Fake News
Deception Detection from Data mining perspective

jan.skacel@fs.cvut.cz

Jan Skácel
January 2018

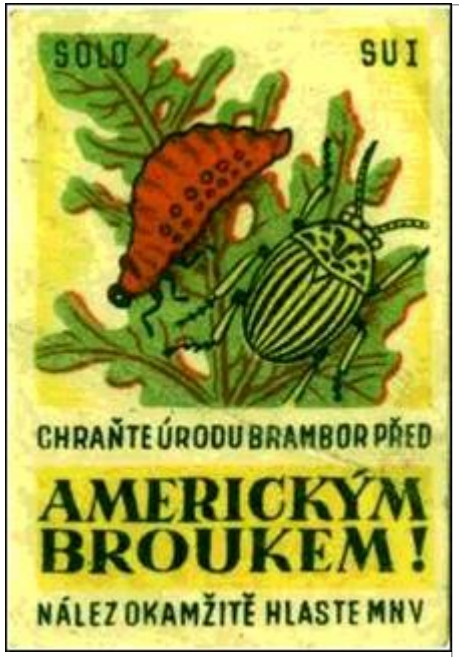
Outline

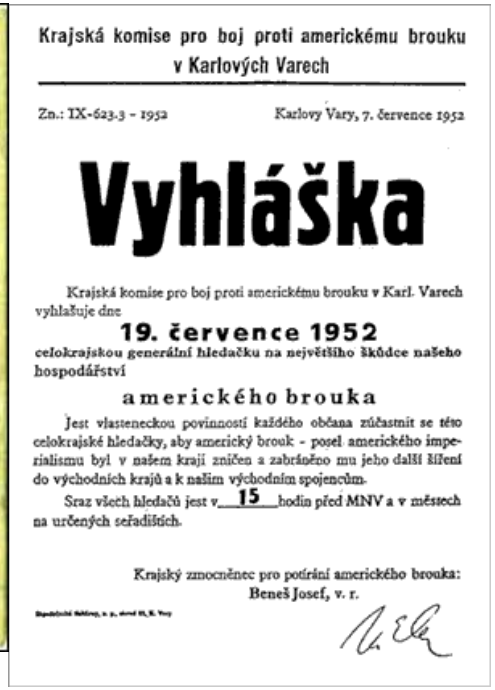
- Deceptive news - Social context
- Disinformation in 21. century
  - Spreading on social media
  - Publisher-news-user relationship
- Detecting Fake news
  - Strategies, Feature selection and Models
- Computational Fact checking
  - Using knowledge graph – Wikipedia
  - Embedding Framework Publisher – News - User
- Automated News Generation

…fake news is not new…the way and speed it spreads and it's power is….

Yesterday:
Isolate => spread FN/propaganda
Back Channels v.s. reputable sources
Propaganda vs. free media

Today:
Overwhelm with excess news incl. FN
62% of adults in US get news from social media
500 mil. tweets every day





Krajská komise pro boj proti americkému brouku v Karlových Varech

Zn.: IX-623.3 - 1952 Karlovy Vary, 7. července 1952

Vyhláška

Krajská komise pro boj proti americkému brouku v Karl. Varech vyhlašuje dne

19. července 1952

celokrajskou generální hledačku na největšího škůdce našeho hospodářství

amerického brouka

Jest vlasteneckou povinností každého občana zúčastnit se této celokrajské hledačky, aby americký brouk - posel amerického imperialismu byl v našem kraji zničen a zabráněno mu jeho další šíření do východních krajů a k našim východním spojencům.

Sraz všech hledačů jest v 15 hodin před MNV a v městech na určených seřadištích.

Krajský zmocněnec pro potírání amerického brouka:
Beneš Josef, v. r.

Pope Francis endorsed Donald Trump for president.

July 2016 – above claim @ web site *WTOE 5 News (then 2 weeks old)*
By November 8, the story had picked up 960,000 Facebook engagements, according to Buzzfeed.

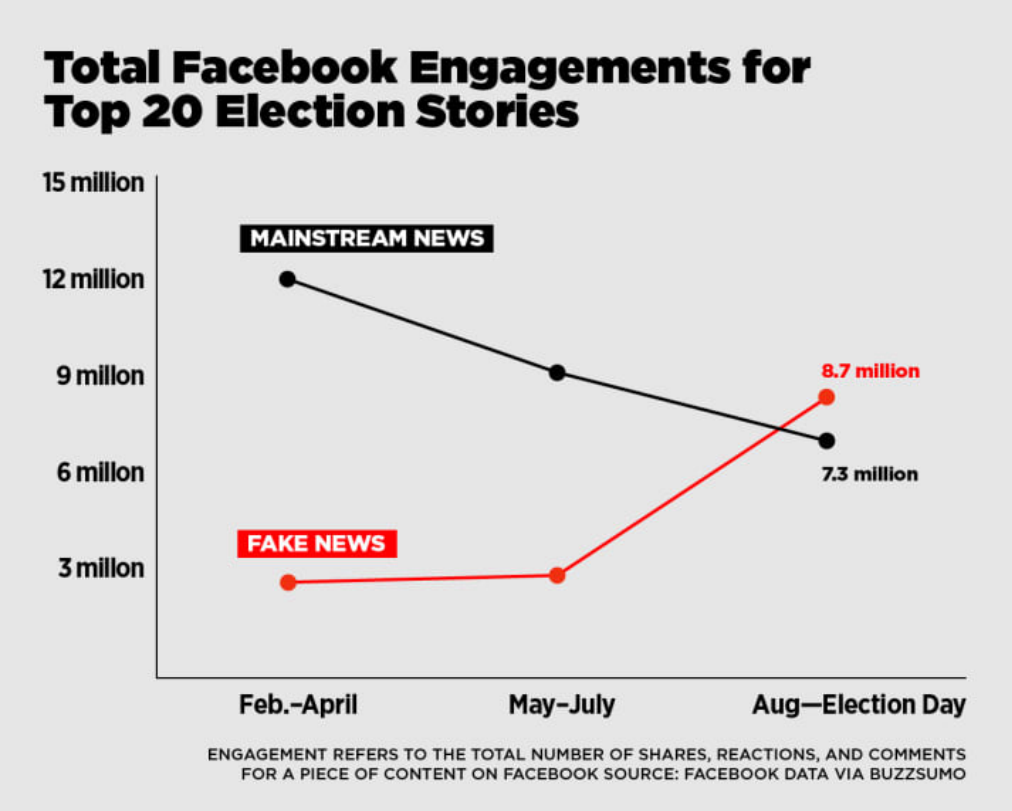
Fake News Motivation & Dissemination

Definition: **Fake news is a news article that is intentionally and verifiably false**
FN is not: satire, rumors, conspiracy theories, unintentional misinformation & fun hoaxes

