

TOWARDS AUTOMATED FACT-CHECKING FOR DETECTING AND VERIFYING CLAIMS

ROMCIR 2021, 1st April 2021

Arkaitz Zubiaga

WHO AM I

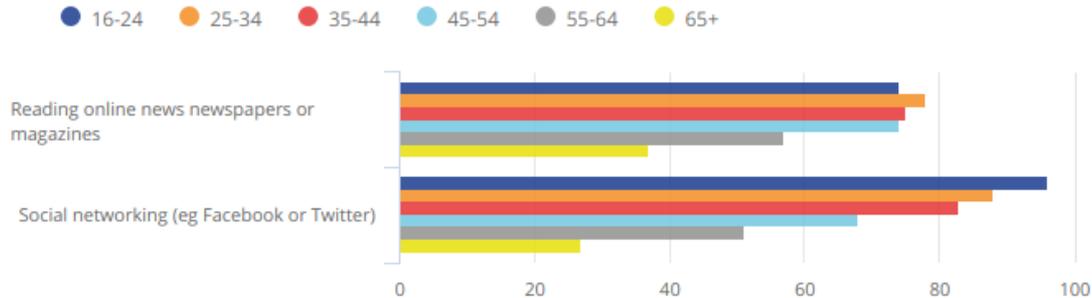
- Lecturer, Queen Mary University of London.
- Worked on misinformation research since 2012.
- Currently focusing on a number of related areas:
 - Hate speech detection.
 - Automated fact-checking.
 - Stance detection.



SOCIAL MEDIA & ONLINE NEWS READERSHIP

- UK (2017) Facebook or Twitter used by ([ONS](#)):
 - 66% of population.
 - 96% of 16-24 year olds.

Figure 4: Internet activities by age group, 2017, Great Britain



WE CAN GET REAL-TIME, EXCLUSIVE UPDATES

Imane @lmrh1 Follow

OIII THE WHOLE OF GRENFELL IS ON FIRE

Likes **3**

0:34 AM - 14 Jun 2017 0:34

Fabio Bebber @biobber Follow

Fire consuming Grenfell Tower. People screaming for their lives. Horrible #London #GrenfellTower

RETWEETS **463** LIKES **190**

0:41 AM - 14 Jun 2017 0:41

EBajgora @Ebjgora17 Follow

Fire fighters have begun to battle the blaze at Grenfell tower

Retweets **2,108** Likes **2,474**

1:03 AM - 14 Jun 2017 1:03

LBC Breaking @lbcbreaking Follow

We're receiving calls about a large fire at Grenfell Tower in Shepherd's Bush, west London - updates to follow

Retweets **110** Likes **134**

1:09 AM - 14 Jun 2017 1:09

1:43



BBC Breaking News ✓
@BBCBreaking

BUT NOT EVERYTHING IS TRUE



frick u
@wtfrickyou



This shark was found in front of someone's house in New Jersey. #ohmygod #Sandy
pic.twitter.com/1B0JqQUr

← Reply ↻ Retweet ★ Favorite ⋮ More



1,267
RETWEETS

246
FAVORITES

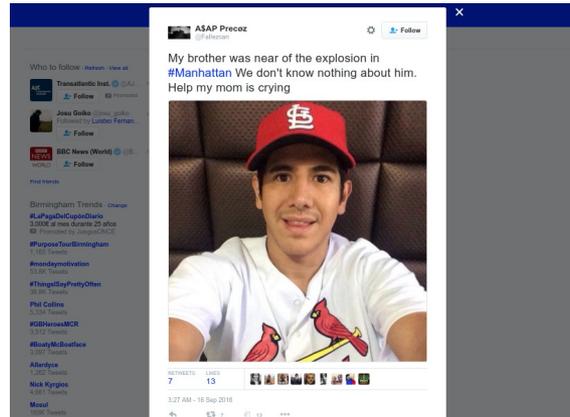
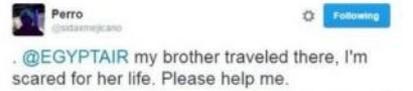


2:37 AM - 30 Oct 12

Flag media



BUT NOT EVERYTHING IS TRUE



When news of the missing EgyptAir flight emerged on Thursday morning people began sharing details of the unfolding story - but not everything was genuine.

Flight MS804 was an airliner from Paris to Cairo with 66 passengers and crew when

AUTOMATED FACT-CHECKING IS CHALLENGING

- **Fact-checking** social media content is **challenging**.
 - Huge volume of content, where not everything needs verification.
 - Widely studied as: fake vs real classification.
 - But what is the input to this classifier?

