Exposing Paid Opinion Manipulation Trolls

Preslav Nakov

Qatar Computing Research Institute, HBKU

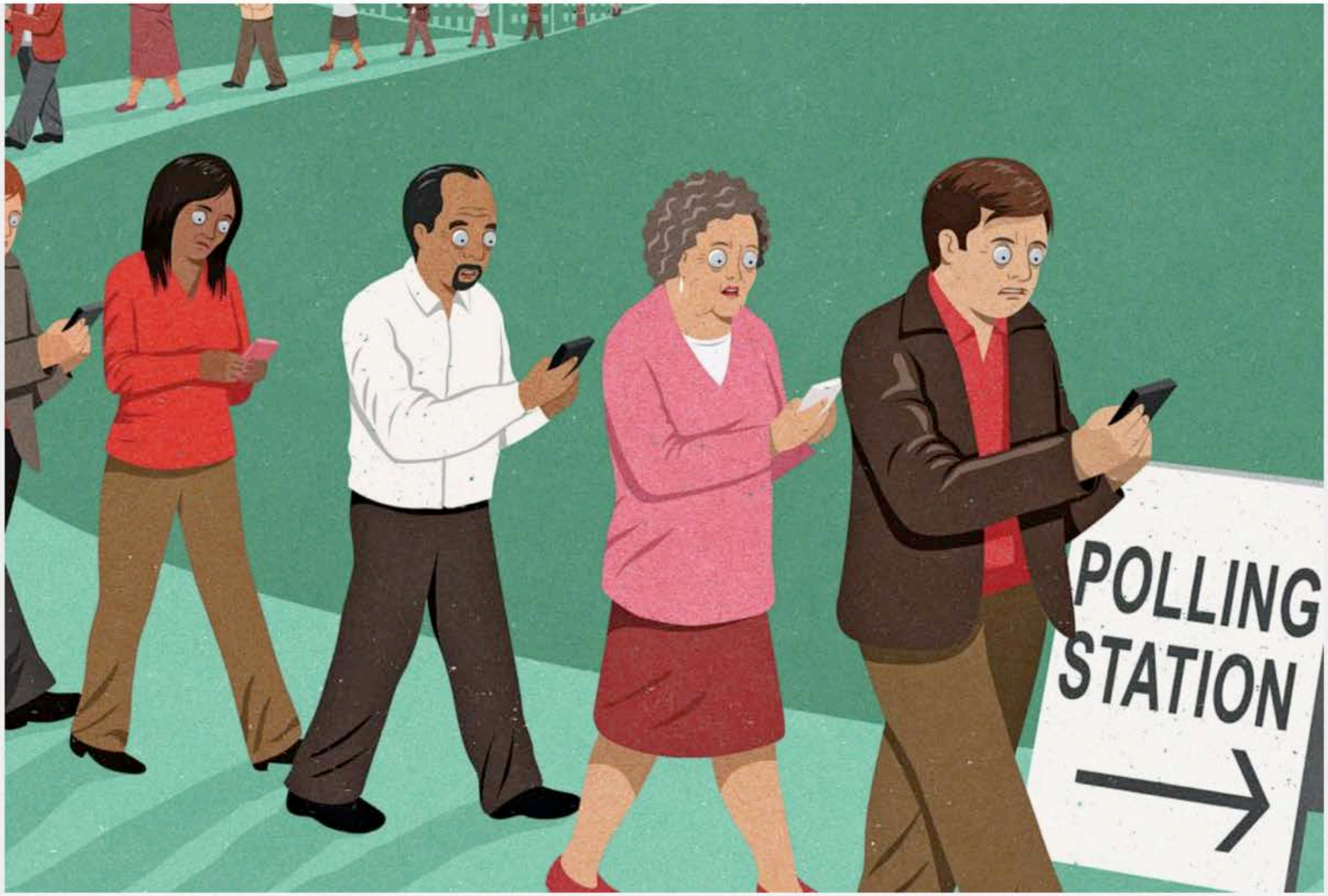
Keynote talk at PAN@CLEF 2019
September 10, 2019
Lugano, Switzerland

“Fake News”: A Weapon of Mass Deception

NEWSWEEK MAGAZINE

How Big Data Mines Personal Info to Craft Fake News and Manipulate Voters

BY NINA BURLEIGH ON 6/8/17 AT 1:01 PM



Fake news handed Brexiteers the referendum – and now they have no idea what they're doing

'Would we have won without immigration? No. Would we have won without...the NHS? All our research and the close result strongly suggests no. Would we have won by spending our time talking about trade and the single market? No way'

Andrew Grice | @IndyPolitics
Wednesday 18 January 2017 16:45 | 157 comments

Like Click to follow The Independent Voices

0 Notifications
https://www.facebook.c

SPIEGEL ONLINE SPIEGEL+

Sign in

Menu | Politik Meinung Wirtschaft Panorama Sport Kultur Netzwelt Wissenschaft more

INTERNATIONAL

Schlagzeilen | DAX 11.792,85 | Abo

English Site > Germany > Alternative for Germany > Germany: AfD Populists Dominate on Facebook

Facebook Frenzy

How the German Right Wing Dominates Social Media

A comprehensive analysis has revealed the degree to which German right-wing populists from the Alternative for Germany (AfD) party are dominating the social media landscape. They might be getting help from abroad.

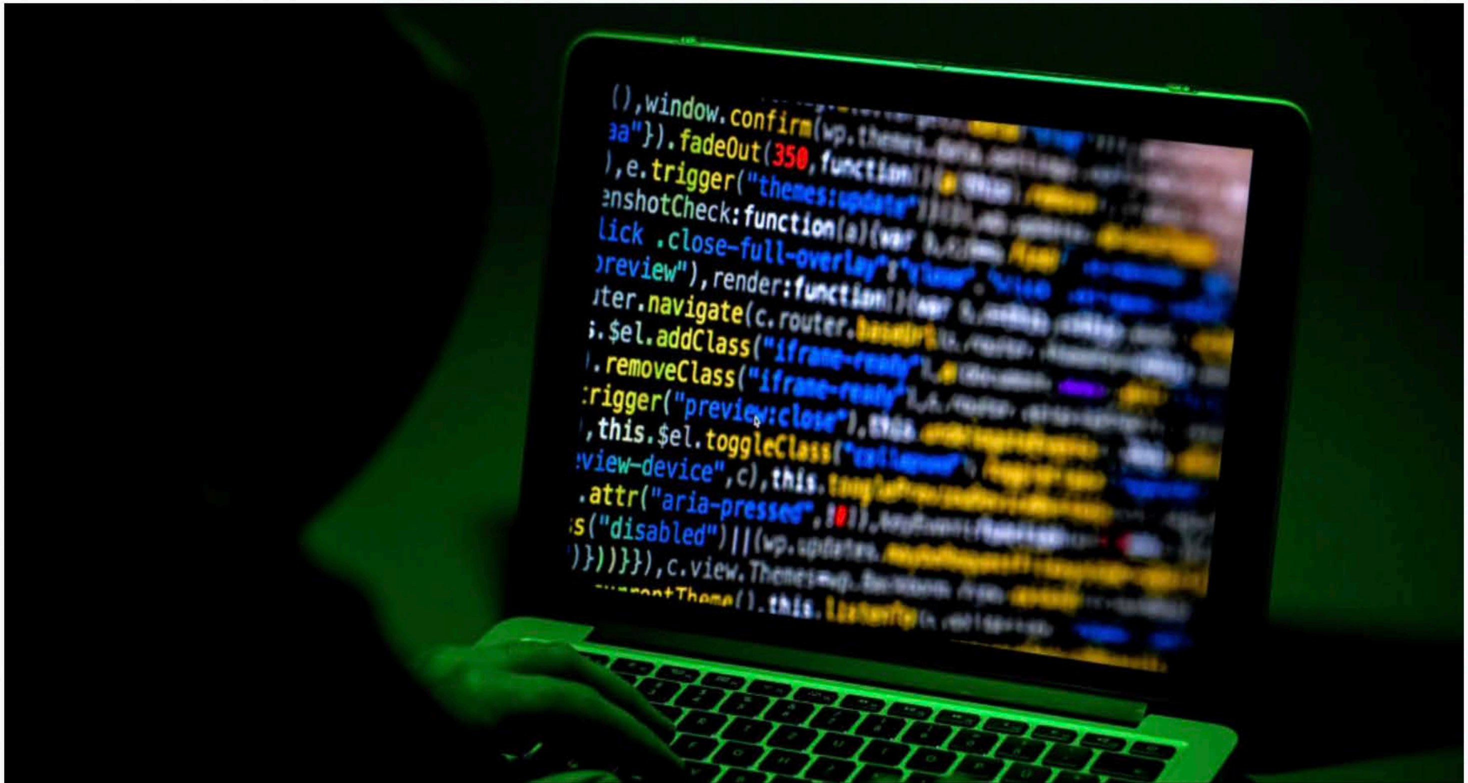
By Jörg Diehl, Roman Lehberger, Ann-Katrin Müller and Philipp Seibt

Monika Skolimowska/ DPA

https://www.spiegel.de/international/germany/germany-afd-populists-dominate-on-facebook-a-1264933.html

European Elections

Polling from across Europe. Updated daily.



Half of European voters may have viewed Russian-backed 'fake news'

Kremlin-backed campaigns are promoting extremist views and amplifying them to sow discord, says cyber firm.

By **MARK SCOTT** | 5/7/19, 12:45 PM CET | Updated 5/8/19, 7:00 PM CET

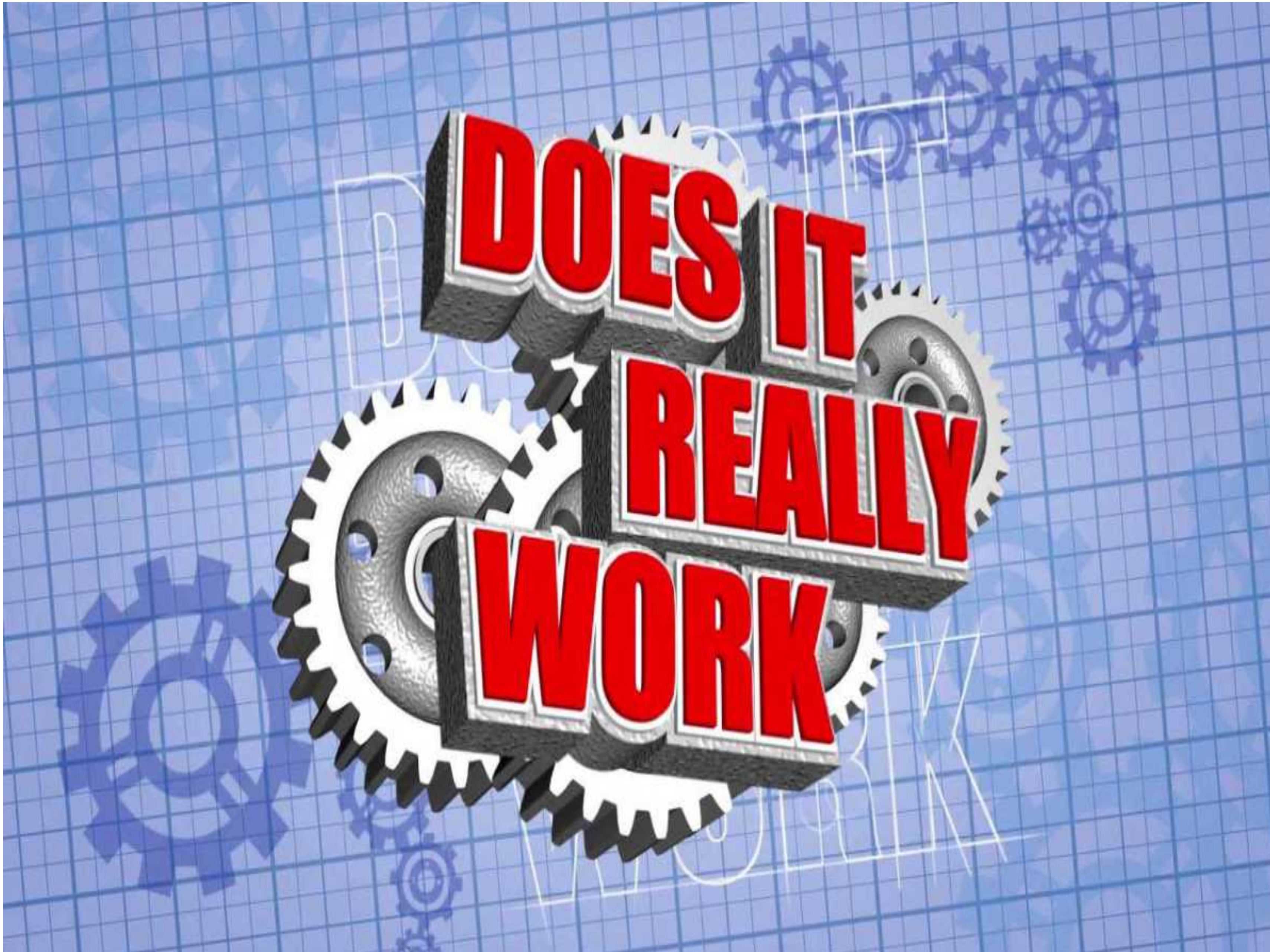
https://www.politico.eu/article/european-parliament-russia-mcafee-safeguard-cyber/

Half of Americans see fake news as bigger threat than terrorism, study finds

Almost 70% of Americans feel fake news has greatly affected their confidence in government institutions, a new study says

▲ Lawmakers have yet to take concrete action against fake news and misinformation. Photograph: Erik McGregor/Pacific/Barcroft

https://www.theguardian.com/us-news/2019/jun/06/fake-news-how-misinformation-became-the-new-front-in-us-political-warfare


DOES IT
REALLY
WORK

FORTUNE

TECH • HILLARY CLINTON

Hillary Clinton Blames the Russians, Facebook, and Fake News for Her Loss

.doubleclick.net...

ISSIE LAPOWSKY BUSINESS 04.04.18 05:43 PM

FACEBOOK EXPOSED 87 MILLION USERS TO CAMBRIDGE ANALYTICA

https://www.wired.com/story/facebook-exposed-87-million-users-to-cambridge-analytica/

The Fix

Donald Trump will be president thanks to 80,000 people in three states

Donald Trump speaks in Washington in 2014. (Jewel Samad/AFP/Getty Images)

https://www.washingtonpost.com/news/the-fix/wp/2016/12/01/donald-trump-will-be-president-thanks-to-80000-people-in-three-states/

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | M

NEWS

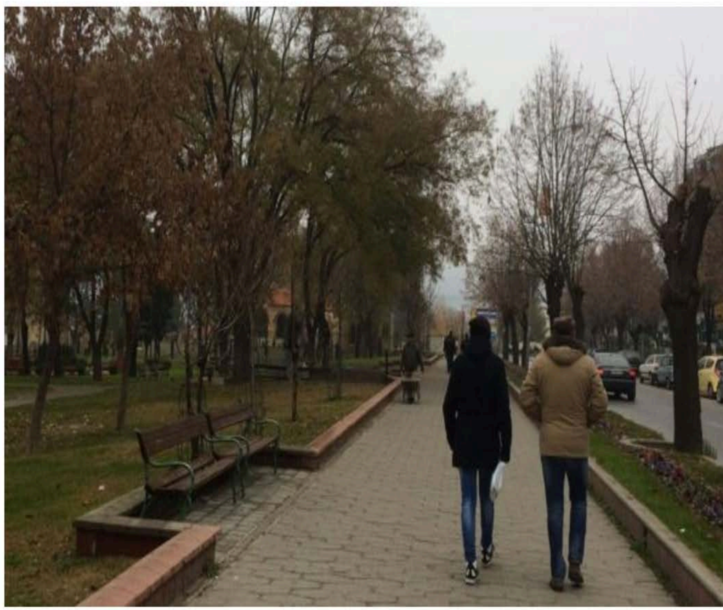
Home | Video | World | UK | Business | Tech | Science | Magazine | Entertainment & Arts

Magazine

The city getting rich from fake news

By Emma Jane Kirby
BBC News

5 December 2016 | Magazine



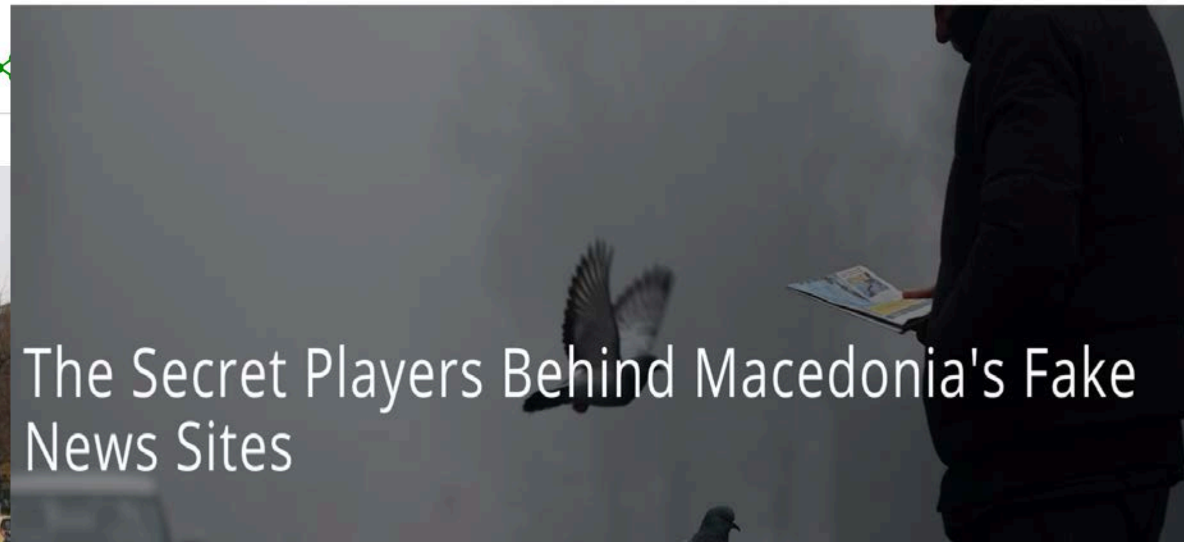
Many of the fake news websites that sprang up during the US election campaign have been traced to a small city in Macedonia, where teenagers are pumping out sensationalist stories to earn cash from advertising.

In today's Magazine

My partner vanished without warning. I had to

OCCRP

SPOOKS AND SPIN: INFORMATION WAR IN THE BALKANS · THE SECRET PLAYERS BEHIND MACEDONIA'S FA... Русский



The Secret Players Behind Macedonia's Fake News Sites

Credit: Nake Batev / Getty Images

by Saska Cetkovska, Aubrey Belford, Craig Silverman, and J. Lester Feder
18 July 2018

Donate

A joint investigation by the Organized Crime and Corruption Reporting Project (OCCRP) and partners has uncovered new information that rewrites the story of the fake news boom in the Macedonian town of Veles.

A week before Election Day in 2016, BuzzFeed News revealed that young men and teens in Veles were running over a hundred websites that pumped out often false viral stories that supported Donald Trump.

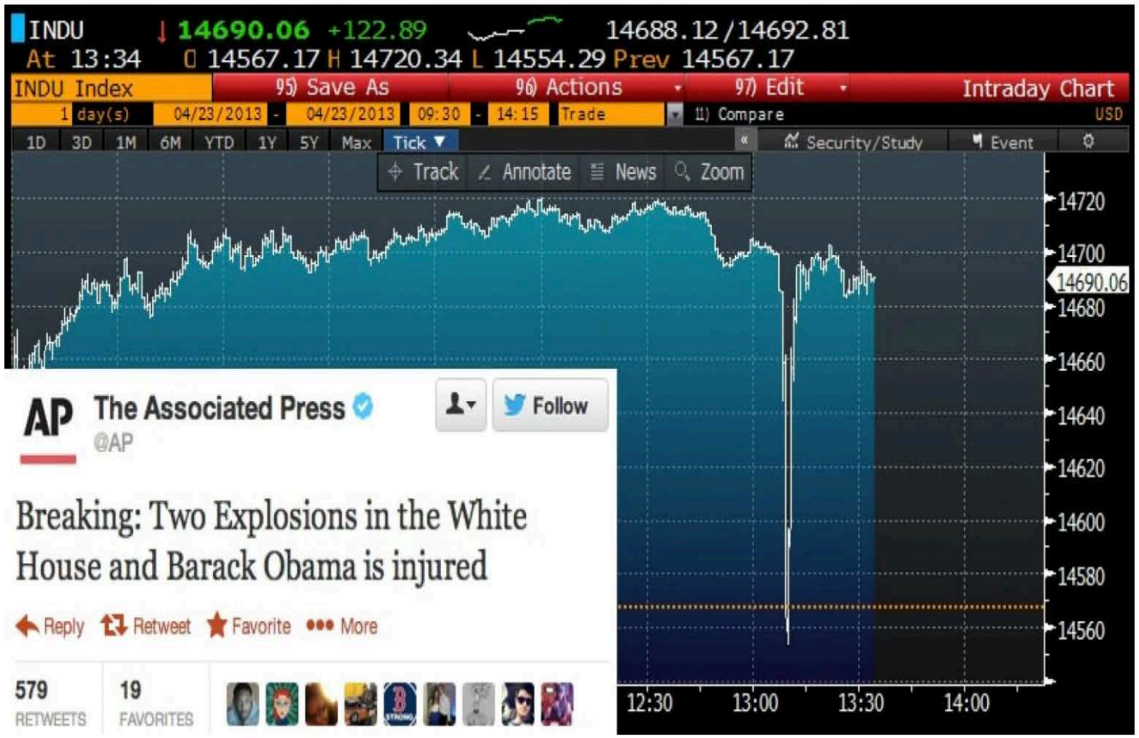
Media outlets from around the world descended upon Veles to tell the story of how the so-called fake news teens — many of whom had a shaky understanding of English — made large sums of money from digital ads shown next to their misleading stories

Above: A man reads a magazine near a newsstand in Skopje, Macedonia, 2017.
Credit: Nake Batev / Getty Images

WorldViews

Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism?

By **Max Fisher**
April 23, 2013



This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, inset at left.

At 1:07 p.m. on Tuesday, when the official Twitter account of the Associated Press sent a tweet to its nearly 2 million followers that warned, "Breaking: Two Explosions in the White House and Barack Obama is injured," some of the people who momentarily panicked were apparently on or near the trading floor of the New York Stock Exchange.


Година VII, бр. 88 (2333) • Събота - неделя, 13-14 април 2019 г. 1.00 лв.
Антиваксърът отец Евгений Янакиев:
ЗЪЛ ДУХ СЕ ВСЕЛИ в дъщеря ми след ваксина
стр. 8
800 ЕКСКЛУЗИВНИ ГОСТИ НА ПРЕМИЕРАТА НА „ЛОШО МОМИЧЕ“
стр. 7
Сензационно разследване на сръбския вестник „Курир“:
стр. 4-5

UNICEF blames anti-vaxxers for the 300% spike in global measles outbreaks

Published: Apr 25, 2019 3:49 p.m. ET

'The ground for the global measles outbreaks we are witnessing today was laid years ago,' report says

Measles cases have spiked 300% around the world over the past year.

Measles: Four European nations lose eradication status

https://www.bbc.com/news/health-49507253

6 hours ago | 100

Share

It's a numbers game... if some people are not vaccinated, it can cause a big problem for us all

Measles has returned to four European nations previously seen as free of the illness, according to the World Health Organization (WHO).

The disease is no longer considered eradicated in Albania, the Czech Republic, Greece and the UK.

