



---

# Module 6: Summarisation and Rumour Detection





## About this tutorial

---

- This tutorial will be a mixture of explanation, demos and hands-on work
- Things for you to try yourself are in red
- It assumes basic familiarity with the GATE GUI and with ANNIE and JAPE; no Java expertise



# Summarising Social Media

---



---

## Do we suffer from information overload?

- Home users don't feel overloaded.. (Hargittai 2012)
- But they can't keep up with everything and need help with filtering noise (Bontcheva 2013)

## When overload arises (Hargittai 2012):

- **Time sensitivity:** Limitation of time for reviewing available information
- **Decision requirement:** time constraints on actual decision making
- **Structure of information:** The extent to which information is structured, to help retrieval of relevant information
- **Quality of information:** Filter failure (Clay Shirky) or signal-to-noise ratio



---

## What is Text Summarisation

Summary	Purpose
Journal abstract	Indicative
News report	Informative
Movie review	Critical
Novel blurb	Indicative
Football highlights	Informative



---

## Types of Text Summarisation

- What is being summarised:
  - Single document summarisation
  - Multi-document summarisation
- What is in the summary
  - Extractive summarisation
  - Abstractive summarisation



# How are extractive summaries created

---

1. Score textual units (e.g. sentences) according to some representation of the document or document set
2. Generate summaries by selecting high scoring textual units until some desired compression ratio has been achieved.



# Scoring methods

---

1. Frequency based methods
2. Sentence position in the document
3. Centrality scores



# SUMMA: Text Summarisation in GATE

---



- Implemented by Horacio Saggion (now at UPF)
- GATE Plugin available from <http://www.taln.upf.edu/pages/summa.upf/>
- Both single- and multi-document extractive summarisation



# Single document summ. example

---

BBC News - Fees cut as part of emergency passport backlog measures

Home Secretary Theresa May has announced measures to help clear the backlog in passport applications - and scrapped charges for urgent renewals.

People renewing their UK passports from overseas will be given a 12 month extension to their existing passport.

Those applying for passports overseas on behalf of their children will be given emergency travel documents.

Labour said Mrs May had lost her grip and called for an apology.

Mrs May was being grilled on the situation by Labour's Yvette Cooper in the Commons - two days after denying claims of a crisis.

Labour MP Geoffrey Robinson called the timing of this change - at the pre-summer peak applications season - "idiotic".

But Mrs May and other Home Office ministers have blamed the problems on a surge in applications which they say are running at a 12 year high.

'Huge turnaround'

Mrs May stressed that security would not be compromised by the changes, telling MPs that parents issued with emergency travel documents for their children would still have to "provide comprehensive proof that they are the parents before we will issue these documents".

In the longer term, the passport office could be stripped of its agency status and brought directly under Home Office control "in line with other parts of the immigration system", she told MPs.

She has also ordered a review by her department's top civil servant, Mark Sedwell, into how the agency could be run more efficiently, saying it was not just a question of "throwing more staff at the problem".

Nick de Jong fears his family might not make his brother's wedding in Italy because of passport delays

The Public and Commercial Services union said it would support the move of the Passport Office into the Home Office, saying staff were currently paid less than equivalent Home Office staff.



## Summary of 4 news articles: Example

---

- Multi-document summarisation example
- 4 news articles (BBC, Guardian, Independent, Telegraph) on the passport crisis to be summarised
- 10% of all sentences to be selected

On Wednesday Prime Minister David Cameron said up to 30,000 passport applications had been hit by delays. Passport Office chief executive Paul Pugh - who will be questioned next week by MPs on the Commons Home Affairs Select Committee - said there had been "exceptional" summer demand but that extra staff had been brought in to handle applications.

Applying for a passport

A premium service, costing £128, means passports can be collected within four hours of being approved. Under the fast-track service, costing £103, a passport is returned within a week of the application being approved. For over-16s applying for their first adult passport, the Passport Service says people should allow for at least six weeks to receive it.

He said almost three million passports had been issued for UK customers in 2014, including over one million issued in the eight weeks since the start of April.