AUTOMATED FACT-CHECKING IS CHALLENGING

- Let's build automated fact-checking systems that:
 - 1) Detect pieces of information **needing verification** (checkworthy).
 - 2) **Make judgements** on checkworthy pieces of info.
 - 3) Use this to **assist humans**.

AUTOMATED FACT-CHECKING IS CHALLENGING

- Example of Twitter timeline:
 - I want to have a coffee now.
 - Yesterday there were 5,000 new cases of COVID-19 in the UK.
 - I hate COVID-19 and the lockdown.
 - Good morning everyone!
 - Today is Thursday.

AUTOMATED FACT-CHECKING IS CHALLENGING

- Example of Twitter timeline:
 - I want to have a coffee now. **[not checkworthy]**
 - Yesterday there were 5,000 new cases of COVID-19 in the UK. **[checkworthy]**
 - I hate COVID-19 and the lockdown. **[not checkworthy]**
 - Good morning everyone! **[not checkworthy]**
 - Today is Thursday. **[??]**

CONFLATION OF TERMS

- **Misinformation.**
- **Disinformation.**
- **Hoaxes.**
- **Fake News.**
- **Rumours.**

CONFLATION OF TERMS

- **Misinformation:** inaccurate, no intent to deceive.
- **Disinformation:** inaccurate, there is intent to deceive.
- **Hoaxes:** false story used to masquerade the truth, originating from the verb hocus, meaning “to cheat”
- **Fake News:** not 100% clear; fabricated news articles, parody, etc. (?)
- **Rumours:** piece of information that starts of as an unverified statement. Might be eventually resolved.

CONFLATION OF TERMS

- **Misinformation.**

- **Disinformation.**

- **Hoaxes.**

- **Fake News.**

} Always false.

- **Rumours.**

→ Starts off as unverified. Can be proven true / false, or remain unverified.

THREE STUDIES

- I'll be discussing three studies:
 - 1) Assessing the ability of people to verify social media.
 - 2) Detecting rumours needing verification.
 - 3) Attempting to predict the veracity of viral social media stories.

STUDY 1: VERIFICATION BY (UNTRAINED) HUMANS

HUMAN VERIFICATION

- How well would humans do in verifying social media content?
- What are the factors that lead to optimal verification by humans?

EPISTEMOLOGY RESEARCH

- Fallis (2004): we put together multiple factors when determining if something is true.
- For example:
"The Empire State Building, located in San Francisco, has 102 floors."

EPISTEMOLOGY RESEARCH

- Fallis (2004): we put together multiple factors when determining if something is true.
- For example:
"The Empire State Building, located in San Francisco, has 102 floors."
- The Empire State Building is in NYC, so.. is **the number of floors correct?**
- It actually is, but most probably **we wouldn't trust.**

HUMAN VERIFICATION

- Fallis (2004) stated that the key factors we relying on include:
 - Authority.
 - Plausibility and Support.
 - Independent Corroboration.
 - Presentation.

HUMAN VERIFICATION

- Our dataset included 332 popular pictures (34.9% fake) that were tweeted while the hurricane was in the NYC area.
- Through crowdsourcing, we asked workers to determine the veracity of pictures.
- We showed them different features, e.g.:
 - Only user info (picture not shown) -> Authority.
 - Multiple tweets with the same tweet -> Independent corroboration.
 - Etc.

HUMAN VERIFICATION

	P	R	F1
Authority	0.849	0.546	0.665
Plausibility	0.748	0.880	0.809
Picture	0.825	0.829	0.827
Corroboration	0.739	0.903	0.813
Presentation	0.674	0.583	0.625
Tweet	0.838	0.931	0.882
Random	0.651	0.5	0.565

HUMAN VERIFICATION

	P	R	F1
Authority	0.849	0.546	0.665
Plausibility	0.748	0.880	0.809
Picture	0.825	0.829	0.827
Corroboration	0.739	0.903	0.813
Presentation	0.674	0.583	0.625
Tweet	0.838	0.931	0.882
Random	0.651	0.5	0.565

- Overall **best** when they looked at the **entire tweet**: image + text + user.
- Best **precision**, however, when looking at the **user info** only.
- **Repetition bias**: seeing multiple tweets for the same image (corroboration) leads to a tendency to believe that more cases are real.

HUMAN VERIFICATION

- **Great** about the Twitter interface:
 - **We see all** tweet text + timestamp + basic user info together.

- **Not so great** about the Twitter interface:
 - We see **very limited user info!**
 - More user info needed, e.g. number of followers, user bio.

DETECTING RUMOURS

DATA COLLECTION IS CHALLENGING

- How to identify rumour data?
- How to make the dataset representative?
- How to build a sufficiently large dataset?
- How to get reliable data and labels?