4:13

+ QUEUE

DOWNLOAD

EMBED

WORLD

Viral WhatsApp Messages Are Triggering Mob Killings In India

July 18, 2018 · 9:12 AM ET

LAUREN FRAYER

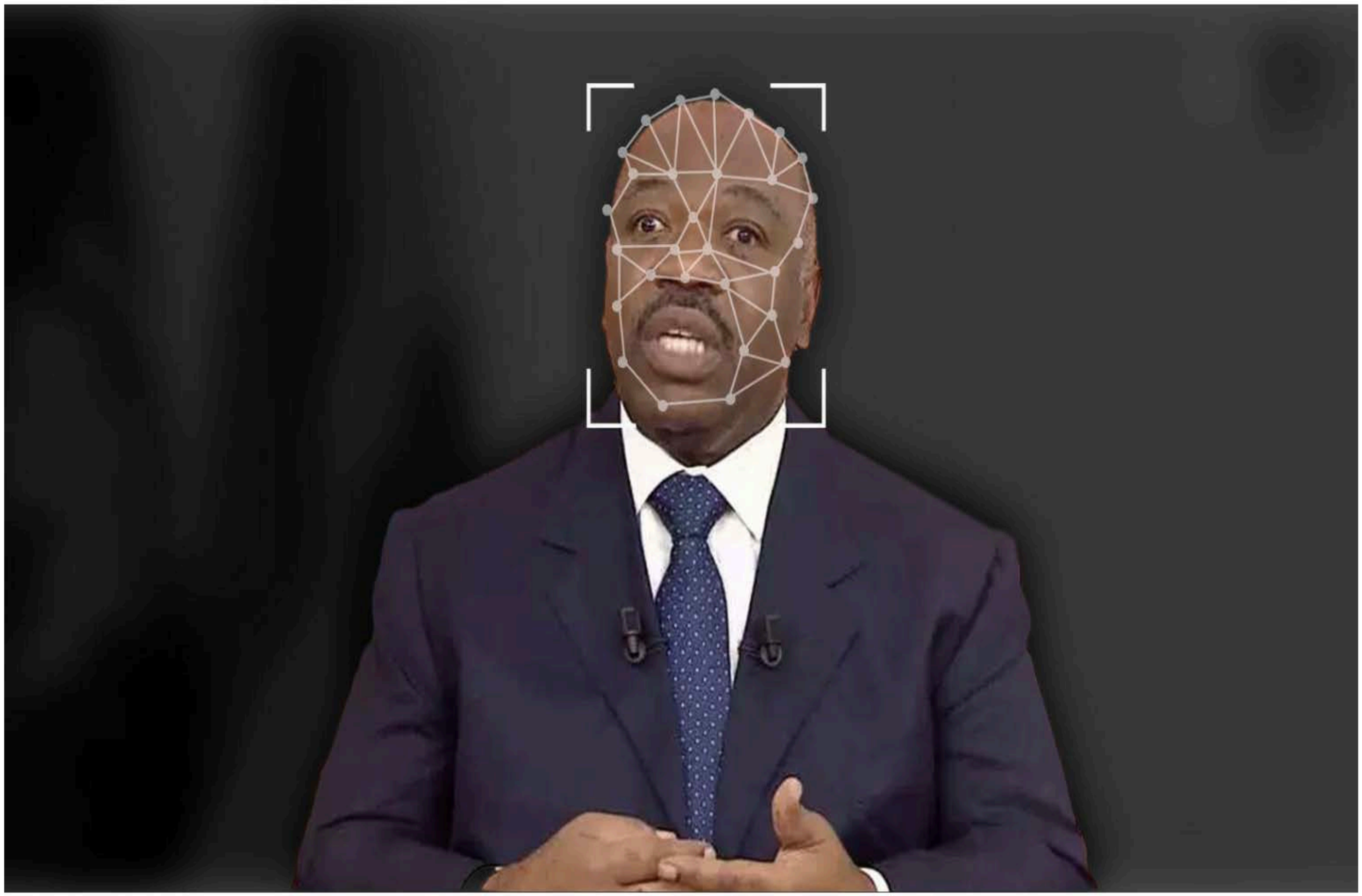
U.N. Fact Finders Say Facebook Played a 'Determining' Role in Violence Against the Rohingya

A girl looks out of her tent in the Balukhali refugee camp in Cox's Bazar, Bangladesh on Jan. 17, 2018. U.N. fact finders have pointed to the role that hate speech on social media has played in fueling anti-Rohingya sentiment and violence

A Military Coup in Gabon Inspired by a Potential Deepfake Video is Our Political Future

Ty Joplin

Published May 8th, 2019 - 09:54 GMT



Ali Bongo (Youtube, Rami Khoury/Al Bawaba)

A video of Gabon's President, Ali Bongo, appeared on Jan 1, 2019, sparking a wave of confused reactions inside the country.

https://www.albawaba.com/news/military-coup-gabon-inspired-potential-deepfake-video-our-political-future-1284760

SECTIONS | WATCHLIST

MarketWatch

Web creator Tim Berners-Lee blasts Facebook, saying it makes his invention easy to ‘weaponize’

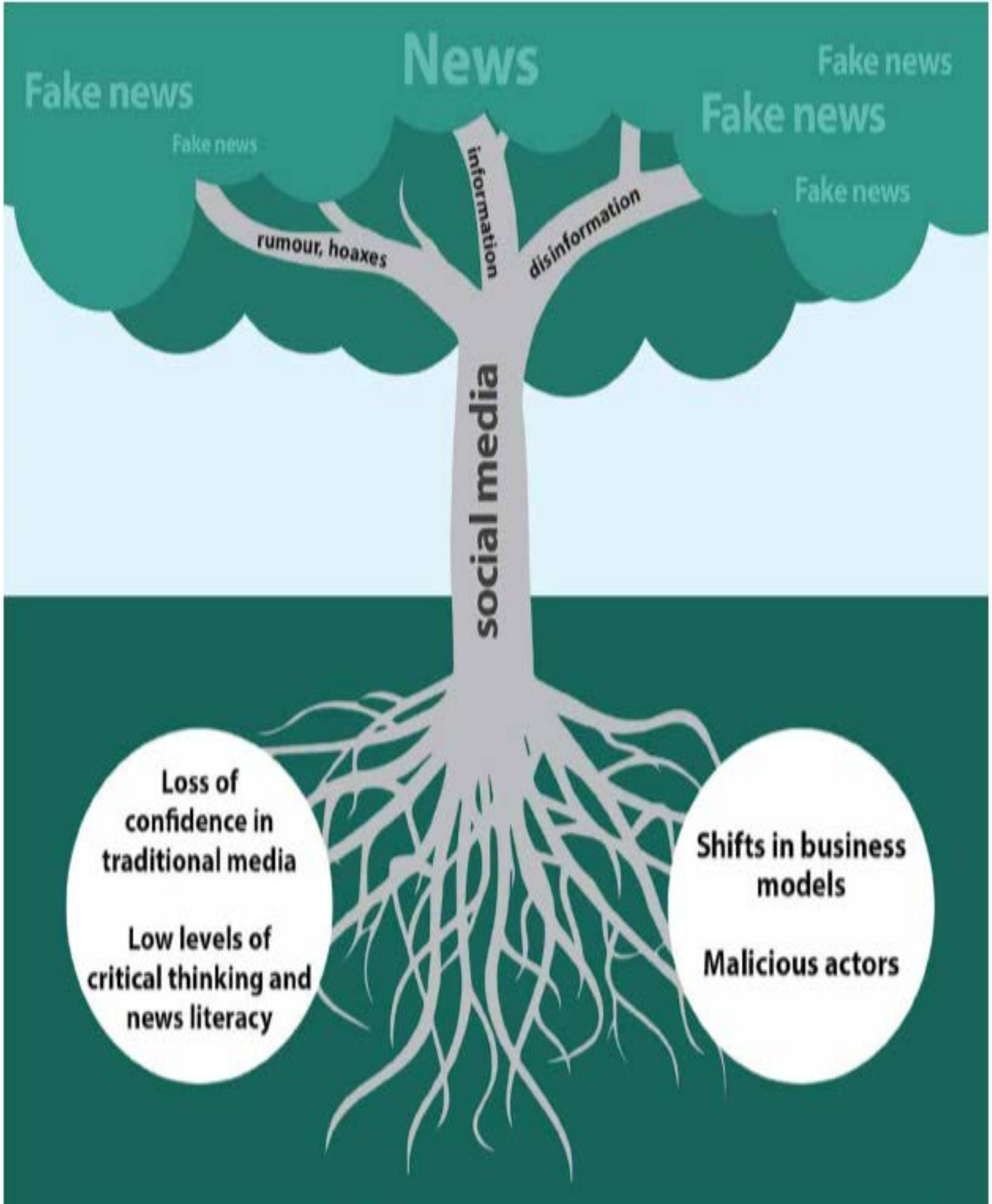
Published: Mar 19, 2018 4:40 a.m. ET

17

Aa

On World Wide Web’s 29th birthday, Tim Berners-Lee criticizes its “gatekeepers” such as Facebook and suggests more regulation

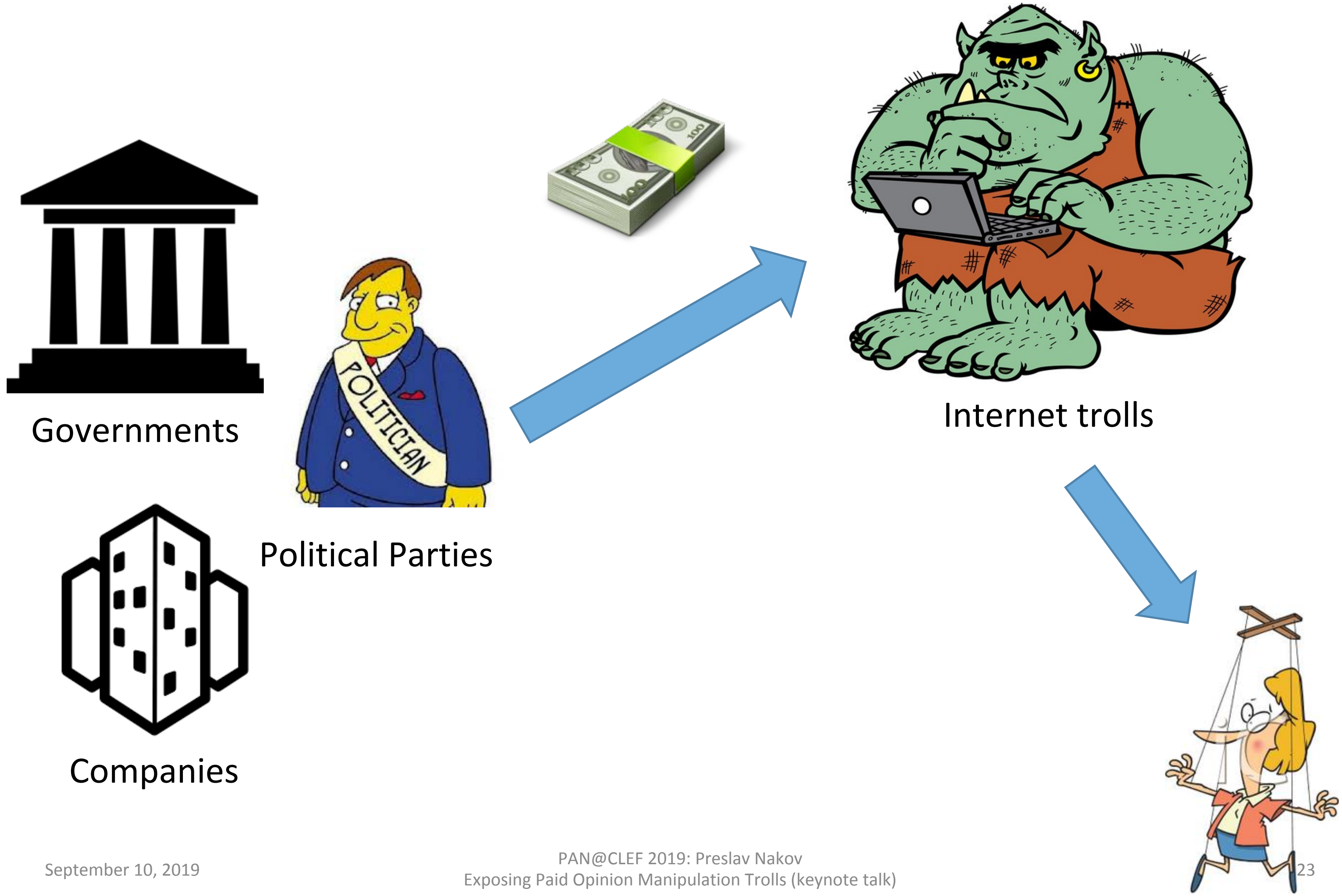
Computer scientist Tim Berners-Lee, the “father of the World Wide Web.”



By UNESCO - World Trends in Freedom of Expression and Media Development Global Report 2017/2018, CC BY-SA 3.0-igo, https://commons.wikimedia.org/w/index.php?curid=70286494

The Role of Trolls

Problem


Governments
Political Parties
Companies
Internet trolls

Trolling in Politics: Astroturfing

- In political science, it is defined as the [**process of seeking electoral victory]** or legislative relief for grievances by [**helping political actors find and mobilize a sympathetic public]**, and is designed to create the image of public consensus where **[there is none]**.

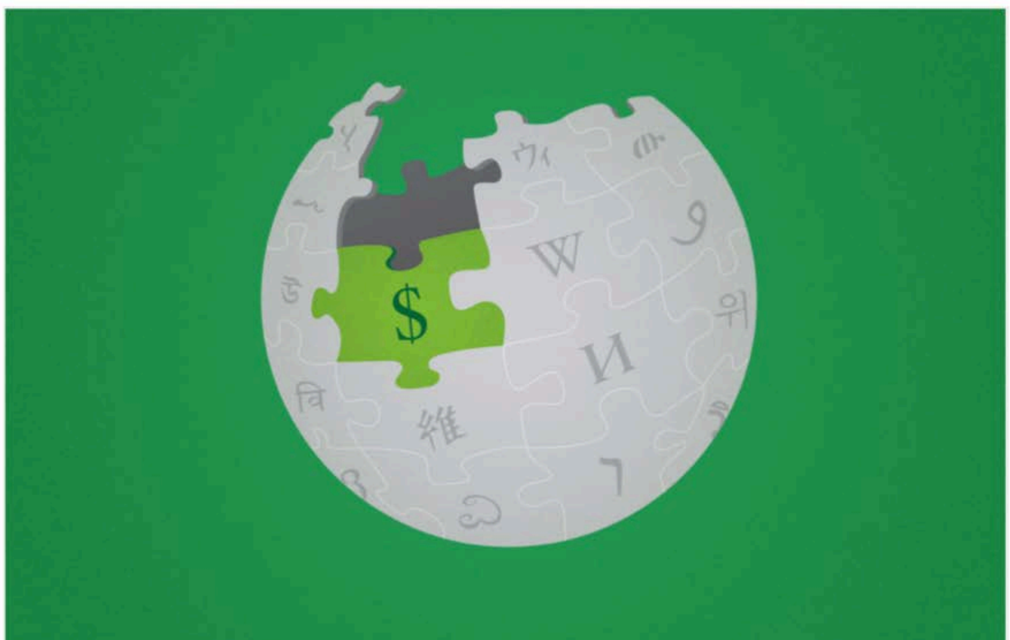
Paid Manipulation Internet Trolls

The Atlantic

BUSINESS

The Covert World of People Trying to Edit Wikipedia—for Pay

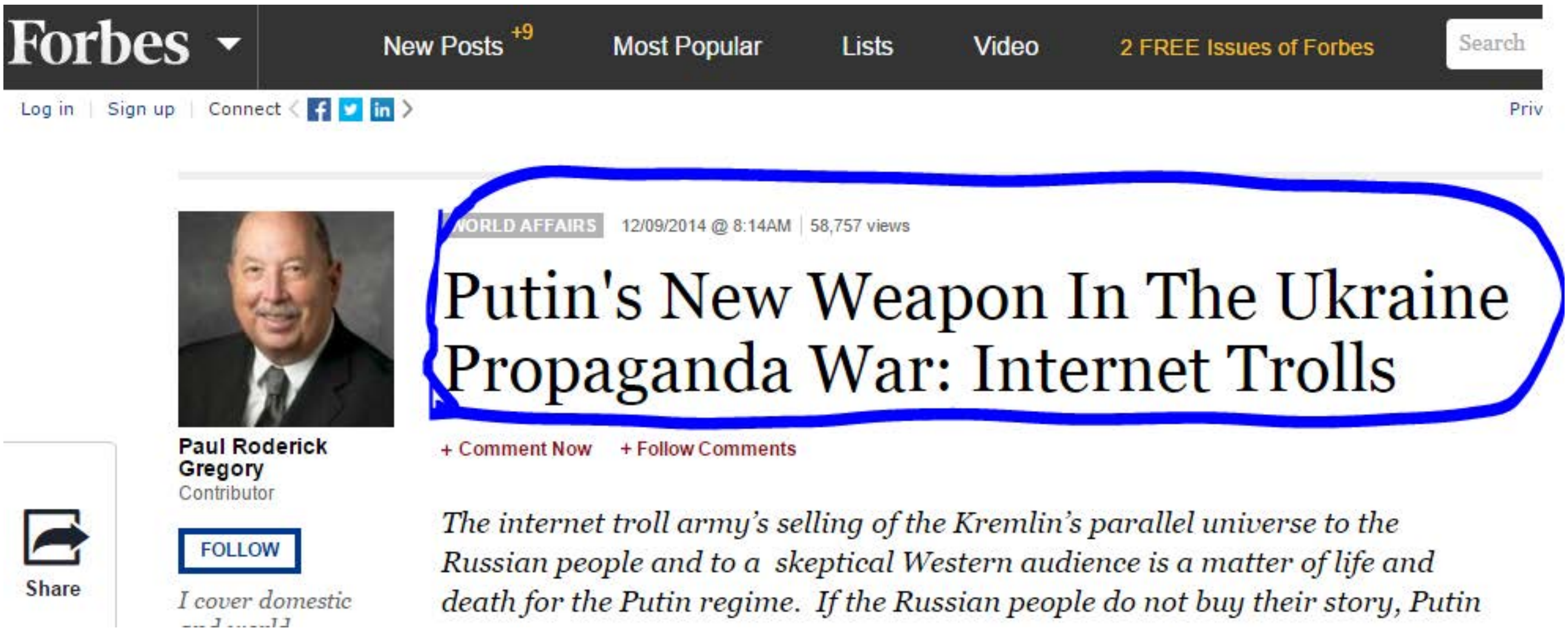
Can the site's dwindling ranks of volunteer editors protect its articles from the influence of money?



Adapted from Wikimedia / Lauren Giordano / The Atlantic

JOE PINSKER | AUG 11, 2015

Propaganda Internet Trolls



Forbes New Posts +9 Most Popular Lists Video 2 FREE Issues of Forbes Search

Log in | Sign up | Connect

Paul Roderick Gregory

Contributor

FOLLOW

I cover domestic

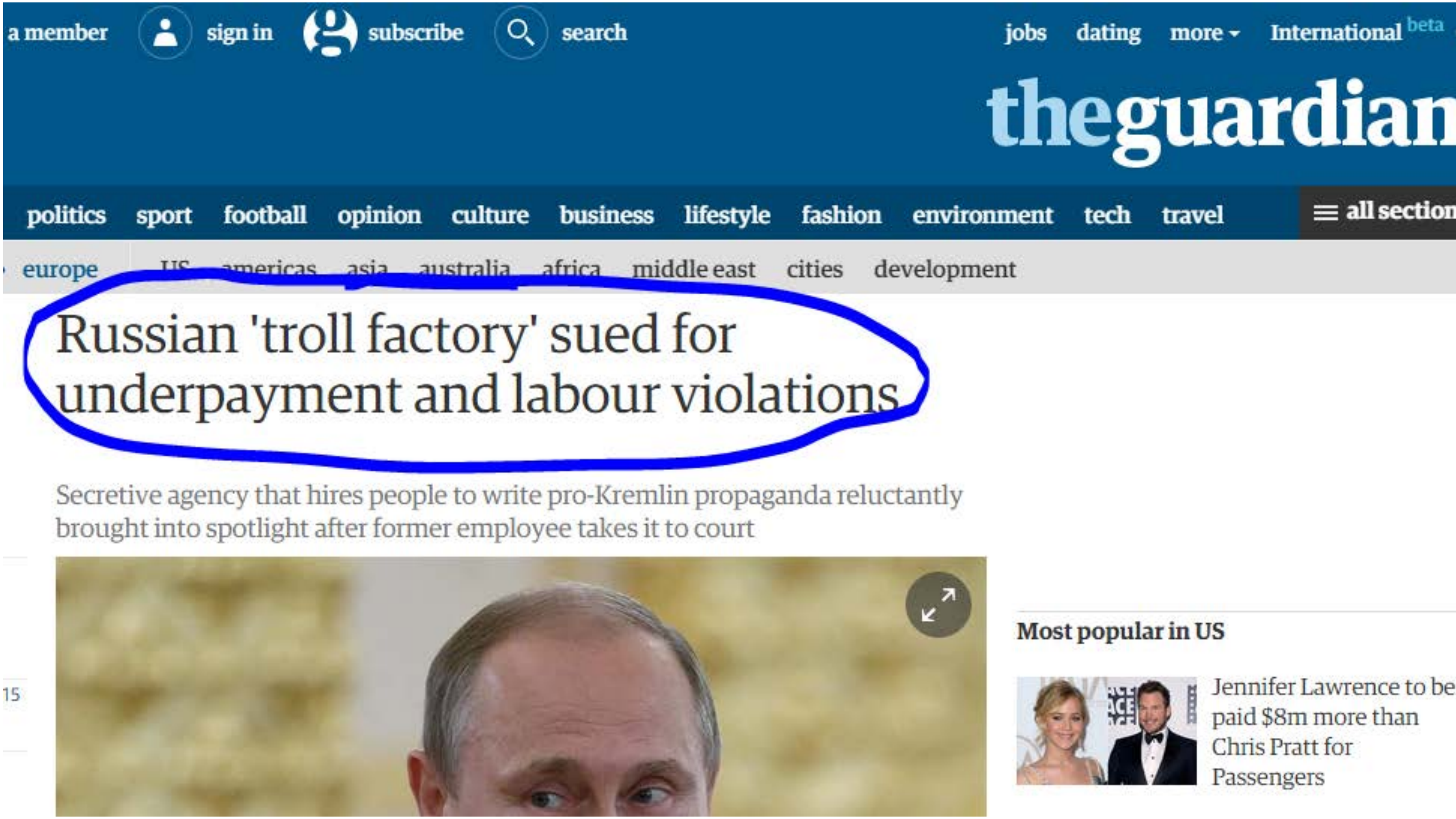
Share

WORLD AFFAIRS 12/09/2014 @ 8:14AM | 58,757 views

Putin's New Weapon In The Ukraine Propaganda War: Internet Trolls

+ Comment Now + Follow Comments

The internet troll army's selling of the Kremlin's parallel universe to the Russian people and to a skeptical Western audience is a matter of life and death for the Putin regime. If the Russian people do not buy their story, Putin


a member sign in subscribe search jobs dating more International beta

theguardian

politics sport football opinion culture business lifestyle fashion environment tech travel all sections

europe US americas asia australia africa middle east cities development

Russian 'troll factory' sued for underpayment and labour violations

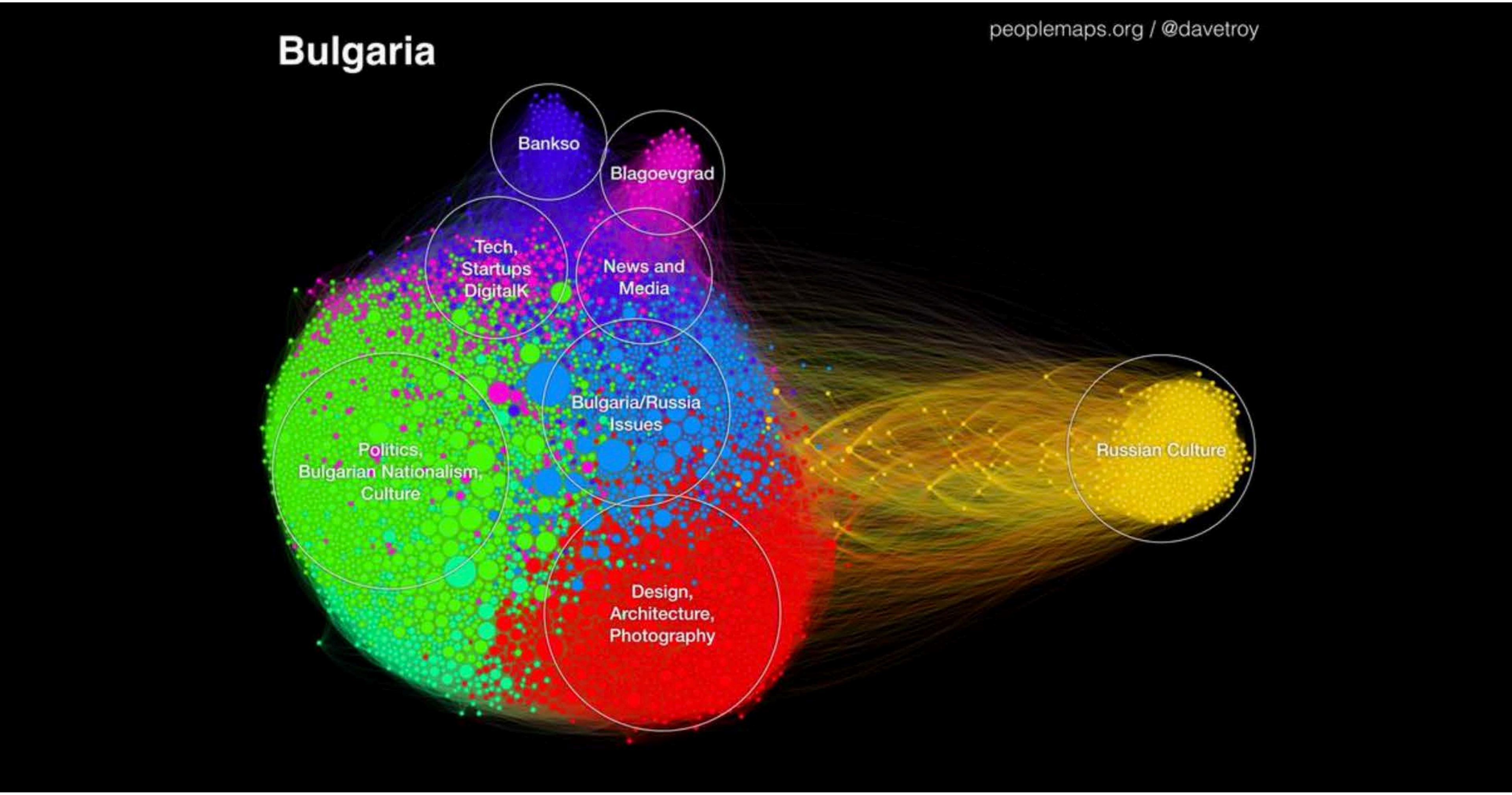
Secretive agency that hires people to write pro-Kremlin propaganda reluctantly brought into spotlight after former employee takes it to court

Most popular in US

Jennifer Lawrence to be paid $8m more than Chris Pratt for Passengers

They are so real...

