Modelling news headline appeal in social media

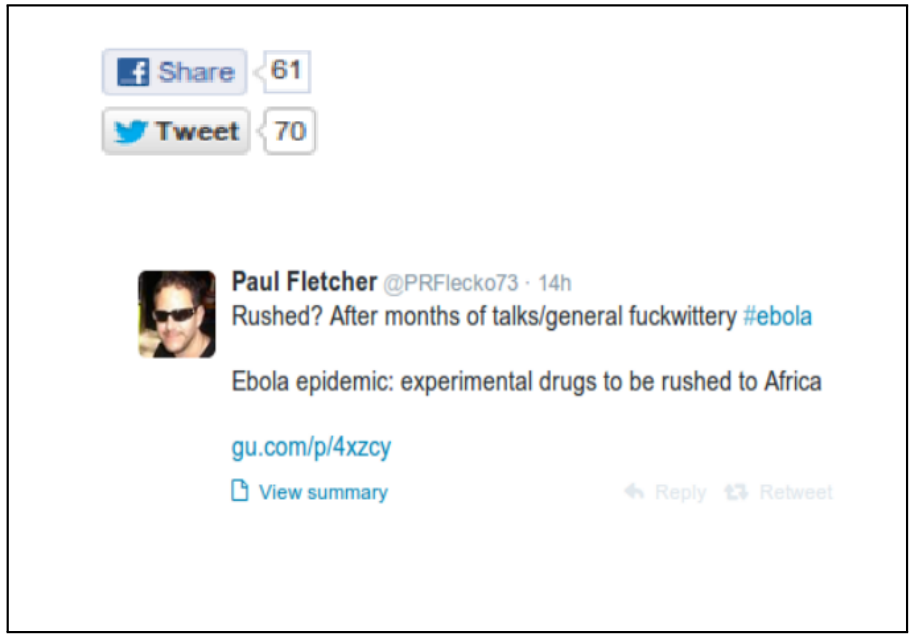
Katja Markert

L3S Research Center
University of Hannover

Hannover 2015

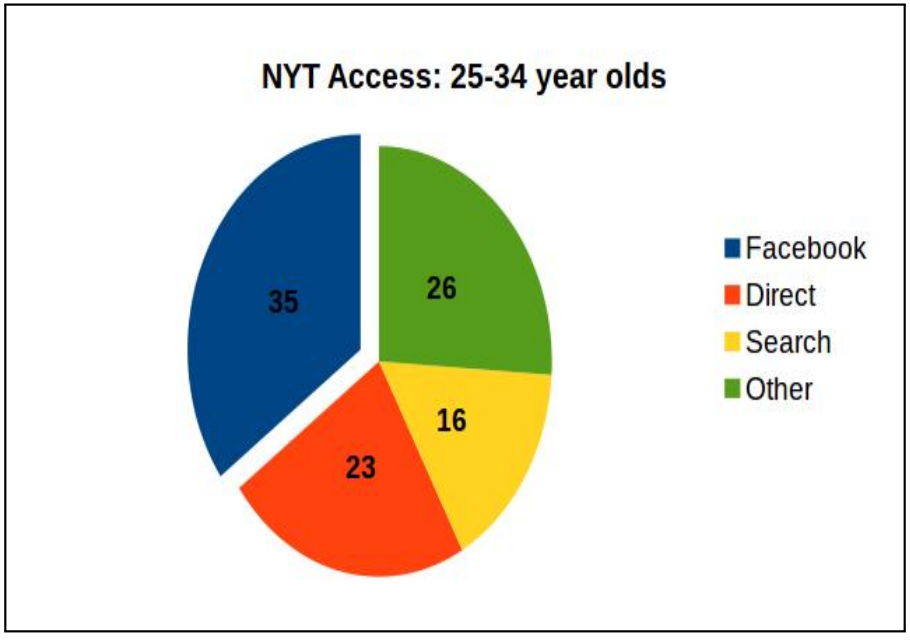
Social media reception of online news

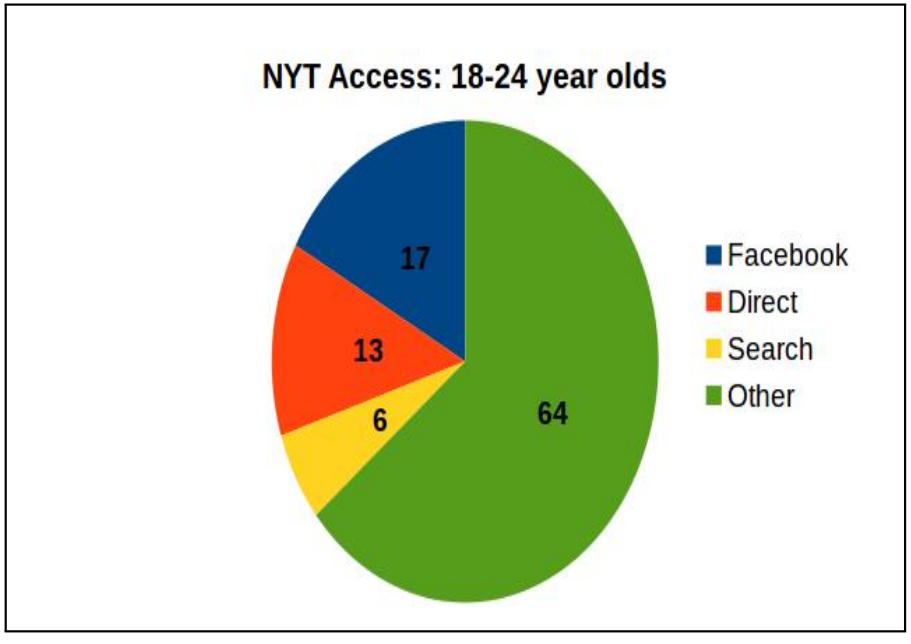
Interaction of two important, inherently temporal web genres