BOTTOM-UP DATA COLLECTION

- Keyword-based data collection led to very large datasets for each datasets:
 - Data needed sampling.
 - We considered different sampling strategies.
 - Ended up choosing a popularity-based sampling strategy, i.e. more than N retweets, assuming that:
 - Rumours will be popular if they garner interest.

ANNOTATION OF RUMOURS

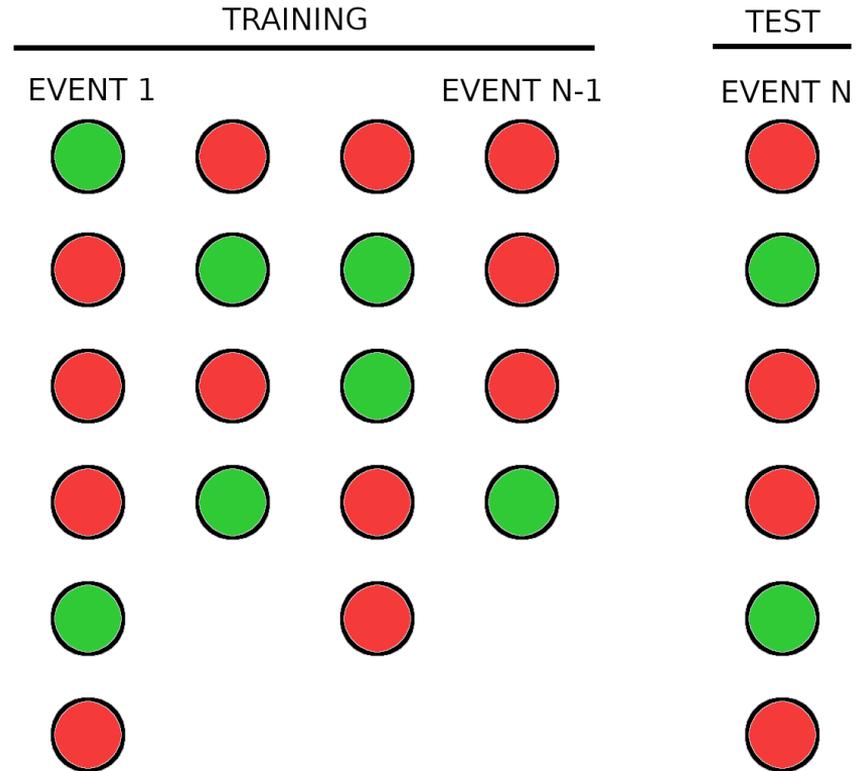
- Annotating rumours (i.e. needing to be checked) vs non-rumours.

✓ X ?		<p>@AFP #BREAKING Hostages held inside Sydney cafe, Islamic flag held up Sun Dec 14 23:50:42 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p> @ABCNews24 There are reportedly 13 hostages in the siege at Martin Place, Sydney CBD - watch us live for continuous updates Sun Dec 14 23:52:59 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@zaidbenjamin Strange flag in #Sydney, not the one affiliated with #ISIS. pic.twitter.com/36m6A3AKSb Sun Dec 14 23:53:45 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@SkyNewsAust Reportedly 13 hostages & 2 or more gunmen inside Lindt cafe in Sydney #sydneysiege (@lizlilley84) Sun Dec 14 23:54:23 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p> @Leroy_Lynch If you are in Sydney near Martin Place, do NOT tweet police movements, or mention them anywhere else. Sun Dec 14 23:55:44 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p> @tomsteinfort Latest reports are that there 13 hostages being held in the centre of Sydney, by two armed Islamic terrorists. Black Islamic flag in window Sun Dec 14 23:57:05 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@GrayConnolly I repeat=>if you are in Sydney, do NOT tweet or post news of what Police are doing around Martin Place. Assume jihadists follow social media Sun Dec 14 23:57:07 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@CBCNews Hostages being held in Sydney, Australia cafe ift.tt/1AaeOMX pic.twitter.com/wyDBmMzs0N Sun Dec 14 23:57:37 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@Sandra_Sully Sydney Opera House has been completely evacuated and Sydney airspace closed. #siege Sun Dec 14 23:58:10 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@TRobinsonNewEra What's happening in Sydney will happen in the uk it's inevitable, and our political leaders will have blood on their hands Sun Dec 14 23:58:52 +0000 2014 - see tweet on Twitter</p>
✓ X ?		<p>@MajorNews911 UPDATE: Hostages in Sydney cafe brandishing black Shahada flags, identical to those used by ISIS. - @7NewsSydney pic.twitter.com/xap6sGvEdT Sun Dec 14 23:58:53 +0000 2014 - see tweet on Twitter</p>

RUMOUR DETECTION

- **Task definition:** Given a stream of tweets linked to an event (e.g. breaking news), determine if each of these tweets constitutes a rumour or non-rumour.
- **Motivation:** rumours, as unverified pieces of information, need flagging as such.

