Automatic Detection of Rumours in Social Media: Methods and Challenges

Kalina Bontcheva

K.Bontcheva@dcs.shef.ac.uk

@kbontcheva

Veracity: the 4th V of Big Data

PHEME focuses on a fourth crucial, but hitherto largely unstudied, challenge: veracity

- 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*"

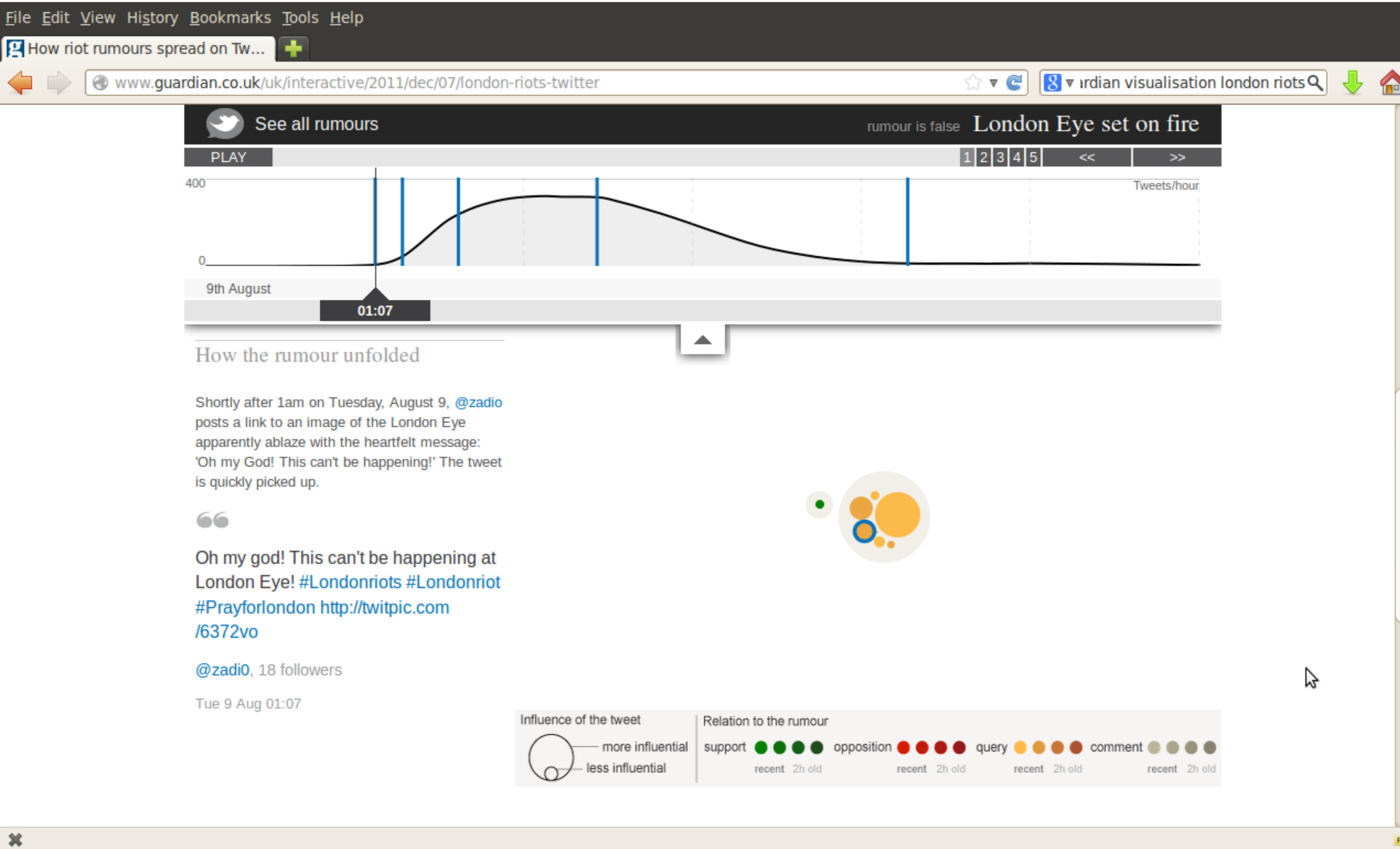
http://en.wikipedia.org/wiki/Pheme

What is a Rumour?

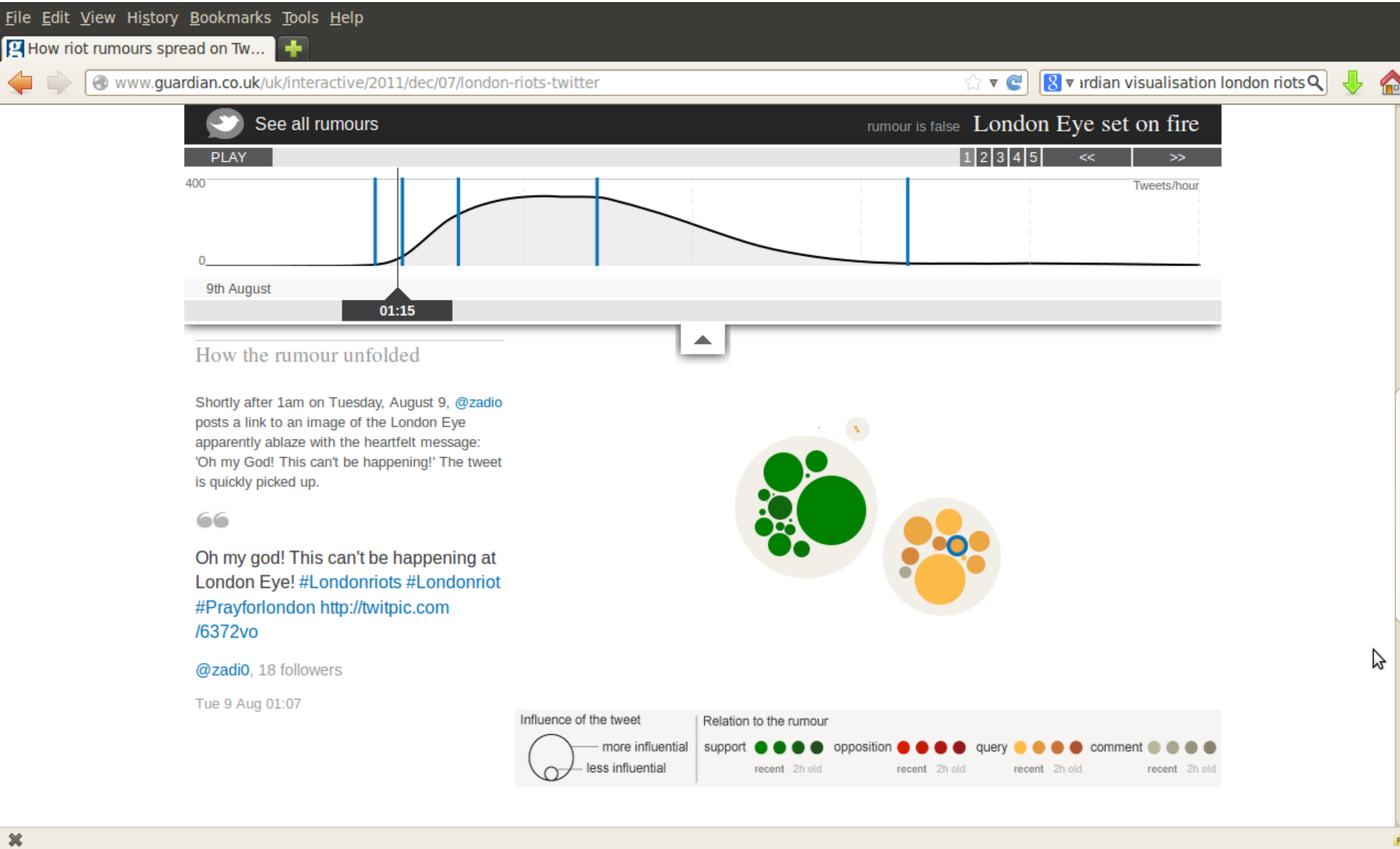
- Our definition
  - “a circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety”

