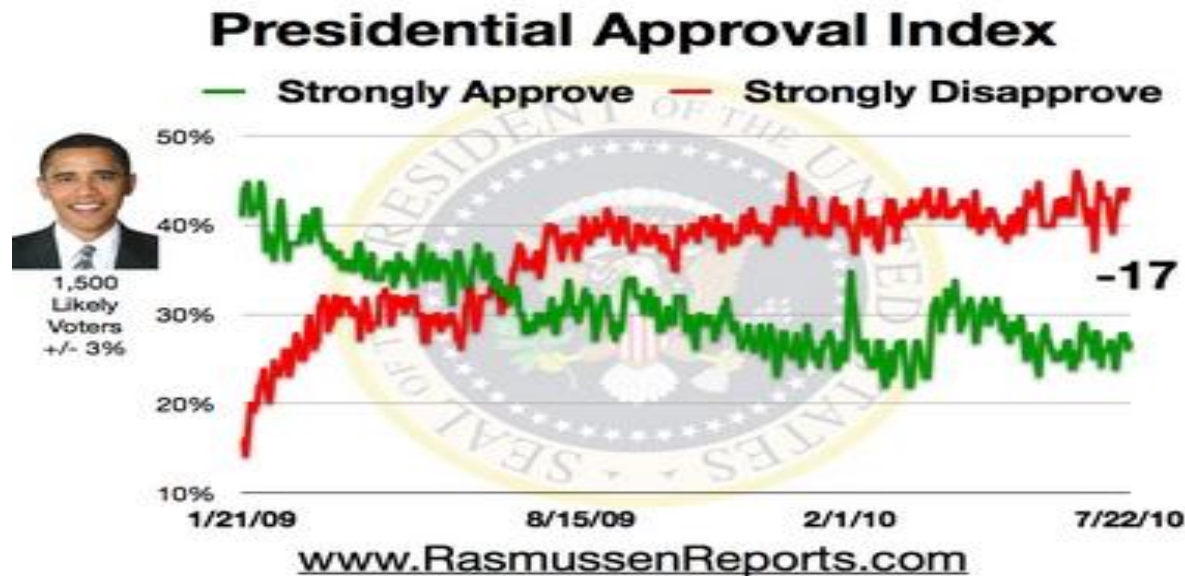
Motivation

- Polling enables us to understand what public thinks about a certain topic.
- Yet polling does not provide clues about how public opinion changes through time, and **why**?



Motivation

- Polling enables us to understand what public thinks about a certain topic.
- Yet polling does not provide clues about how public opinion changes through time, and **why**?