The Bulgarian Twitter Space



Dave Troy @davetroy · May 21
New visualization: Bulgaria Twitter Users. Presented this morning at #digitalk2015 in Sofia, Bulgaria!

14 13

Can We Stop the Trolls?

Option 0: Launch Your Own Trolls

KyivPost
INDEPENDENCE. COMMUNITY. TRUST

Search

Daily Beast: Baltic elves take on pro-Kremlin trolls in online propaganda war

Baltic | Cyberspace | trolling | Mar. 21, 2016 15:32

Share | Tweet | Print | email

Option 1: Gentlemen's Agreement
(e.g., Bulgarian Political Parties Against Paid Trolls)

НАЧАЛО | БЪЛГАРИЯ

Партиите тържествено си обещаха да не ползват тролове в интернет

Share 13 Share 0 Tweet

18:30 | 20 май 2015 | **5 коментара**

Реформаторският блок, ГЕРБ, БСП, ПФ и ПГ на БДЦ подписаха декларация да не използват умишлено платени коментатори (тролове) в интернет пространството. АБВ и ДПС се възмутиха, че не са ги поканили.

Партиите няма да използват и партийно или фирмено организирани тролове с цел разпространение на заблуди и клевети, насаждане на омраза или манипулативни внушения за политически опоненти и други участници в публичния живот, което да се публикува в интернет, обещаха помежду си парламентарните формации.

Инициативата за пожелателното споразумение е на депутата от РБ Антони Тренчев, а до парафирането на "документа" се стигна след проведена кръгла маса по темата миналата седмица.

Под декларацията, че няма да ползват тролове не са се подписали от АБВ, ДПС и "Атака".

"Това е първа крачка за разрешаване на проблема и за регулацията му в интернет", посочи Тренчев.

Option 2: Expose Widely the Known Trolls

Todor Yalamov
1 hr ·

Не се чудете като четете как ЕС забранява шкембето, розовия домат, кръщенетата... Точно същите истории ги разпространяваха и у нас, ако помните. Не четете тези медии и хора сред приятелите си, които пишат врели-некипели... или се опитайте да им обясните защо не трябва да вярват на медии, които си измислят - като случая със забраната на кръщенетата, които не умеят да различат не-новините style от нормална медия. Прекъснете разпространението на фалшивите новини - иначе участвате в хибридната война...

A Powerful Russian Weapon: The Spread of False Stories

Using both conventional media and covert channels, the Kremlin relies on disinformation to create doubt, fear and discord in Europe and the United States.

NYTIMES.COM | BY NEIL MACFARQUHAR

Option 3: Block the Trolls

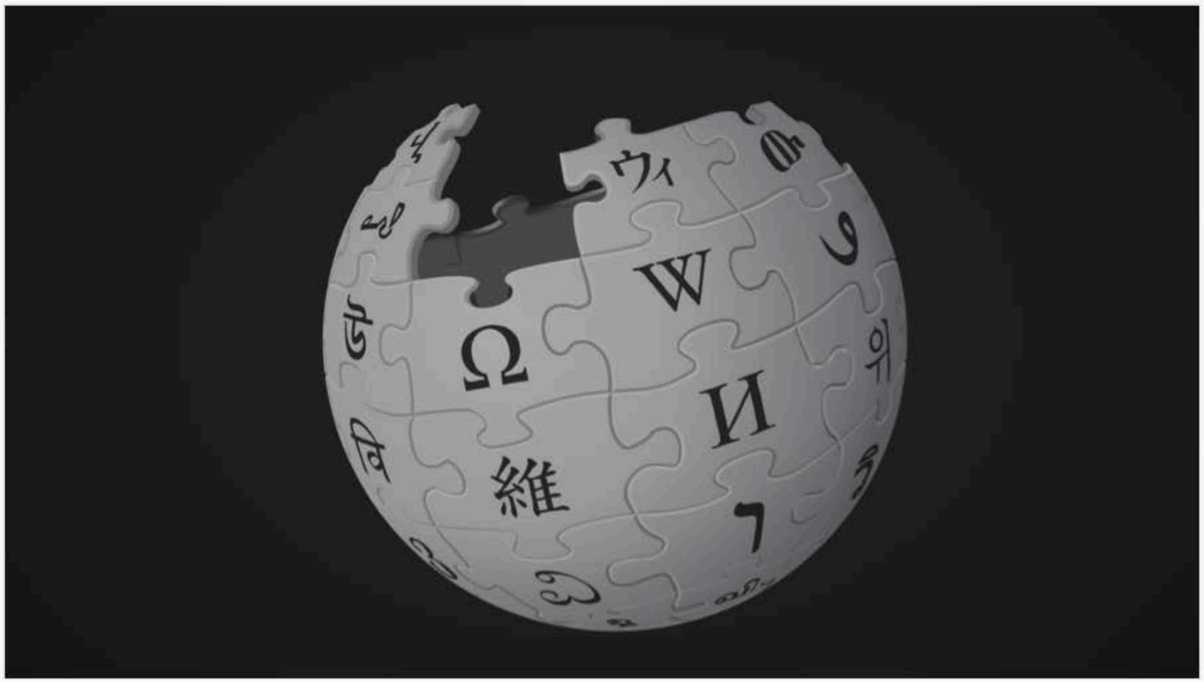
TC NEWSLETTERS

Get TechCrunch News Delivered To Your Inbox **Sign Up Here**

Wikipedia Bans Hundreds Of "Black Hat" Paid Editors Who Created Promotional Pages On Its Site

Posted 13 hours ago by Sarah Perez (@sarahintampa)

603 SHARES



Sometimes Wikipedia's reliance on volunteers to craft its online content comes back to bite

TC News Video Events CrunchBase

:ors on the English

g in "undisclosed

paid advocacy." In other words, they were posting promotional articles to the user-editable online encyclopedia, without revealing that they were paid to do so.

Option 4: Sue the Trolls

me a supporter | subscribe | search | dating | more | International

theguardian

football opinion culture business lifestyle fashion environment tech travel | browse all sections

Amazon sues 1,000 'fake reviewers'

Online retailer files lawsuit in US against people whose names it says it does not know, claiming they offer reviews for sale

What we Need is to Automatically Find & Filter Trolls and Their Comments

theguardian

Comments on articles are valuable. So how to weed out the trolls?

Joseph Reagle

It's the question facing every website that allows comments: how to curb abuse without neutering the conversation

Sunday 17 April 2016 14.02 BST

Last month, the technology news site Engadget announced it was "shutting down our comments ... see you next week". The deployment of a new comment system hadn't worked as hoped.

Its community manager noted that a good comments section has "users who feel a sense of duty and kinship, who act as a community"; an exceptional one "informs its readers, corrects authors and provides worthwhile insights in a polite and constructive manner".

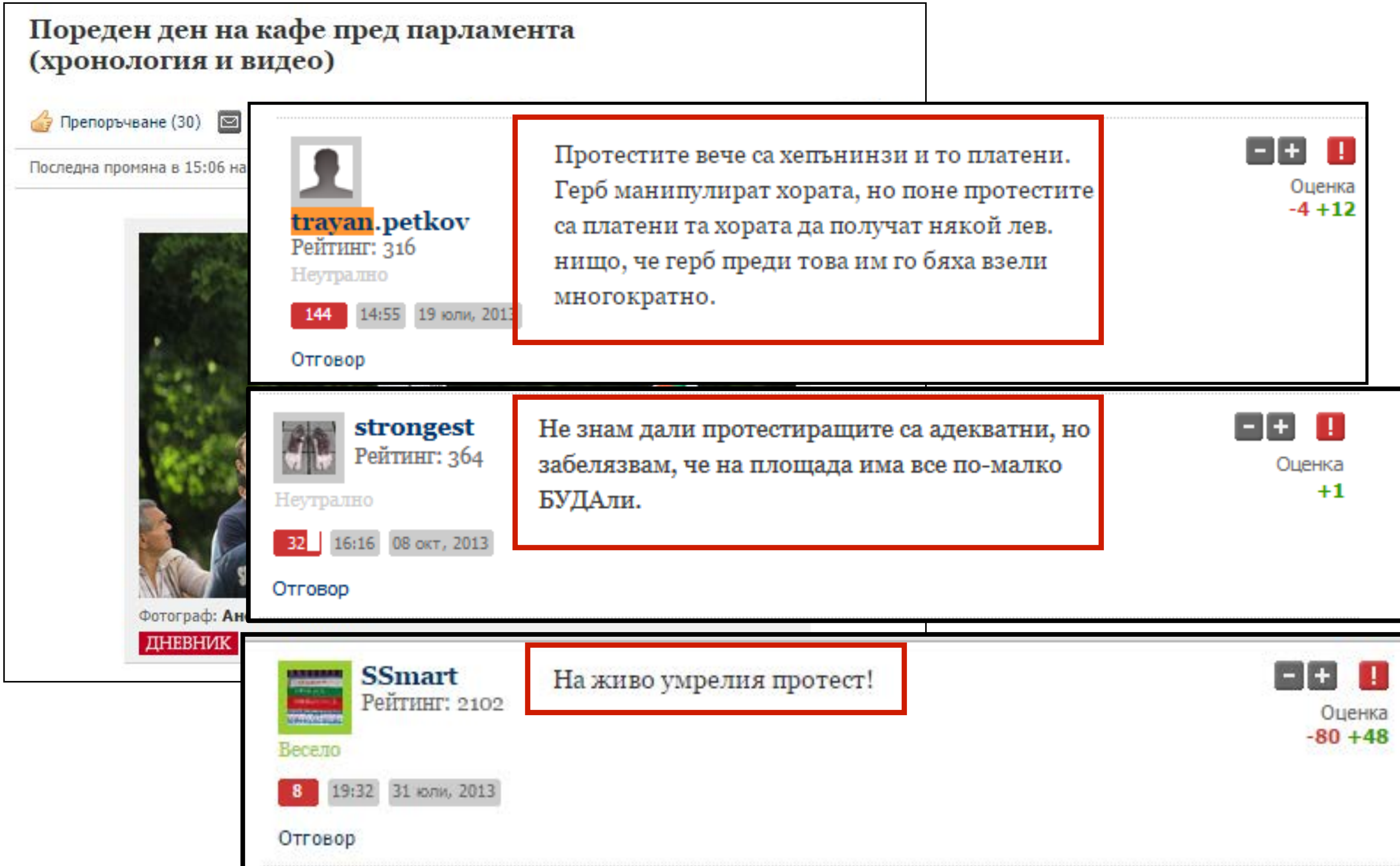
Can we Find Trolls Automatically?

Probably, if we have training data…

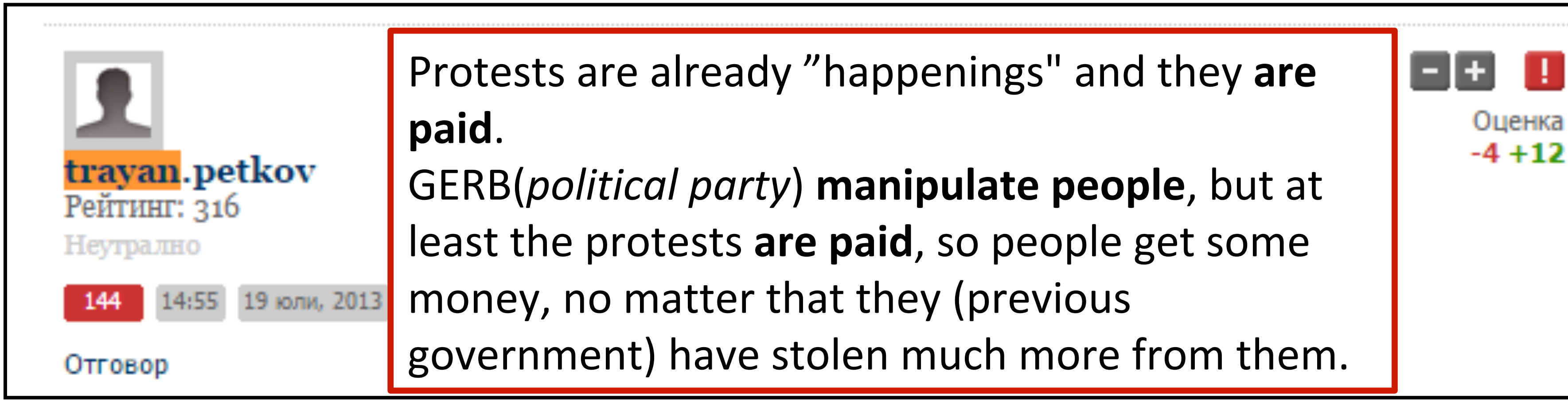
Political Trolls in Bulgaria: Historical Context

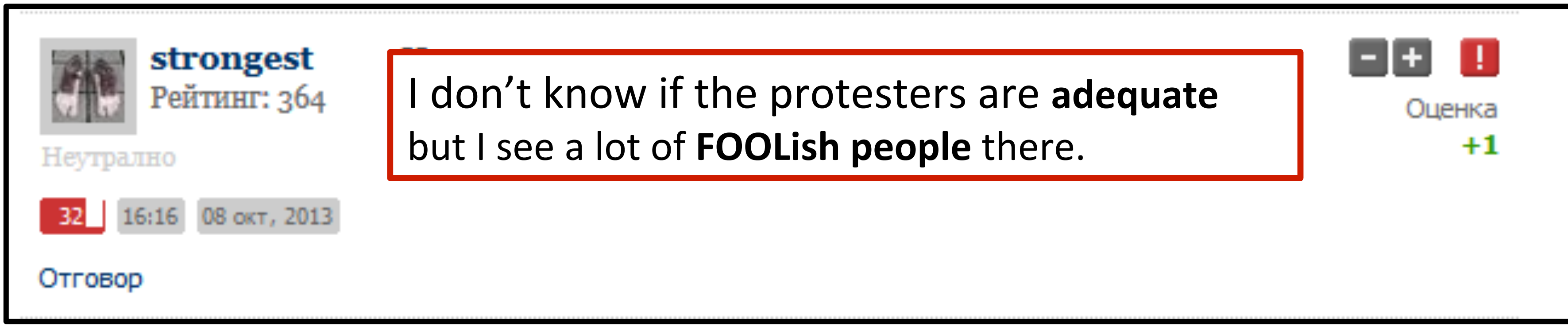
Protests against the government in Bulgaria (2013-2014)

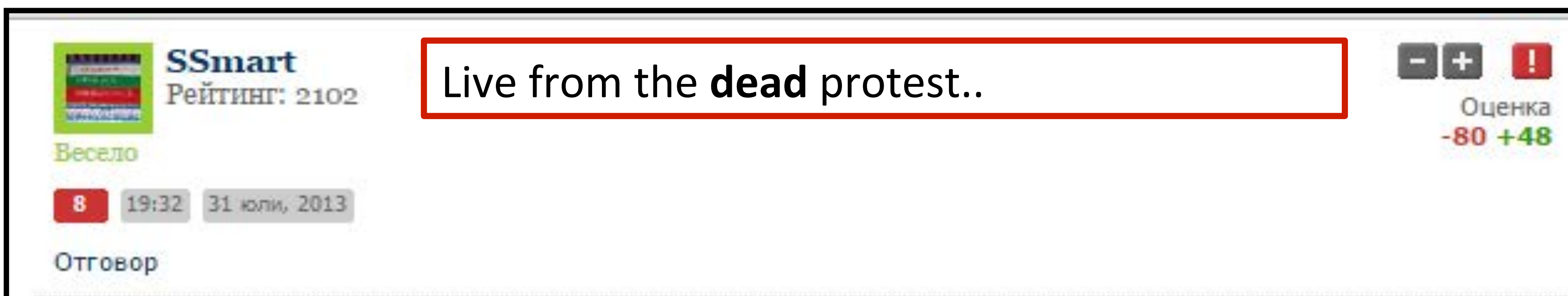
In the Forums…



In the Forums…

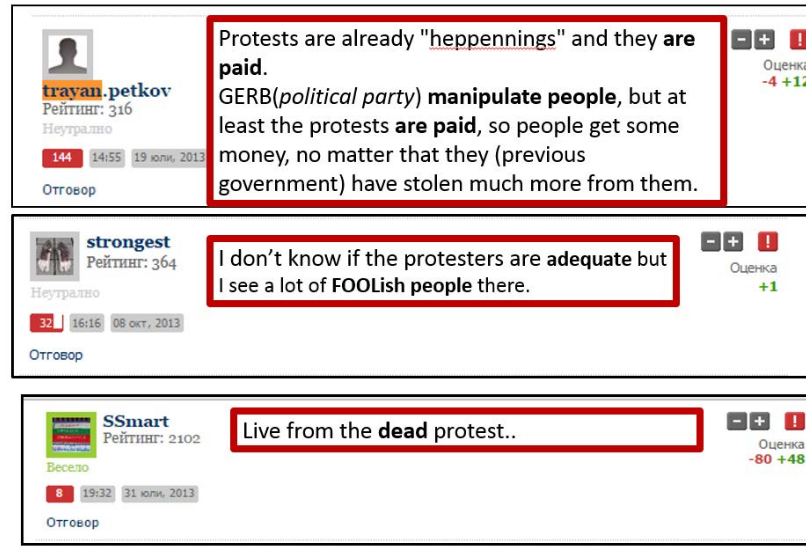






50,000+ People on the Streets

vs.



Notable presence of government supporters in the Web forums

Accusations of Using Paid Trolls

БИВОЛЪ *Съ рогата напредъ!*

Илияна Йотова плащала за тролове по 1500 лв. на месец

Автор: Биволъ   Дата на публикуване май 23, 2014   В: Досиетата Бъ, Потеря, Тролгейт

Accusations of Using Paid Trolls

BULGARIAN SOCIALIST MEP PAID 1 500 LEVS (€ 750) PER MONTH FOR INTERNET TROLLS

BULGARIAN SOCIALIST MEP PAID 1 500 LEVS (€ 750) PER MONTH FOR INTERNET TROLLS

Posted By: Биволъ   Posted date: May 23, 2014   In: Investigations, The B-files, Trolls gate

Print

Capture_2014-02-15_a_22.52.48

The July 24, 2013, offer of the company "Leadway Media Solutions" to Bulgarian MEP **Iliyana Yotova**, from the group of the **Party of European Socialists** (PES), for online reputation management proposes the use of internet trolls against 1

https://bivol.bg/en/trolls-yotova-1500-english.html

Extract from a "Reputation Management" Contract

„Публикуване на 250 коментара на месечна база в Интернет пространството от виртуални потребители с разнообразни, типизирани и еволюиращи профили от различни (неповтарящи се) IP адреси с цел информиране, промотиране, балансиране или противодействие. Интензивността на осъществяваното онлайн присъствие ще е адекватно разпределена и ще съответства на политическата ситуация в страната."

„ Monthly posting online of 250 comments by ***virtual users*** *with varied,* ***typical and evolving profiles*** *from different (non-recurring) IP addresses to* ***inform, promote, balance or counteract****. The intensity of the provided online presence will be adequately distributed and will correspond to the political situation in the country."*

Leaked Reports

LEADWAY media solutions

10/11/2013

Антон Кутев

ИНТЕРНЕТ АКТИВНОСТИ

ОТЧЕТ

18.10.-31.10.2013

INTERNET ACTIVITIES REPORT

ОБЕМ НА СЕДМИЧНАТА АКТИВНОСТ

I. През отчетния период са публикувани **1200 коментара** (1058 във форуми и 142 във Facebook), позиционирани в **38 онлайн медии.** Коментарите, които се генерират на база медийните изяви на представители на управляващата партия не просто подкрепят политиката на правителството, но се стремят да разясняват конкретните секторни политики, като се обяснява очаквания ефект от предприетите управленски мерки. Активно разпространяваме в онлайн пространството тезите на партията-мандатоносител по всички актуални теми, провокиращи разнопосочни реакции в обществото. Целта е до аудиторията системно да достигат аргументи в подкрепа на наложителността и правилността на стъпките на правителството.

http://bsptrolls.bivol.bg/media/bsptrolls/2013-10-18-2013-10-31.pdf

Leaked Reports = Golden Data?

Comment content

LEADWAY media solutions

10/11/2013

Original URL

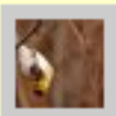
User

	Original URL	User	Comment content
47	http://www.blitz.bg/index.php?news=229757#commentsform	Краси	Толкова глупаво дело няма никъде - за загубени листи. Виж друго си е да те съдят за корупция или подпомагане на организираната престъпност, като някои видни гербери...
48	http://www.investor.bg/ikonomika-i-politika/332/a/prokuraturata-poiska-imuniteta-na-sergei-stanishev,160299/#comment-newest	kroki	точно така си беше - Станишев сам поиска да си изчисти името и да се прекрати случая, щото винаги се намират зложелатели да сравняват тези кокошкарски обвинения с корупционните дела на видни гербери...
49	http://www.dnevnik.bg/bulgaria/2013/10/25/2168726_glavniiat_prokuror_poiska_imuniteta_na_stanishev_po/?p=0#addcomment	transa_23	Прокуратурата е обективна и си чисти казусите. Ама като накрая се окаже много шум за нищо, дали ще млъкнат клакьорите на герберите....

10,150 paid comments: ~**2,000 in Facebook**, and ~**8,150 in news community forums**

More Troll Accusations

Example Accusation of Trolling



Да бе да
Рейтинг: 4584
Неутрално

44 16:59 21 мар, 2014

Отговор

До коментар [#1] от "Gianpiero":
До коментар [#2] от "comtec":
До коментар [#4] от "comtec":
До коментар [#3] от "Night Rider":

Първата редица тролчета
седнали на столчета,
чукат по слвишите,туй, що други са им
написали!
И аз съм от пртестната мрежа!
И аз стоя зад Асен и другите!

Който иска - търси начин. Който Не - търси оправдание.

Оценка
-5 +29

Posts by Exposed Trolls

Article:

БЪЛГАРИЯ 21 мар 2014, 16:15

Тази статия е част от темата "Формулата на КТБ: пари-медии-власт".

"Протестна мрежа": Занимавате ли се с политика, г-н Василев?

Comments:

Gianpiero
Рейтинг: 356
Неутрално
1 16:18 21 мар, 2014
Отговор

Да, занимава се, точно колкото се занимава с политика и самата Протестна мрежа. И в това е големия проблем, че разни такива необособени субекти се занимават с политика!

Оценка -156 +33

comtec
Рейтинг: 625
Неутрално
2 16:18 21 мар, 2014
Отговор

Тия па кво пищят?

Оценка -169 +28

Night Rider
Рейтинг: 363
Неутрално
3 16:20 21 мар, 2014
Отговор

Псевдоморалисти тръгнали да четат морал. С нищо не могат да ме убедят че преследват безкористни цели.

Оценка -163 +26

comtec
Рейтинг: 625
Неутрално
4 16:20 21 мар, 2014
Отговор

И каква е тая фантомна организация "Протестна мрежа"? Фирма ли е, фондация ли е, творческо дружество ли е? И въобще кой се занимава с подобни анонимници?

Оценка -149 +26

Our Data

Our Dataset (from Dnevnik.bg)

Object	Number*
Publications	34,514
Comments	1,930,818
- Comment replies	897,806
Users	14,598
Topics	232
Keywords/tags	13,575

*Politics topics, Period 01.01.2013 – 01.04.2015, Dnevnik.bg

Plus the 10,150 paid comments from Leadway

Experiment 1: Finding Mentioned Opinion Manipulation Troll Users

Todor Mihaylov, Georgi Georgiev, Preslav Nakov. *Finding Opinion Manipulation Trolls in News Community Forums*. **CoNLL-2015**, pp. 310-314

Method

- Define “trollness”:

  A user who is called a troll by several people is likely to be one.

- Create a labeled dataset.

- Define and extract features.