European Commission
Information Society and Media

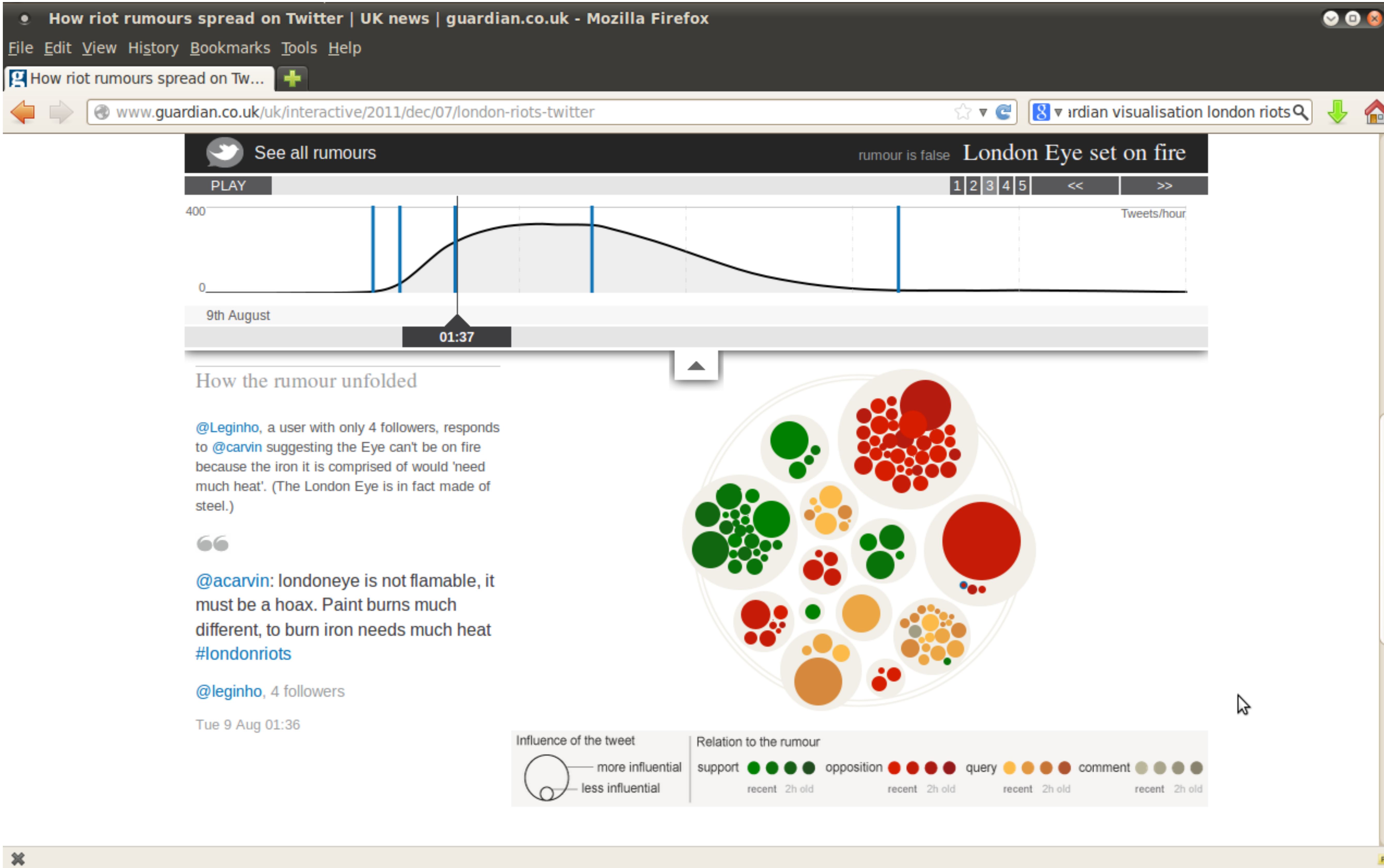
Social Media is Rife with Phemes



Social Media is Rife with Phemes (2)



Social Media is Rife with Phemes (3)



Rumour analysis: The Challenges

- Journalists do it mostly manually
- Rumour analysis is challenging
  - Some rumours could take days, weeks or even months to die out
  - Ill-meaning humans can currently outsmart computers (and humans) and appear genuine
  - Role of conversational threads in rumour analysis
  - Tweet-rot and deletions
  - Real-time analysis of temporally dynamic phemes

What: High Level Goal

- Create a computational framework for automatic discovery and verification of rumours, at scale and fast

- Draw on:
  - NLP: what's said
  - web science: a priori knowledge from Linked Data
  - social science: who spread it, why and how
  - information visualisation: visual analytics

European Commission
Information Society and Media

Content verification

PHEME

- A number of projects focussed on detecting inaccurate reports, which has recently been relabelled as "fake news".
- Most projects deal with what is known as "fact-checking", comparing sources to validate reports, with an assumption that other sources exist.

SUMMA

Reveal

HOAXY beta

Truthy

ClaimBuster

CrossCheck
A collaborative journalism project

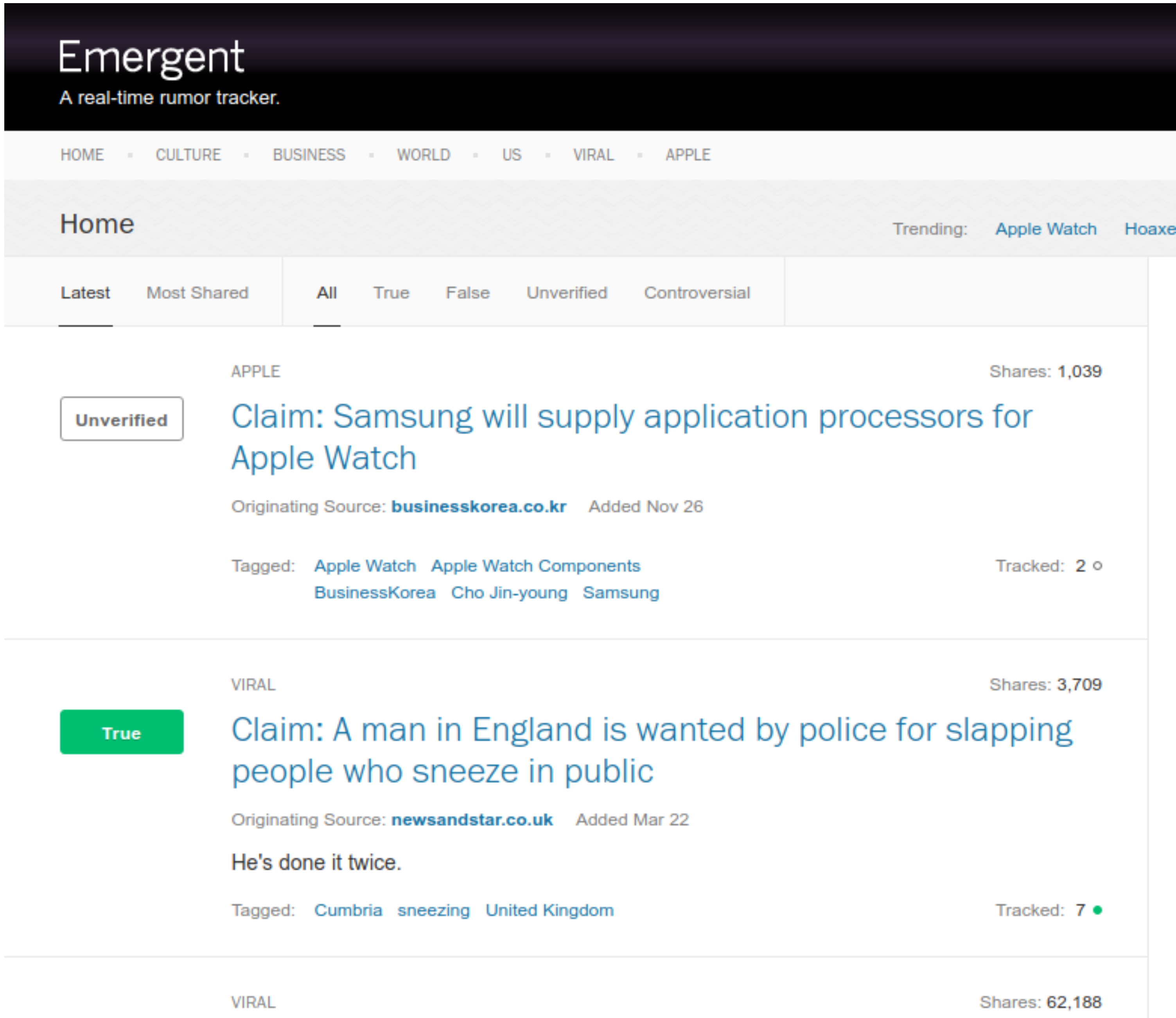
European Commission
Information Society and Media