15,000 interviews per month

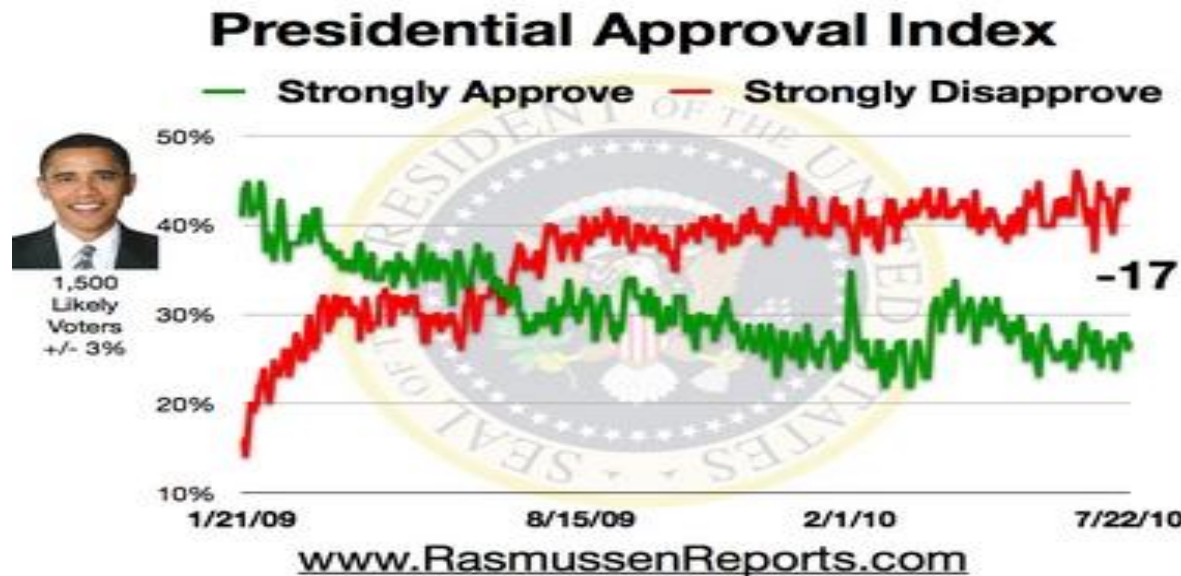
500 per day

2500 minutes, ~40 hours

5 interviewers

Motivation

- Polling enables us to understand what public thinks about a certain topic.
- Yet polling does not provide clues about how public opinion changes through time, and **why**?



15.000 interviews per month

500 per day

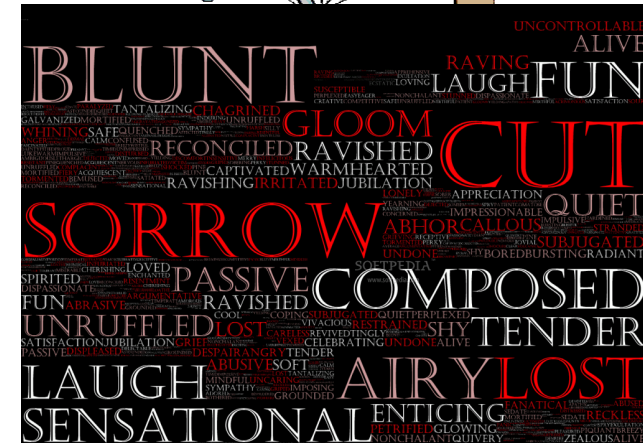
2500 minutes, ~40 hours

5 interviewers

**If you have
~400.000\$**

Problem Definition

- Can we use a microblogging tool (Twitter) to observe public opinion through time?
- **Efficient:** Tumasjan et al. [1] predicted German elections only by counting party mentions in tweets.



- Can we find the events that changed the public opinion?

Observations: What signals Opinion Change?

- 1- An opinion changing event brings about a new **emotion pattern** in tweets. This new emotion pattern is repeated in the following time period(s).
 - 2- In the time periods following the event, public still post about the issue with quite **similar wording**.
- To claim a change in the public opinion, the emotion pattern and the word pattern must change according to these observations.

Finding emotion(s) in Tweets

- Mining each tweet for emotions.
- Employing multi-grained sentiment analysis with eight emotion classes:
 - Anger, Fear, Sadness, Love, Joy, Shame, Surprise, Disgust

Finding emotion(s) in Tweets

- Mining each tweet for emotions.
- Employing multi-grained sentiment analysis with eight emotion classes:
 - Anger, Fear, Sadness, Love, Joy, Shame, Surprise, Disgust
- Mapping a tweet to an eight dimensional emotion vector:

Tweet: *I was on main street in Norfolk when I heard about tiger woods updates and it made me feel angry, on 2009-12-11.*
Emotion vector: *(1, 0, 0, 0, 0, 0, 0, 0).*
- A tweet can have more than one emotion.

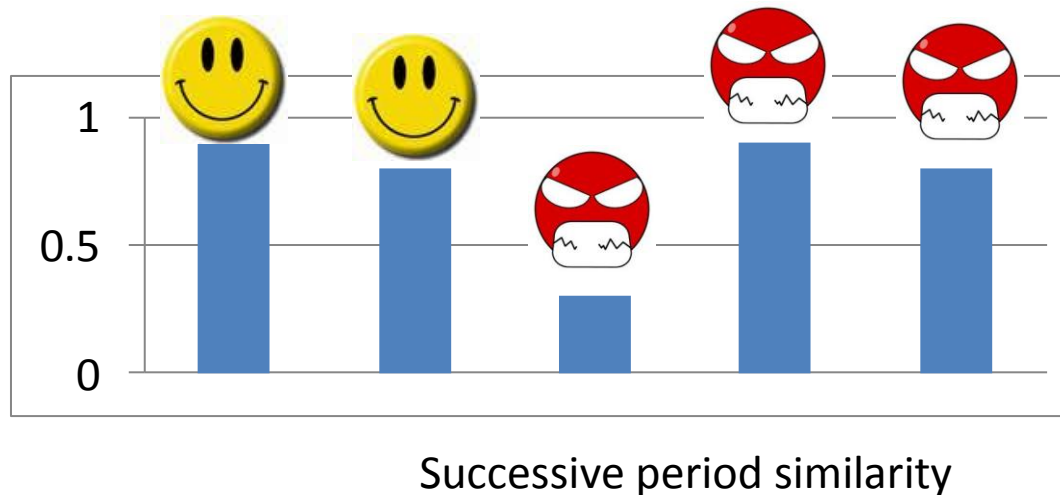
Cosine Similarity for detecting emotion pattern change