62% of adults in US get news from social media
Twitter: 500 mil. tweets every day

Fake news:
intentionally written to mislead readers – difficult to detect based on news content

US 2016 presidential election



Fake news motivation:
Profit – click bait – advertising revenue
Political gain – influence public opinion, distract from other topics
Cover-up – distort inconvenient truth, “alternative facts”
Undermine – beliefs in what’s right or wrong, media, democracy

Fake news detection

Publisher credibility - Correlation between the partisan bias of publisher and news contents veracity (p1 left, p2 right, p3 neutral)

User credibility - malicious accounts or users vulnerable to fake news, are more likely to spread fake news

Confirmation bias – people preferentially believe information that fit to their views, form social relationships with like-minded people

Fake news detection – based on:
Content – truth?
Source – credible?
Dissemination pattern

Motivation	Publisher	Consumer
Fake news	Short-term - profit - # consumers reached	Psychology – concurring news
Objective news	Long-term - reputation	Information – true news

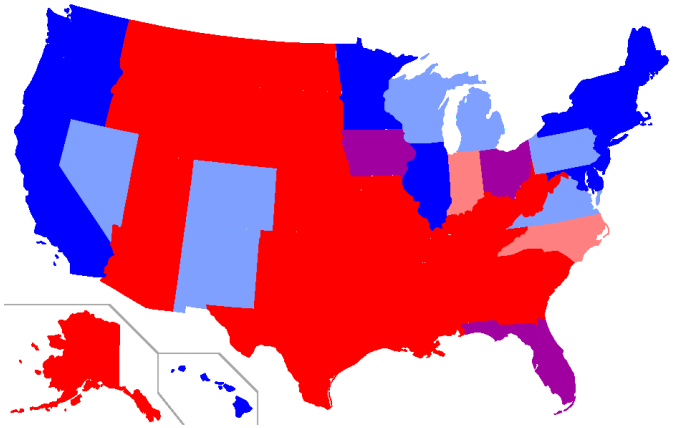
Fake News Enablers

Malicious accounts

Bots – 19 million bot accounts on Twitter before U.S. 2016 elections

Trolls – humans 1000 paid trolls before U.S. election

Cyborgs –hybrid – combination of bot and human - hard to detect



US Red (Republican) and Blue (Democrats) states - polarization, echo chamber

Echo Chamber

Users follow like-minded people, homepage news feed algorithm predominantly selects concurring news – keeping user's attention

Results - loss of perspective e.g. "everyone I know is red"

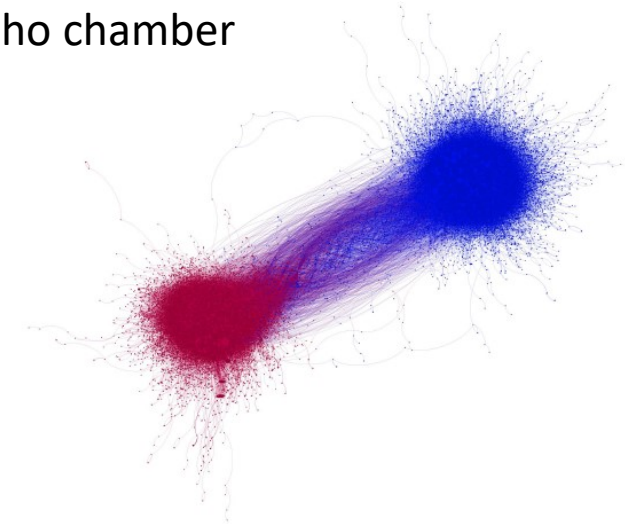


Figure 1: Network of retweets showing democrats (in blue) and republicans (in red) divided into two distinct communities. What is the impact of such polarization in what is perceived as "fake news"?

Psychological effects:

Social credibility – if my friends believe it, I believe it too

Frequency heuristic – if I hear it many time I believe it

Fake news detection - Feature Extraction

Content based – traditional media

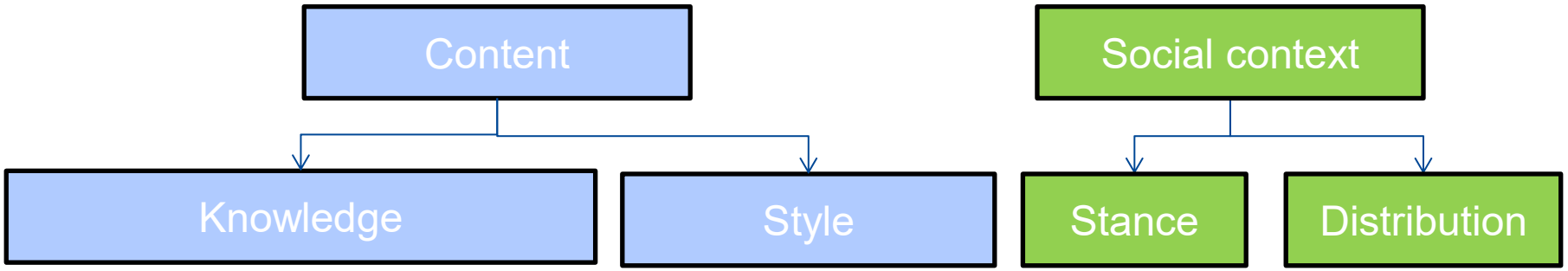
Raw data - Author/publisher, headline, article body, pictures
Linguistic - writing style - FN catchy headline – clickbait, deceptive content , references, sources, unique words, total words
Visual - Pictures- features, count, properties, clarity score, coherence score, similarity distribution histogram, diversity score, and clustering score, count, image ratio, hot image ratio etc.