Working with rumours

- Rumours spawn and circulate as plausible pieces of information that are unverifiable when they emerge.
- One needs to further explore information to find sources that may not have yet been identified and often wait for new evidence to be released.
- A rumour detection system has to:
  - Identify pieces of information that are unverified and require corroboration.
  - Track what different sources are saying about the rumour.
  - The system progressively updates likely veracity of a rumour, ideally identifying supporting evidence.

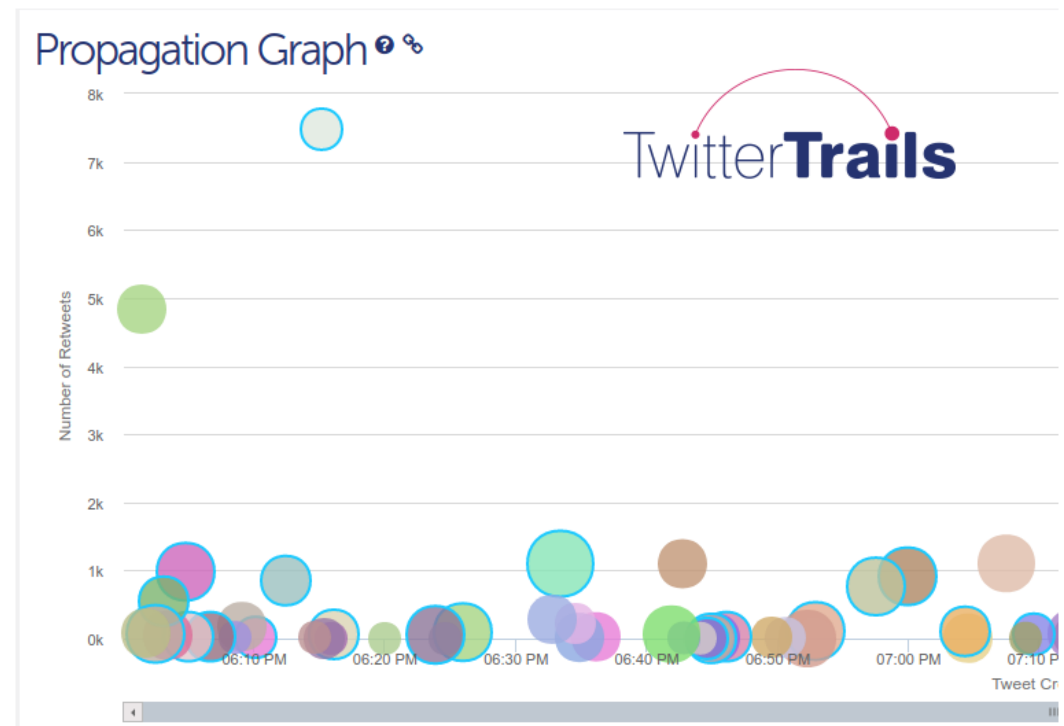
Work by others on rumours

- Emergent.info - the most similar project to PHEME - tracking unverified reports for debunking/confirmation
- The project has however been discontinued
- Owing to the project being totally manual and lacking automation, maintenance was not viable

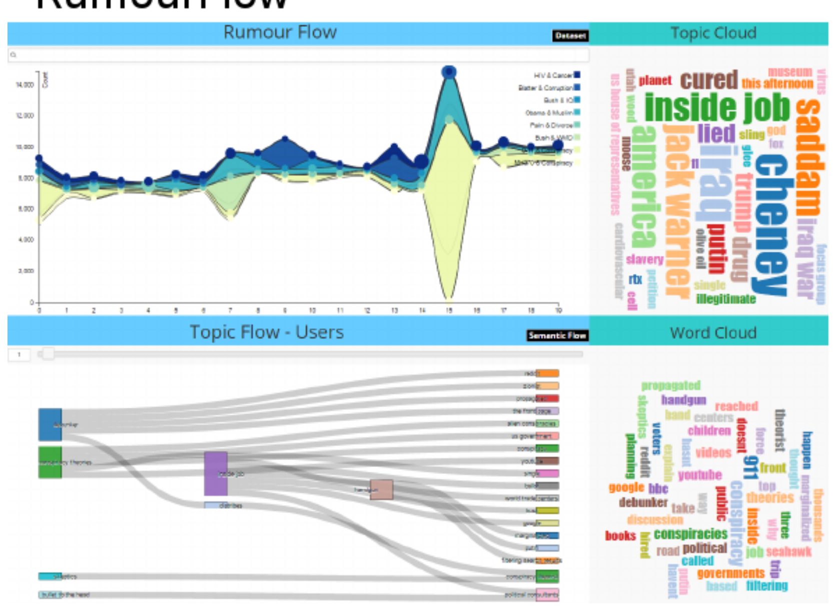


Work by others on rumours

- Other projects on rumours largely focussed on visualisation, tracking rumours input by the user:



RumourFlow



PHEME Rumour Datasets

Events:

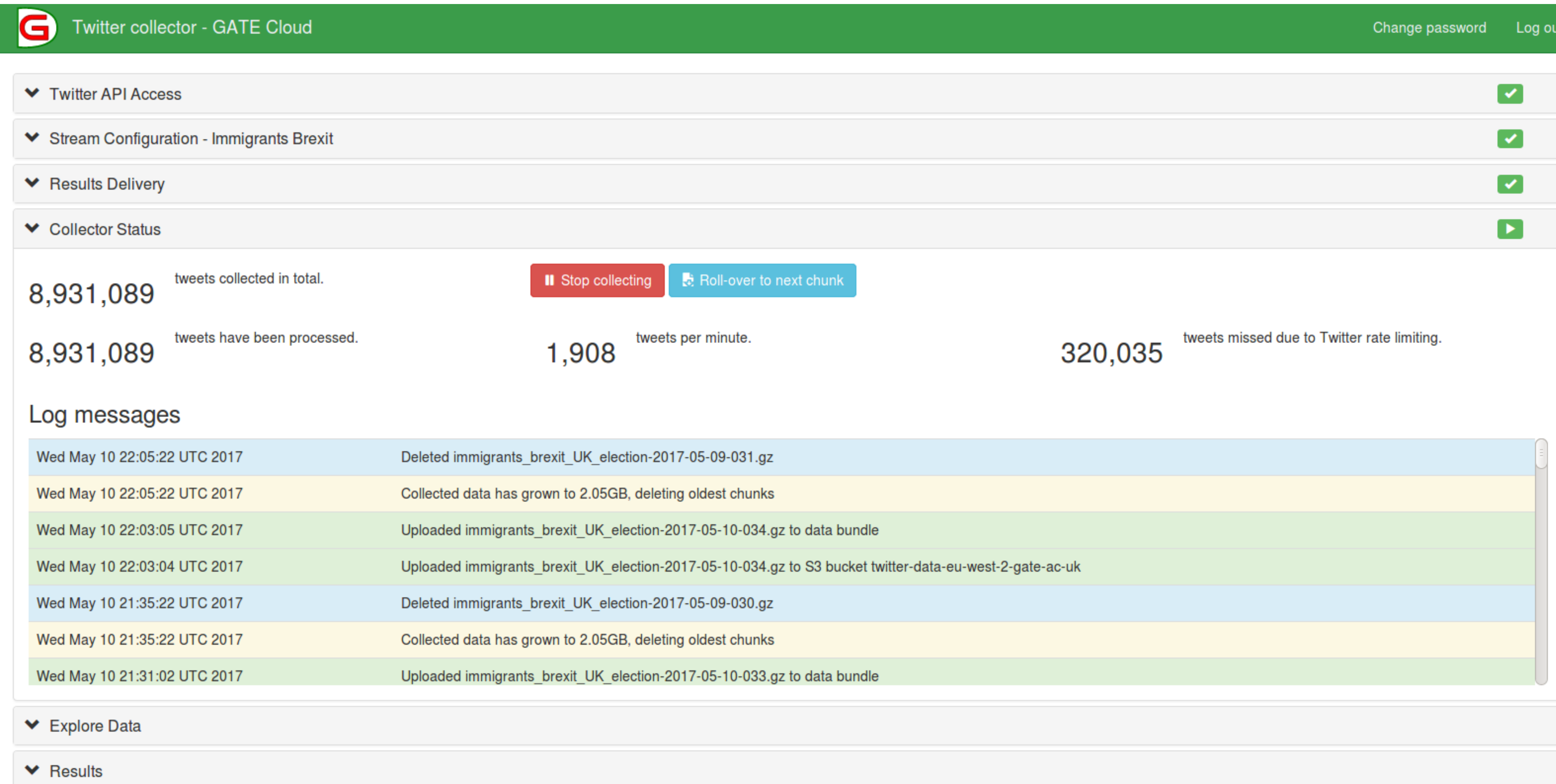
- Ferguson unrest
- Ottawa shooting
- Sydney seige
- Charlie Hebdo shooting
- German Wings crash

Specific rumours:

- Putin missing
- Prince concert
- Michael Essien
- Gurlit collection

Collect Tweet Datasets

Twitter Collector as a Service



Docs & YouTube videos:

https://cloud.gate.ac.uk/info/help/twitter-collector.html

Live Stats and Fine Tuning

^ Explore Data

The following reports are automatically updated in real time based on your data.

Reset Reports

See top hashtags.

See a count of tweet frequency by hour.

See a bar chart of top topics according to our lists of terms.

See most mentioned users.

See frequency for each tracked term.

See top words of each kind (such as noun, adjective and so on, English only).

Show Hashtags

Toggle bar chart/cloud Save as SVG Get JSON

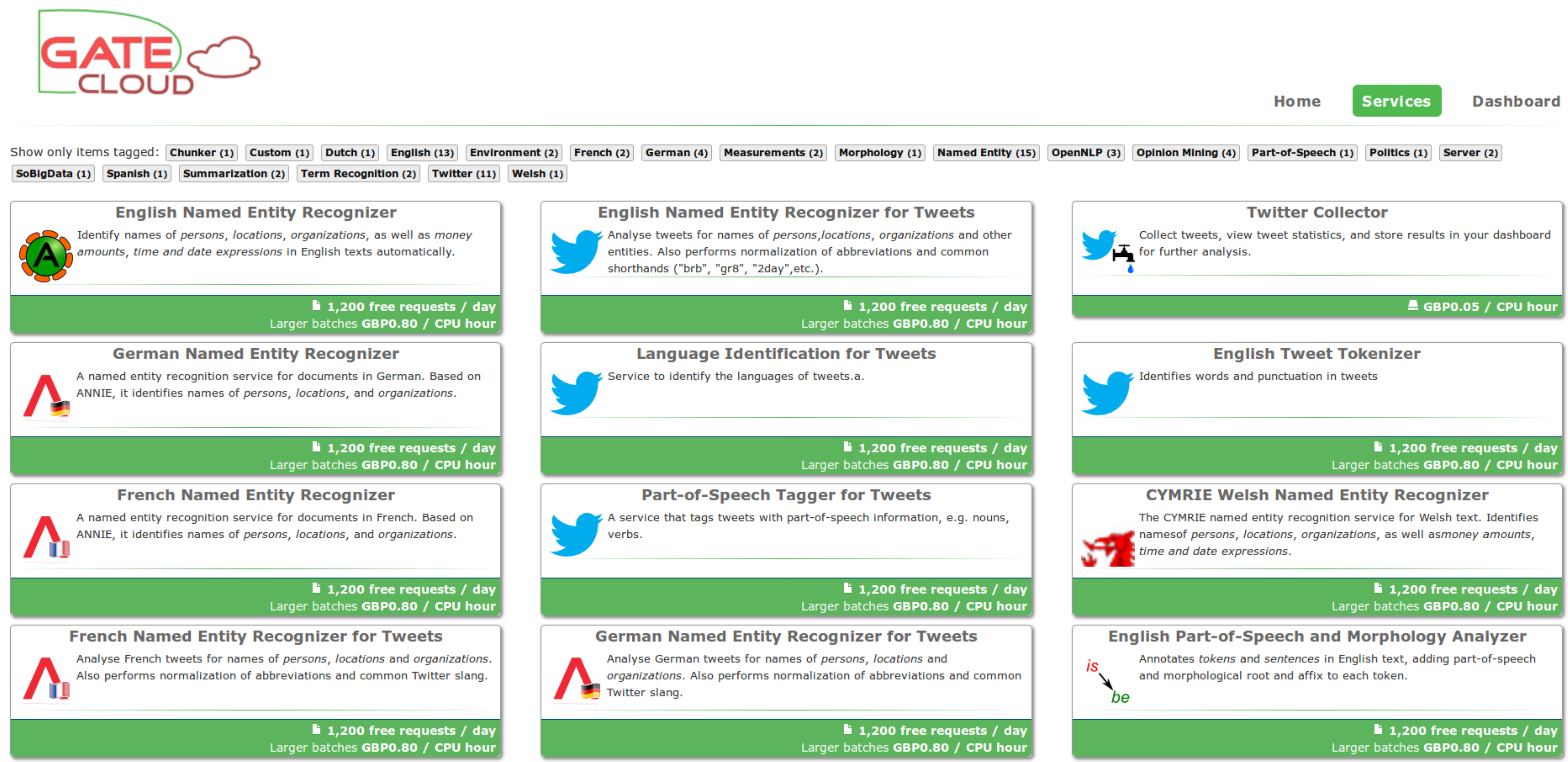
	Count	Topic
#btsbbmas	944032	
#ge2017	754503	
#bbmas	675078	
#news	611818	
#generalelection	365412	
#ge17	270783	
#brexit	252393	
#votelabour	138793	
#theresamay	114951	
#trumprussia	111415	
#toryelectionfraud	96264	
#iran	76887	
#votesnp	69378	
#trump	65950	
#almalki	64802	
#pmqs	57897	
#labour	57366	
#generalelection2017	53085	
#council17	50513	
#election	48432	
#marr	48068	
#france	46240	
#maga	45903	