- Train and evaluate an SVM classifier with different
  - “troll” class definitions
  - feature groups

Users expose other users as trolls
… after the publication of the leaked documents

*“To comment from "Rozalina": You, **trolls**, are so funny :) I saw the same signature under other comments:)‘”*

*“To comment from "Historama": **Murzi**, you know that you cannot manipulate public opinion, right?”*

Labels & Data

- **Trolls:** users called *troll* or *murzilka* by at least 5 distinct users
- **Non-trolls:** users (w/ 100+ posts) that have never been called so

X 317

X 964

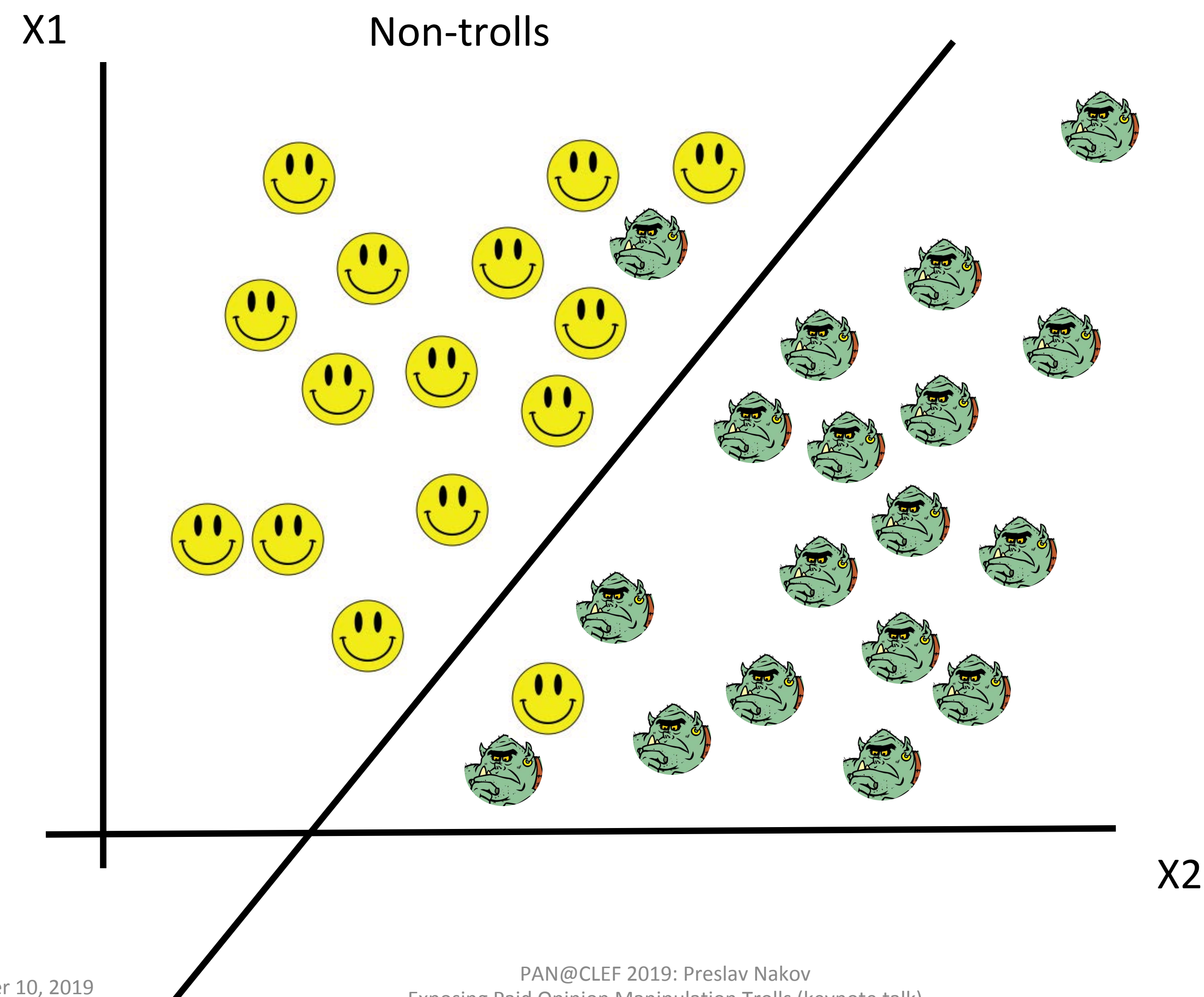
Feature Types *(w/ Our Assumptions)*

- Vote-based features. *People vote troll comments low.*
- Comment-to-publication similarity. *Trolls like to change topic.*
- Comment order-based features. *Trolls like to comment first, to get maximum attention.*
- Top loved/hated comments. *Troll comments are often hated.*
- Comment replies-based features. *Trolls tend to provoke or to engage in discussions.*
- Time-based features. *Paid trolls tend to write during working days and working times.*

All the features are scaled, i.e., divided by the number of comments, the number of days in the forum, the number of days with more than one comment.

Total: 338 scaled features

SVM to Classify Users as Trolls vs. Non-trolls



Results: Ablation Excluding Some Feature Groups

Features	Accuracy	Diff
AS + Non-scaled	94.37(+3.74)	19.13
AS - total comments	91.17(+0.54)	15.93
AS - comment order	91.10(+0.46)	15.85
AS - similarity	91.02(+0.39)	15.77
AS - time day of week	90.78(+0.15)	15.53
AS - trigg rep range	90.78(+0.15)	15.53
AS - time all	90.71(+0.07)	15.46
All scaled (AS)	90.63	15.38
AS - top loved/hated	90.55(-0.07)	15.30
AS - time hours	90.47(-0.15)	15.22
AS - vote u/down rep	90.47(-0.15)	15.22
AS - similarity top	90.32(-0.31)	15.07
AS - triggered cmnts	90.32(-0.31)	15.07
AS - is rep to has rep	90.08(-0.54)	14.83
AS - vote up/down all	89.69(-0.93)	14.44
AS - is reply	89.61(-1.01)	14.36
AS - up/down votes	88.29(-2.34)	13.04

Results: Performance of Individual Feature Groups

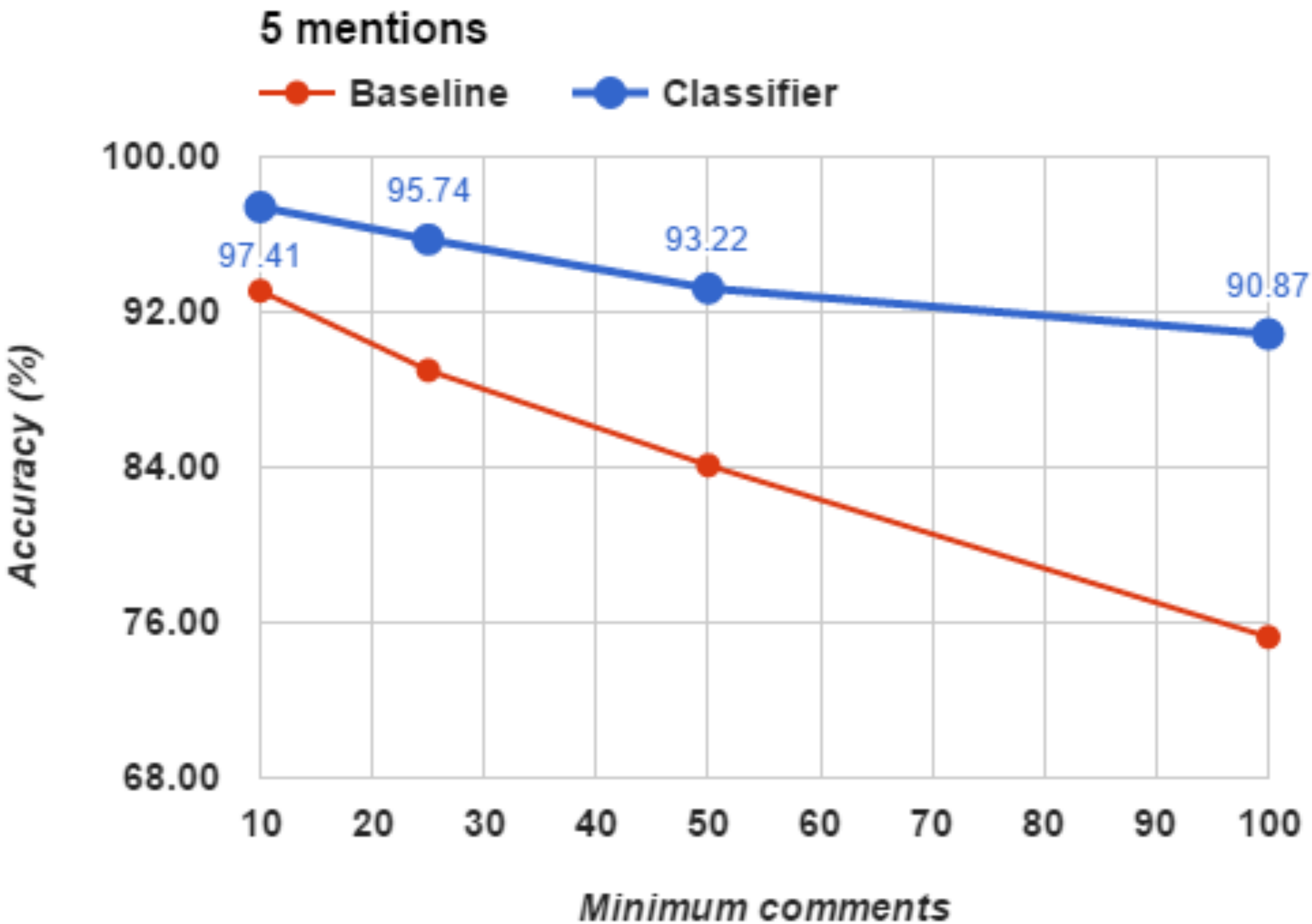
Features	Accuracy	Diff
All Non-scaled	93.21	17.95
Only vote up/down	87.67	12.41
Only vote up/down totals	87.2	11.94
Only reply up/down voted	86.1	10.85
Only time hours	84.93	9.68
Only time all	84.31	9.06
Only is reply with rep	82.83	7.57
Only triggered rep range	82.83	7.57
Only day of week	82.28	7.03
Only total comments	82.28	7.03
Only reply status	80.72	5.46
Only triggered replies	80.33	5.07
Only comment order	80.09	4.84
Only top loved/hated	79.39	4.14
Only pub similarity top	75.25	0
Only pub similarity	75.25	0

Results: Impact of Number of Mentions

Min mentions	3	4	5	6
Trolls	545	419	317	260
Non-troll	964	964	964	964
Accuracy	85.49	87.85	90.87	92.32
Diff	21.60	18.15	15.61	13.56

Improvement over the majority class baseline

Results: Impact of the Minimum Number of Comments



Summary

- Experimented with a large number of features
  - both scaled and non-scaled
- Achieved accuracy of 82-95%

- The nature of our features means that our troll detection works best for ``elder trolls'' with at least 100 comments in the forum.

Discussion

- As the minimum number of comments increases, the improvement of our classifier over the baseline also increases.
- The more we know about a user, the better we can predict whether s/he will be seen as a troll by other users.
- The results of experiments with different features groups show that most of our assumption were confirmed.

Experiment 2: Finding <u>**Paid**</u> Opinion Manipulation Troll <u>**Users**</u>

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, Preslav Nakov. *Exposing Paid Opinion Manipulation Trolls*. **RANLP-2015**, pp. 443-450.

Method

- Define “trollness”:

  A user who is called a troll by several people is likely to be one.

- Create a labeled dataset:
  - Mentioned trolls
  - Non-trolls
  - Paid trolls

- Extract features.
- Train an SVM classifier: *mentioned trolls* vs *non-trolls*
- Test the classifier: *paid trolls* vs. *non-trolls*

Labels & Data

- **Trolls:** users called *troll* or *murzilka* by at least 5 distinct users
- **Non-trolls:** users (w/ 100+ posts) that have never been called so
- **Paid trolls**: from Bivol/Leadway

X 314

X 964

X 15

Feature Types

- Vote-based features. *People vote troll comments low.*
- Comment-to-publication similarity. *Trolls like to change topic.*
- Comment order-based features. *Trolls like to comment first, to get maximum attention.*

Same as before

- Time-based features. *Paid trolls tend to write during working days and working times.*

All the features are scaled, i.e., divided by the number of comments, the number of days in the forum, the number of days with more than one comment.

Total: 338 scaled features

Results: Ablation Excluding Some Feature Groups

Features (Bottom is better)	Acc	P	R	F
All Scaled (AS)	0.88	1.00	0.75	0.86
AS - comment order(Scaled-S)	0.88	1.00	0.75	0.86
AS - is reply (S)	0.88	1.00	0.75	0.86
AS - is reply to has reply (S)	0.88	1.00	0.75	0.86
AS - similarity (S)	0.88	1.00	0.75	0.86
AS - similarity top (S)	0.88	1.00	0.75	0.86
AS – top loved hated (S)	0.88	1.00	0.75	0.86
AS - total comments (S)	0.88	1.00	0.75	0.86
AS - trigg replies range (S)	0.88	1.00	0.75	0.86
AS - trigg replies total (S)	0.88	1.00	0.75	0.86
AS - vote up/down total (S)	0.88	1.00	0.75	0.86
AS - time (S)	0.75	1.00	0.50	0.67
AS - time hours (S)	0.75	1.00	0.50	0.67
AS - vote up/down reply stat(S)	0.75	1.00	0.50	0.67
AS - time day of week (S)	0.63	1.00	0.25	0.40
AS + Non Scaled (NS)	0.63	1.00	0.25	0.40
AS - vote up/down all (S)	0.38	0.00	0.00	0.00

Results: Performance of Individual Feature Groups

Features (Top is better)	Acc	P	R	F
only day of week (S)	0.88	0.80	1.00	0.89
only reply status (S)	0.75	0.75	0.75	0.75
only time hours (S)	0.75	0.75	0.75	0.75
only top loved hated (S)	0.75	1.00	0.50	0.67
only comment order (S)	0.63	0.67	0.50	0.57
only vote up/down is reply (S)	0.63	0.67	0.50	0.57
only similarity top (S)	0.63	1.00	0.25	0.40
only triggered replies range (S)	0.63	1.00	0.25	0.40
only is reply to has reply (S)	0.50	0.50	0.25	0.33
only similarity (S)	0.50	0.50	0.25	0.33
only time (S)	0.50	0.50	0.25	0.33
only total comments (S)	0.50	0.50	0.25	0.33
only triggered replies total (S)	0.50	0.50	0.25	0.33
only vote up/down all (S)	0.50	0.50	0.25	0.33
only vote up/down total (S)	0.50	0.50	0.25	0.33
All Unscaled	0.50	0.00	0.00	0.00

Testing on Mentioned vs. Paid Trolls

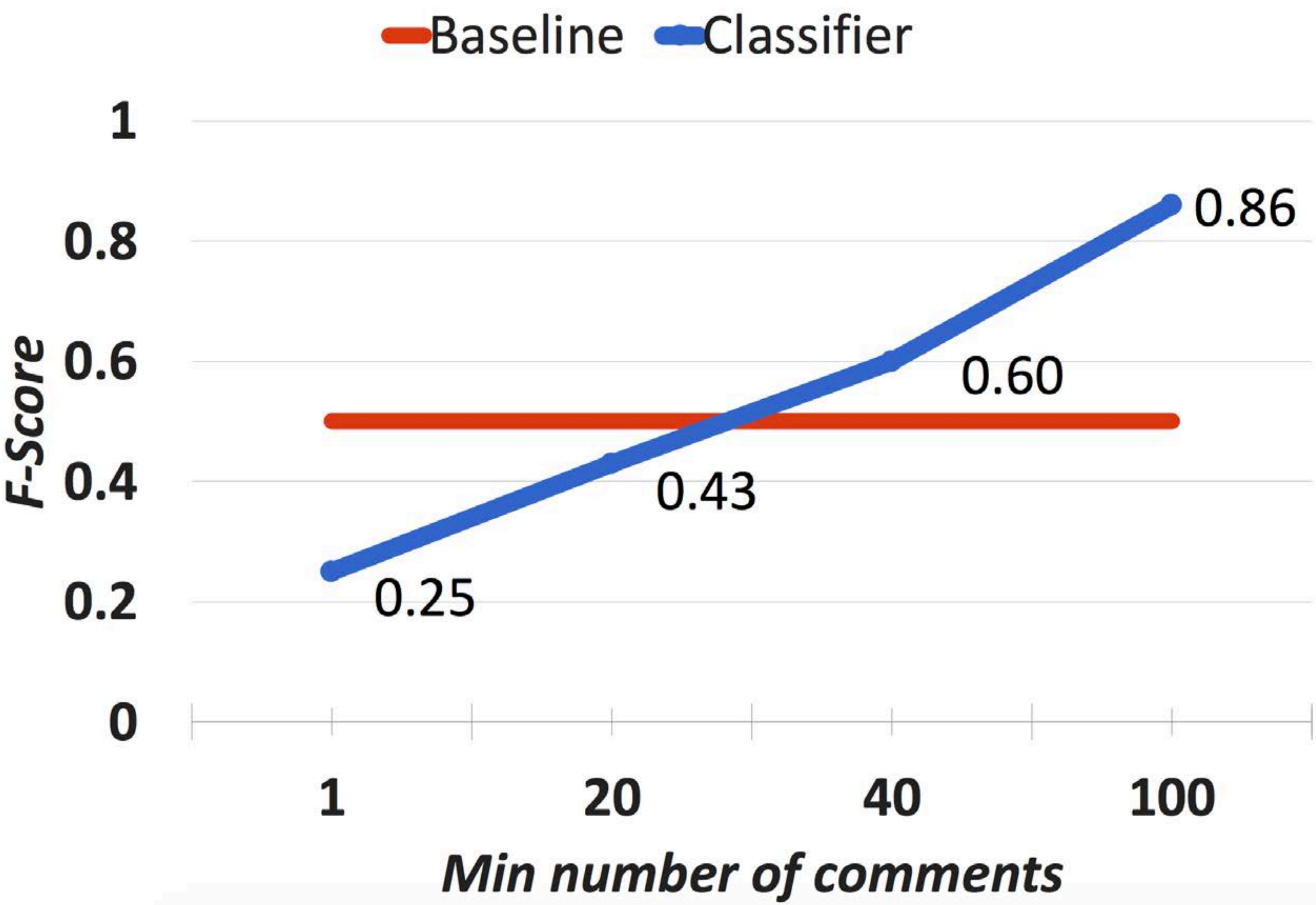
Min mentions	3	4	5	6
Mentioned Trolls	536	416	314	269
Non-troll	536	416	314	269
Accuracy	0.75	0.88	0.88	0.75
F-score	0.67	0.86	0.86	0.67

Finding paid trolls with 100+ mentions (4~trolls + 4 non-trolls).Training with AS features, and users with 150+ comments and varying minimum number of mentions as a troll.

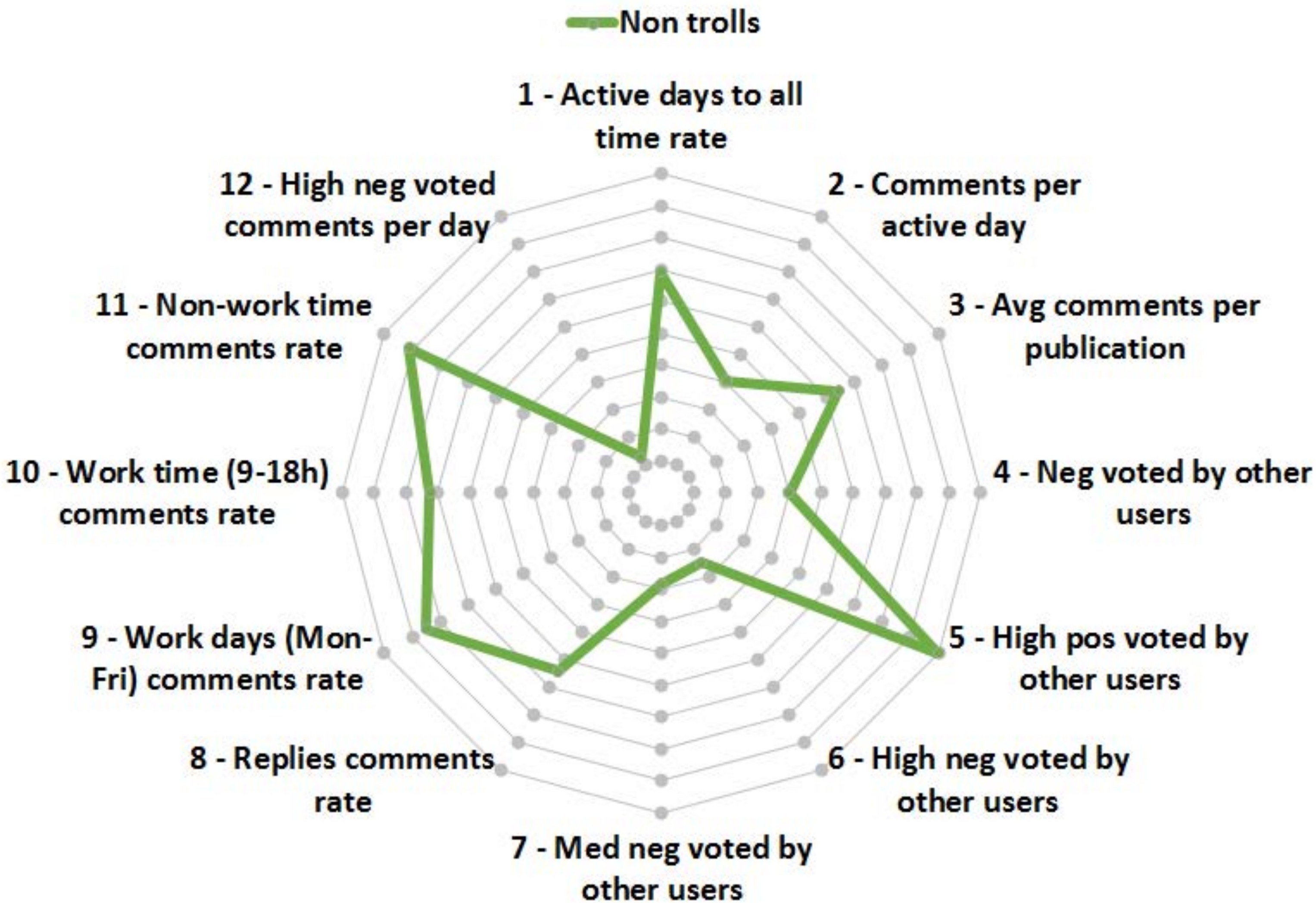
Min mentions	3	4	5	6
Mentioned Trolls	536	416	314	269
Non-troll	536	416	314	269
Accuracy	0.83	0.87	0.91	0.92
F-score	0.83	0.87	0.91	0.92

Finding “mentioned” trolls (cross-validation on the training dataset). Training with AS features, and users with 150+ comments and varying minimum number of mentions as a troll.