Social context – how is news spread - additional features from social networks
Users level – statistic, #followers/followees, verification – idea bots, trolls different to humans
Posts – user feedback on news – agree/disagree, stance, changes in time, credibility
Networks – FN dissemination – echo chamber – numerous networks
stance network – nodes tweets relevant to the news – edge – similarity in stance
co-occurrence network – counts user engagements in the news
diffusion – trajectory of the spread of the news among users
friendship – follower/followee structure

After these networks are assembled, relations can be drawn, e.g.
degree and clustering coefficient been diffusion and friendship network

Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang, Huan Liu, 2017: Fake News Detection on Social Media: A Data Mining Perspective

Fake news detection - Model Construction



Fact-checking:

- Expert – human checking (PolitiFact, Snopes)
- Crowdsourcing – crowd annotates news article (Fiskkit)
- Computational information extraction from *open source* or *knowledge graph*
  i) identifying check-worthy claims
  ii) discriminating the veracity of fact claims

Deception

- style, syntax
- Catchy headlines
- Title and text not corresponding

Objectivity

One sided arguments, partisan journalism

Like/dislike

User stance - from post whether the user is in favor of, neutral toward, or against some target entity

interrelations of relevant social media posts

Assumption: credibility of news related to credibility of related posts

PageRank-like credibility propagation

Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang, Huan Liu, 2017: Fake News Detection on Social Media: A Data Mining Perspective

Computational Fact Checking from Knowledge Networks

- statement of fact represented by a subject-predicate-object triple, e.g., “Socrates” - “is a” - “person”
- A set of such triples can be combined to produce a knowledge graph (KG), where
  nodes denote entities, i.e. subjects or objects of statements
  edges denote predicates

Fact Checking:
Given a set of statements extracted from a knowledge repository e.g. Wikipedia
resulting KG network represents all factual relations among entities mentioned in those statements.

New statement: TRUE if it exists as an edge of the KG, or if there is a short path linking its subject to its object within the KG
FALSE – no link or short path exists

Limitations:
- statements - only relevant to positive SPO objects
- knowledge source – limited scope, may not cover recent news

Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F, et al. (2015) Computational Fact Checking from Knowledge Networks. PLOS ONE 10(6): e0128193. https://doi.org/10.1371/journal.pone.0128193
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128193

Using Wikipedia to Fact-check Statements

(a) Wikipedia Knowledge Graph (WKG) from Wikipedia's "infoboxes" - 3 million entity nodes linked by approximately 23 million edges

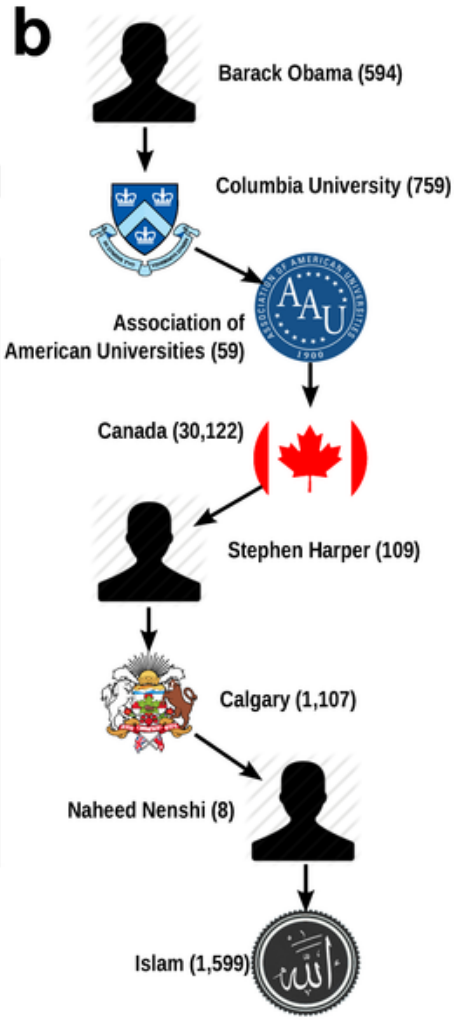
(b) Computing the truth value of a subject-predicate-object statement "*Barack Obama is a muslim"* finding shortest path between subject and object entities. Numbers in () indicate the degree of the nodes.
The path traverses high-degree nodes representing generic entities, e.g. Canada, and is assigned a low truth value.

Semantic value

$$\mathcal{W}(P_{s,o}) = \mathcal{W}(v_1 \ldots v_n) = \left[1 + \sum_{i=2}^{n-1} \log k\,(v_i)\right]^{-1}$$

k(v) is the degree of node v, P is predicate
n is path length, e is statement e=(o,p,s)
Truth τ(e)=maxW, $\tau \in [0,1]$

"Barack Obama is a muslim"



Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F, et al. (2015) Computational Fact Checking from Knowledge Networks. PLOS ONE 10(6): e0128193. https://doi.org/10.1371/journal.pone.0128193
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128193

Fake News Detection using A Tri-Relationship Embedding Framework Publisher – News - User