RUMOUR DETECTION



RUMOUR DETECTION

- **Intuition:** whether or not a tweet is a rumour depends on context, i.e. what is being reported in preceding tweets.
- **Proposed method:** Conditional Random Fields for sequential modelling.

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
MaxEnt	0.338	0.442	0.383
CRF	0.667	0.556	0.607

RUMOUR DETECTION WITH USER INFO

- **Intuition:** some users are more likely to spread rumours, so user info can be useful to detect rumours.
- **Problem:** many users in test data are new, unseen in training data.

RUMOUR DETECTION WITH USER INFO

- **Intuition:** some users are more likely to spread rumours, so user info can be useful to detect rumours.
- **Problem:** many users in test data are new, unseen in training data.
- **Proposed solution:** based on the theory of homophily, users will follow others like them, i.e. if a user follows others who spread rumours in the spread, they're likely to spread rumours themselves.

RUMOUR DETECTION WITH USER INFO

- **CRF:** no user info.
- **CRF + RR:** user's own rumour ratio (from user's history, how many rumours vs non-rumours they posted).
- **CRF + HP:** average rumour ratio of followed users.

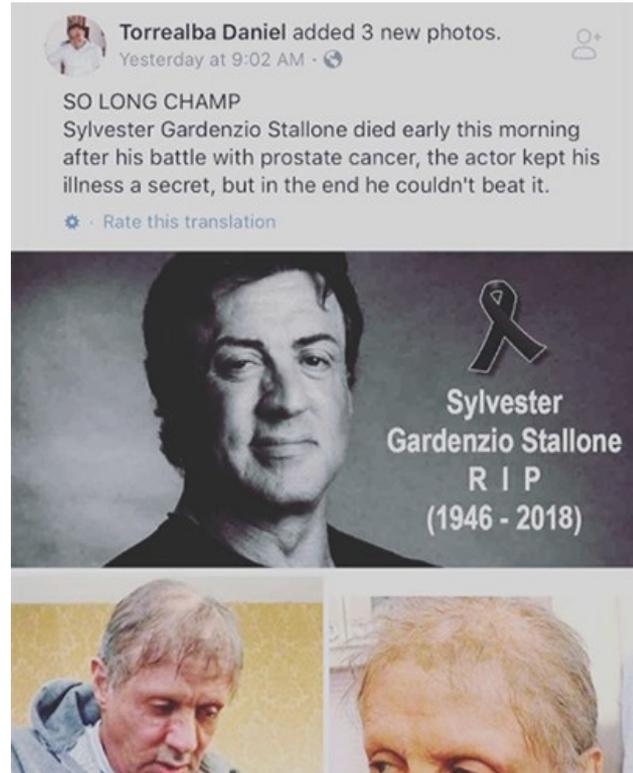
RUMOUR DETECTION WITH USER INFO

- **CRF**: no user info.
- **CRF + RR**: user's own rumour ratio (from user's history, how many rumours vs non-rumours they posted).
- **CRF + HP**: average rumour ratio of followed users.

Classifier	P	R	F1
CRF	0.667	0.556	0.607
CRF + RR	0.654	0.593	0.622
CRF + HP	0.633	0.635	0.634

AUTOMATED VERIFICATION OF VIRAL STORIES

CELEBRITY DEATH HOAXES



COLLECTION OF SOCIAL MEDIA HOAXES

- Collection of **death reports (RIP + person name)**, e.g.:
 - *"RIP Elizabeth II, she was so inspiring."*
 - *"RIP Elizabeth II oh dear :("*
 - *"Sad to hear about the passing of RIP Elizabeth II"*
 - *"Those posting RIP Elizabeth II, stop it!"*

COLLECTION OF SOCIAL MEDIA HOAXES

- Easy to verify (post hoc) using Wikidata.

- "RIP Elizabeth II, she"
- "RIP Elizabeth II oh d"
- "Sad to hear about t"
- "Those posting RIP E"



The screenshot shows the Wikidata page for Elizabeth II (Q9682). It includes a portrait of her in a pink hat and coat, the Wikidata logo, and a table of properties and values:

Property	Value
date of birth	21 April 1926 <i>Gregorian</i> 6 references
place of birth	Mayfair located at street address: 17 Bruton Street, London (British English) 1 reference
father	George VI

FAKE!