- We will define time periods, and compare the emotion patterns of two successive time periods.
- For every $n=12$ hour period, find a **centroid of the emotion vectors from tweets** .



Cosine Similarity for detecting emotion pattern change

- For every two successive periods, calculate cosine similarity of **centroids**.
 - The newly emerging **emotion pattern** is detected here.



Jaccard Similarity for detecting word pattern change

- For every n=12 hour period, get a bag of words (take all **unique** words from tweets).
- We will use set based Jaccard similarity.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

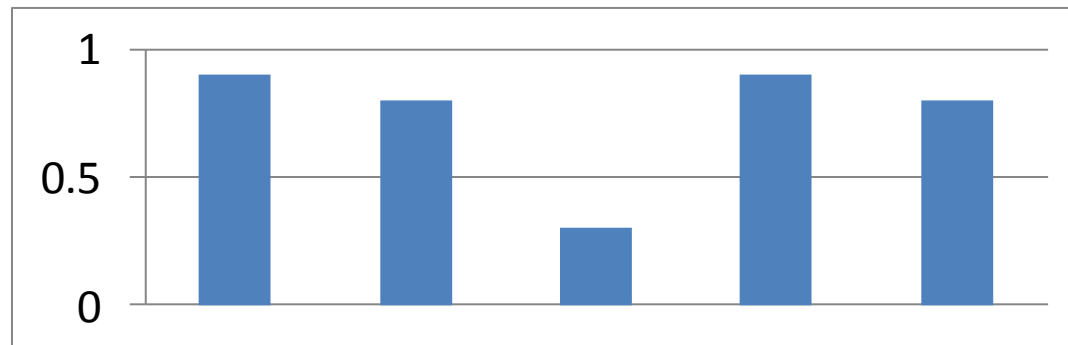


Jaccard Similarity for detecting word pattern change

- For every two successive periods, calculate the Jaccard similarity.
 - The newly emerging **word pattern** is detected here.