The Public and Commercial Services Union has said it does not know how many applications have been delayed. It has claimed the loss of a tenth of the agency's workforce in the past five years and the closure of local passport offices have contributed to "major problems".

# Google Product Review Summaries




Google product search



Nikon D90 Digital SLR Camera with Nikon AF-S DX 18-105mm lens

\$852 online

★★★★★ 1,013 reviews  41 people +1'd this

August 2008 - Nikon - SLR - 12.3 megapixel - Optical Viewfinder - Crop Sensor - 5.8 x optical zoom - CMC SD - SDHC - Pop-up Flash - 22.4 ounce - ISO 6400








[« Back to overview](#)

## Reviews

Summary - Based on 1,013 reviews



What people are saying

- pictures**  "Easy to use and the picture quality is superb"
- zoom/lens**  "Very good and versatile lens."
- features**  "Lots of cool features."
- design**  "Great pictures and easy controls."
- video**  "easy to use, video is clear"
- screen**  "Fast, easy to use, great lcd, etc, etc..."
- value**  "A low cost benefit rating."



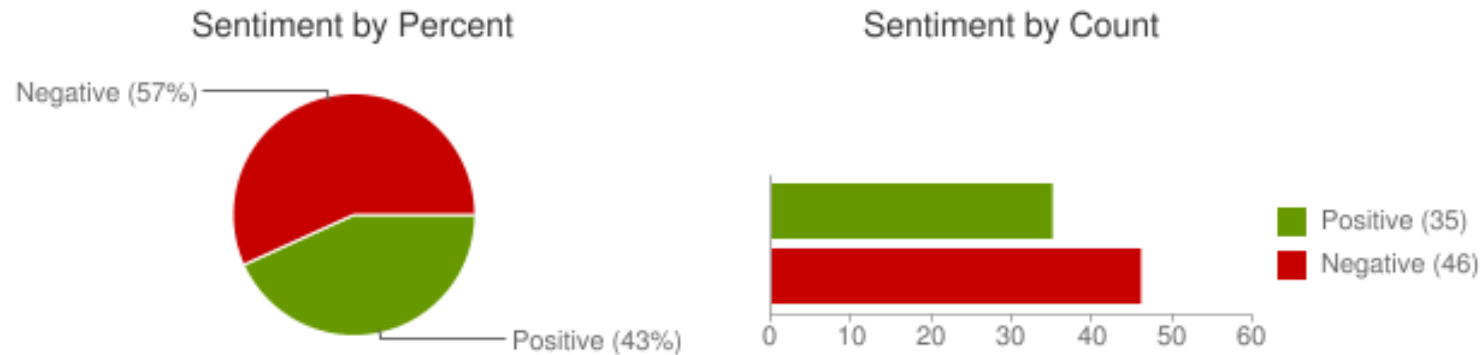
## Twitter Sentiment

Type in a word and we'll highlight the good and the bad

Search

[Save this search](#)

### Sentiment analysis for iPod



### Tweets about: iPod



# Sentiment Summarisation in Twitter

## Twitter Sentiment

Tweet < 273

Like < 319

+1 < 20

"Whitney Houston"

Search

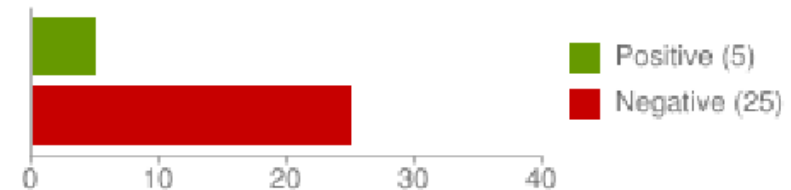
[Save this search](#)

### Sentiment analysis for "Whitney Houston"

Sentiment by Percent



Sentiment by Count





# Micropinions (1)

---

- Concise opinion snippets
  - Non-redundant, readable, representative

## Micropinion Summarization

Please enter one tweet per line:

```
t seen the limitations that are implied all over the web .
in the real world , this works great i now must sound like a fanboy , and i guess i just might be now .
when a product exceeds your expectations , and works so well , it .
s hard not to appreciate it .
bought at the apple store while on vacation , so i could spend my vacation learning to play with the new phone .
apple has superb customer service .
```