Importance of social media take-up

Example: 7% of NYT online access via Facebook





Source: Mitchell et al. 2014, Pew Research Center.

▸ Guardian access data

Our task

Model which news articles are appealing on social media

Guardian headlines from spring 2014:

- *Merkel urged to press Obama on NSA scandal ahead of Washington talks*
- *O'Farrell's downfall will put lobbyists under microscope, campaigners say*

Our task

Model which news articles are appealing on social media

Guardian headlines from spring 2014:

- *More than 100 hate crime murders linked to single website*
- *Vodafone lifted by positive JP Morgan comments*

Overview

Related work: Computational reception prediction

- Network methods after publication and early votes (Lerman und Hogg, 201; Hsieh et al, 2013; Castillo et al. 2014; Tatar et al. 2014)
- Very few papers on pre-publication prediction (Berger und Milkman 2012; Bandari et al. 2012, Arapakis et al. 2014):
  1. Hard to improve over just using article source (Bandari et al. 2012, Arapakis et al. 2014)
  2. Use of whole article

Little relationship to insights in media studies and journalism community

Related work: Journalism and media studies

News values

Factors which influence which news should be reported and how much room it should be given. Inherent event factors (such as event size) as well as style.

- Walter Lippman (1922), **Galtung and Ruge (1965)**, Gans (1979), Harcup and O'Neill (2001), Bednarek and Caple (2012)
- Mostly **manual** validation: Operationalisation?
- Explanation of **editorial selection**
- **Not** tested for prediction of **social media reception**.

Our goals

- Analysis **prior to** publication
  1. Is the traditional **theory of news values** valid for text reception in social media?
  2. Does **writing style** influence article popularity?
- **Source control**
- **Headlines** only input
  1. Headlines "main entry point" for online News (Leckner 2012)
  2. Often only visible signal in access via search, RSS etc.

- Collaboration with Vania Dimitrova and Alicja Piotrokowizcz
- Funding: 365media.com, EPSRC
- Papers: submitted to *Transaction of the ACL*

Sources

- One of most popular online UK newspapers (www.pressgazette.co.uk))
- Train: 11,980 HL, Apr. 14
- Test: 13,806 HL, July 14

- Most read digital in US (auditedmedia.com))
- Train: 2,301 HL, Oct. 14
- Test: 2,540 HL, Nov. 14

Social Media: Tweets and Likes/Shares after 1 and 3 days (T1, T3, F1, F3)

News Values: Prominence

Elite

- Big names better news than nobodies (Golding and Elliott 79)
- Long term and short term prominence curves in Wikipedia and news