Publishers set $P = \{p_1, p_2, \ldots p_l\}$ l publishers

News articles set $A = \{a_1, a_2, \ldots a_n\}$ n news

Users set $U = \{u_1, u_2, \ldots u_m\}$ m users

Feature matrix $X \in R^{n \times t}$

User adjacency $A \in \{0,1\}^{m \times m}$

News engagement matrix $W \in \{0,1\}^{m \times n}$

focus on user stance – agree with news

Publisher –news relation matrix $B \in R^{l \times n}$

B_{kj} =1 publisher p_k publishes a_j

Publisher bias $o \in \{-1, 0, 1\}^{l \times 1}$ left, neutral, right

assume partisan labels available

User credibility score $c = \{c_1, c_2, \ldots, c_m\}$

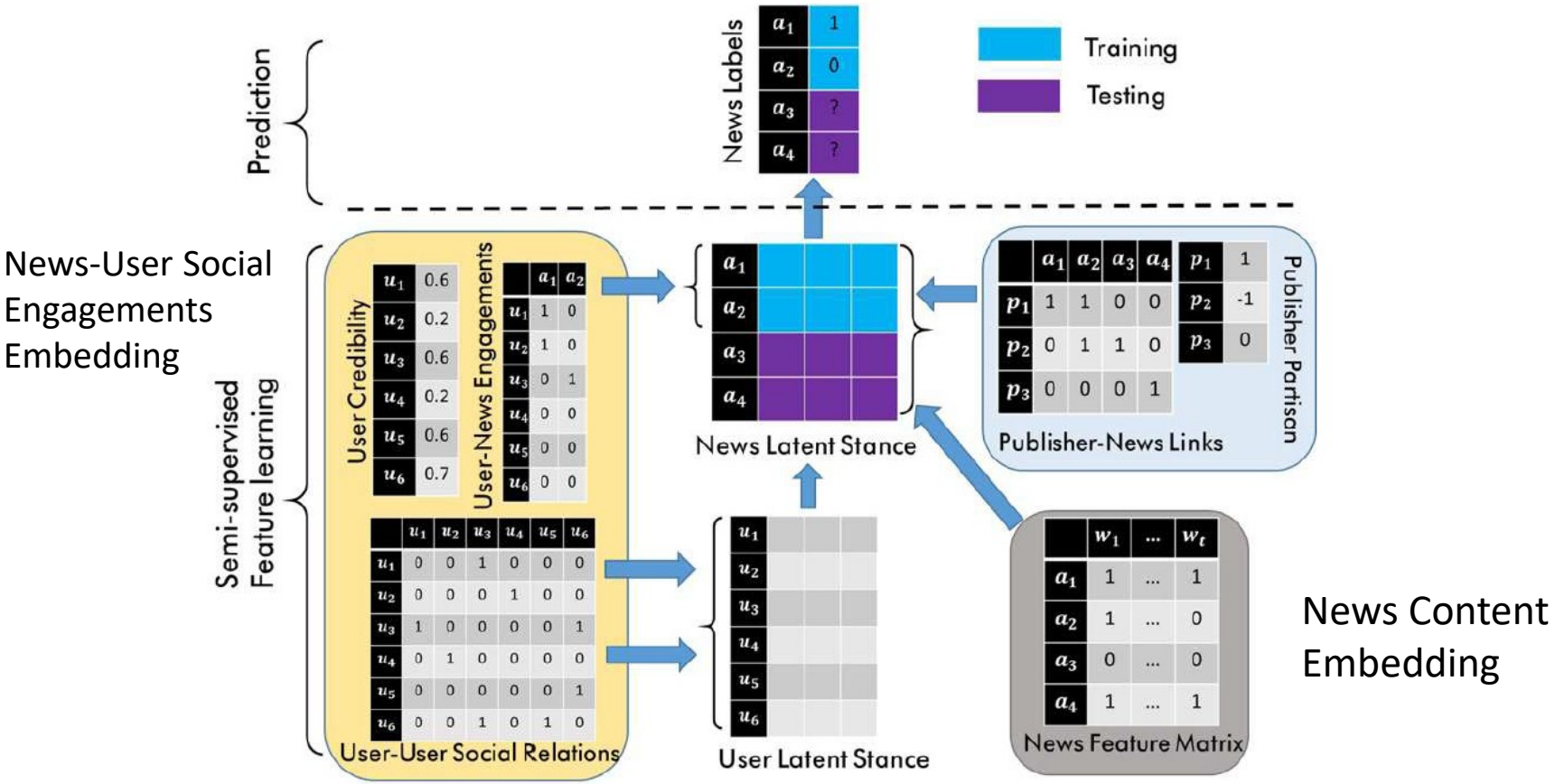
Veracity news label $y = \{y_1, y_2, \ldots y_n\} \in R^{n \times 1}$

Fake news – binary problem y=1 fake, y=-1 true

Kai Shu, Suhang Wang and Huan Liu, 2017: Exploiting Tri-Relationship for Fake News Detection

Fake News Detection using A Tri-Relationship Embedding Framework Publisher – News - User

Goal: **Predict unlabeled news y_U** given news matrix X, user engagement matrix A, user news engagement matrix W, publisher-news publishing matrix B, publisher partisan label vector o, partial labeled news vector y_L



Kai Shu, Suhang Wang and Huan Liu, 2017: Exploiting Tri-Relationship for Fake News Detection

Tri-Relationship Embedding Framework Verification

Table 1: The statistics of datasets

Platform	BuzzFeed	PolitiFact
# Candidate news	182	240
# True news	91	120
# Fake news	91	120
# Users	15,257	23,865
# Engagements	25,240	37,259
# Social Links	634,750	574,744
# Publisher	9	91

Table 3: Summary of the detection methods for comparison

Method	News Content	Social Engagements	Publisher Partisan
RST (*28*)	✓		
LIWC (*93*)	✓		
Castillo (*10*)		✓	
RST+Castillo (*38*)	✓	✓	
LIWC+Castillo (*103*)	✓	✓	
TriFN	✓	✓	✓

RST extracts news style-based features

LIWC extracts the lexicons falling into psycholinguistic categories and capture the deception features from a psychology perspective

Castillo predicts news veracity using social engagements

Table 2: Performance comparison for fake news detection

Datasets	Metric	RST	LIWC	Castillo	RST+Castillo	LIWC+Castillo	TriFN
BuzzFeed	Accuracy	0.610 ± 0.023	0.655 ± 0.075	0.747 ± 0.061	0.758 ± 0.030	0.791 ± 0.036	$\mathbf{0.864 \pm 0.026}$
	Precision	0.602 ± 0.066	0.683 ± 0.065	0.735 ± 0.080	0.795 ± 0.060	0.825 ± 0.061	$\mathbf{0.849 \pm 0.040}$
	Recall	0.561 ± 0.057	0.628 ± 0.021	0.783 ± 0.048	0.784 ± 0.074	0.834 ± 0.094	$\mathbf{0.893 \pm 0.013}$
	F1	0.555 ± 0.057	0.623 ± 0.066	0.756 ± 0.051	0.789 ± 0.056	0.802 ± 0.023	$\mathbf{0.870 \pm 0.019}$
PolitiFact	Accuracy	0.571 ± 0.039	0.637 ± 0.021	0.779 ± 0.025	0.812 ± 0.026	0.821 ± 0.052	$\mathbf{0.878 \pm 0.020}$
	Precision	0.595 ± 0.032	0.621 ± 0.025	0.777 ± 0.051	0.823 ± 0.040	0.856 ± 0.071	$\mathbf{0.867 \pm 0.034}$
	Recall	0.533 ± 0.031	0.667 ± 0.091	0.791 ± 0.026	0.792 ± 0.026	0.767 ± 0.120	$\mathbf{0.893 \pm 0.023}$
	F1	0.544 ± 0.042	0.615 ± 0.044	0.783 ± 0.015	0.793 ± 0.032	0.813 ± 0.070	$\mathbf{0.880 \pm 0.017}$

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Precision = TP / (TP+FP)

Recall = TP / (TP+FN)

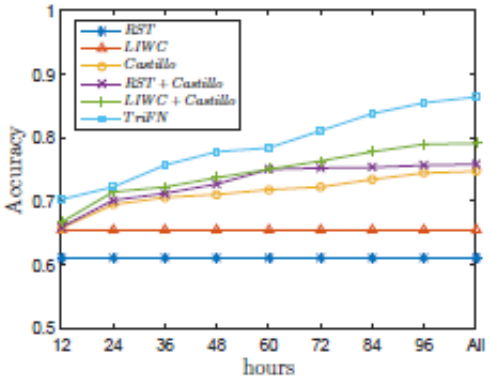
F1 = 2 · Precision·Recall / (Precision+Recall)

Kai Shu, Suhang Wang and Huan Liu, 2017: Exploiting Tri-Relationship for Fake News Detection