WIKIDATA ENTRY

```
{"id": "8023",  
  "name": "Nelson Mandela",  
  "birth": {"date": "1918-07-18", "precision": 11},  
  "death": {"date": "2013-12-05", "precision": 11},  
  "description": "former President of South Africa, anti-apartheid activist",  
  "aliases": ["Nelson Rolihlahla Mandela", "Mandela", "Madiba"]}
```

 Names to match

 Death date to compare with

COLLECTION OF SOCIAL MEDIA HOAXES

- 1) Collection of tweets with **keyword 'RIP'** in it for 3 years (Jan 2012 – Dec 2014).
- 2) Sample tweets matching the 'RIP person-name' pattern.
- 3) **Sampling**, i.e. names with 50+ occurrences on a given day.
- 4) Semi-automated **labelling**.
- 5) 4,007 death reports (13+ million tweets):
 - 2,301 real deaths.
 - 1,092 commemorations.
 - 614 death hoaxes.

RESULTS

	0	1'	2'	5'	10'	15'	30'	60'	120'	300'
social	.427	.495	.509	.510	.510	.528	.535	.577	.594	.591
w2v	.641	.655	.658	.663	.667	.670	.680	.696	.699	.698
social+w2v	.612	.634	.661	.671	.671	.677	.675	.709	.709	.724
gw2v	.556	.565	.574	.608	.612	.618	.623	.645	.648	.664
social+gw2v	.569	.590	.599	.616	.633	.647	.663	.679	.688	.686
infersent	.637	.640	.653	.664	.683	.681	.697	.722	.734	.759
social+infersent	.643	.655	.670	.678	.691	.688	.698	.731	.748	.767
multiw2v*	.669	.676	.691	.703	.714	.722	.723	.721	.738	.741
social+multiw2v*	.647	.677 ‡	.696 ‡	.707 ‡	.716 ‡	.725 ‡	.724 †	.744 †	.752	.748

Proposed methods indicated with a star (*). Best method highlighted in bold and second-best method for different types of features highlighted in italic. ‡: statistically significant at $p < .01$, †: statistically significant at $p < .05$.

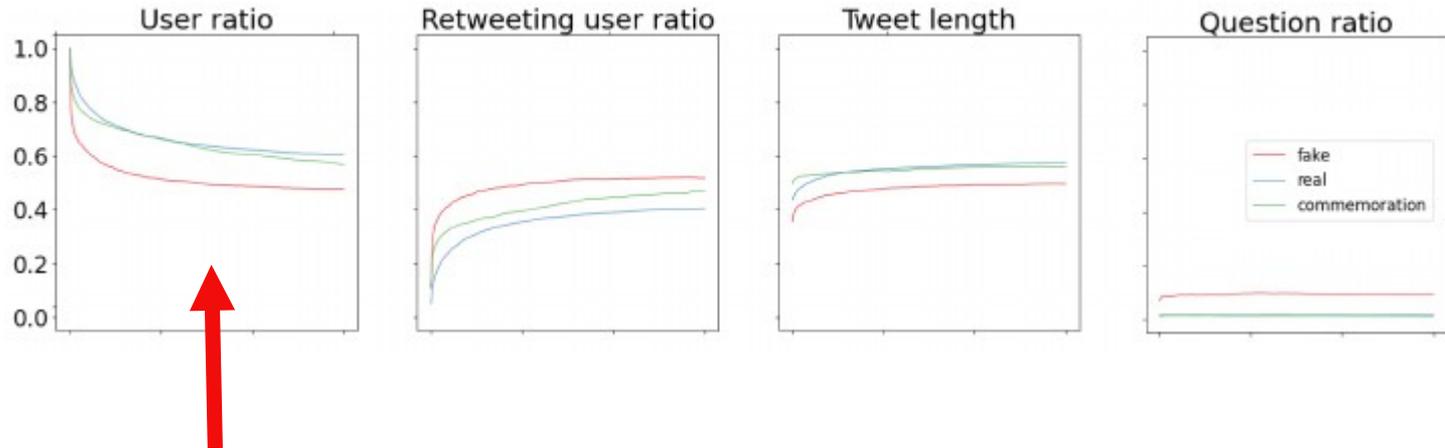
RESULTS: USING SLIDING WINDOWS

- What if we use the last few minutes of data only?

window	0	1'	2'	5'	10'	15'	30'	60'	120'	300'
0.1	.647	.385	.399	.413	.423	.442	.452	.459	.466	.514
0.25	.647	.422	.468	.476	.478	.519	.522	.547	.582	.617
0.5	.647	.228	.284	.369	.537	.544	.575	.589	.642	.673
0.75	.647	.253	.319	.396	.554	.580	.598	.626	.671	.718
1.0	.647	.677	.696	.707	.716	.725	.724	.744	.752	.748