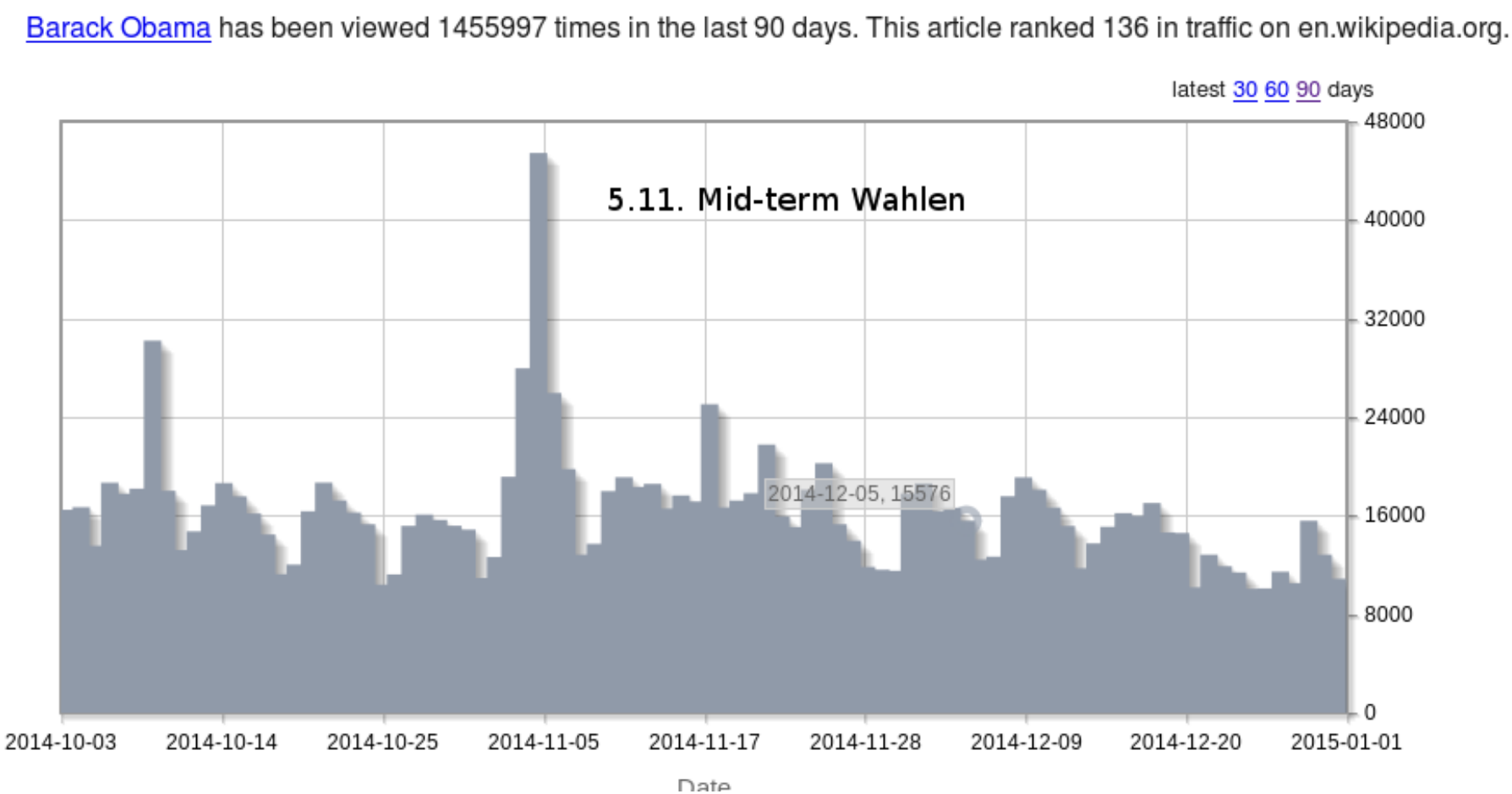
$$\sum_{e \in S} med_{365\ days}(Wikipediaviews(e)) \quad (1)$$

Merkel urged to press Obama on NSA scandal ahead of Washington talks (with T1= 152)

Name	Entity	Median Wiki Longterm
Merkel	Angela Merkel	2949
Obama	Barack Obama	18928
NSA	National Security Agency	3143
Washington	Washington D.C.	187
All	All	25107

Long term vs. short term prominence

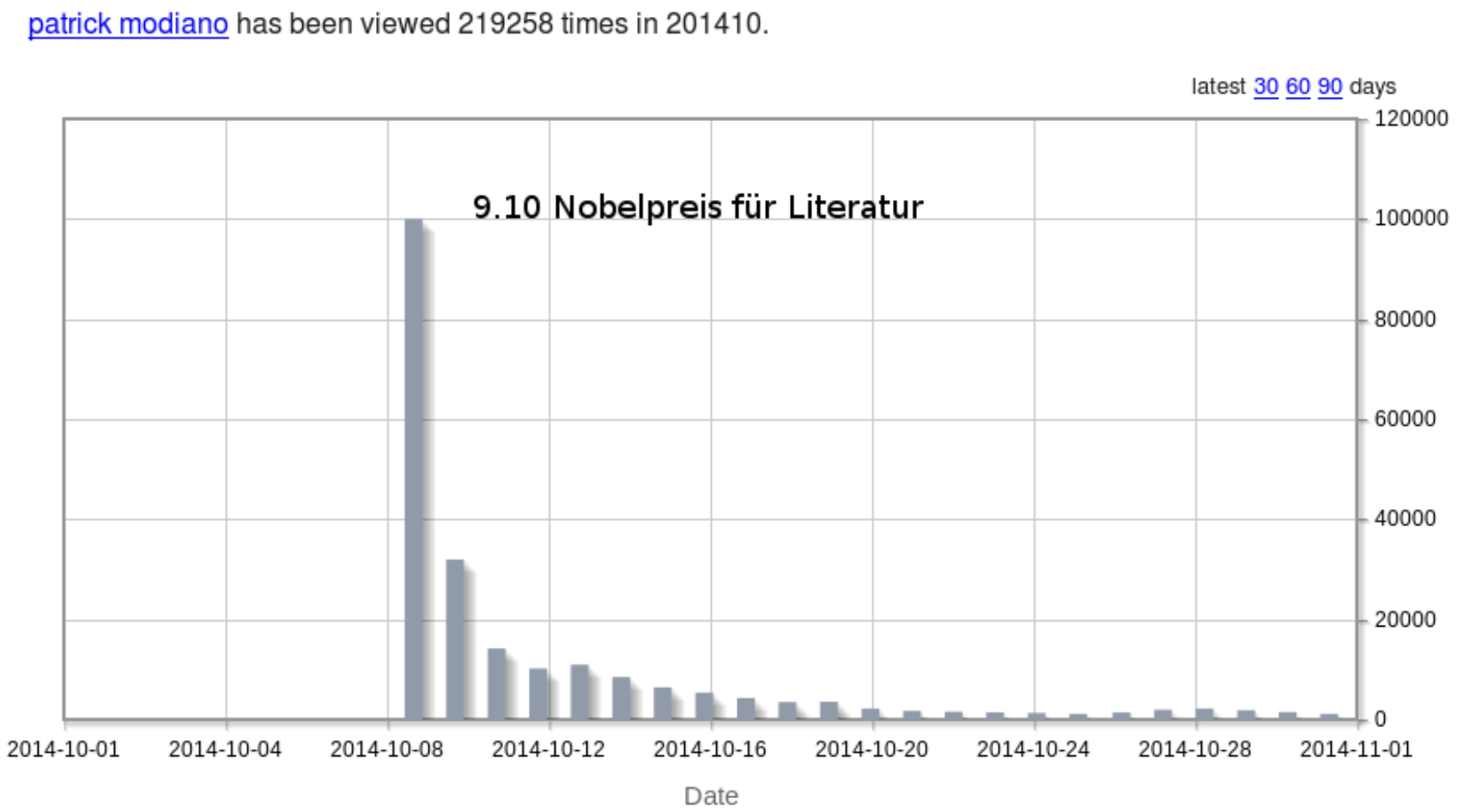
Long term prominence for entities that are always elite