Tweets collected in Amazon buckets for Download

Processing on GATE Cloud

GATE Cloud services

- 24 services, multiple languages



Rumour Annotation

Dataset Overview

Data collection and sampling

- Twitter public stream
- Retweet threshold to establish salience
- SWI identified rumourous source tweets and grouped them into stories

Event name	Rumour stories	Annotated threads	Rumour threads	Non-rumour threads
Sydney Siege	61	1321	535	786
Ottawa Shooting	51	901	475	426
Charlie Hebdo	61	2169	474	1695
Germanwings	19	1022	332	690
Ferguson	42	1183	291	892
Prince to play in Toronto	6	241	237	4
Gurlitt	3	386	190	196
Putin missing	6	266	143	123
Essien has Ebola	1	18	18	0
TOTAL	250	7507	2695	4812

en-charliehebdo-553558982476828674

[go back to rumours by accuracy] [go back to rumours by acceptability]

Links to Other Media

[news-media] [for] http://jpupdates.com/2015/01/09/breaking-gunman-takes-hostage-kosher-gocery-store-paris/

Social Media Conversation

@AvitalLeibovich [FR: 2, verified, ~~journalist~~]
On Israeli TV, an Israeli woman speaks of her nephew,a mother with 6 month old baby, held hostage in the supermarket in Paris. #JeSuisJuif

Annotation: certainty: certain, support: supporting, evidentiality: witnessed,

@talziv [FR: 1, ~~verified~~, ~~journalist~~]
@AvitalLeibovich link?

Annotation: certainty: uncertain, evidentiality: unverifiable-source-quoted, responsetype-vs-source: appeal-for-more-information,

@AvitalLeibovich [FR: 2, verified, ~~journalist~~]
@talziv @Channel2News

Annotation: certainty: underspecified, responsetype-vs-source: comment, responsetype-vs-previous: agreed, evidentiality: url-given,

@talziv [FR: 1, ~~verified~~, ~~journalist~~]
@AvitalLeibovich @Channel2News actually, mako.co.il/news

Annotation: responsetype-vs-source: comment, responsetype-vs-previous: comment,

@jessicaelgot [FR: 1, verified, ~~journalist~~]
@AvitalLeibovich do you have a link?

Annotation: certainty: uncertain, evidentiality: no-evidence, responsetype-vs-source: appeal-for-more-information,

Rumour Analysis

Analysis Stages

1) **Rumour detection:** identify pieces of information that need to be verified.

2) **Rumour stance classification:** classify public stance towards the rumour.

3) **Veracity classification:** determine if a rumour is true, false, or remains unverified.

European Commission
Information Society and Media

Rumour detection

- Classification problem
  - Tweet level
  - Classes: rumour or non-rumour

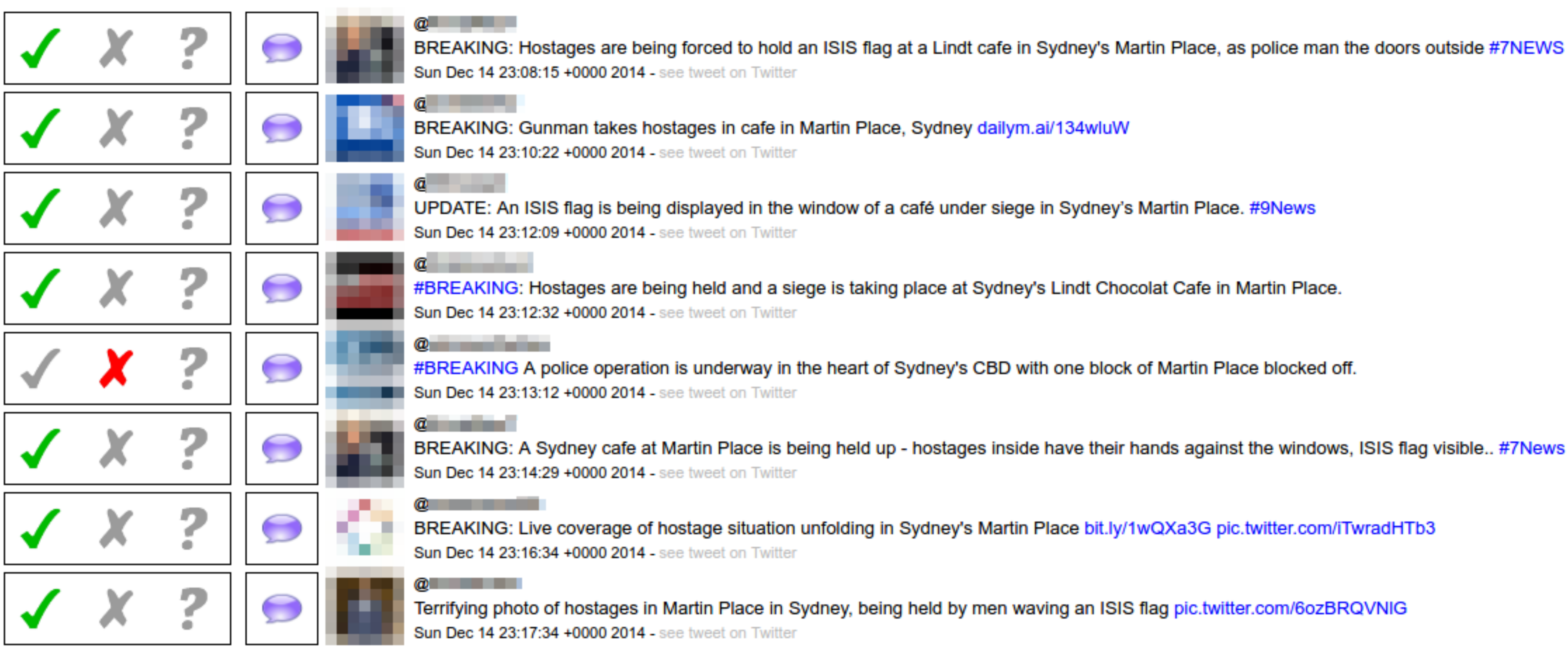
- Why
  - Identify tweets reporting unverified information.
  - Necessary for subsequent stance and veracity classification.

– "Michael Brown was involved in a robbery before being shot by police" -> needs context

European Commission Information Society and Media

Rumour detection (2)

Can we then leverage context accumulated throughout the event?



Rumour detection (3)

- Conditional Random Fields.
- Comparison vs:
  - Baseline by Zhao et al. (2015) at WWW.
  - Maximum Entropy.
  - SVM, Random Forest, NB.
- Features:
  - Content and social

A. Zubiaga, M. Liakata, R. Procter . Exploiting Context for Rumour Detection in Social Media. International Conference on Social Informatics, 109-123. 2017.