Results: Impact of Minimum Number of Comments

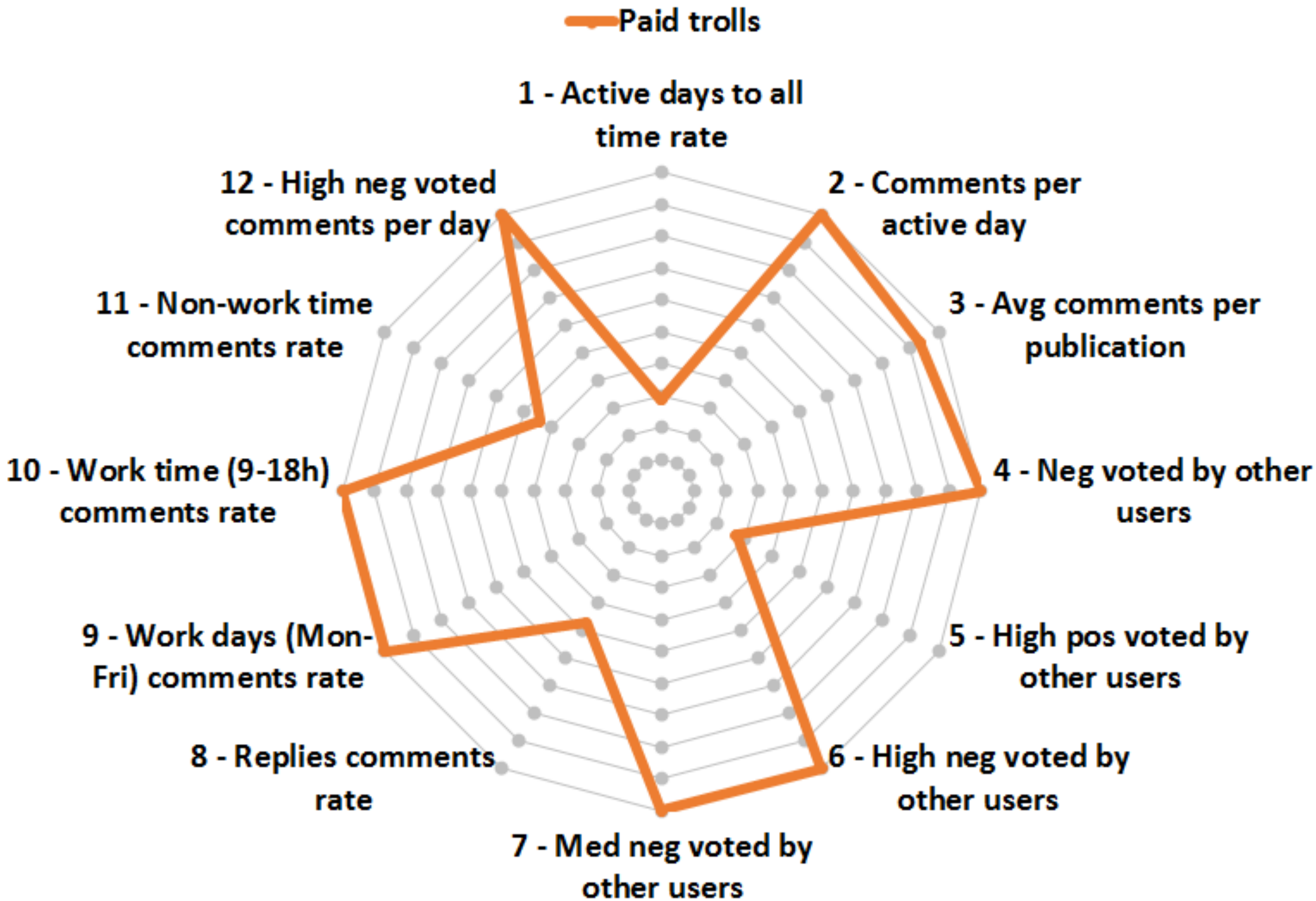


Non Trolls

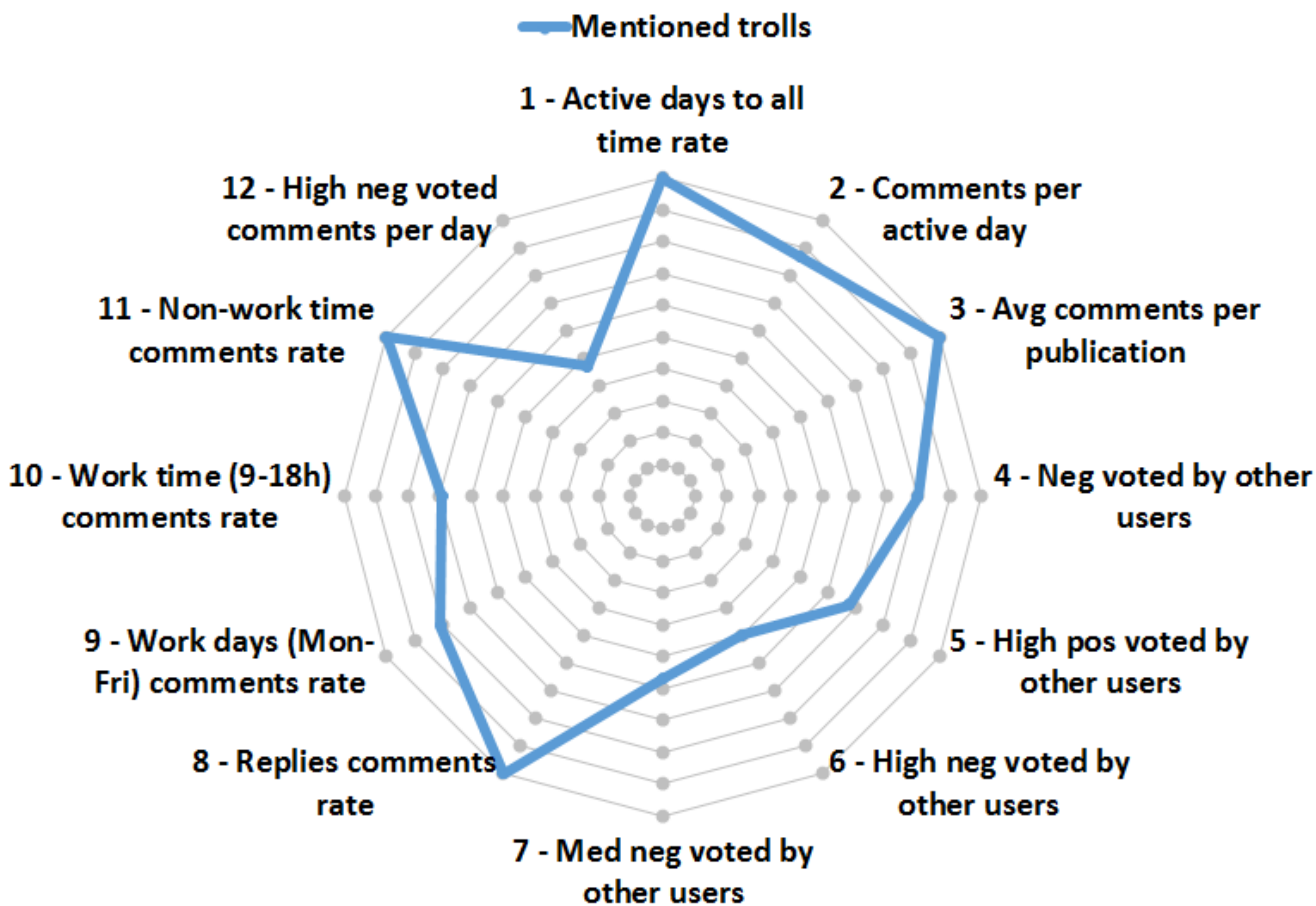


Non trolls
1 - Active days to all time rate
2 - Comments per active day
3 - Avg comments per publication
4 - Neg voted by other users
5 - High pos voted by other users
6 - High neg voted by other users
7 - Med neg voted by other users
8 - Replies comments rate
9 - Work days (Mon-Fri) comments rate
10 - Work time (9-18h) comments rate
11 - Non-work time comments rate
12 - High neg voted comments per day

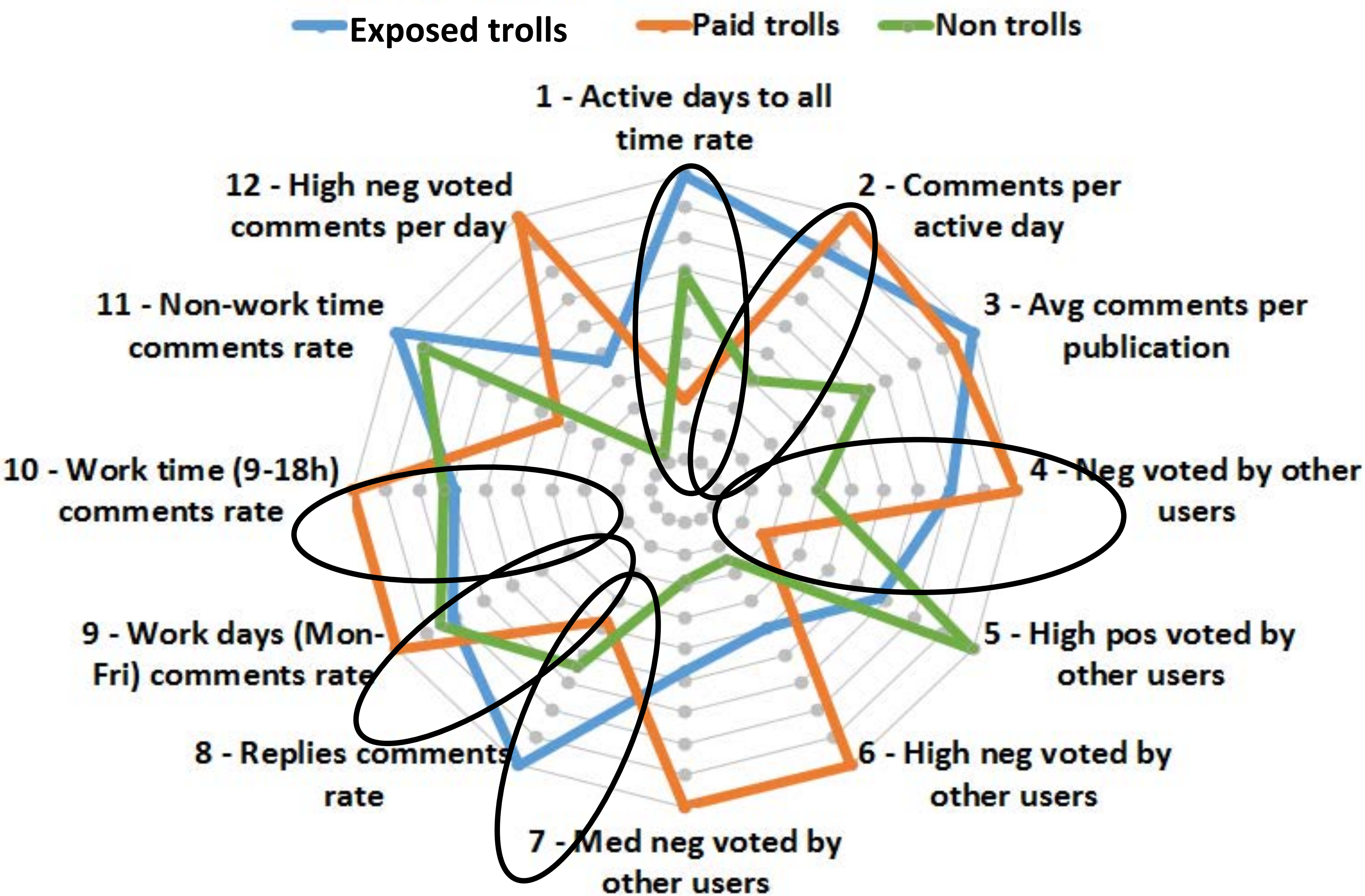
Paid Trolls


Paid trolls
1 - Active days to all time rate
2 - Comments per active day
3 - Avg comments per publication
4 - Neg voted by other users
5 - High pos voted by other users
6 - High neg voted by other users
7 - Med neg voted by other users
8 - Replies comments rate
9 - Work days (Mon-Fri) comments rate
10 - Work time (9-18h) comments rate
11 - Non-work time comments rate
12 - High neg voted comments per day

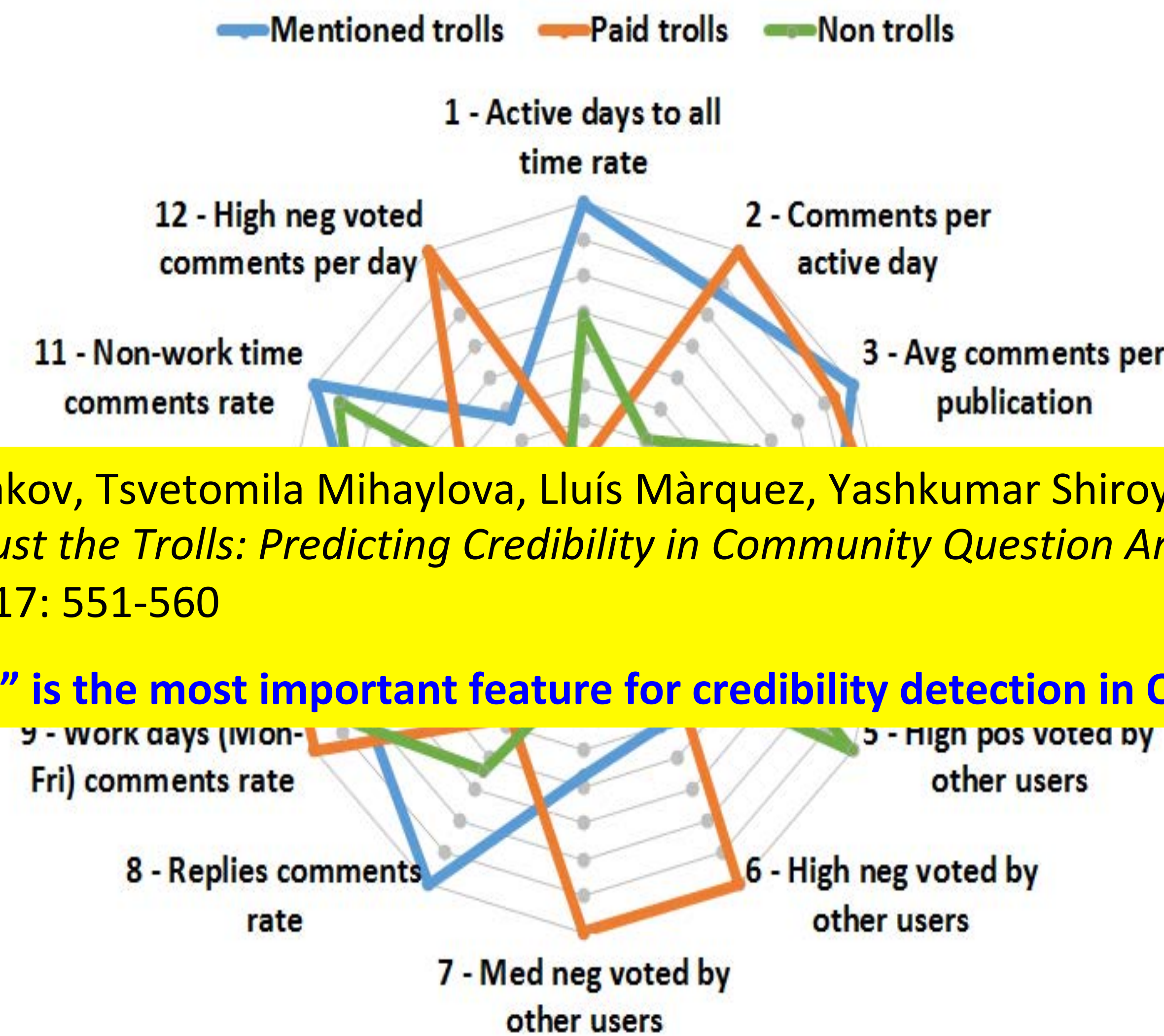
Mentioned Trolls



User behavior



Mentioned vs. Paid vs. Non Trolls



Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, Ivan Koychev: *Do Not Trust the Trolls: Predicting Credibility in Community Question Answering Forums.* RANLP 2017: 551-560

"Trollness" is the most important feature for credibility detection in Qatar Living!

Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, Ivan Koychev: *The dark side of news community forums: opinion manipulation trolls.* Internet Research 28(5): 1292-1312 (2018)

Discussion

- As the minimum number of comments increases, the improvement of our classifier over the baseline also increases.
- The more we know about a user, the better we can predict whether s/he will be seen as a troll by other users.
- Paid Trolls are similar to Mentioned Trolls, but there are also differences.
- Unfortunately, we have too few paid trolls…

Experiment 3: Finding Opinion Manipulation Troll **Comments**

Todor Mihaylov, Preslav Nakov. *Hunting for Troll Comments in News Community Forums*. **ACL-2016.**

Method

- Define "trollness":

Same as before (but focusing on **<u>comments</u>**)

  - *<u>mentioned</u> trolls* vs. *non-trolls*
  - *<u>paid</u> trolls* vs. *non-trolls*

Manual Checking of Troll Accusations
(whether they are accusations; not whether true)

- Annotate comments
  - 1,140 comments containing the words “troll” or “murzilka”
  - two annotators: Kappa = 0.82
- Found
  - agree on 578 actual accusations
- A simple classifier can find actual accusations
  - Bag of words, Word N-grams, Stemmed BoW
  - F-score = 0.85

Labels & Data
(comments, not users!)

- **Trolls:** users called *troll* or *murzilka* by at least 5 distinct users
- **Non-trolls:** users (w/ 100+ posts) that have never been called so
- **Paid trolls**: from Bivol/Leadway

X 650

X 650

X 578

X 578

Feature Types (1)

- Bag of words
- Bag of stems
- Word n-grams
- Char n-grams
- Word prefix
- Word suffix
- POS tag distribution: coarse- and fine-grained
- Named entities

Feature Types (2)

- Emoticons
- Punctuation counts
- Word2Vec clusters
- Sentiment scores: from lexicons
- Bad words: from lexicons + word2vec expansion
- Mentions of Bulgarian politicians and their nicknames.
- Metadata: comment rank, time and day of posting.

Mentioned vs Non-troll Comments: Ablation Excluding Some Feature Groups

Features	F	Acc
All − char n-grams	79.24	78.54
All − word suff	78.58	78.20
All − word preff	78.51	78.02
All − bow stems	78.32	77.85
All − bow with stop	78.25	77.77
All − bad words	78.10	77.68
All − emoticons	78.08	77.76
All − mentions	78.06	77.68
All	78.06	77.68
All − (bow, no stop)	78.04	77.68
All − NE	77.98	77.59
All − sentiment	77.95	77.51
All − POS	77.80	77.33
All − w2v clusters	77.79	77.25
All − word 3-grams	77.69	77.33
All − word 2-grams	77.62	77.25
All − punct	77.29	76.90
All − metadata	70.77	70.94

Paid trolls vs Non-troll Comments: Ablation Excluding Some Feature Groups

Features	**F**	**Acc**
All – char n-grams	81.08	81.77
All – word suff	81.00	81.77
All – word preff	80.83	81.62
All – bow with stop	80.67	81.54
All – sentiment	80.63	81.46
All – word 2-grams	80.62	81.46
All – w2v clusters	80.54	81.38
All – word 3-grams	80.46	81.38
All – punct	80.40	81.23
All – mentions	80.40	81.31
All	80.40	81.31
All – bow stems	80.37	81.31
All – emoticons	80.33	81.15
All – bad words	80.09	81.00
All – NE	80.00	80.92
All – POS	79.77	80.69
All – (bow, no stop)	79.46	80.38
All – metadata	75.37	76.62

Mentioned vs Non-troll Comments: Individual Feature Groups

Features	F	Acc
All	78.06	77.68
Only metadata	84.14	81.14
Sent,bad,pos,NE,meta,punct	77.79	76.73
Only bow, no stop	73.41	73.79
Only bow with stop	73.41	73.44
Only bow stems	72.43	72.49
Only word preff	71.11	71.62
Only w2v clusters	69.85	70.50
Only word suff	69.17	68.95
Only word 2-grams	68.96	69.29
Only char n-grams	68.44	68.94
Only word 3-grams	64.74	67.21
Only POS	64.60	65.31
Sent,bad,pos,NE	63.68	64.10
Only sent,bad	63.66	64.44
Only emoticons	63.30	64.96
Sent,bad,ment,NE	63.11	64.01
Only punct	63.09	64.79
Only sentiment	62.50	63.66
Only NE	62.45	64.27
Only mentions	62.41	64.10
Only bad words	62.27	64.01

Paid trolls vs Non-troll Comments: Individual Feature Groups

Features	**F**	**Acc**
All	80.40	81.31
Sent,bad,pos,NE,meta,punct	78.04	78.15
Only bow, no stop	75.95	76.46
Only word 2-grams	75.55	74.92
Only bow with stop	75.27	75.62
Only bow stems	75.25	76.08
Only w2v clusters	74.20	74.00
Only word preff	74.01	74.77
Sent,bad,pos,NE	73.89	73.85
Only metadata	73.79	72.54
Only char n-grams	73.02	74.23
Only POS	72.94	72.69
Only word suff	72.03	72.69
Only word 3-grams	69.20	68.00
Only punct	66.80	65.00
Only NE	66.54	64.77
Sent,bad,ment,NE	66.04	64.92
Only sentiment	64.28	62.62
Only mentions	63.28	61.46
Only sent,bad	63.14	61.54
Only emoticons	62.95	61.00
Only bad words	62.22	60.85

Discussion

- Paid trolls' comments similar to mentioned trolls'
  - ✓Paid trolls vs non-trolls comments: 80-81% accuracy
  - ✓Mentioned troll vs. non-troll comments: 79-80% accuracy

- We cannot directly compare mentioned troll and paid troll comments as they were posted in different time spans.
  - ✓ Mentioned troll accusation went viral after the documents were leaked.

- Non-troll comments are not gold.

Summary

Summary

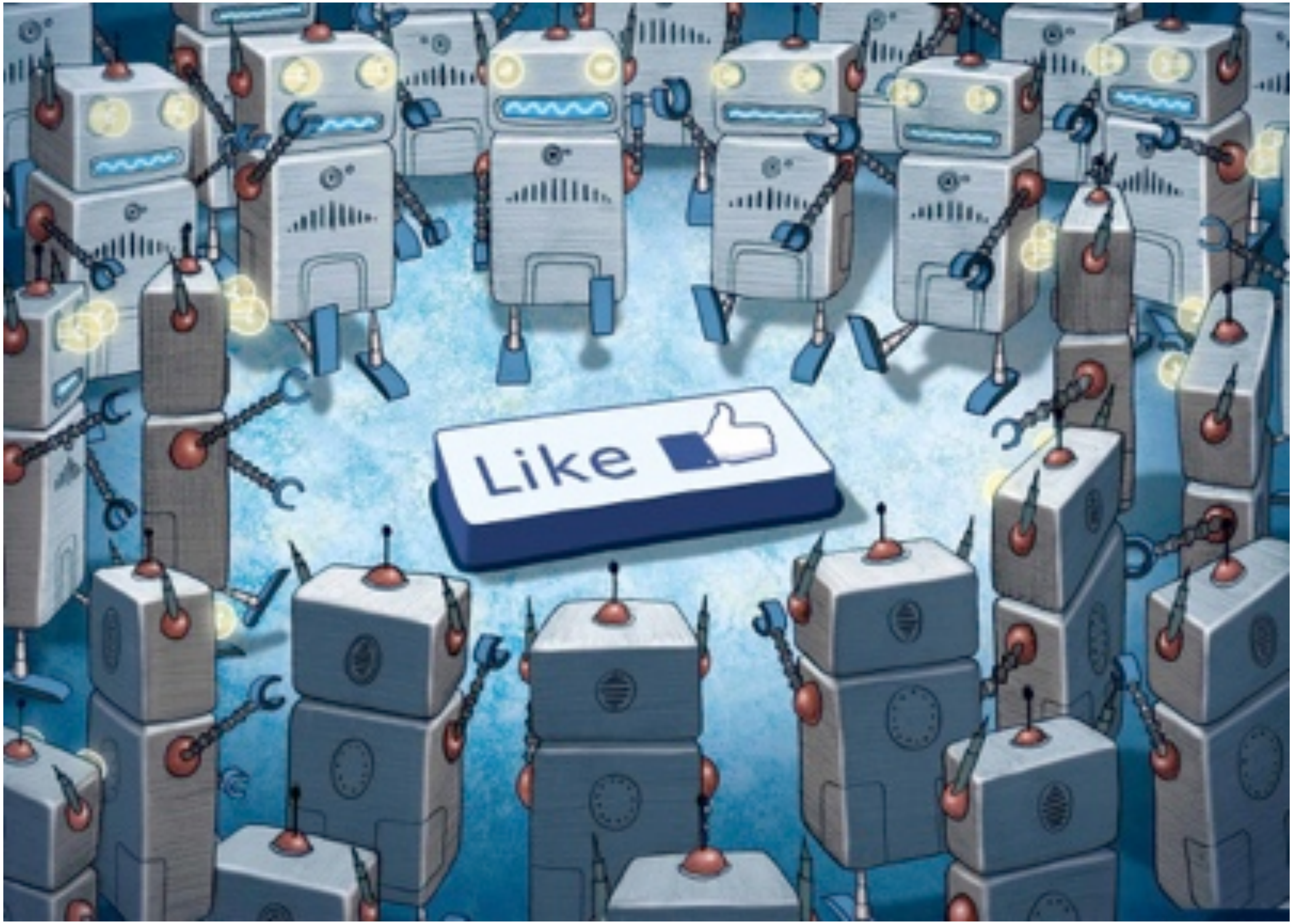
- **Conclusion**
  - ✓**New, useful definition:** ***A user who is called a troll by several people is likely to be one.***
  - ✓New datasets with troll data (Bulgarian)
  - ✓Evaluated several feature groups
  - ✓Experiments
    - Troll profile behavior detection – **90-96% accuracy**
    - Troll comment detection – **80-84% accuracy**

- **Future Work**
  - ✓Combine user- and comment-level features
  - ✓Apply to other community forums and other languages.
    - ✓(work-in-progress) trollness – top features for credibility (English)

Typology
of Manipulative Users

Social Bots

Social bots: accounts that are programmatically controlled to produce content and to interact with other users

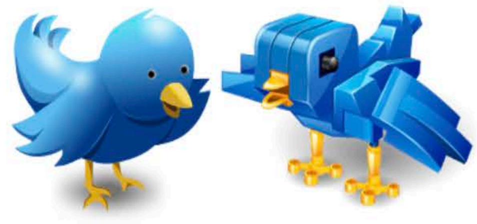


Finding Social Bots (2)

https://botometer.iuni.iu.edu/

Botometer®

An OSoMe project (bot•o•meter)



Botometer (formerly BotOrNot) checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot. Higher scores are more bot-like.

Use of this service requires Twitter authentication and permissions. (Why?)