Via stats.grok.se

Long term vs. short term prominence

Short term prominence for trends and unique events



Via stats.grok.se

burstiness

News Values: Sentiment

Sentiment

"If it bleeds, it leads" (Johnson-Cartee 2005)

100 hate crime murders linked to single website: T1 =274

Vodafone lifted by positive JP Morgan comments: T1 = 7

Several features using positivity and negativity scores in large sentiment dictionaries

News values: Size

Size

"stories that are perceived as sufficiently significant either in the numbers of people involved or in potential impact" (Harcup and O'Neill, 2001)

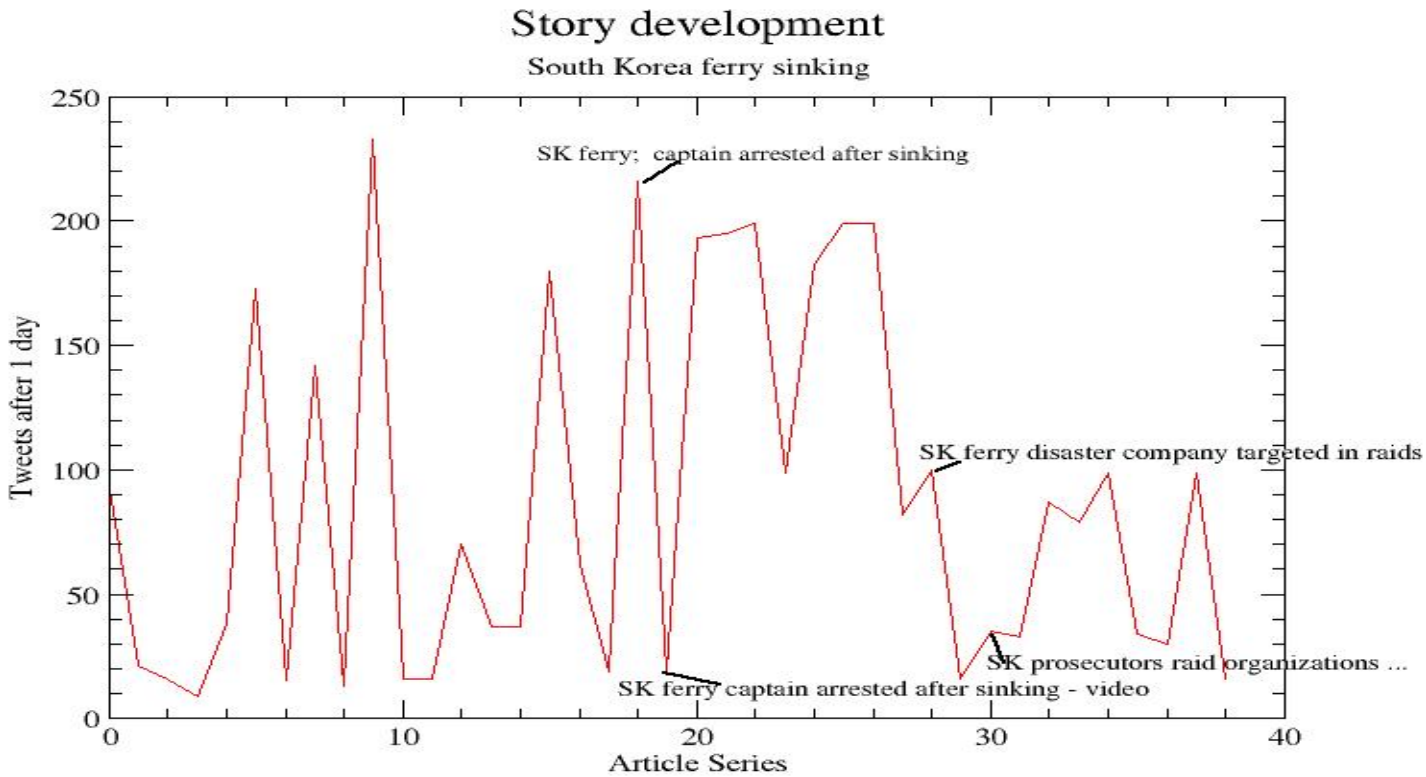
- Modelling event size needs a lot of world knowledge
- Current approximation: **Verbal emphasis** of event size
- Comparatives/superlatives, 248 intensifiers, 39 downtoners