European Commission
Information Society and Media

Rumour detection (4)

- Experiment methodolgy:
  - Leave-one-event-out: train on n-1 event, and test on the n-th event. Cross-validation
  - P/R/F1 evaluation.

A. Zubiaga, M. Liakata, R. Procter . Exploiting Context for Rumour Detection in Social Media. International Conference on Social Informatics, 109-123. 2017.

Rumour detection: Results

Classifier	P	R	F1
SVM	0.337	0.483	0.397
Random Forest	0.275	0.099	0.145
Naive Bayes	0.310	**0.723**	0.434
Maximum Entropy	0.338	0.442	0.383
CRF	**0.667**	0.556	**0.607**

State-of-the-art Baseline

Classifier	P	R	F1
(Zhao et al., 2015)	0.410	0.065	0.113

A. Zubiaga, M. Liakata, R. Procter . Exploiting Context for Rumour Detection in Social Media. International Conference on Social Informatics, 109-123. 2017.

European Commission
Information Society and Media

Rumour Stance classification

- Classification Problem
  - Tweet level
  - Classes: support, deny, question, comment
- Why
  - Show journalists the unfolding online debates on a given pheme
  - Aid veracity classification

European Commission
Information Society and Media

Stance classification

[depth=0] **u1:** These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada –PICTURE– **[support]**

[depth=1] **u2:** @u1 Apparently a hoax. Best to take Tweet down. **[deny]**

[depth=1] **u3:** @u1 This photo was taken this morning, before the shooting. **[deny]**

[depth=1] **u4:** @u1 I don't believe there are soldiers guarding this area right now. **[deny]**

[depth=2] **u5:** @u4 wondered as well. I've reached out to someone who would know just to confirm that. Hopefully get response soon. **[comment]**

[depth=3] **u4:** @u5 ok, thanks. **[comment]**

Rumour Stance & Veracity

- Manual Analysis [Mendoza et al. 2010]
- True rumours: 95% "support", 4% "questioning" and only 0.4% were "denies"
  → True rumours are largely supported by other Twitter users
- False rumours: 38% "denies" and 17% "questioning"
  → Stance is an important factor for verification step

Rumour Stance Data

PHEME

Dataset	Rumours	Supporting	Denying	Questioning	Commenting	Large
Ottawa shooting	58	161	76	64	481	✓
Ferguson riots	46	192	83	94	685	✓
Prince in Toronto	12	19	7	11	59	×
Charlie Hebdo	74	236	56	51	710	✓
Ebola Essien	2	6	6	1	21	×
Germanwings crash	68	177	12	28	169	✓
Putin missing	9	17	7	5	33	×
Sydney siege	71	89	4	99	713	✓

European Commission Information Society

RumourEval

PHEME

SemEval-2017 Task 8

RumourEval: Determining rumour veracity and support for rumours

RumourEval: Determining rumour veracity and support for rumours

European Commission
Information Society and

LSTM won RumourEval

Team	Score
DFKI DKT	0.635
ECNU	0.778
IITP	0.641
IKM	0.701
Mama Edha	0.749
NileTMRG	0.709
Turing	0.784
Uwaterloo	0.780

European Commission
Information Society and Media

Features: Stance Classification

- Syntactic: #negations, POS
- Lexical: e.g. bag of words, key words
- Lexicon: word embeddings, brown clusters
- Indicator: e.g. "this is nonsense", "is a hoax"
- User metadata: e.g. verified, #followers/followees
- Semantic: e.g. emoticons, sentiment, NEs
- Message metadata: e.g. length of the post, URL
- Content formatting: e.g. capital words ratio
- Punctuation: e.g. ?, …, !

A. Aker, L. Derczynski, K. Bontcheva. Simple Open Stance Classification for Rumour Analysis. Recent Advances in Natural Language Processing, 2017.

European Commission
Information Society and Media

USFD on RumourEval Data

- Applying traditional ML
- Focus on "problem specific" features
  - Capturing *Surprise*, *Doubt*, *Certainty* and *Support* towards rumourous tweets
  - Similarity to initial tweet
- Random Forest (best performing traditional ML) with features reported by related work yields to 76.54 accuracy
- Adding "problem specific" features yields 79.02 accuracy (LSTM 78.4).

A. Aker, L. Derczynski, K. Bontcheva. Simple Open Stance Classification for Rumour Analysis. Recent Advances in Natural Language Processing, 2017.

European Commission
Information Society and Media

Stance Classification: Findings

- Supporting tweets are more likely to include links.
  - i.e. provide evidence
- Looking at the temporal dimension, S/D/Q tend to occur in early stages of a rumour, and then mostly comments later
- We've looked at persistence, finding that supporting users tend to post more tweets, i.e., further insisting

https://gate.ac.uk/wiki/pheme-stance.html

Veracity classification

- Classification Problem
  - Rumour level
  - Classes: true, false or unverified
- Why
  - Help journalists and the public determine if report is accurate.
  - Flagging false rumours to warn users.

Features

- Features similar to stance classification
- NB: those features need to be computed for the entire rumour instead of individual posts
  - E.g. adding feature values over all posts within the rumour and performing normalisation.
- Stance of individual posts is an important feature

A. Aker, L. Derczynski, K. Bontcheva. Simple Open Stance Classification for Rumour Analysis. Recent Advances in Natural Language Processing, 2017.

European Commission
Information Society and Media

Veracity classification

- Evaluation: determine, after each tweet, ability to classify veracity of rumour.

Classifier	Without Stance	With stance over time
J48	64.61	69.84
Random Forest	68.20	70.76
K-NN	63.69	69.84

PHEME's impact

- Fake News Challenge adopted PHEME's methodology, i.e., a 4-way typology akin to {support, deny, query, comment}

FAKE NEWS CHALLENGE STAGE 1 (FNC-I): STANCE DETECTION

Input

A headline and a body text - either from the same news article or from two different articles.

Output

Classify the stance of the body text relative to the claim made in the headline into one of four categories:

1. **Agrees**: The body text agrees with the headline.
2. **Disagrees**: The body text disagrees with the headline.
3. **Discusses**: The body text discuss the same topic as the headline, but does not take a position
4. **Unrelated**: The body text discusses a different topic than the headline

Journalism dashboard

? Brexit vote may have been hacked. If Russia hacked Brexit, safe to assume they hacked vote totals in other elections.?

OVERVIEW

Topic: UK triggers Article 50

Number of images: 0

Number of URLs: 4

Verified authors: No

Controversiality score: Low

Veracity: False (100% confidence)

ACTIVITY

Size: 132 tweets

Average activity: 0.16 tweets/second

Oldest tweet: *5:35 AM - 12 Apr 2017*

Last activity: *5:48 AM - 12 Apr 2017*

EXAMPLE TWEET

@DustinGiebel

"Brexit vote site may have been hacked, MPs say in report" I know some people they should look at.? https://t.co/EWp4HeEDyg

3:22 AM - 12 Apr 2017 open in twitter

#WillLondonFall FALL?!?? I AIN'T A SEASON - LONDON. #brexit #NYTimes #UKParliament #sixwordstory

OVERVIEW

Topic: UK triggers Article 50

Number of images: 0

Number of URLs: 1

Verified authors: No

Controversiality score: Low

Veracity: False (100% confidence)

ACTIVITY

Size: 122 tweets

Average activity: 2.30 tweets/second

Oldest tweet: *9:19 AM - 12 Apr 2017*

Last activity: *9:20 AM - 12 Apr 2017*

EXAMPLE TWEET

@michelepaschal1

RT @SixWrd: #WillLondonFall FALL?!?? I AIN'T A SEASON - LONDON. #brexit #NYTimes #UKParliament #sixwordstory https://t.co/R7bOGPBxr1 https?