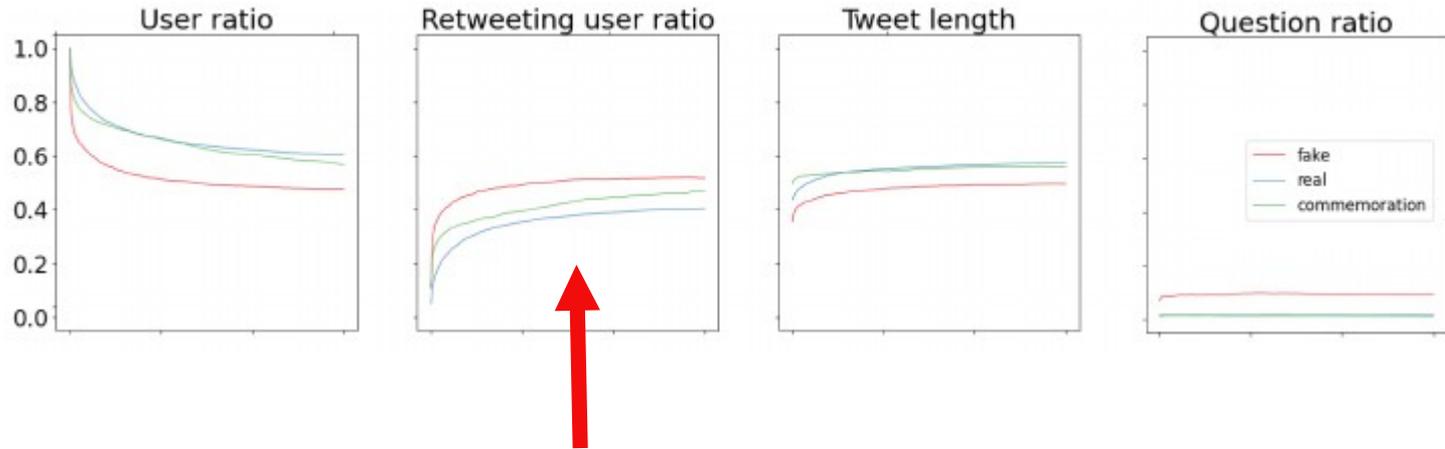
- Limited capacity when verification is done without linking to evidence.
 - Verification linked to evidence is showing better performance.

ANALYSIS OF FEATURES



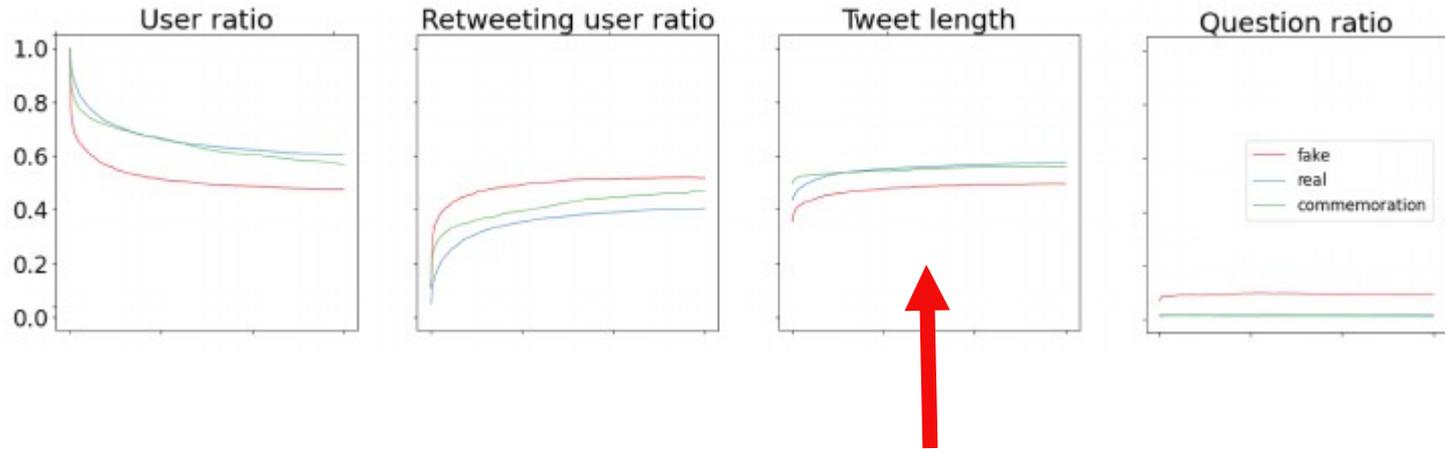
- Hoaxes tend to have fewer distinct users posting them.