↓ Summarize ↓

- many happy people
- new free phone
- nice new phone
- best free apps
- best mobile device
- same new version

## Micropinions (2)

---



- How:
  - Find high frequency unigrams
  - From these generate all possible bigrams
  - Merge overlapping bigrams to produce longer phrases as long as they are:
    - Readable
    - Representative





# Micropinions: Summarising Opinions

---

- Concise opinion snippets (Ganesan et al, 2012)
  - Non-redundant, readable, representative
- How:
  - Find high frequency unigrams
  - From these generate all possible bigrams
  - Merge overlapping bigrams to produce longer phrases iff they are readable and representative
- Example:

Mp3 Player Y	Restaurant X
Very short battery life. Big and clear screen. (8 words)	Good service. Delicious soup dishes. Very noisy at nights. (9 words)



# Summarising Tweets

---

## The challenge: making sense of streams

Automotive RDFa (a horribly researched SEO article on RDFa/Microformats): <http://ow.ly/5JSoS> #somanerrorsitsfunny.

'Impossible for police force to meet Govt target of doing more for less' - Home Affairs Cttee chair responds to 34,000 jobs cuts by 2015.

Call for Papers: Sentiment Elicitation from Natural Text for Information Retrieval and Extraction <http://t.co/93UnxJF> Dec 10 Vancouver

Browsers used for month's visits to @SentimentSymp site: Mozilla 61%, Safari 20%, Internet Explorer 15%; Google driving ~25% of traffic.

Sony Music appoints Nick Gatfield to head its British operations <http://bit.ly/pkqXy0>

Twitter's plan: Take a chunk of sales tweets? <http://bit.ly/oKd8lQ>

Daily Number: 12% of adults owned an e-reader as of May, double the number last November <http://t.co/ew5NlW8>

Hitler's deputy Rudolf Hess exhumed <http://bbc.in/q8E6g2>

Nokia Posts Huge Quarterly Loss, Sees Better Times Ahead - <http://on.mash.to/nCSh4i>

'Public will notice difference' as police forces 'transform' to meet job cuts challenge - ACPO's Chief Const Chris Sims <http://bbc.in/p0NJ1Y>

Fox News lags in coverage of News Corp scandal <http://t.co/97x5s7D> ;how the Journal has changed under Murdoch \n<http://t.co/Fybmqne> H9eR

Fox News lags in coverage of News Corp scandal <http://t.co/97x5s7D> ;how the Journal has changed under Murdoch \n<http://t.co/elfsd00>

Fifth patient has died at Stepping Hill Hospital, Gtr Manchester after receiving contaminated saline <http://bbc.in/o0LkLD>

I think I can stop banging my head against the wall now ...

Just stumbled across #Google's What Do You Love #search site -- <http://t.co/1r1tV4d> -- aggregates results of various types.

#Google's What Do You Love #search has an interesting results-window slider on the left of the SERP <http://t.co/1r1tV4d>

'We are determined to identify & bring to justice person responsible' for saline deaths at Stepping Hill Hospital - Asst Chief Const Hopkins

RT @datastore (Gdn data blog) #phonehacking: what happened when? Visualised <http://bit.ly/puFtby> - you've got to see this.

~~News has to be subsidised - cheap and free! - Clay Shirkly <http://bit.ly/edeKUP>~~



# Why is Social Media Summarization Hard?

---

Short messages (microtexts), URLs, #tags

Noisy:

- Unusual spelling (2moro)
- Bad capitalisation
- Emoticons
- Idiosyncratic abbreviations (ROFL)
- Large variance in styles

Temporal

Social context

- Our relationship to the tweet's author influences importance

User-generated

- Gender
- Location
- Age

Multi-lingual: fewer than 50% of tweets are English



## Key Questions in Tweet Summarization

---

- What is a summary for a set of tweets?
- Select phrases/entities from the tweets, as a high level overview
- Extract opinions and summarise these
- Select the most representative subset
- Temporal aspect:
  - Are more recent tweets on a topic more important?
- How to make use of natural topic groupings like user lists and hash tags