7:19 AM - 12 Apr 2017 open in twitter

NATO to reschedule meeting to accommodate Tillerson

OVERVIEW

Topic: Nato

Number of images: 1

Number of URLs: 4

Verified authors: Yes

Controversiality score: Low

Veracity: True (100% confidence)

ACTIVITY

Size: 71 tweets

Average activity: 0.04 tweets/second

Oldest tweet: *4:44 AM - 23 Mar 2017*

Last activity: *5:11 AM - 23 Mar 2017*

MOST SHARED IMAGE

EXAMPLE TWEET

@thehill

NATO to reschedule meeting to accommodate Tillerson https://t.co/DGrYB6t5nz https://t.co/xuqa5LUIr7

2:49 AM - 23 Mar 2017 open in twitter

By the time Germany was admitted into NATO the world had literally changed. Germany was no longer a military threat to France, UK, or the US because a

OVERVIEW

Topic: Nato

Number of images: 0

Number of URLs: 0

Verified authors: Yes

Controversiality score: High

Veracity: uncertain

ACTIVITY

Size: 70 tweets

Average activity: 0.00 tweets/second

Oldest tweet: *6:24 PM - 19 Mar 2017*

Last activity: *6:45 PM - 20 Mar 2017*

EXAMPLE TWEET

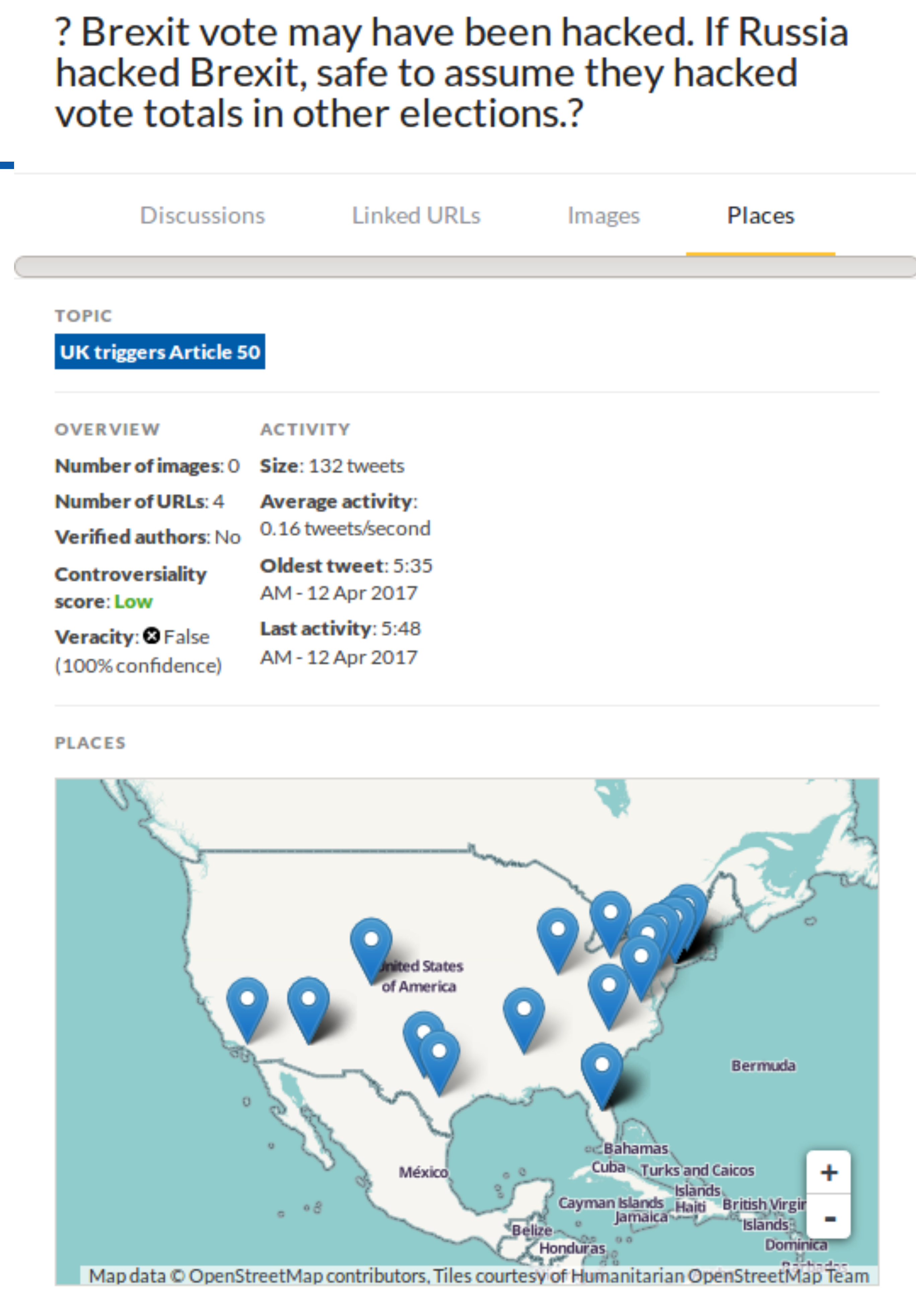
@Bazuka125

There's a difference between "they should pay fair share to NATO" and "They should pay us"