Scientists name world's 100 most unique and endangered birds (T1 = 313)

News Values: Timeliness and Recency

Timeliness

Yesterday's news quickly seizes to be news at all (Montgomery, 07)



Measure headline similarity to previous headlines

Topics

International Press Telecommunications Council (IPTC) topics

Guardian frequency	T1	F1
Culture 3071	Environment 66	Social Issues 116
Sport 2206	Education 66	Environment 87.5
Politics 1463	S&T 65	Education 72
Economy 1373	Social Issues 62	Politics 69
S&T 704	Politics 54	Religion 54.5
Social Issues 489	Health 43	Health 46
Environment 442	Religion 42	S&T 38
Religion 440	Weather 42	Crime 37.5
Crime 312	Crime 40.5	Weather 37
Health 307	Sport 39	Culture 33
Education 263	Economy 37	Economy 24
Weather 189	Culture 23	Sport 16

Difference editorial selection and social media reception (see also Toledo-Bastos 2014)

Style

Some examples (overall 18 features):

Group	**Merkmal**	**Implementierung**
Length	Length in words	# Words
Simplicity	Entropy	n-gram-model
Simplicity	"Headlinese"	consecutive nouns
Simplicity	Use verbs	$\frac{\#\ Verbs}{\#Words}$
Ambiguity	word senses	Median # senses

- Longer headlines more popular $\longrightarrow$ Change from printed to online headlines?
- **Having at least one verb** makes the HL **more popular**