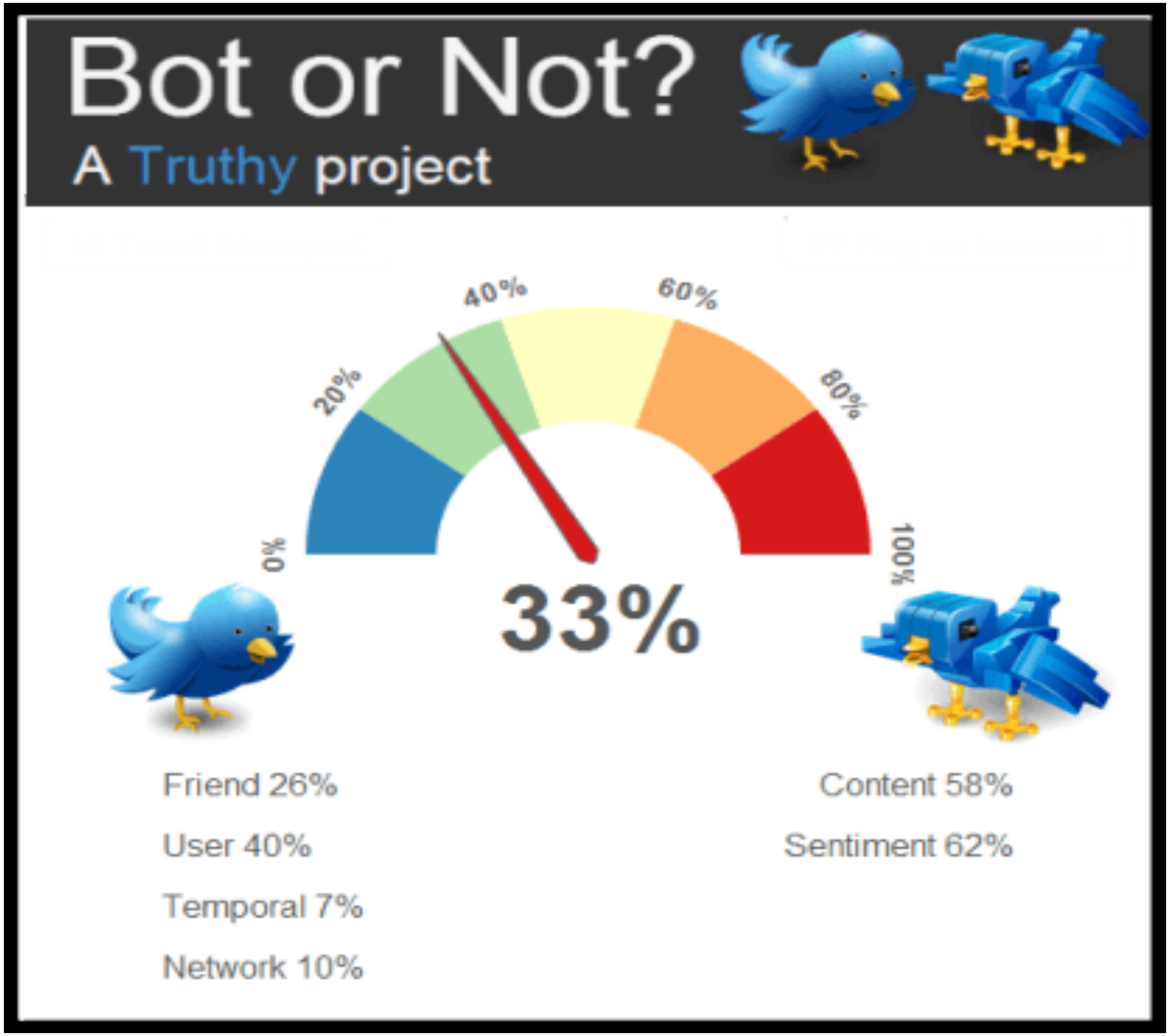
If something's not working or you have questions, please contact us only after reading the FAQ.

Botometer is a joint project of the Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS) at Indiana University.

@ScreenName | Check user | Check followers | Check friends

Finding Social Bots (2)

https://botometer.iuni.iu.edu/



Trolls

Trolling: malicious online behavior that is intended to disrupt interactions, to aggravate interacting partners, and to lure them into fruitless argumentation in order to disrupt online interactions and communication. (BUT can also mean opinion manipulation)

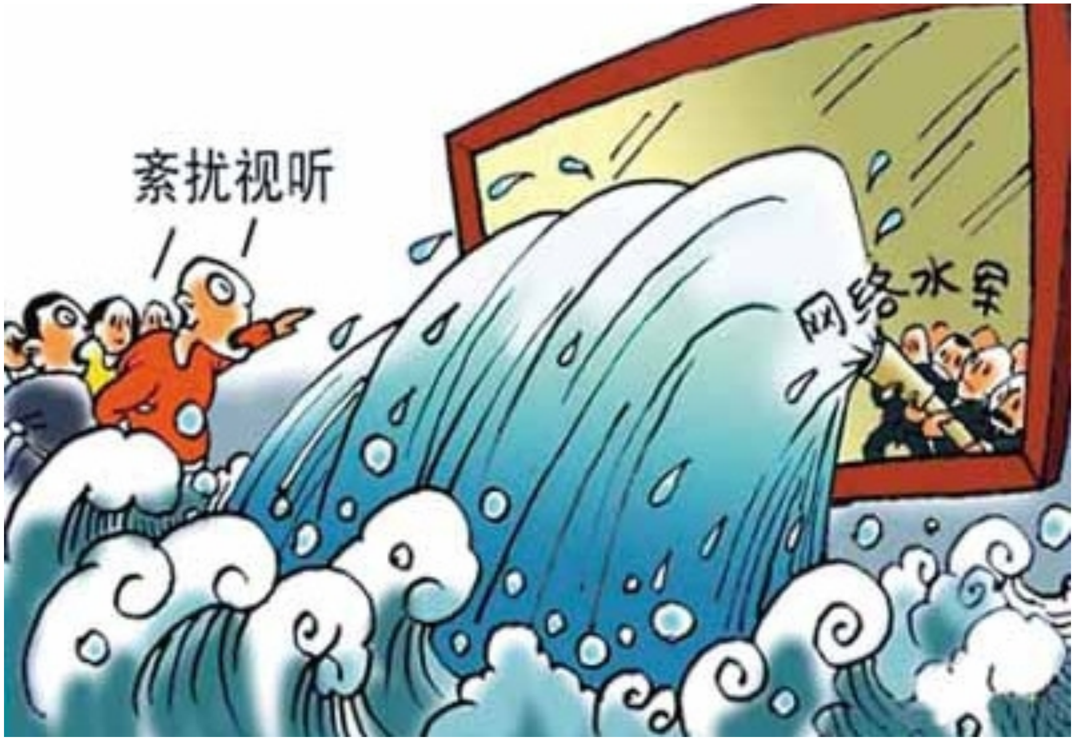
Sock Puppets

Sockpuppets: people who assume a false identity in an Internet community and then speak to/about themselves while pretending to be another person.

Astroturfers

Internet Water Army

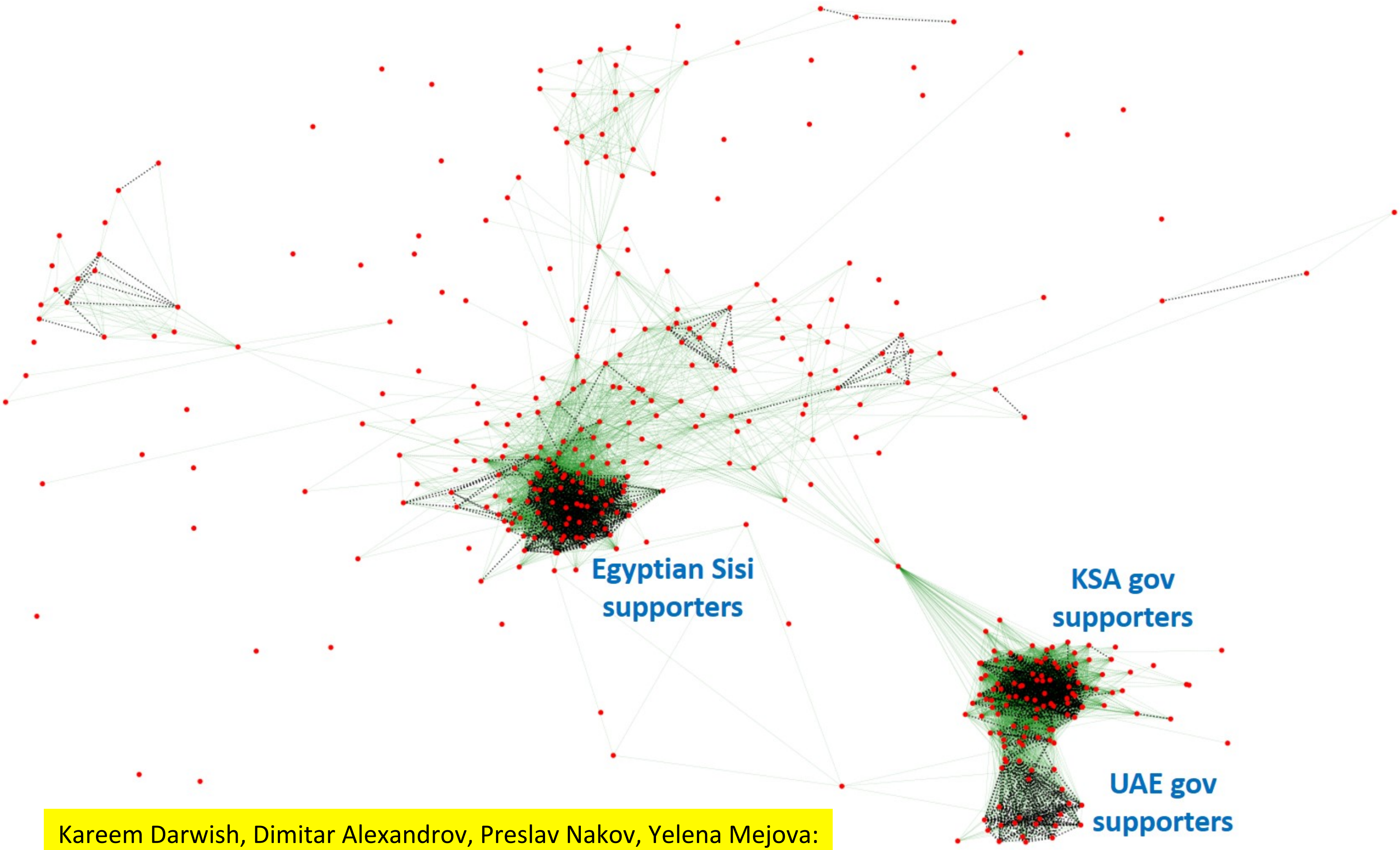
Internet Water Army: a large number of people who are well organized to flood the Internet with purposeful comments and articles.



“Seminar Users”

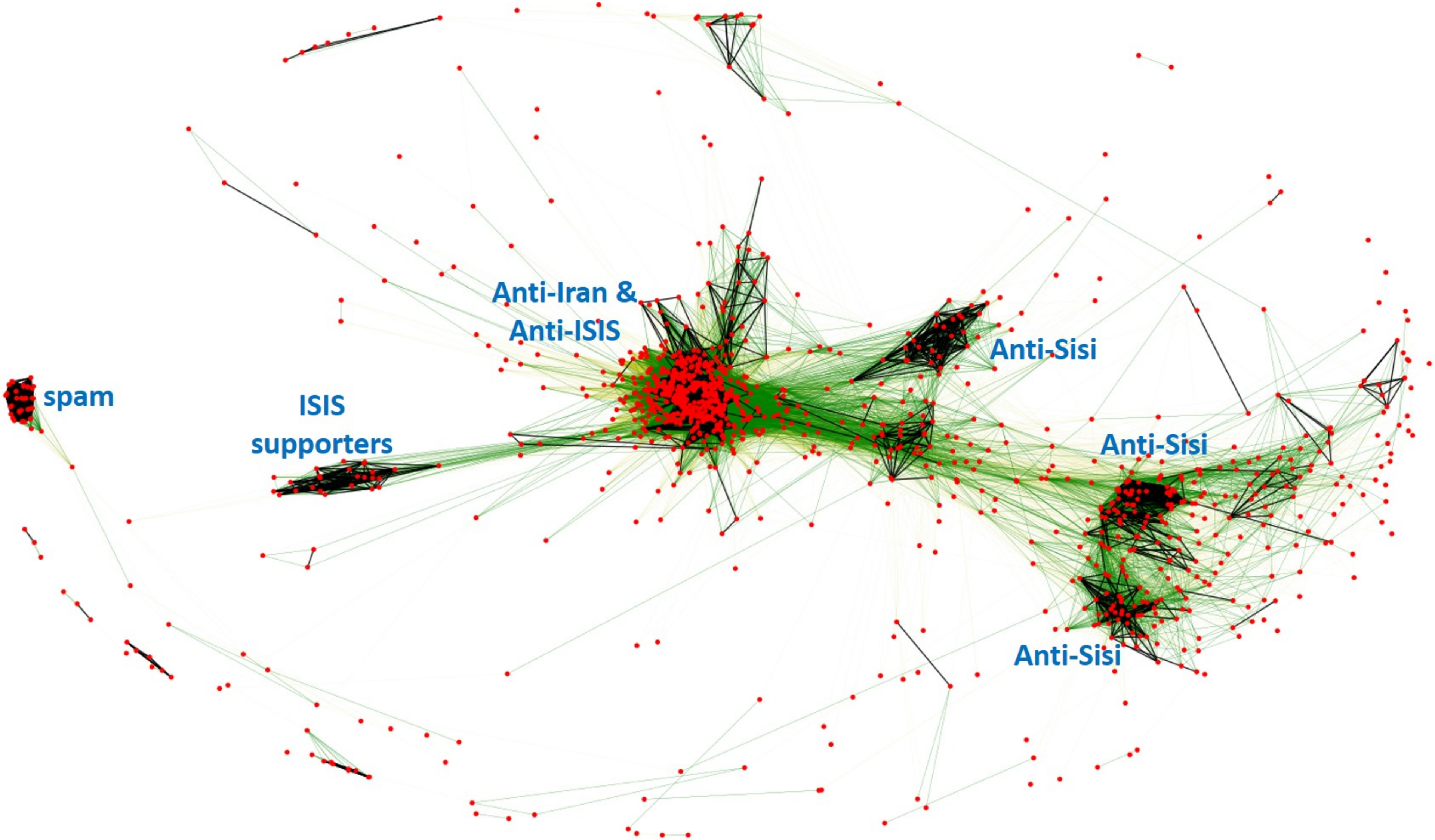
“Seminar Users”: social media users engaged in propaganda in support of a political entity.

Pro-Sisi “Seminar Users”



Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, Yelena Mejova: *Seminar Users in the Arabic Twitter Sphere.* SocInfo (1) 2017: 91-108

Anti-Sisi "Seminar Users"



Understanding the Role of Political Trolls in Social Media

Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov:

Understanding the Roles of Political Trolls in Social Media. **CoNLL 2019**

THE DAILY NEWSLETTER
Sign up to our daily email newsletter

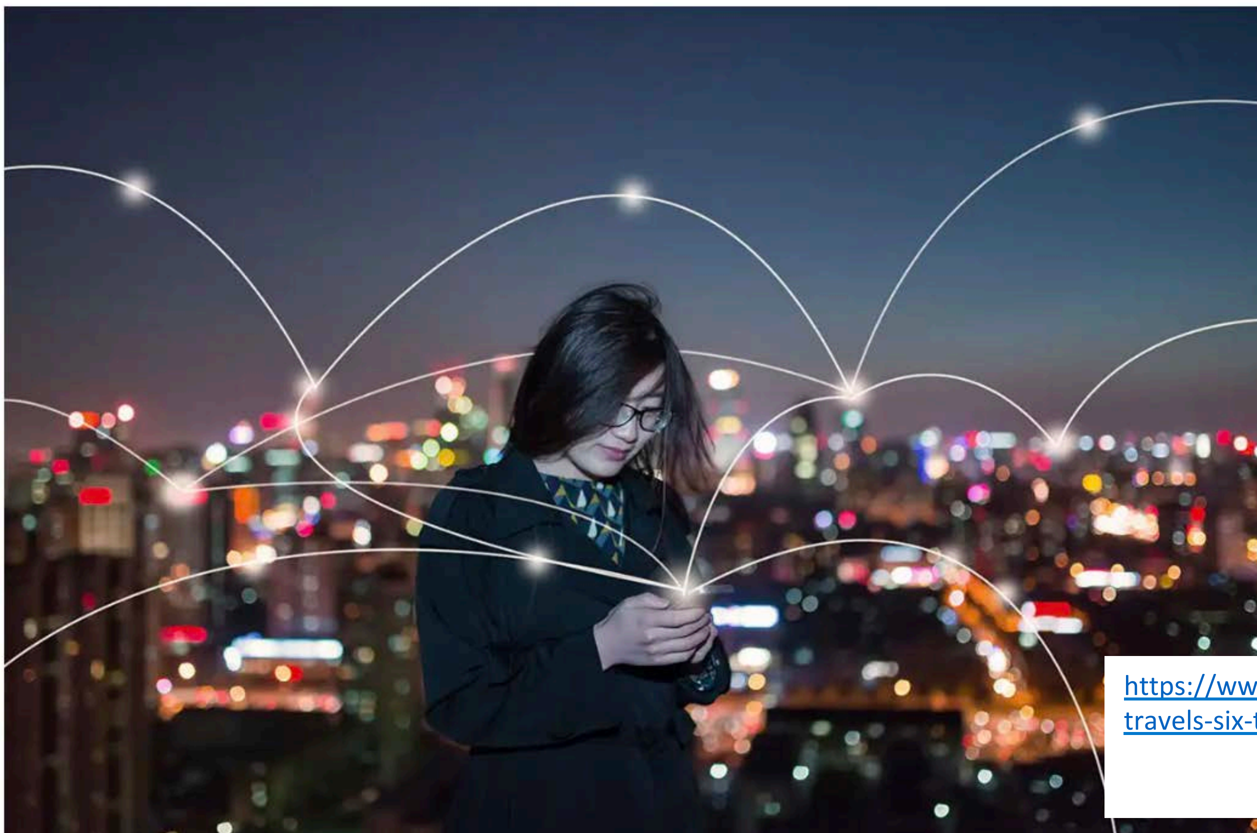
NewScientist

News Technology Space Physics Health Environment Mind Video | Tours Events Jobs

Fake news travels six times faster than the truth on Twitter

TECHNOLOGY 8 March 2018

By Chris Stokel-Walker

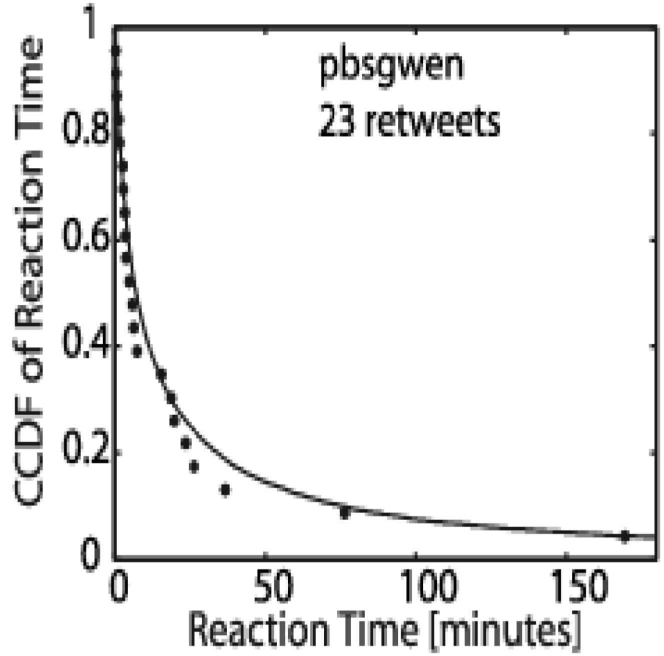


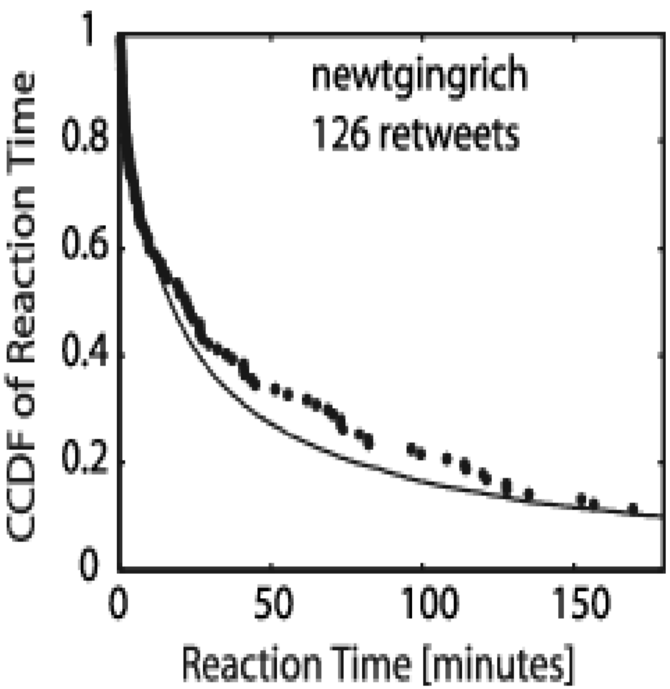
News travels fast, especially when it's wrong

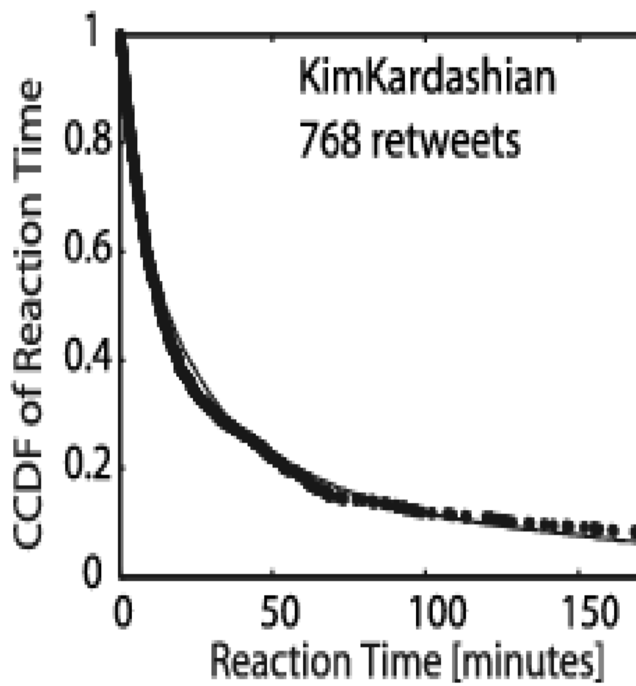
https://www.newscientist.com/article/2163226-fake-news-travels-six-times-faster-than-the-truth-on-twitter/

they

50% of the spread of "fake news" on Twitter: <10 minutes







Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow
A Bayesian approach for predicting the popularity of tweets.
Ann. Appl. Stat. 8(3):1583-1611, 2014.
https://projecteuclid.org/euclid.aoas/1414091226

To Understand the Trolls Strategy, We Need to Understand Their Role

Understanding the Roles of Political Trolls in Social Media

Role	Users	Tweets	User Example	Tweet Example
Left	233	427,141	@samirgooden	@MichaelSkolnik @KatrinaPierson @sames-fandiari Trump folks need to stop going on CNN.
Right	630	711,668	@chirrmorre	BREAKING: Trump ERASES Obamas Islamic Refugee Policy! https://t.co/uPTneTMNM5
News Feed	54	598,226	@dailysandiego	Exit poll: Wisconsin GOP voters excited, scared about Trump #politics

Table 1: Summary statistics for the IRA Russian Trolls Tweets (IRA) dataset.