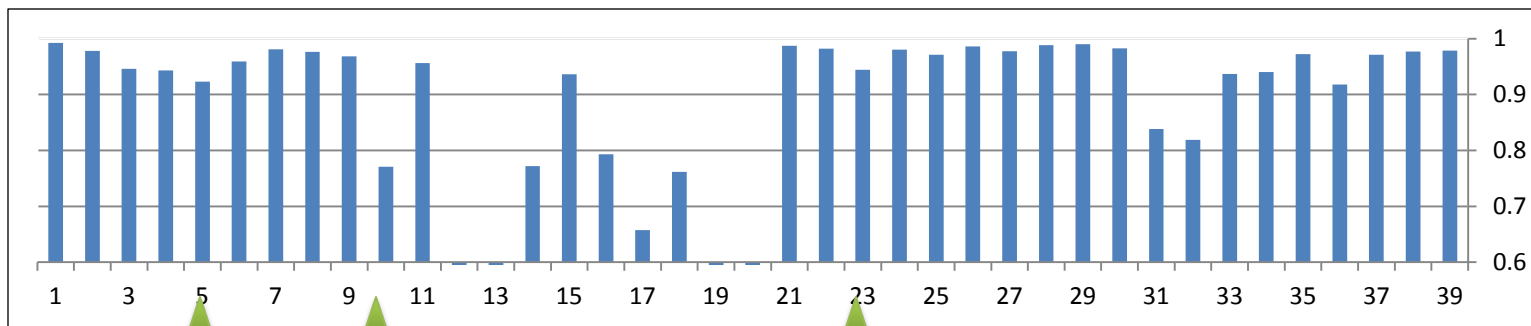
Jaccard similarity values



Successive period similarity

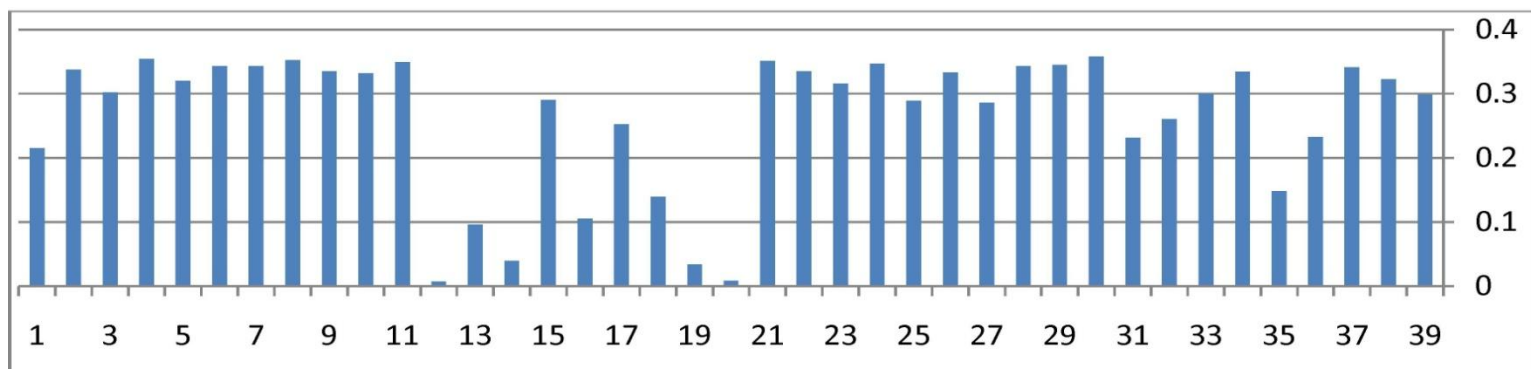
Combining the two similarity methods

Cosine
Similarity



Breakpoint
False Positive

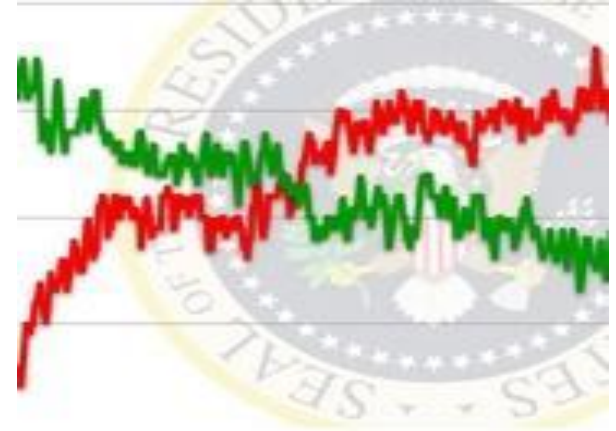
Jaccard
Similarity



Similarity graphs for Tiger Woods case

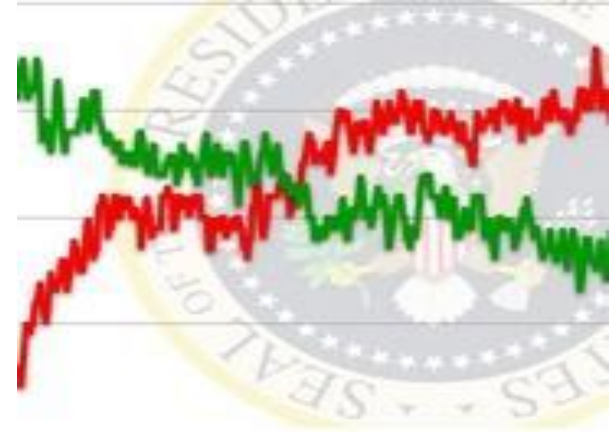
How do we account for the breakpoints?

- Can we summarize details of the event(s) that brought about that change?



How do we account for the breakpoints?

- Can we summarize details of the event(s) that brought about that change?
- Breakpoint Representation: Extracting keywords to describe breakpoints.



How do we account for the breakpoints?