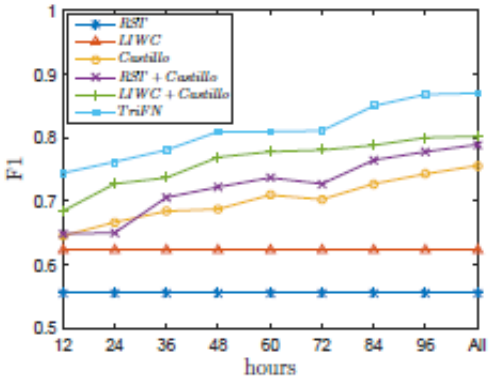
Tri-Relationship Embedding Framework Early Fake News Detection

Early detection to give early alert
- limited social engagements
- delay time in 12 to 96 hours
- detection improves with increased delay - prove that engagements on social media provide additional information

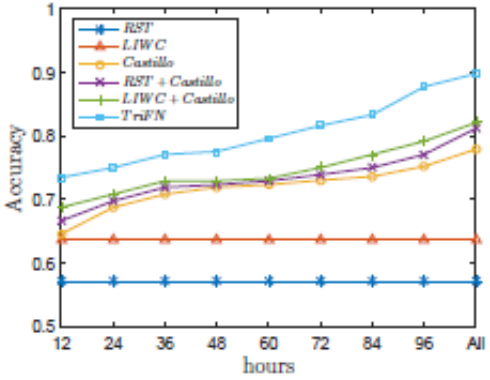
Proposed TriFN always achieve best performance - shows importance of modeling user-user relation and news-user relations to capture effective feature representations



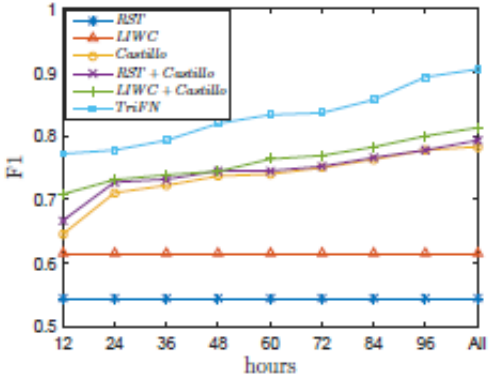
(a) Accuracy on BuzzFeed



(b) F1 on BuzzFeed



(c) Accuracy on PolitiFact



(d) F1 on PolitiFact

The performance of early fake news detection on BuzzFeed and PolitiFact in terms of Accuracy and F1

Kai Shu, Suhang Wang and Huan Liu, 2017: Exploiting Tri-Relationship for Fake News Detection

Automated True News Generation - Reuters Tracer

News agencies - Reuters

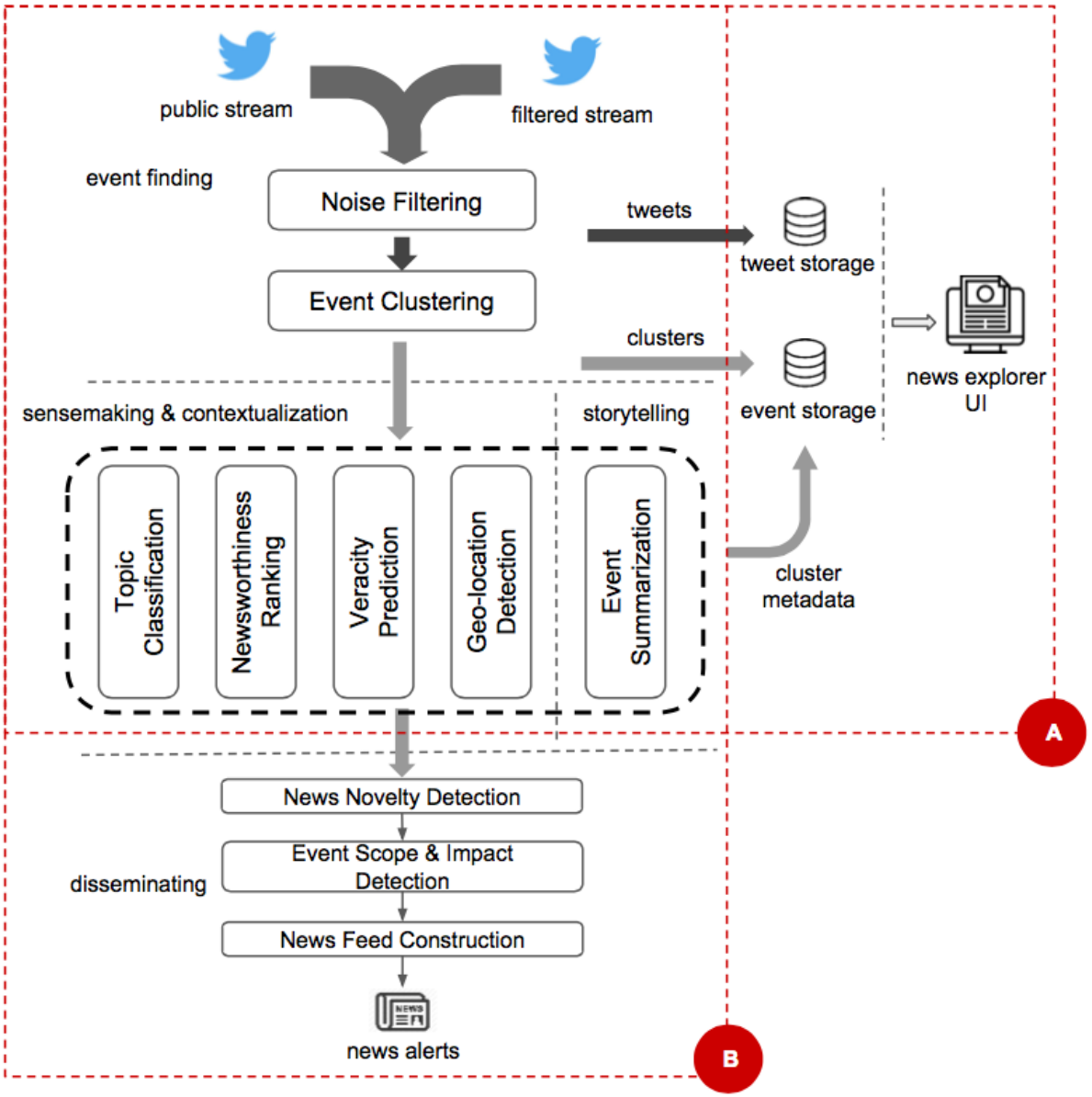
500 mil. tweets/day

Tracer's system architecture for two use cases:
(A) news exploration UI;
(B) automated news feeds