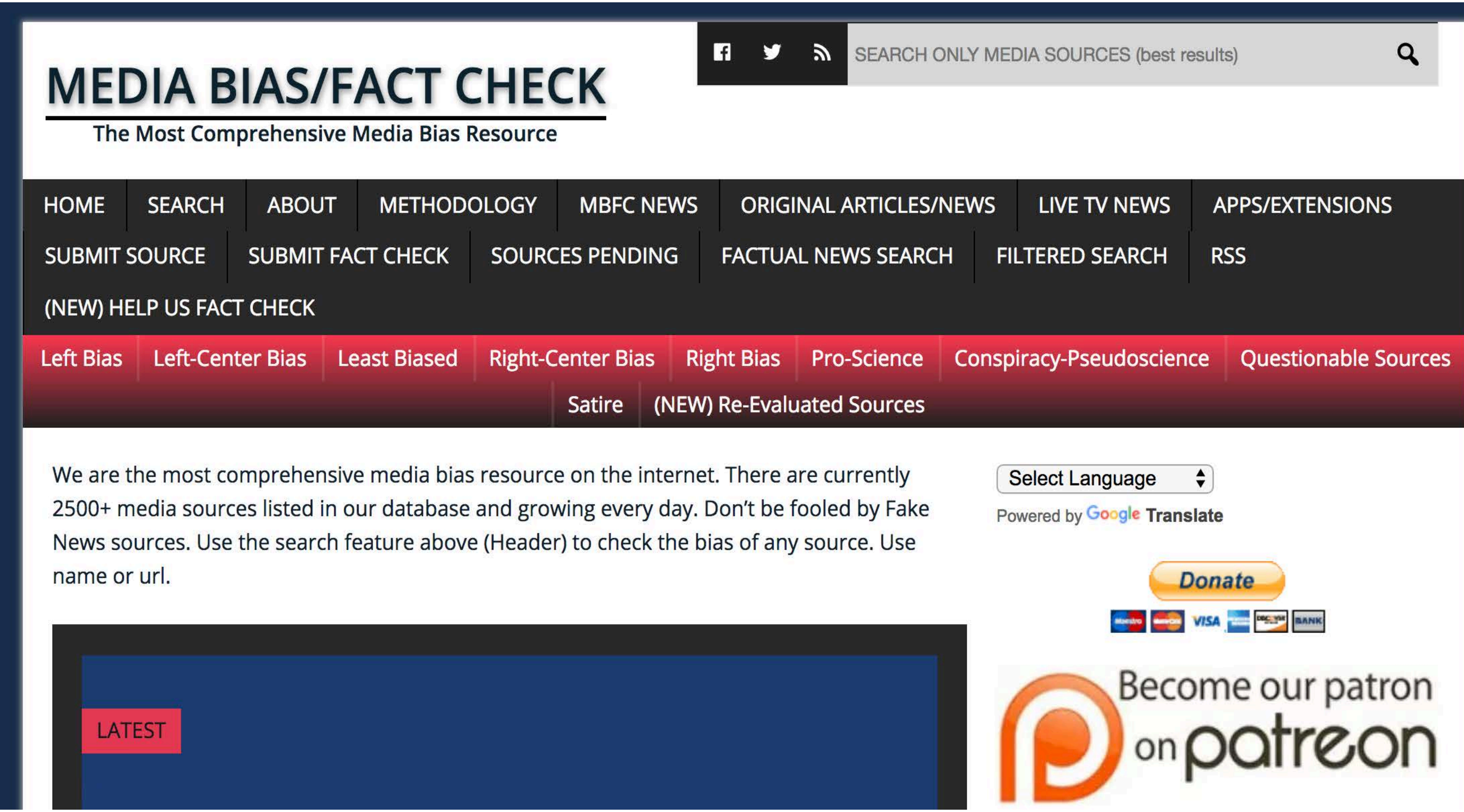
- 2,973,371 tweets
- 2,848 Twitter users
- February 2012 to May 2018
- linked to the Internet Research Agency (IRA), according to the US House Intelligence Committee

Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: *Understanding the Roles of Political Trolls in Social Media.* CoNLL 2019

Understanding the Roles of Political Trolls in Social Media

Inferred the troll's role by projecting information about media the trolls cited, among other things

Bias	Count	Example
LEFT	341	www.cnn.com
CENTER	372	www.apnews.com
RIGHT	619	www.foxnews.com

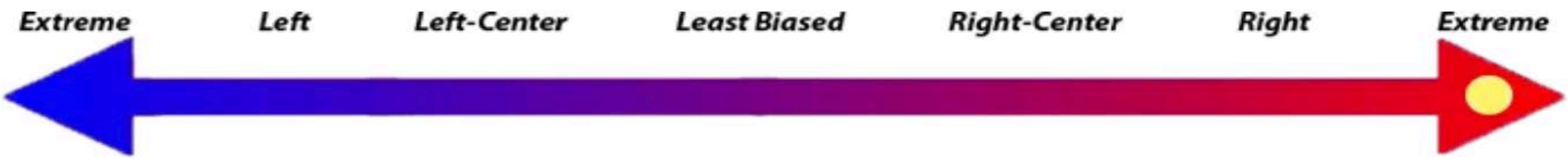

SEARCH ONLY MEDIA SOURCES (best results)
MEDIA BIAS/FACT CHECK
The Most Comprehensive Media Bias Resource
HOME
SEARCH
ABOUT
METHODOLOGY
MBFC NEWS
ORIGINAL ARTICLES/NEWS
LIVE TV NEWS
APPS/EXTENSIONS
SUBMIT SOURCE
SUBMIT FACT CHECK
SOURCES PENDING
FACTUAL NEWS SEARCH
FILTERED SEARCH
RSS
(NEW) HELP US FACT CHECK
Left Bias
Left-Center Bias
Least Biased
Right-Center Bias
Right Bias
Pro-Science
Conspiracy-Pseudoscience
Questionable Sources
Satire
(NEW) Re-Evaluated Sources
We are the most comprehensive media bias resource on the internet. There are currently 2500+ media sources listed in our database and growing every day. Don't be fooled by Fake News sources. Use the search feature above (Header) to check the bias of any source. Use name or url.
LATEST
Select Language
Powered by Google Translate
Donate
Become our patron on patreon

The Angry Patriot

Has this Media Source failed a fact check? **LET US KNOW HERE.**

Share:

137



QUESTIONABLE SOURCE

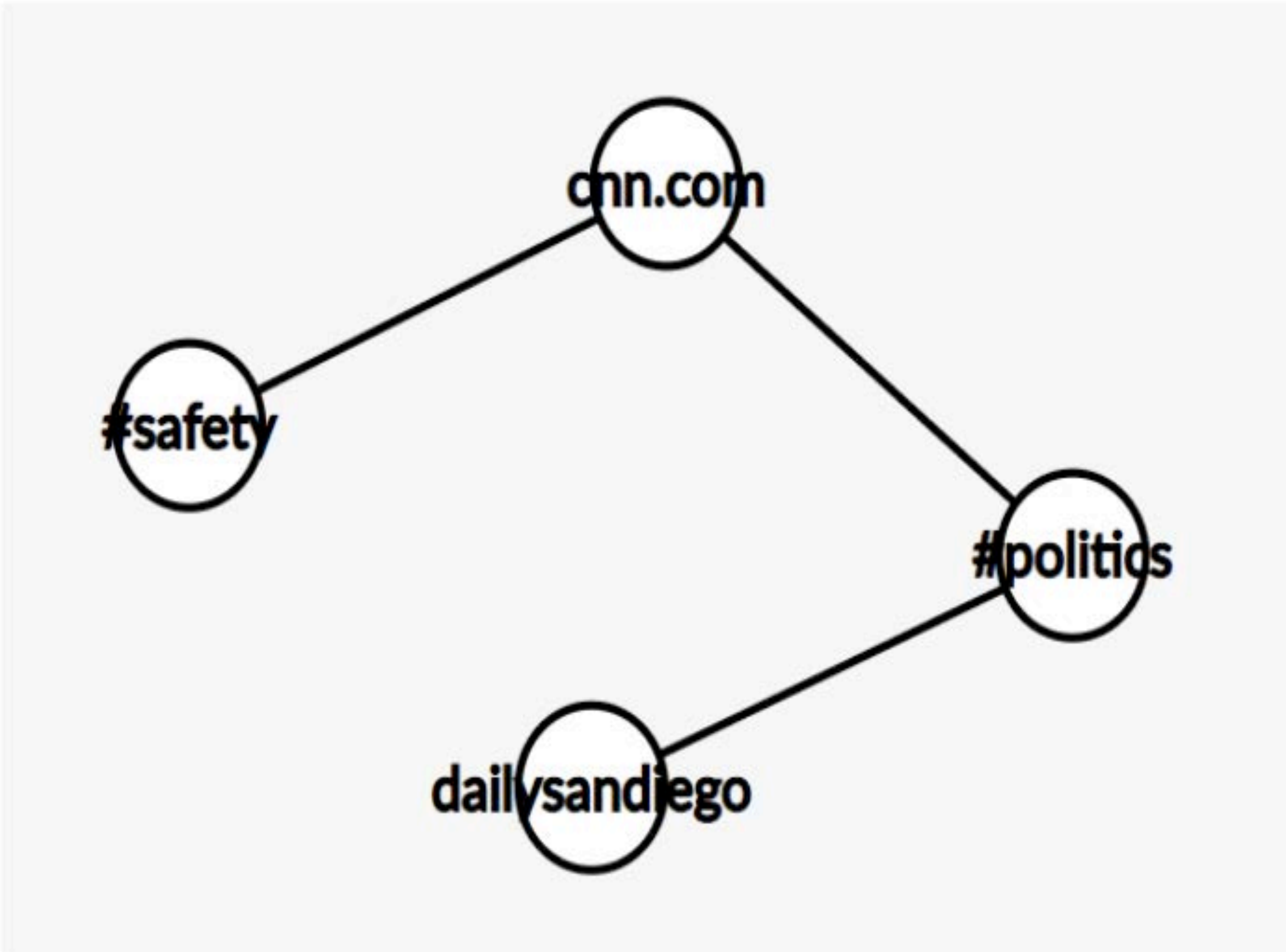
http://www.angrypatriotmovement.com/

A questionable source exhibits *one or more* of the following: extreme bias, overt propaganda, poor or no sourcing to credible information and/or is fake news. Fake News is the *deliberate attempt* to publish hoaxes and/or disinformation for the purpose of profit or influence (Learn More). Sources listed in the Questionable Category *may* be very untrustworthy and should be fact checked on a per article basis. Please note sources on this list *are not* considered *fake news* unless specifically written in the notes section for that source. See all Questionable sources.

Bias: **Extreme Right, Conspiracy, Propaganda, Some Fake News**

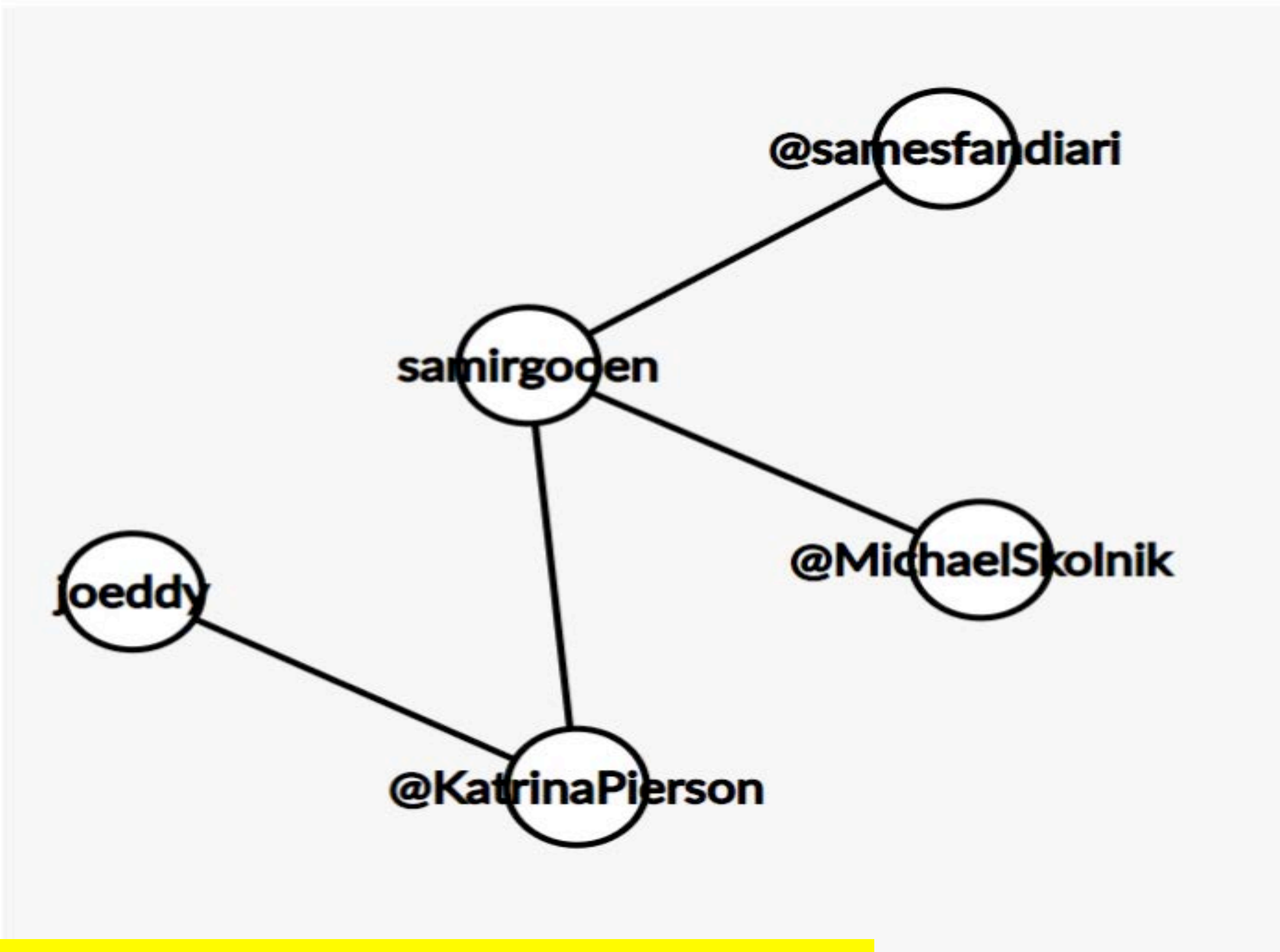
Notes: The Angry Patriot is a very right wing biased website with many conspiracies and fake news stories. A very untrustworthy source that was placed on Politifact's fake news source list. (10/9/2016) Updated (6/20/2017)

U2H: User to Hashtag Graph



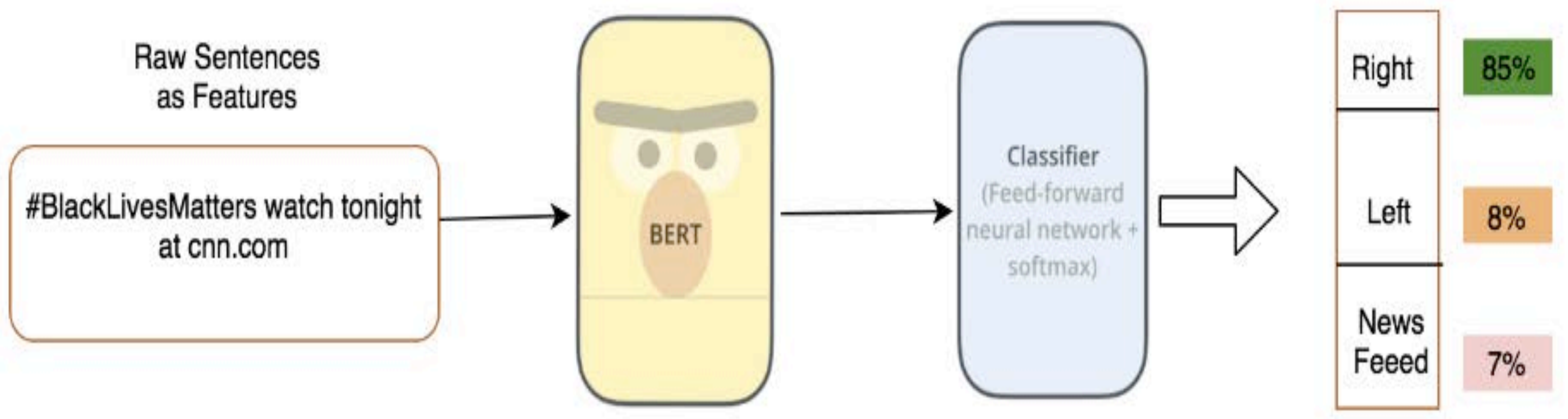
Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: *Understanding the Roles of Political Trolls in Social Media.* CoNLL 2019

U2M: User to User Mention Graph



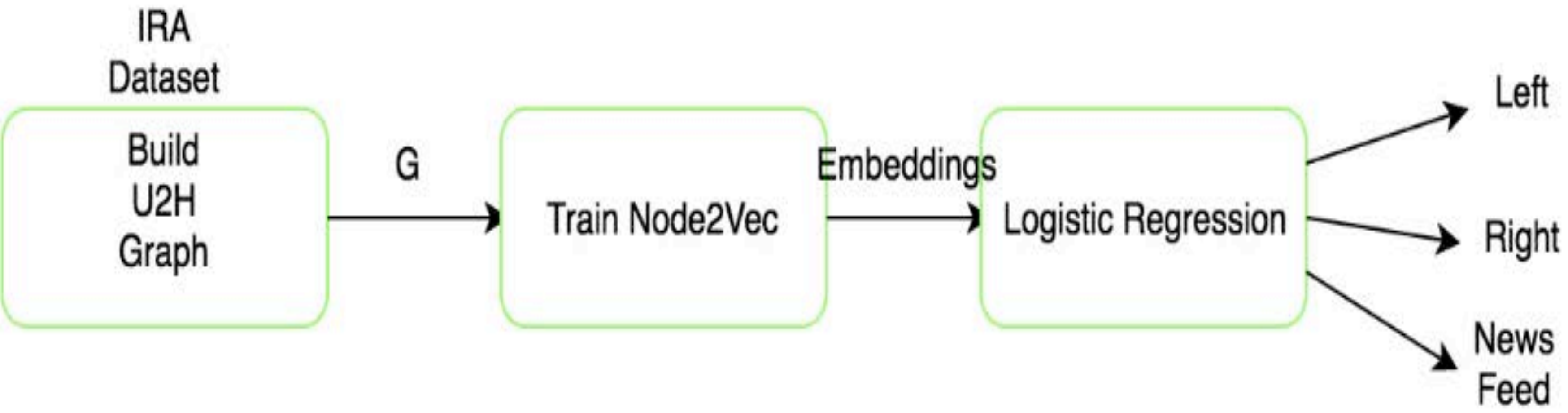
Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov:
Understanding the Roles of Political Trolls in Social Media. CoNLL 2019

BERT



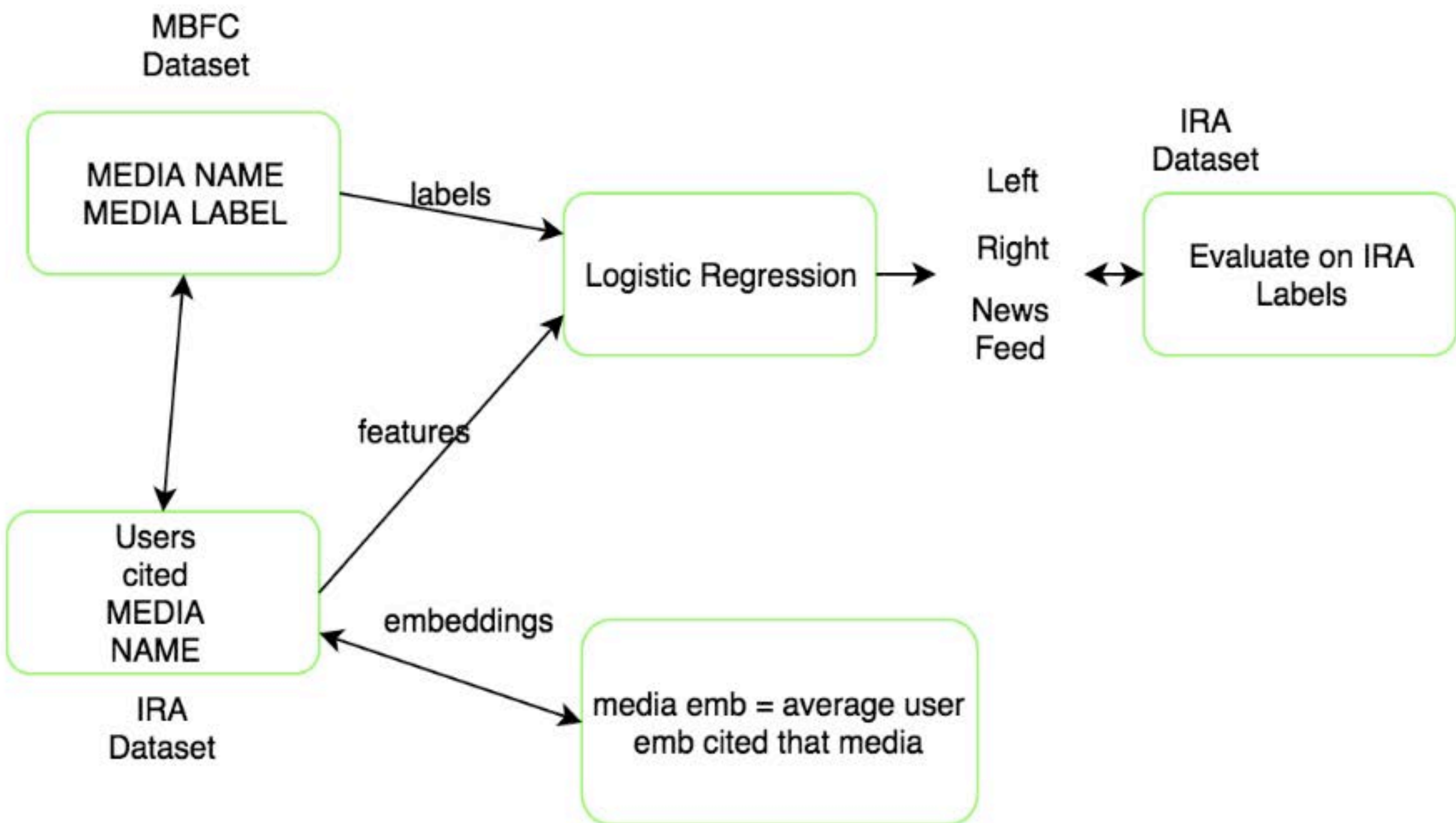
Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: *Understanding the Roles of Political Trolls in Social Media.* CoNLL 2019

Supervised Learning



Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: *Understanding the Roles of Political Trolls in Social Media.* CoNLL 2019

Distant Supervision: Projecting a Label from a Medium



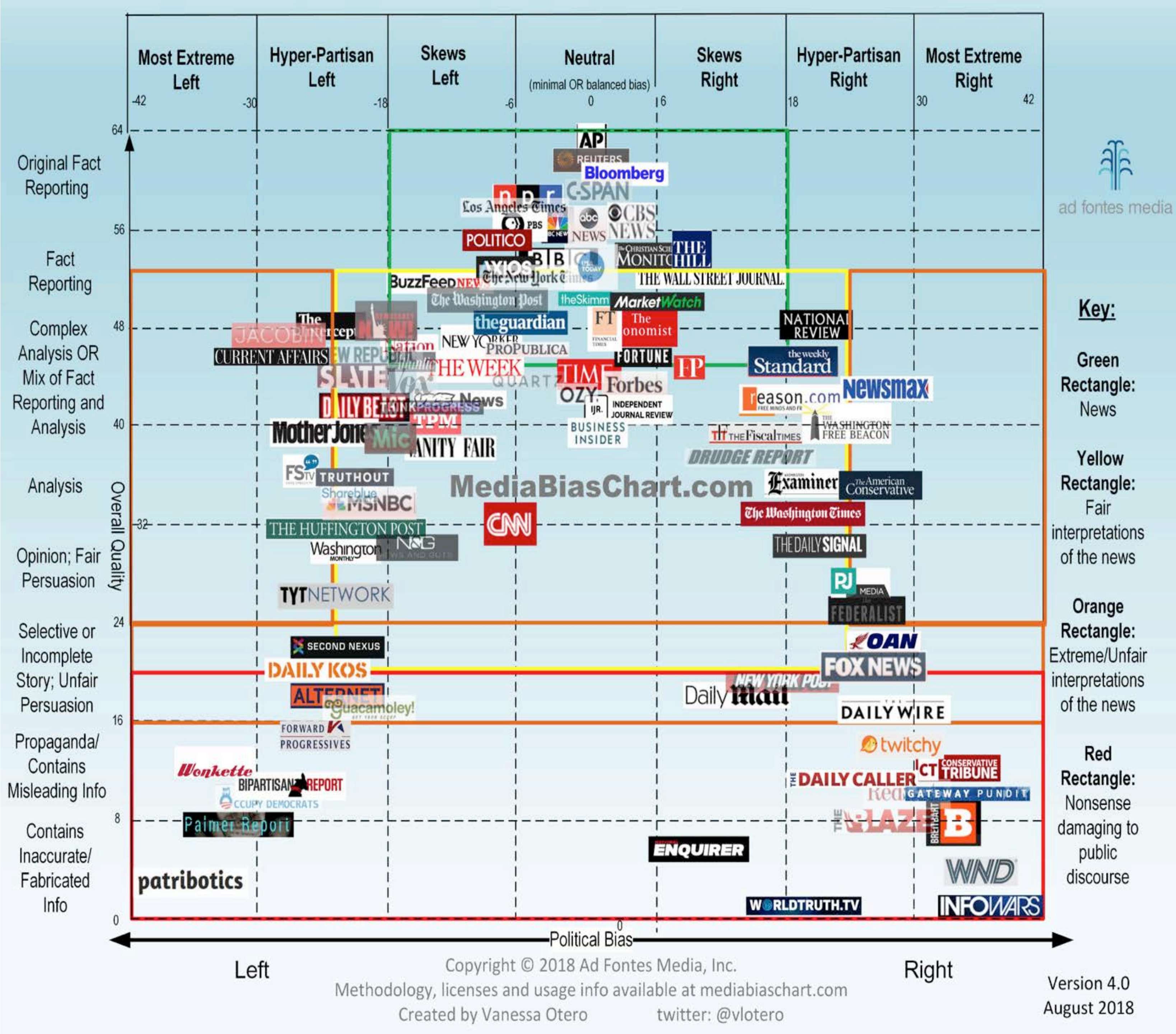
Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: *Understanding the Roles of Political Trolls in Social Media.* CoNLL 2019