- The **term count** in the given document is simply the number of times a given term appears in that document.
- This count is usually normalized to prevent a bias towards longer documents.
- For normalization, we use **Term Frequency, Tf**.
- The **Inverse Document Frequency, IDF**, is a measure of the general importance of the term for a document.
- Obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

Streaming TfIdf

- We created an algorithm that is dynamic, i.e, takes into account incoming documents.
- Each new period is examined in the **initial phase** of the algorithm.
- The algorithm detects breaking news, and pinpoints the first interval that the news emerged.
- In the **document update phase**, we re-calculate the importance of a word when the word keeps appearing in the new periods.

Breakpoint Representation

- We are detecting $p=5$ prominent words to summarize the breakpoint.
- The Streaming TfIdf Algorithm: **Initial Phase**

$$Tf_{x,D_n} = \frac{Count_{x,D_n}}{\sum_m Count_{m,D_n}}$$


$$Idf_{x,D_n} = \log\left(\frac{n}{|\{\forall k, k \leq n : x \in D_k\}|}\right)$$

Breakpoint Representation

- We are detecting $p=5$ prominent words to summarize the breakpoint.
- The Streaming TfIdf Algorithm: **Initial Phase**

$$Tf_{x,D_n} = \frac{Count_{x,D_n}}{\sum_m Count_{m,D_n}}$$

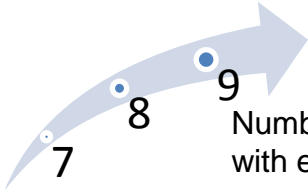
In traditional cases, number of documents, n , is known beforehand, and does not change



$$Idf_{x,D_n} = \log\left(\frac{n}{|\{\forall k, k \leq n : x \in D_k\}|}\right)$$

Breakpoint Representation

- We are detecting $p=5$ prominent words to summarize the breakpoint.
- The Streaming TfIdf Algorithm: **Initial Phase**

$$Tf_{x,D_n} = \frac{Count_{x,D_n}}{\sum_m Count_{m,D_n}}$$


Number of documents increasing with every new period

$$Idf_{x,D_n} = \log\left(\frac{n}{|\{\forall k, k \leq n : x \in D_k\}|}\right)$$

- n is increasing with every new period.

Breakpoint Representation

- The Streaming TfIdf Algorithm: **Document Update Phase**

$$tf_{x,D_o} = tf_{x,D_o} + F(D_o, D_n) \times tf_{x,D_n}$$

$$F(D_o, D_n) = 1/(n - o)$$

- Sort all words of a period according to their TfIdf values, and set top-5 words as the prominent words of that document.

Breakpoint Representation

- **Document Update Phase:** How it works

$$tf_{x,D_o} = tf_{x,D_o} + F(D_o, D_n) \times tf_{x,D_n}$$

$$F(D_o, D_n) = 1/(n - o)$$

- Cheat
- Accident

1

- Golfer
- Tournament

5

- Cheat
- Tournament

9


Breakpoint Representation

- Document Update Phase: How it works


$$tf_{x,D_o} = tf_{x,D_o} + F(D_o, D_n) \times tf_{x,D_n}$$

$$F(D_o, D_n) = 1/(n - o)$$

- Cheat
- Accident



- Golfer
- Tournament



- Cheat
- Tournament



$$tf_{\text{Cheat}, 1} = tf_{\text{Cheat}, 1} + F(\text{1}, \text{9}) \times tf_{\text{Cheat}, 9}$$

Breakpoint Representation

- Prominent words for the *Tiger Woods* case:

Period	Prominent words
1	crash, florida, injur, golf, accident
6	crash, wife, accident, mistress, golf
11	voicemail, wife, f***, golf, cheat
24	drop, stop, santa, claus, gatorade

Breakpoint Representation

- Prominent words for the *Tiger Woods* case:

Period	Prominent words
1	crash, florida, injur, golf, accident
6	crash, wife, accident, mistress, golf
11	voicemail, wife, f***, golf, cheat
24	drop, stop, santa, claus, gatorade