## Term/Entity Clouds as Frequency-based Summaries

---

### Pros:

- Give a high level topic and entity summary/overview of disparate tweets
- Easily understood and widely used
- Good starting point for interactive summarisation

### Cons:

- Do not show opinions
- Frequency  $\neq$  Interesting/Important

Took a random sample of 450 tweets from news agencies (BBC, Guardian, CNN)  
Ran NE recognition and then plotted based on frequency

Afghanistan Andy Schleck Bing China David Cameron France  
Greece India James **James Murdoch** Libya  
Lucian Freud Matt Nixon Matt Nixson **Murdoch**  
OFA Prince Andrew Rebecca Black Rebekah Brooks  
Somalia **UKUS**

Improvements: normalise names (Nixon vs Nixson), handle co-reference (Murdoch vs James Murdoch)

University of Sheffield, NLP

# Tweet Ranking and Summarisation

---



University of Sheffield, NLP

# Tweet User Geolocation

---



University of Sheffield, NLP

# Rumour Detection

---







## Veracity: the 4<sup>th</sup> V of Big Data

The PHEME project focuses on detecting rumours in social media  
<http://pheme.eu>

- We coined the term *phemes*
  - **memes** are thematic motifs that spread through social media in ways analogous to genetic traits
  - **phemes** add truthfulness and deception to the mix
  - named after ancient Greek PHEME, "*embodiment of fame and notoriety, her favour being notability, her wrath being scandalous rumours*"

<http://en.wikipedia.org/wiki/PHEME>



# Social media is rife with phemes

File Edit View History Bookmarks Tools Help

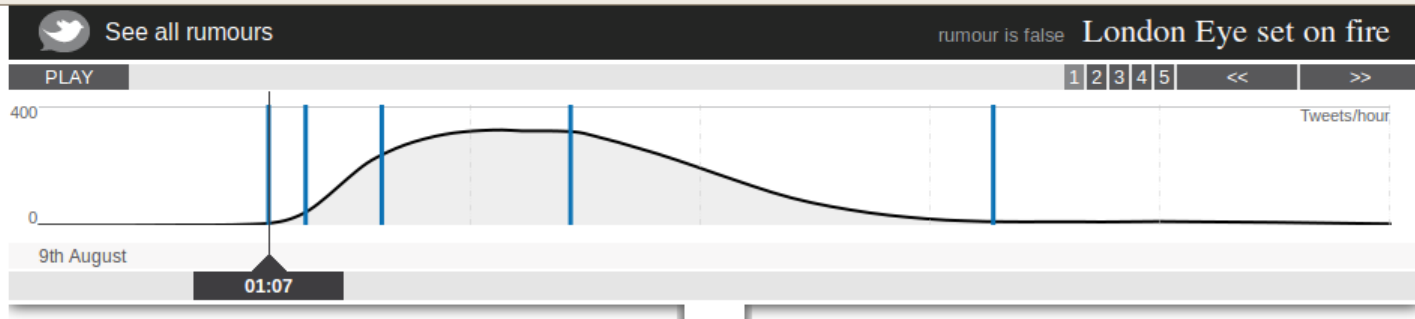
How riot rumours spread on Tw... +

www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter

irdian visualisation london riots

See all rumours rumour is false London Eye set on fire

PLAY 1 2 3 4 5 << >>



400  
0  
Tweets/hour

9th August  
01:07

### How the rumour unfolded

Shortly after 1am on Tuesday, August 9, @zadio posts a link to an image of the London Eye apparently ablaze with the heartfelt message: 'Oh my God! This can't be happening!' The tweet is quickly picked up.