10-20% of news breaks first on Twitter

2% of Twitter data (~12+ million tweets everyday)
1% Random public stream
1% filtered stream

Success rate 70% of daily news reported by
journalists of global news media and agencies such as Reuters,
(95% if using 50+ million tweets)



X. Liu, A. Nourbakhsh, Q. Liy, S. Shahz, R. Martin, J. Duprey, 2017: Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data

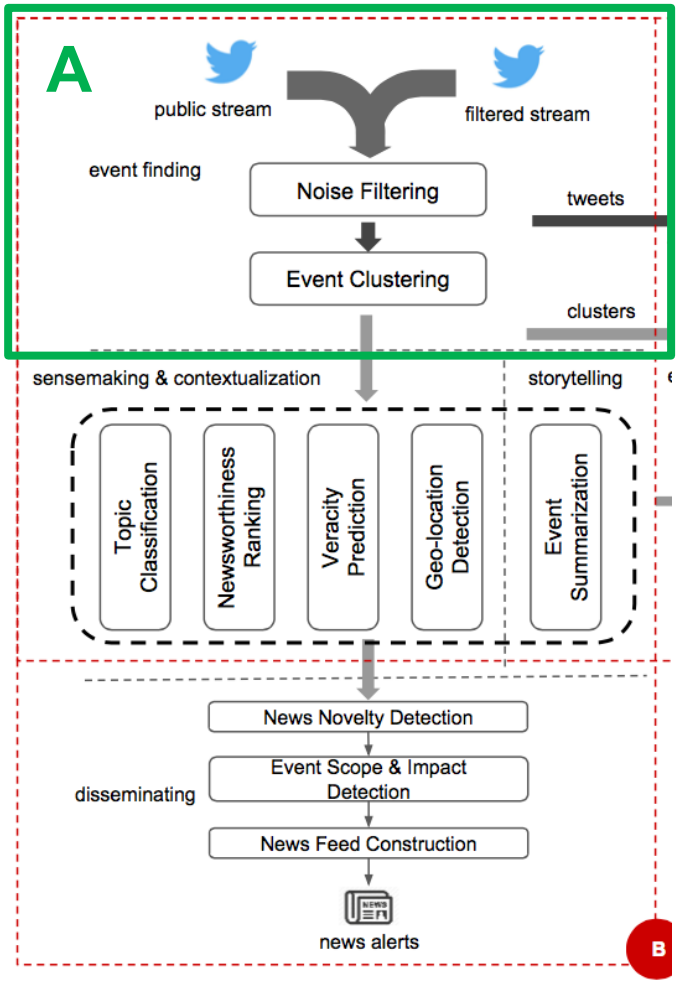
Reuters Tracer – News Detection Algorithm

A Event Discvovery

1/ Noise filtering
Spectrum from noise to news: spam, advertisements, chit-chat, general info, events, normal and breaking news

2/ Event Clustering
event detection via clustering tweets
Idea: if group of people talk about the same subject at a particular time, it is likely to be an event

Two phases: clustering and merging
Unit clusters of min 3 tweets with similar content merged with a pool of existing clusters
No merge => NEW EVENT



X. Liu, A. Nourbakhsh, Q. Liy, S. Shahz, R. Martin, J. Duprey, 2017: Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data

Reuters Tracer – News Detection Algorithm – cont.

<u>B Sense-making & Contextualization</u>

1/ Topic - modification of TD-IDF

Cluster e $\{w_1,\ldots w_m\}$ with m tweets

2/ Newsworthiness Detection

Model of News Topics

- trained from positive dataset – 1 year of Twitter feed from 31 reputable news agencies (AP, CNN, BBC, NYT)

Topic model $Z_i = \{z_1;..; z_n\}$ for n topics

News probability of cluster topics:

$P_T(e)=\Sigma_i\Sigma_{t5}\, p(Z_l|w_i)/m$: top 5 topics, averaged tweets in cluster

Model of News Object

name entities $S_i = \{s_1;..; s_{nk}\}$ extracted from training dataset

Object frequency distribution as news probabilities

Probability of cluster news object $Po_T(e)=\Sigma_i\Sigma_{t2}\, p(S_l|w_i)/m$

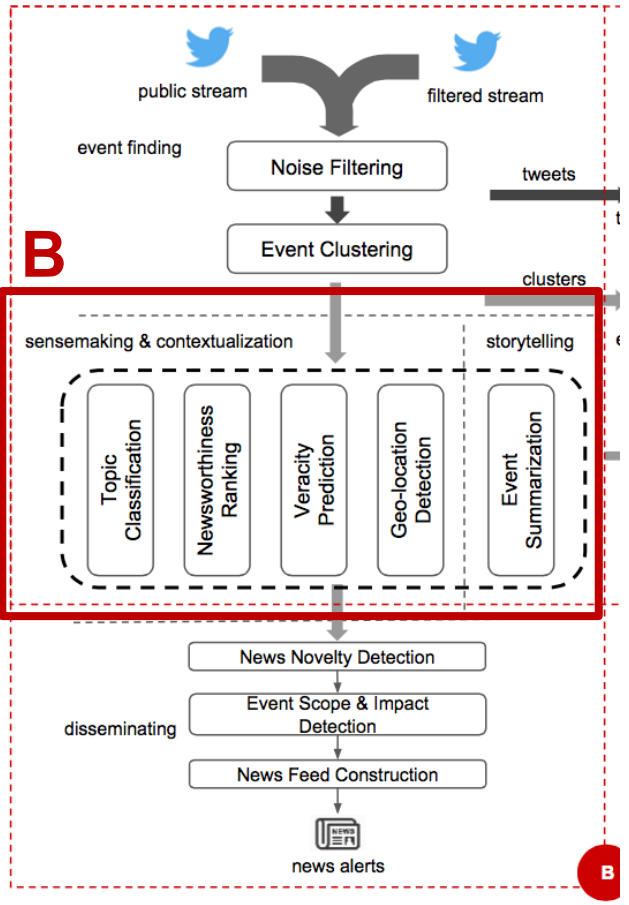
Model of Public Attention

Impact of cluster size ~ public engagement

$P_A(e)=\log_{10}||e|| / \log_{10}S$ where $||e||$ is cluster size, S is 500, $P_A(e) = 1$ if cluster has 500+ tweets

Final newsworthiness score is learned by ordinal regression

3 Categories: **non-news = > partial news => news**



Reuters Tracer – News Detection Algorithm – cont.

<u>B Sense-making & Contextualization</u>

3/ Veracity Prediction

multiple SVM regression models with different features to operate on early and developing stages of an event separately.