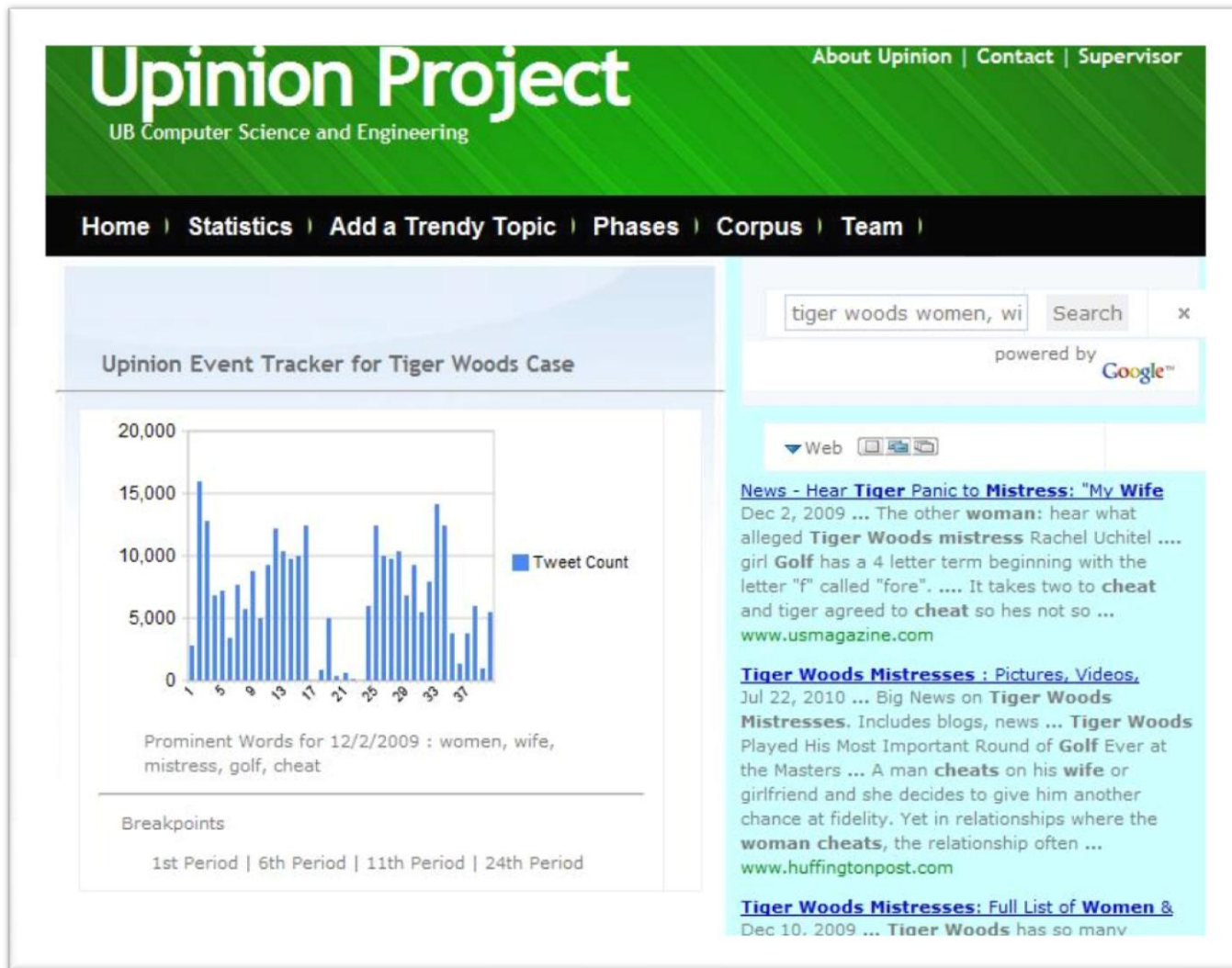
Accident

Talks about a mistress

Woods leaves a voicemail to his mistress

Gatorade ends sponsorship, Twitter users start posting thousands of jokes about Woods

Explaining Breakpoints



Future Work

- So far, we have worked with the emotion words that sociologists handed us.
- But more words are used to convey emotion. We have to find and classify these words. We can do this on Mechanical Turk.
- Classifying opinion changing events according to how much public opinion sways.
- Implementing a real time system for every user entered topic.

Thanks

- My advisor Professor Demirbas, and Professor Qiao.
- Murat, Onur and Yavuz for their time and effort.
- I thank all my colleagues from the Ubiquitous Computing lab.

References

1. Predicting Elections with Twitter:What 140 Characters Reveal about Political Sentiment

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner,
Isabell M. Welp. *AAAI 2010*