ANALYSIS OF FEATURES



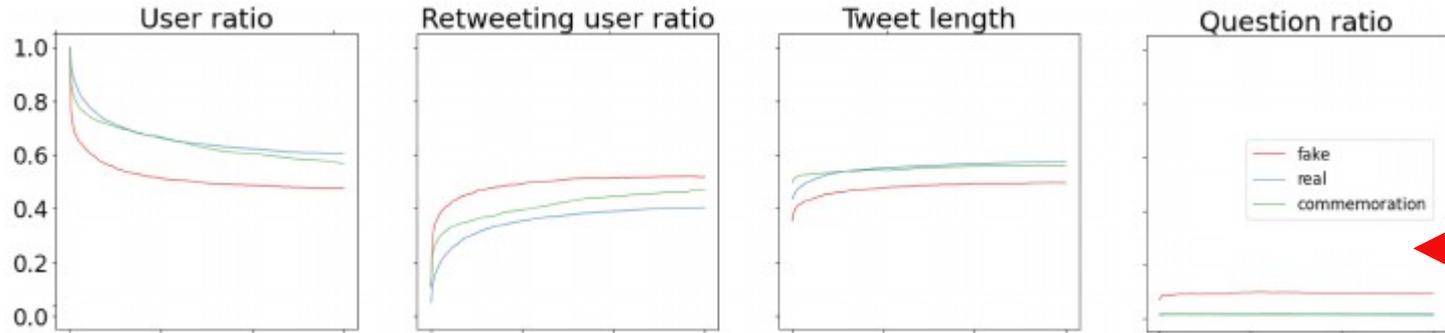
- Hoaxes tend to have fewer distinct users posting them.
 - BUT they are retweeted by more distinct users!

ANALYSIS OF FEATURES



- Hoaxes tend to be shorter in length, not as carefully crafted as true stories?
 - They tend to lack links and pictures.
 - Presumably less evidence linked to them?

ANALYSIS OF FEATURES



- And hoaxes tend to spark more questions!

DATA & PAPER AVAILABLE

Twitter Death Hoaxes dataset

Version 3  Dataset posted on 25.03.2019, 18:55 by [Arkaitz Zubiaga](#)

This is a dataset of death reports collected from Twitter between 1st January, 2012 and 31st December, 2014. It was collected by tracking the keyword 'RIP', and matching those tweets in which a name is mentioned next to RIP. Matching names were identified by using Wikidata as a database of names. For more details, please refer to the paper: <https://arxiv.org/abs/1801.07311>

RESEARCH-ARTICLE

Early Detection of Social Media Hoaxes at Scale



Authors:  Arkaitz Zubiaga,  Aliqi Liang [Authors Info & Affiliations](#)

Publication: ACM Transactions on the Web • August 2020 • Article No.: 18 • <https://doi.org/10.1145/3407194>

99   23



Abstract

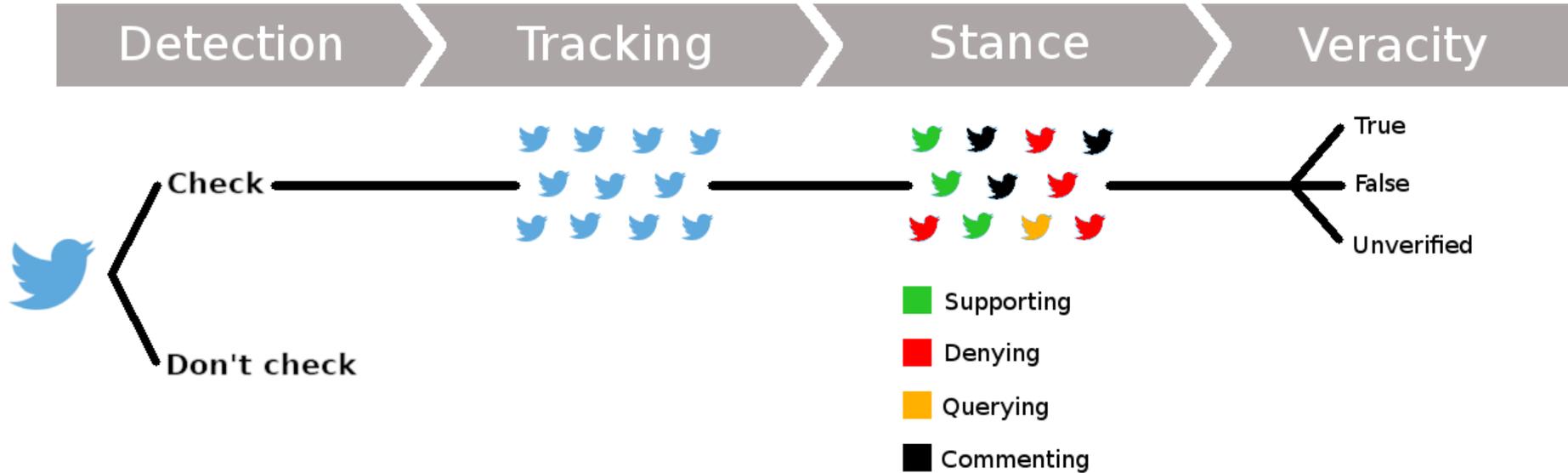
The unmoderated nature of social media enables the diffusion of hoaxes, which in turn jeopardises the credibility of information gathered from social media platforms. Existing research on automated detection of hoaxes has the limitation of using relatively small datasets, owing to the difficulty of getting labelled data. This, in turn, has limited research