“

Oh my god! This can't be happening at London Eye! #Londonriots #Londonriot #Prayforlondon <http://twitpic.com/6372vo>

@zadi0, 18 followers

Tue 9 Aug 01:07

Influence of the tweet

more influential  
less influential

Relation to the rumour

support recent 2h old opposition recent 2h old query recent 2h old comment recent 2h old

# Social Media is Rife with Phemes (2)

File Edit View History Bookmarks Tools Help

How riot rumours spread on Tw... +

www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter

irdian visualisation london riots

See all rumours rumour is false London Eye set on fire

PLAY 1 2 3 4 5 << >>

400 Tweets/hour

0

9th August 01:15

### How the rumour unfolded

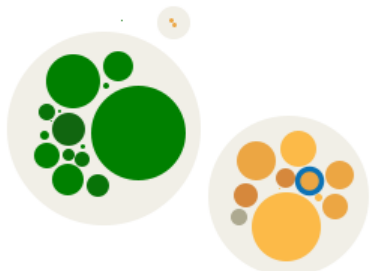
Shortly after 1am on Tuesday, August 9, @zadio posts a link to an image of the London Eye apparently ablaze with the heartfelt message: 'Oh my God! This can't be happening!' The tweet is quickly picked up.

“

Oh my god! This can't be happening at London Eye! #Londonriots #Londonriot #Prayforlondon <http://twitpic.com/6372vo>

@zadio, 18 followers

Tue 9 Aug 01:07



Influence of the tweet

more influential

less influential

Relation to the rumour

support recent 2h old

opposition recent 2h old

query recent 2h old

comment recent 2h old

# Social Media is Rife with Phemes (3)



How riot rumours spread on Twitter | UK news | guardian.co.uk - Mozilla Firefox

File Edit View History Bookmarks Tools Help

How riot rumours spread on Tw... +

www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter

See all rumours rumour is false London Eye set on fire

PLAY 1 2 3 4 5 << >>

9th August 01:37

### How the rumour unfolded

@Leginho, a user with only 4 followers, responds to @carvin suggesting the Eye can't be on fire because the iron it is comprised of would 'need much heat'. (The London Eye is in fact made of steel.)

“

@acarvin: londoneye is not flammable, it must be a hoax. Paint burns much different, to burn iron needs much heat #londonriots

@leginho, 4 followers

Tue 9 Aug 01:36

Influence of the tweet: more influential (larger circle), less influential (smaller circle)

Relation to the rumour: support (green), opposition (red), query (yellow), comment (grey). Legend also includes 'recent' and '2h old' for each category.



## The UK riots study

---

- 2.6M tweets harvested from Twitter ‘fire hose’ matching specified #tags.
- 700,000 individual accounts.
- What the corpus can reveal about:
  - Reactions to events, both general and specific
  - How information flows through social media
  - Kinds of ‘actors’ involved and how they shape discourse
  - How social media used to inform, organise, etc

Procter, R., Vis, F. and Voss, A. (2013).  
Reading the riots on Twitter: methodological  
innovation for the analysis of big data.  
International Journal of Social Research  
Methodology, Special Issue on Computational  
Social Science: Research Strategies, Design &  
Methods.

Procter, R., Crump, J., Karstedt, S., Voss, A., &  
Cantijoch, M. (2013). Reading the riots: what were  
the police doing on Twitter?. Policing and Society, 1-

### **3. Rumours**

**3.1 Claim without evidence**

**3.2 Claim with evidence**

**3.3 Counterclaim without evidence**

**3.4 Counterclaim with evidence**

**3.5 Appeal for more information**

**3.6 Comment**



# Rumour analysis

---

- Technological challenges
  - Analysis is post-hoc, on 7 known rumours
  - Analysis and visualisations took months of researcher and programmer effort
- Rumours are challenging
  - Some rumours could take days, weeks or even months to die out
  - Ill-meaning humans can currently outsmart computers and appear genuine



## Available Datasets

---

### Observations

- Misinformation and disinformation tend to be questioned more than facts, attract more affirmations and denials/refutations, and result in deeper conversation threads (Mendoza et al, 2011)

### Datasets:

- Small set of tweets annotated with 5 rumours, classified as confirm, deny, and question (Qazvinian et al, 2011)
  - Many of the tweets have since been deleted (!)
  - 30% for one of the rumours, which makes results replication hard
- The seven rumours from the London riots tweets



## Qazvinian et al 2011 & our ongoing work

---

- Used POS tags, n-grams, URL features, retweets, etc.
- We added new features: document-intrinsic(stylistic, personality, sentiment, entities)
- Experimenting on the London riots data too





---

QUESTIONS?