Veracity score [-1; 1] to indicate degree of veracity

Early Verification - Identify news source

(1) if an event tweet is a retweet, the original tweet is source

(2) if it cites a URL, the cited webpage is the source

(3) the algorithm issues a set of queries to the

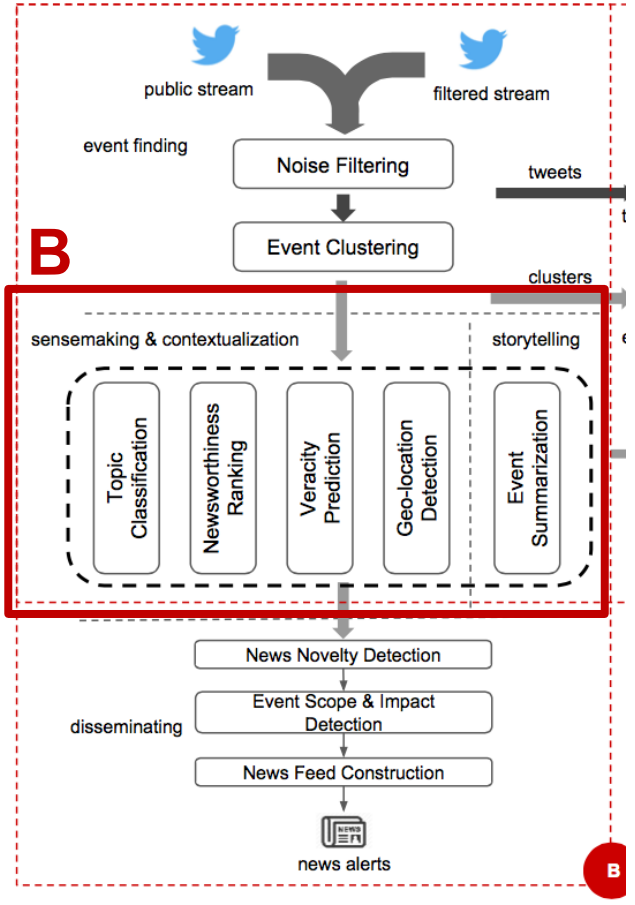
- Twitter search to find the earliest tweet related to the event
- credibility and identity of the source

Developing Verification

event gains momentum, Tracer cluster collects more tweets.

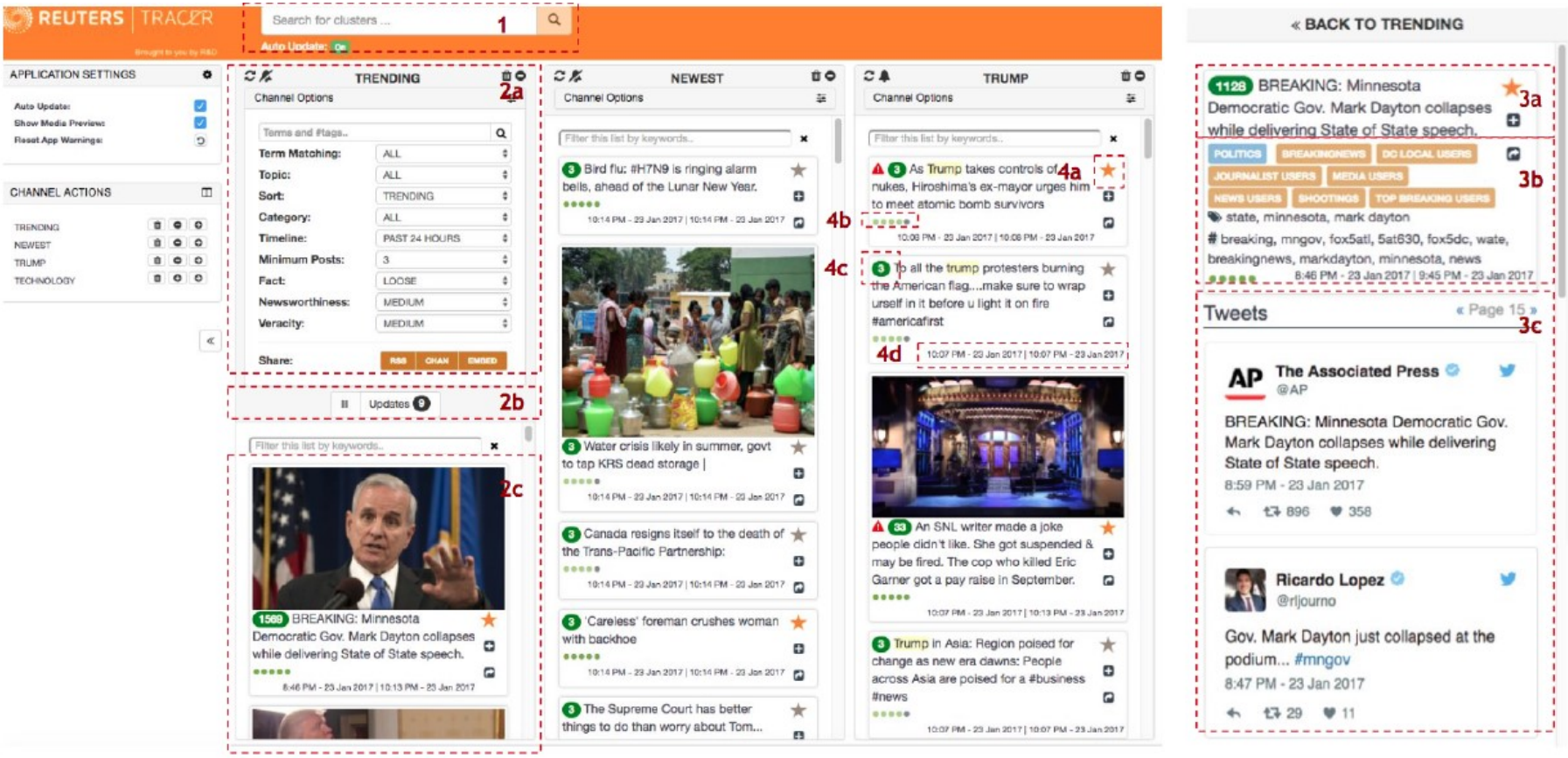
Stance assessed - negation (e.g. “this is a hoax"), question (e.g. “is this real?"), support (e.g. “just confirmed") and neutrality (mostly retweets)

Veracity prediction conceptualize as a “debate" between two sides. Whichever side is more credible wins the “debate”



X. Liu, A. Nourbakhsh, Q. Liy, S. Shahz, R. Martin, J. Duprey, 2017: Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data

Automated True News Generation - Reuters Tracer



Tracer's news exploration UI. (1) Global search; (2) News channel (2c) with editable channel options (2a) and live updates (2b); (3) News cluster with its summary (3a) and metadata (3b) as well as its associated tweets (3c); (4) Cluster metadata including newsworthy indicator (4a), veracity indicator (4b), cluster size (4c), and created & updated times (4d).