https://figshare.com/articles/Twitter_Death_Hoaxes_dataset/5688811

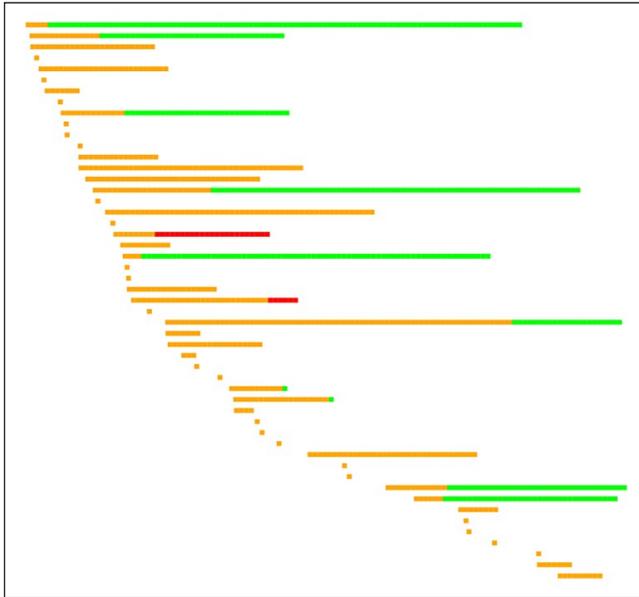
<https://dl.acm.org/doi/10.1145/3407194>

AUTOMATED VERIFICATION PIPELINE

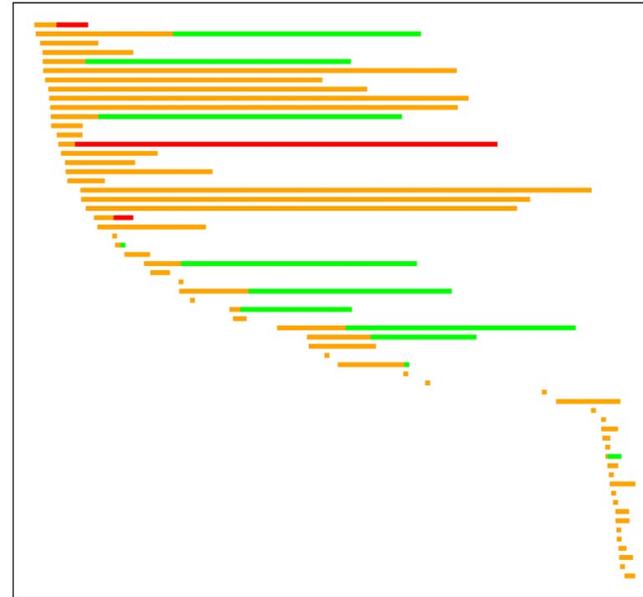


SOCIAL MEDIA STORY TIMELINES

Ottawa shooting



Sydney siege



- **Orange** while story is still **unverified**.
- **Green / red** indicate story has been proven **true / false**.

DISCUSSION

- Automated fact-checking is very **challenging**.
- More research needed considering the **entire pipeline**, i.e. starting from detecting check-worthy claims.
- Stories can be **unverified**, i.e. lacking evidence for verification.
 - We need to consider this in models.
- More research needed in **verification by linking claims to evidence**.
 - Automatically finding evidence is however challenging.

STAY TUNED

Automatic fact checking and misinformation detection

Leon Derczynski
IT University of Copenhagen

Arkaitz Zubiaga
Queen Mary University of London

Kalina Bontcheva
University of Sheffield

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES
#13



MORGAN & CLAYPOOL PUBLISHERS

QUESTIONS?

Zubiaga, A., & Ji, H. (2014). Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, 4(1), 163.

Zubiaga, A., Liakata, M., & Procter, R. (2017, September). Exploiting context for rumour detection in social media. In *International Conference on Social Informatics* (pp. 109-123). Springer, Cham.

Zubiaga, A., & Jiang, A. (2020). Early detection of social media hoaxes at scale. *ACM Transactions on the Web (TWEB)*, 14(4), 1-23.

Lathiya, S., Dhobi, J. S., Zubiaga, A., Liakata, M., & Procter, R. (2020). Birds of a feather check together: Leveraging homophily for sequential rumour detection. *Online Social Networks and Media*, 19, 100097.