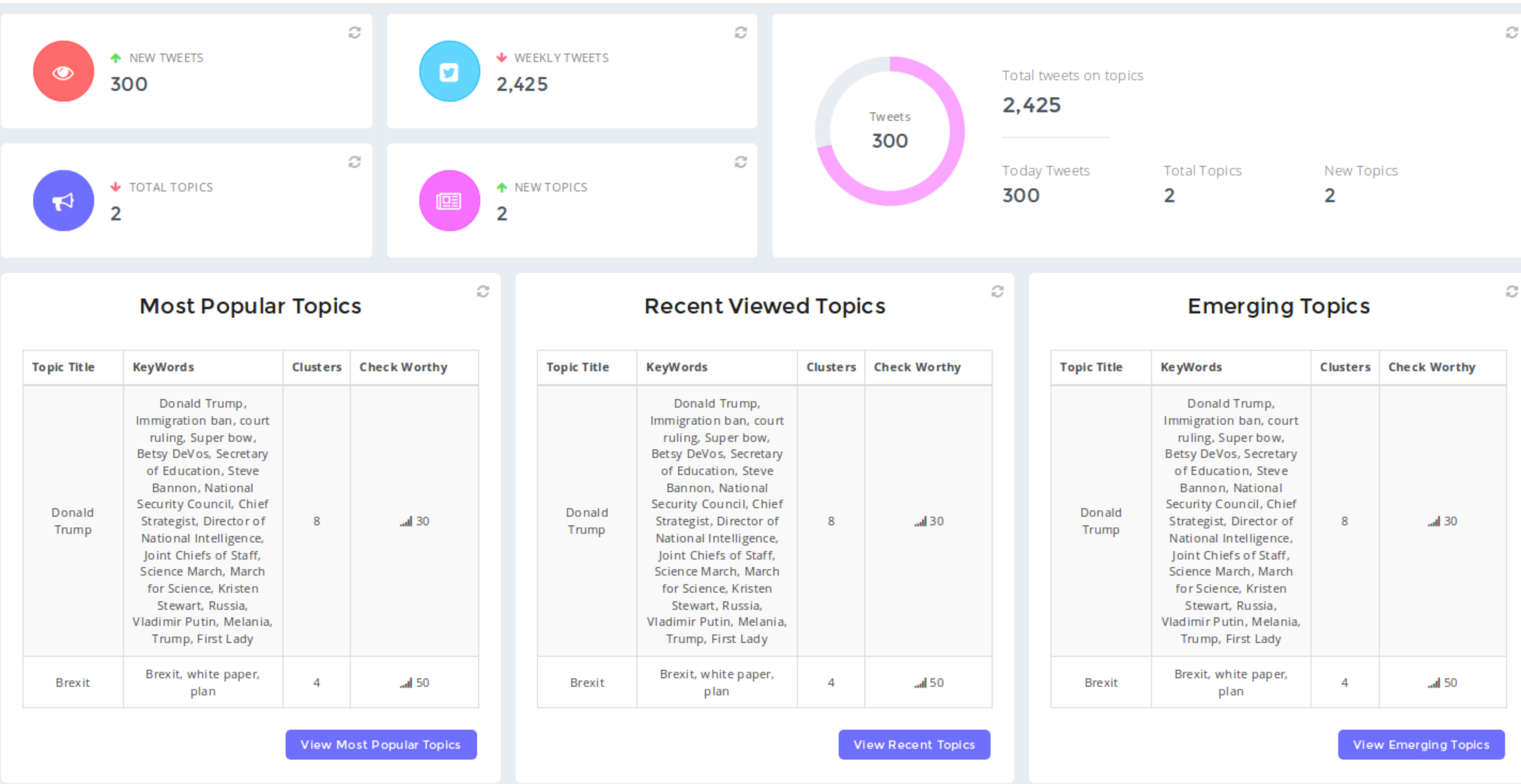
4:24 PM - 19 Mar 2017

Tweet Mapping

- For geo-tagged tweets in cluster, map their origin



Hercule: Fact Checking Dashboard



Most Popular Topics

Topic Title	KeyWords	Clusters	Check Worthy
Donald Trump	Donald Trump, Immigration ban, court ruling, Super bow, Betsy DeVos, Secretary of Education, Steve Bannon, National Security Council, Chief Strategist, Director of National Intelligence, Joint Chiefs of Staff, Science March, March for Science, Kristen Stewart, Russia, Vladimir Putin, Melania, Trump, First Lady	8	30
Brexit	Brexit, white paper, plan	4	50

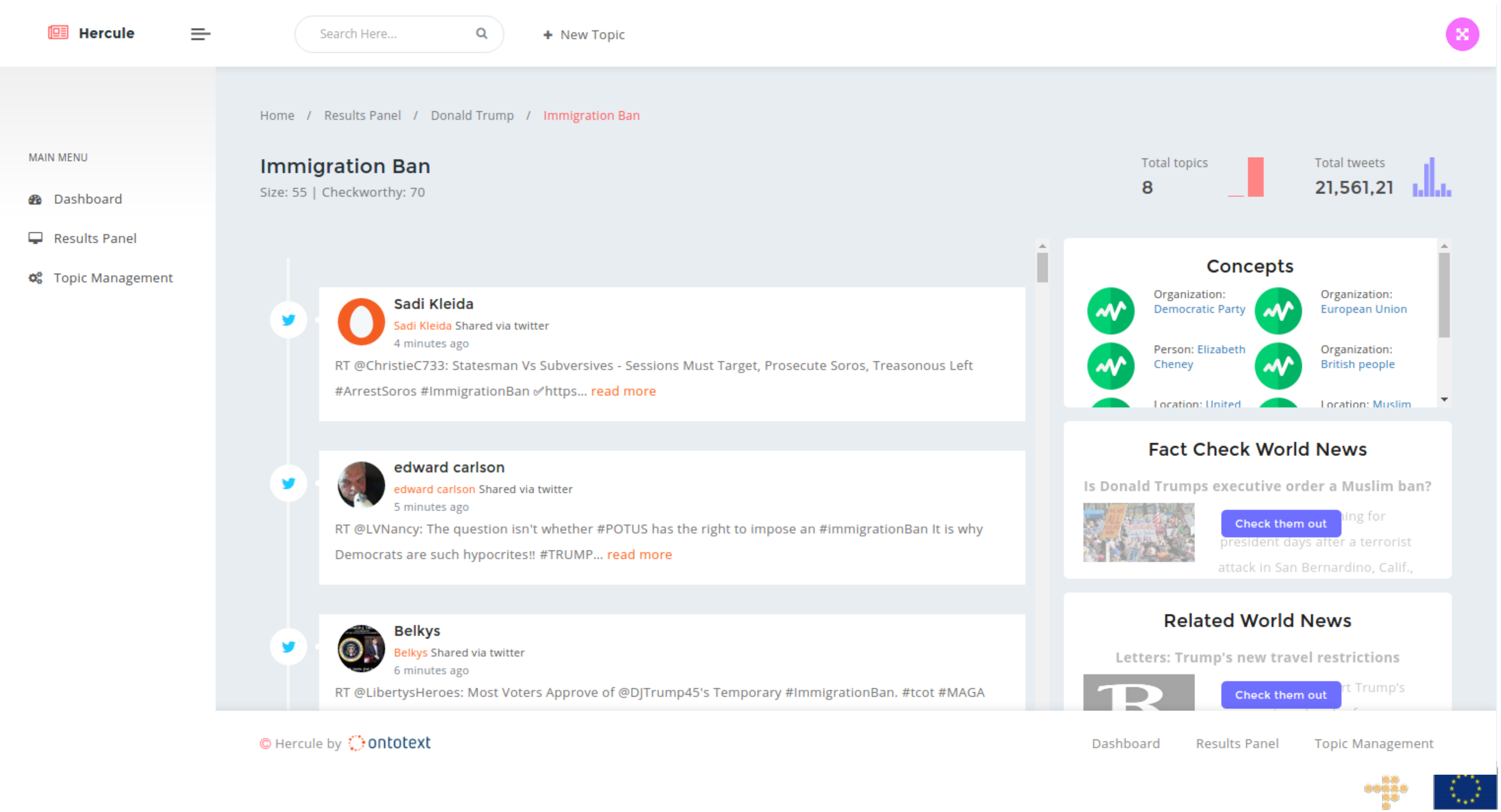
Recent Viewed Topics

Topic Title	KeyWords	Clusters	Check Worthy
Donald Trump	Donald Trump, Immigration ban, court ruling, Super bow, Betsy DeVos, Secretary of Education, Steve Bannon, National Security Council, Chief Strategist, Director of National Intelligence, Joint Chiefs of Staff, Science March, March for Science, Kristen Stewart, Russia, Vladimir Putin, Melania, Trump, First Lady	8	30
Brexit	Brexit, white paper, plan	4	50

Emerging Topics

Topic Title	KeyWords	Clusters	Check Worthy
Donald Trump	Donald Trump, Immigration ban, court ruling, Super bow, Betsy DeVos, Secretary of Education, Steve Bannon, National Security Council, Chief Strategist, Director of National Intelligence, Joint Chiefs of Staff, Science March, March for Science, Kristen Stewart, Russia, Vladimir Putin, Melania, Trump, First Lady	8	30
Brexit	Brexit, white paper, plan	4	50

Hercule: Fact Checking UI



Challenges

- Capturing temporal, streaming dynamics
- Improving further:
  - Rumour detection – F1 of 60.7% still too low to be useful in practice
  - Rumour stance – overall accuracy, but especially on the deny class, needs further improvement
  - Veracity – needs significant improvement, for journalists to find it reliable in practical settings
- Provide journalists with the reasons behind a given classification decision

European Commission
Information Society and Media

Acknowledgements

Project partially supported by the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D (FP7), grant PHEME (611233)"