X. Liu, A. Nourbakhsh, Q. Liy, S. Shahz, R. Martin, J. Duprey, 2017: Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data

Automated True News Generation - Reuters Tracer **Veracity Prediction Verification**

Sample: 300 news - 100 from each newsworthiness category (non-news, partial news, news)
Verified manually by journalists: 4 categories 1-True 2-likely true 3-False 4-likely false
Binary classification: True 1: 1+2 False 0: 3+4

Dataset	True Ratio	Metric	Fair		Strict		Loose	
			3	30	3	30	3	30
Pred. News	99%	Prec.	0.98	0.99	0.99	0.99	0.98	0.99
Pred. Part. News	91%	Prec.	0.96	0.97	0.97	0.98	0.93	0.95

	False Ratio	Metric	Fair		Strict		Loose	
			3	30	3	30	3	30
Pred. Rumors	62%	Prec.	0.63	0.62	0.63	0.63	0.56	0.59
Fake News	100%	Rec.	0.61	0.64	0.54	0.57	0.73	0.76

Precision or recall of veracity prediction
Developing stages - early: cluster is just 3 tweets vs. developing - 30 tweets
Judgement: Fair - uses the 0 score threshold to separate truth/rumors, Strict buffers truth from rumors by a margin (i.e. rumors should fall in the “red” and truth in the “green” region on the UI). Loose judgement includes the yellow indicator in addition

Tracer can verify true stories reliably and debunk false information with decent accuracy on a routine basis. However, when fake news surges such as in political elections, our system can only flag about 65 - 75% rumor clusters. Verified Twitter users can be fooled and help spread fake news.

Automated True News Generation - Reuters Tracer Event Detection Verification

Statistics of tweets processed and events detected by Tracer compared with Reuters journalists, week of 8/11-8/17, 2017

	Tweets			Clusters		Reuters	
	All	Non-Noise	Clustered	All	News.	Alerts	News
Daily	12M	2.6M	624,586	16,261	6,695	3,360	255
Hourly	512,798	107,233	26,024	678	279	140	11

One week data - Tracer processed 12+ million tweets each day, of which 78% were filtered as noise. In the subsequent clustering stage, only 5% of tweets are finally preserved to produce 16,000+ daily event clusters on average yielding 6,600+ events that are potentially newsworthy.
In contrast, Reuters deploys 2,500+ journalists across 200+ world-wide locations.
They bring back 3000+ news alerts to the internal event notification system, resulting in 250+ events on average written as news stories and broadcast to the public daily.
Even though Tracer uses only 2% Twitter data, it can detect significantly more events than news professionals.

Additional news coverage study by Tracer: set of 2,536 news headlines from Reuters, AP and CNN in a week from 05/08/2016 selected and compared to events detected by Tracer. The results indicate Tracer can cover about 70% news stories with 2% free Twitter data.
Cover rate can increase to 95% if 10% of Twitter data is used instead

Automated True News Generation - Reuters Tracer
Unexpected Events October 2017

Event	Mass shooting at Mandalay Bay, Las Vegas Oct 2, 2017	Murder of famed Maltese journalism Oct 16, 2017	Terror attack in NYC Oct 31. 2017
Earliest Tracer cluster	21 Just walked out of Route 91 and heard lots of gunshots. Mandalay is on lockdown. Hope everyone is safe #route91harvest LAW/CRIME lockdown, mandalay, route # route91harvest 1:22 AM - 2 Oct 2017 | 4:02 AM - 2 Oct 2017	9 Daphne Caruana Galizia believed killed in Bidnija car blast CRISIS killed, bomb, panama, malta # panamapapers, malta, daphnecaruanagalizia, pressfreedom, infowars, news, wikileaks, anonymous, nothingtoseehere, bomb 10:25 AM - 16 Oct 2017 | 10:45 AM - 17 Oct 2017	6 Something going on in downtown #NYC with gunshots and lots of sirens LAW/CRIME gunshots # nyc 3:22 PM - 31 Oct 2017 | 3:29 PM - 31 Oct 2017
Earliest disaster feed alert	02-Oct-2017 1:38:02 AM EDT - BREAKING: MULTIPLE REPORTS OF ACTIVE SHOOTER AT MANDALAY BAY IN LAS VEGAS; NUMEROUS INJURED. I Confidence: 89%	16-Oct-2017 10:25:44 AM EDT - DAPHNE CARUANA GALIZIA BELIEVED KILLED IN BIDNIJA CAR BLAST I Confidence: 95%	31-Oct-2017 3:25:05 PM EDT - #BREAKING: AUTHORITIES ARE RESPONDING TO A REPORT OF A SHOOTING IN TRIBECA... MORE SOON. I Confidence: 95%
Earliest Reuters alert	02-Oct-2017 01:49:32 - LAS VEGAS POLICE SAY INVESTIGATING REPORTS OF ACTIVE SHOOTER NEAR MANDALAY BAY CASINO ON FAMOUS STRIP	16-Oct-2017 11:19:33 - INVESTIGATIVE MALTESE JOURNALIST KILLED IN CAR BOMB	31-Oct-2017 15:31:39 - NEW YORK CITY POLICE INVESTIGATING REPORTS OF SHOOTING IN LOWER MANHATTAN - POLICE SPOKESMAN

Timeliness vs. Veracity : earliest cluster published by Tracer => earliest alert published to the disaster feed => earliest news alert by Reuters

Tracer is often able to detect breaking stories by identifying early witness accounts. Disaster feed only report stories with a high Tracer cluster veracity score (indicated by four green dots)

X. Liu, A. Nourbakhsh, Q. Liy, S. Shahz, R. Martin, J. Duprey, 2017: Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data

Related Topics

Truth Discovery

Truth discovery is the problem of detecting true facts from multiple conflicting sources

Clickbait Detection

Clickbait is a term commonly used to describe eye-catching and teaser headlines in online media. Clickbait headlines create a so-called \curiosity gap“ inconsistency between headlines and news contents

Spammer and Bot Detection

Spammer detection on social media, which aims to capture malicious users that coordinate among themselves to launch various attacks

The major challenge brought by social bots is that they can give a false impression that information is highly popular and endorsed by many people, which enables the echo chamber effect for the propagation of fake news.

This is the END

…Thank you!