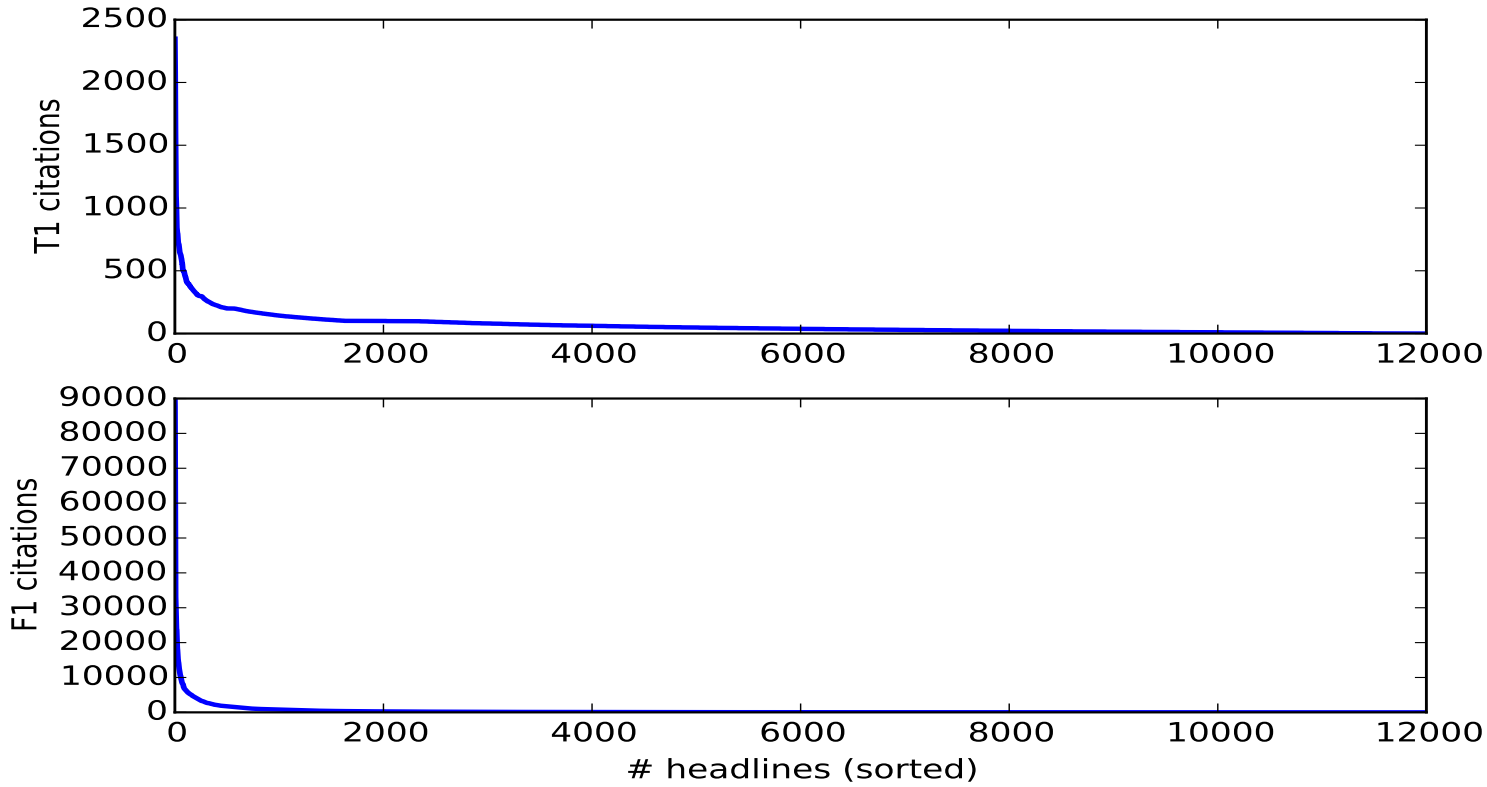
Genres

- *Top 10 drinkers in art* T1=153
- *I'm being distracted by the noisy eating of a new colleague* T1=13

Genre	Frequency	Median T1
News	6092	51
Editorials	1420	61
Reviews	976	14
Interviews	251	34
Letters	187	14
Obitaries	96	12.5
Advice	82	2
Recipes	64	24.5
Announcements	31	0
Lists	23	63
All		38

Zipfian citation distributions

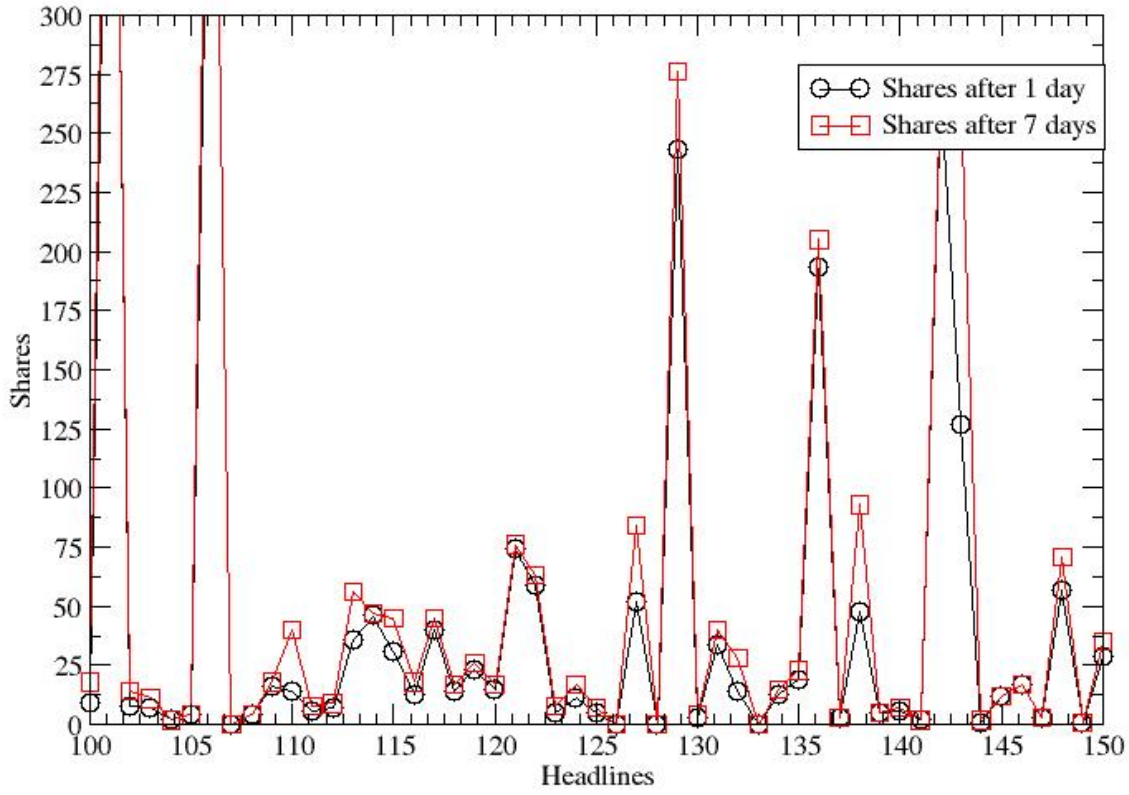
Example Guardian:



Low T1	0-15
Middle T1	16-38
High T1	39-80
Very high T1	81-2349

Citations saturation after 1 day

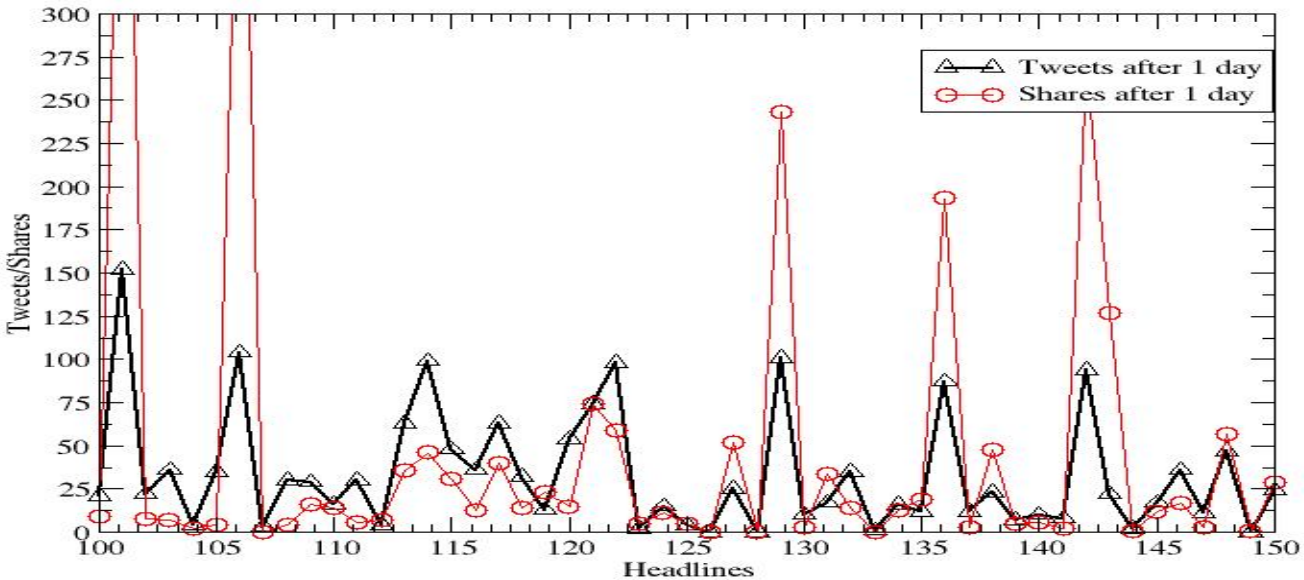
see also Arapakis et al., 2014



Only report T1/F1 results here.

Twitter vs. Facebook

- High correlation between Twitter and Facebook.
- Twitter flatter distribution with higher median (38 vs. 20).



Preliminary Prediction Results

Support vector regression. Kendalls Tau as evaluation.

Guardian: 11,980 HL tr; 13, 806 HLs test		
Model	Tau T1	Tau F1
Unigrams	0.29	0.24
Prom. + Topic	0.25	0.18
Model	**0.36**	**0.30**
NYT: 2,301 HLs tr.; 2,540 HLs test		
Model	Tau T1	Tau F1
Unigrams	0.20	0.21
Prom. + Topic	0.21	0.22
Model	**0.34**	**0.38**

- **Significantly better** than Baseline on two sources
- Best Features Content: Topic, Prominence, Negativity, Timeliness
- Best Features Style: Genre, Length, Verb Phrases

Summary

- First model of news appeal **using headlines only**
- **Transferral of news values theory** from editorial selection to social media reception
- **Style and genre** play a significant role
- Much to do....

Future work

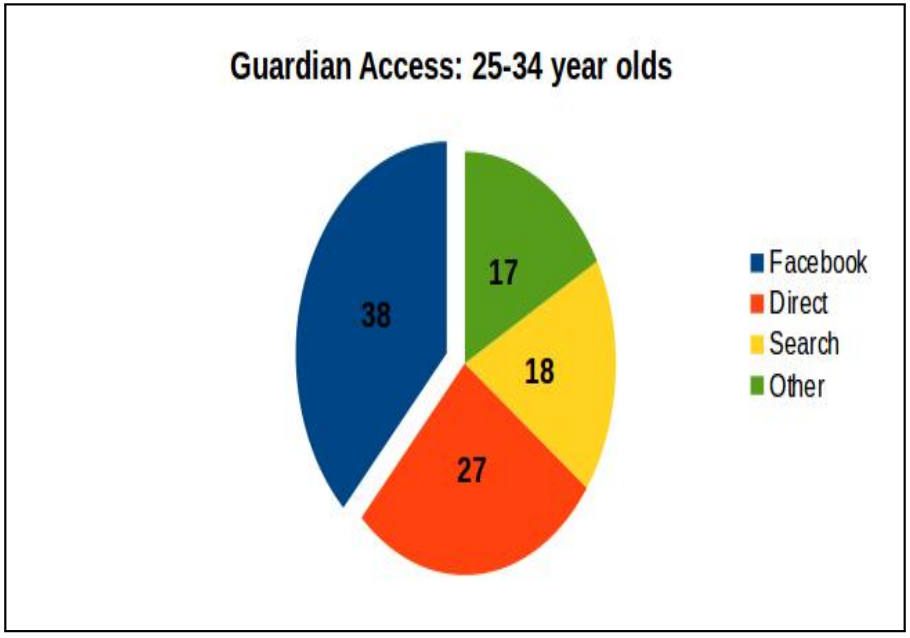
- Better operationalisation of news values (size, timeliness)
- Further news values operationalisation (anomaly, surprise)
- Combination with network methods and user modelling
- Link to past collections and archiving
  - Joint archiving of pages and reception pages
  - Influence of reception on selection for archiving
  - Reception analysis over time

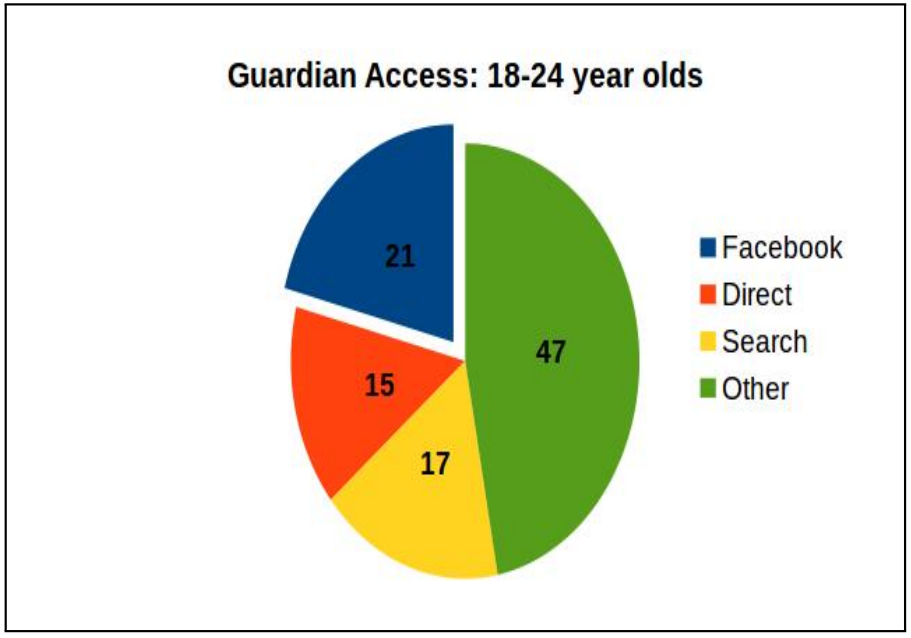
Thank you!

Questions?

Importance of social media take-up: The Guardian

Back



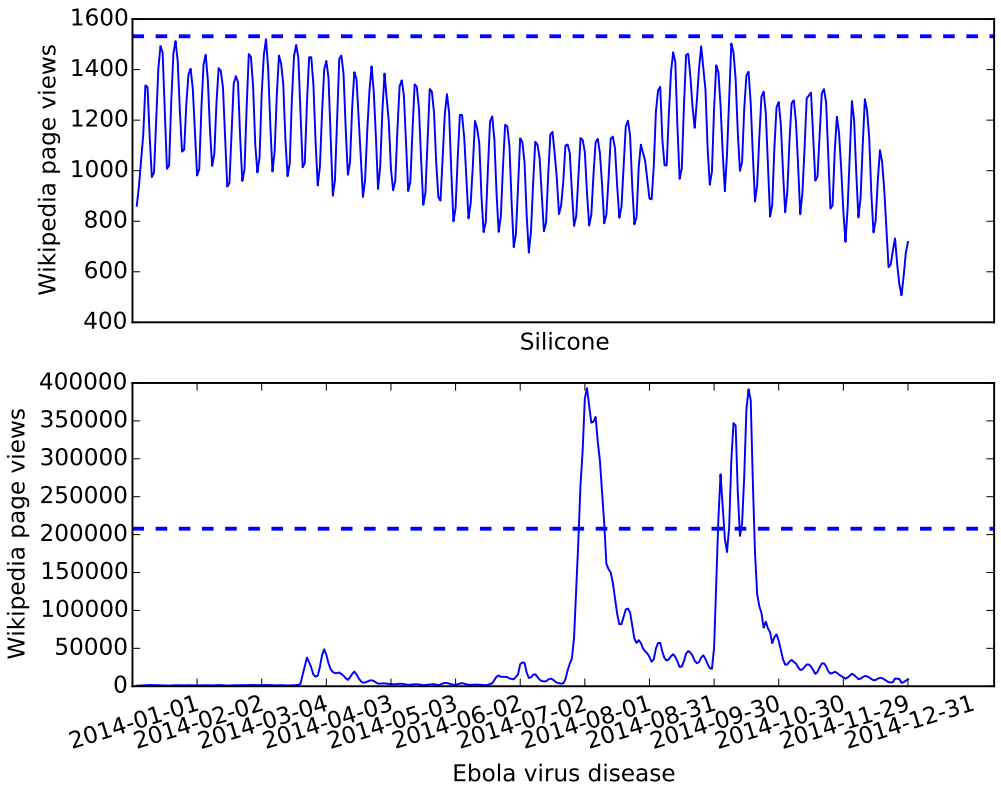


Source: Mitchell et al. 2014, Pew Research Centre.

Back

Burstiness

Measures which entities undergo many prominence variations



- Measure days in burst over last year
- Entities which are burstier increase popularity