Understanding the Roles of Political Trolls in Social Media

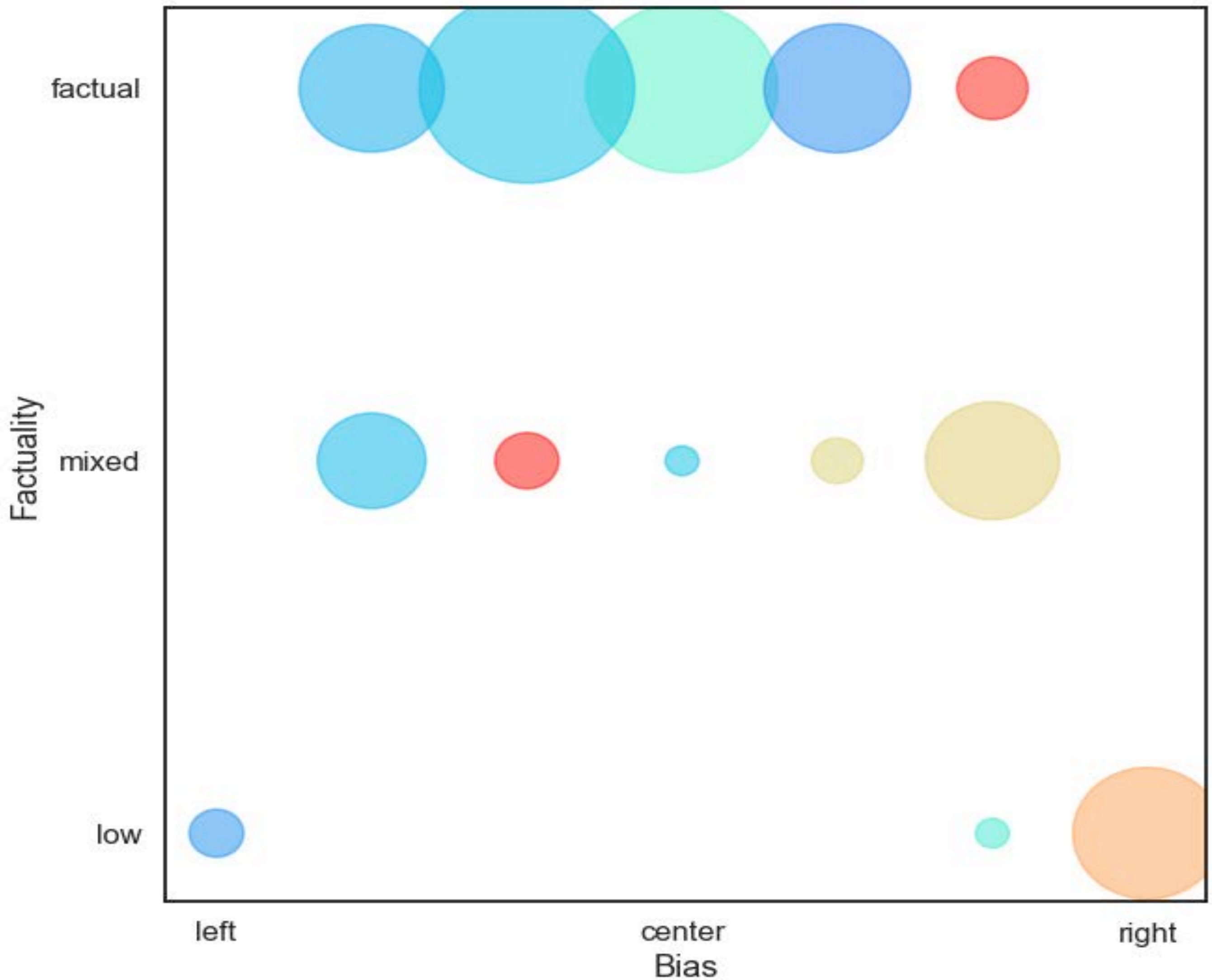
Method	Full Supervision (T1)		Distant Supervision (T2)	
	Accuracy	Macro F1	Accuracy	Macro F1
Baseline (majority class)	68.7	27.1	68.7	27.1
(Kim et al., 2019)	84.0	75.0	N/A	N/A
BERT	86.9	83.1	75.1	60.5
U2H	87.1	83.2	76.3	60.9
U2M	88.1	83.9	77.3	62.4
U2H $\oplus$ U2M	88.3	84.1	77.9	64.1
U2H || U2M	88.7	84.4	78.0	64.6
U2H || U2M || BERT	89.2	84.7	78.2	65.1
U2M $\oplus$ U2H $\oplus$ BERT	89.0	84.4	78.0	65.0
U2M $\oplus$ U2H $\oplus$ BERT + LP1	89.3	84.7	78.3	65.1
U2M $\oplus$ U2H $\oplus$ BERT + LP2	89.6	84.9	78.5	65.7

Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: *Understanding the Roles of Political Trolls in Social Media.* CoNLL 2019

Topical Stance of Media and Twitter Users


Most Extreme Left
Hyper-Partisan Left
Skews Left
Neutral
(minimal OR balanced bias)
Skews Right
Hyper-Partisan Right
Most Extreme Right
-42 -30 -18 -6 0 6 18 30 42
64 56 48 40 32 24 16 8 0
Original Fact Reporting
Fact Reporting
Complex Analysis OR Mix of Fact Reporting and Analysis
Analysis
Opinion; Fair Persuasion
Selective or Incomplete Story; Unfair Persuasion
Propaganda/ Contains Misleading Info
Contains Inaccurate/ Fabricated Info
Overall Quality
Political Bias
Left
Right
MediaBiasChart.com
ad fontes media
Key:
Green Rectangle: News
Yellow Rectangle: Fair interpretations of the news
Orange Rectangle: Extreme/Unfair interpretations of the news
Red Rectangle: Nonsense damaging to public discourse
Copyright © 2018 Ad Fontes Media, Inc.
Methodology, licenses and usage info available at mediabiaschart.com
Created by Vanessa Otero twitter: @vlotero
Version 4.0
August 2018

Bias vs. Factuality in MBFC



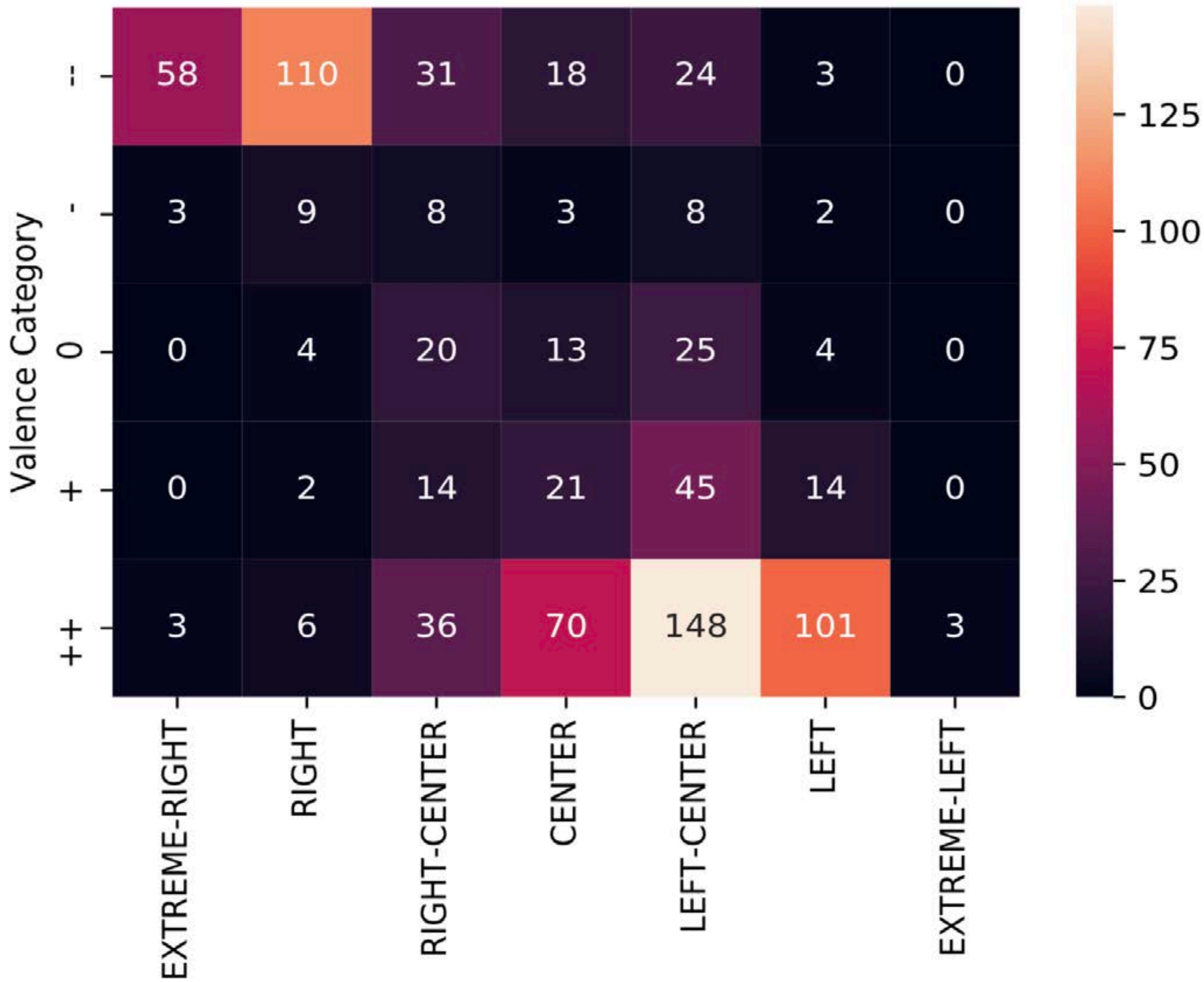
Topical Stance of Media

Topic	Keywords	Date Range	No. of Tweets
Climate change	#greendeal, #environment, #climate, #climatechange, #carbonfootprint, #climatehoax, #climategate, #globalwarming, #agw, #renewables	Feb 25–Mar 4, 2019	1,284,902
Gun control/rights	#gun, #guns, #weapon, #2a, #gunviolence, #secondamendment, #shooting, #massshooting, #gunrights, #GunReformNow, #GunControl, #NRA	Feb 25–Mar 3, 2019	1,782,384
Ilhan Omar remarks on Israel lobby	IlhanOmarIsATrojanHorse, #IStandWithIlhan, #ilhan, #Antisemitism, #IlhanOmar, #IlhanMN, #RemoveIlhanOmar, #ByeIlhan, #RashidaTlaib, #AIPAC, #EverydayIslamophobia, #Islamophobia, #ilhan	Mar 1–9, 2019	2,556,871
Illegal immigration	#border, #immigration, #immigrant, #borderwall, #migrant, #migrants, #illegal, #aliens	Feb 25–Mar 4, 2019	2,341,316
Midterm	midterm, election, elections	Oct 25–27, 2018	520,614
Racism & police brutality	#blacklivesmatter, #bluelivesmatter, #KKK, #racism, #racist, #policebrutality, #excessiveforce, #StandYourGround, #ThinBlueLine	Feb 25–Mar 3, 2019	2,564,784
Kavanaugh Nomination	Kavanaugh, Ford, Supreme, judiciary, Blasey, Grassley, Hatch, Graham, Cornyn, Lee, Cruz, Sasse, Flake, Crapo, Tillis, Kennedy, Feinstein, Leahy, Durbin, Whitehouse, Klobuchar, Coons, Blumenthal, Hirono, Booker, Harris	Sept. 28-30, 2018 & Oct. 6-9, 2018	2,322,141
Vaccination benefits & dangers	#antivax, #vaxxing, #BigPharma, #antivaxxers, #measlesoutbreak, #Antivacine, #VaccinesWork, #vaccine, #vaccines, #Antivaccine, #vaccinestudy, #antivaxx, #provaxx, #VaccinesSaveLives, #ProVaccine, #VaxxWoke, #mykidmychoice	Mar 1–9, 2019	301,209

Kareem Darwish, Michael Aupetit,Peter Stefanov, Preslav Nakov: *Unsupervised User Stance Detection on Twitter.* ICWSM 2020

media	factuality	bias	Average	climate change	gun control	ilhan	immigration	midterm	police & racism	Kavanaugh	vaccine
thehill.com	H	L-C	**+**	0	++	+	+	+	+	++	++
theguardian.com	H	L-C	**++**	++	++	++	++	++	++	++	++
washingtonpost.com	H	L-C	**++**	++	++	++	++	++	++	++	++
breitbart.com	VL	Far R	**−−**	−−	−−	−−	−−	−−	−−	−−	−−
foxnews.com	M	R	**−−**	−−	−−	−−	−−	−−	−−	−−	
nytimes.com	H	L-C	**++**	+	++	+	+	+	++	++	++
cnn.com	M	L	**+**	+	++	+	++	+	+	++	+
apple.news			**+**	0	0	+	0	0	+	+	++
dailycaller.com	M	R	**−−**	−−	−−	−−	−−	−−	−−	−−	
rawstory.com	M	L	**++**	++	++	++	++	++	++	++	++
huffingtonpost.com	H	L	**++**	++	++	++	++	+	++	++	++
truepundit.com	L		**−−**	−−	−−	−−	−−	−−	−−	−−	−−
nbcnews.com	H	L-C	**+**	−−	++	+	++	+	+	++	++
westernjournal.com	M	R	**−−**	−−	−−	−−	−−	−−	−−	−−	
reuters.com	VH	C	**+**	+	++	++	+	+	+	+	++
washingtonexaminer.com	H	R	**−−**	−−	−−	−−	−−	0	−−	−−	
thegatewaypundit.com	VL	Far R	**−−**	−−	−−	−−	−−	−−	−−	−−	
politico.com	H	L-C	**+**	+	+	+	+	++	+	+	++
npr.org	VH	L-C	**+**	0	++	++	++	0	++	++	++
townhall.com	M	R	**−−**	−−	−−	−−	−−	−−	−−	−−	−−
msn.com	H	L-C	**+**	+	+	+	0	++	0	++	0
nypost.com	M	R-C	**−**	−−	0	−	−	+	−−	−	
vox.com	H	L	**++**	++	++	++	++	++	+	++	++
thedailybeast.com	H	L	**++**	++	++	+	++	++	+	++	++
bbc.com	H	L-C	**+**	+	+	++	++	0	+	+	++
independent.co.uk	H	L-C	**++**	++	+	++	++	++	+	++	++
ilovemyfreedom.org	VL	Far R	**−−**	−−	−−	−−	−−	−−	−−	−−	
thinkprogress.org	M	L	**++**	++	++	++	++	++	++	++	++
dailywire.com	M	R	**−−**	−−	−−	−−	−−	−−	−−	−−	++
pscp.tv			**−**	−−	−−	−−	0	−−	0	−	
dailymail.co.uk	VL	R	**−**	−	0	−	−	−	−	−−	−−
msnbc.com	M	L	**++**	++	++	++	++	+	++	++	
dailykos.com	M	L	**++**	++	++	++	++	+	++	++	
bloomberg.com	H	L-C	**+**	+	++	0	++	+	0	+	++
usatoday.com	H	L-C	**+**	+	+	0	+	++	+	0	+

Topical Stance of Media



Kareem Darwish, Michael Aupetit,Peter Stefanov, Preslav Nakov: *Unsupervised User Stance Detection on Twitter.* ICWSM 2020

Propaganda

Why Propaganda?

- “Expression **deliberately** designed to **influence** the opinions /actions of other individuals or groups with reference to predetermined ends.”

  Institute for Propaganda Analysis

- “The rise of the Internet […] has opened the **creation and dissemination of propaganda messages**, which were once the province of states and large institutions, to **a wide variety of individuals and groups**.”

  (Bolsover and Howard, Big Data 5(4))


Weather: Sunny and warm, 75/62 SPORTS • FINAL Monday, September 24, 2007

DAILY NEWS

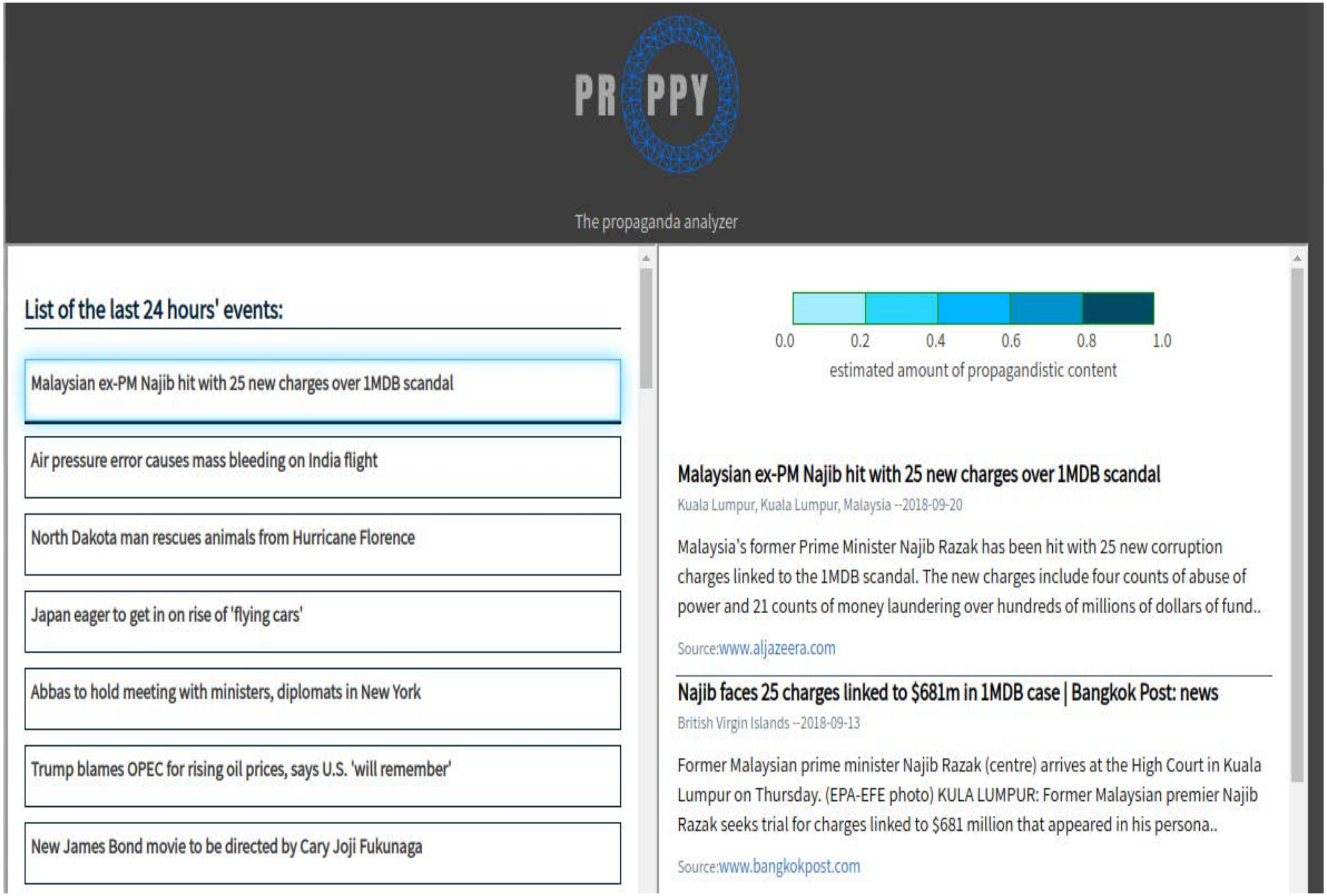
50¢ New York's Hometown Newspaper NYDailyNews.com

THE EVIL HAS LANDED

Mahmoud Ahmadinejad

STORIES ON PAGES 4-7
EDITORIAL PAGE 22

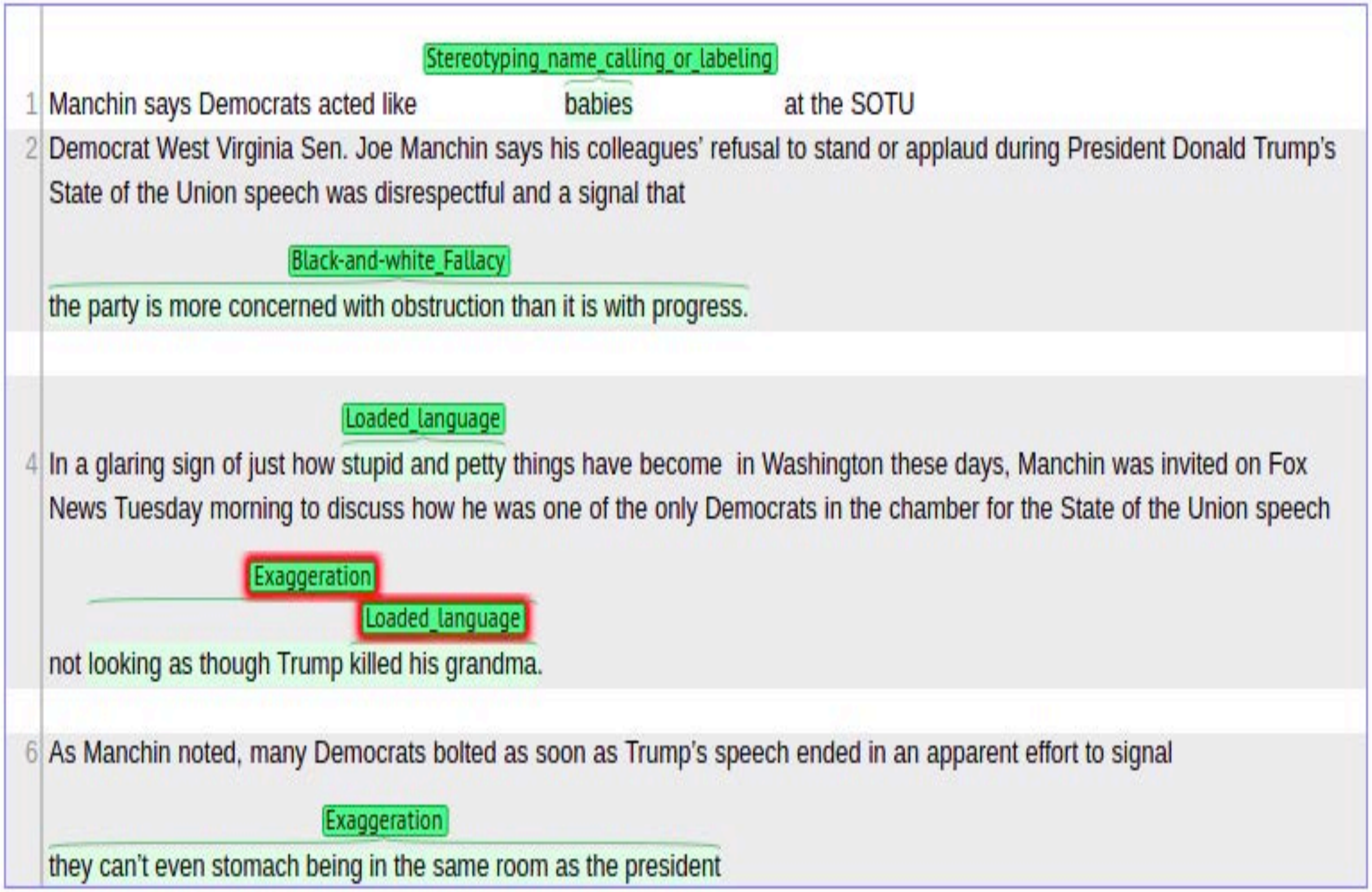
Demo: Proppy http://proppy.qcri.org



Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, Preslav Nakov: *Proppy: A System to Unmask Propaganda in Online News.* AAAI 2019: 9847-9848

reductio ad Hitlerum
whataboutism
thought-terminating cliches
flag-waving
bandwagon
labeling
red herring
causal oversimplification
minimisation
straw men
appeal to authority
obfuscation
exaggeration
name calling
intentional vagueness
black-and-white fallacy
cognitive dissonance
appeal to prejudice
loaded language

Fine-Grained Propaganda Detection



New Dataset

- 18 techniques
- 350k words
- 400 man hours
- 7.3k instances

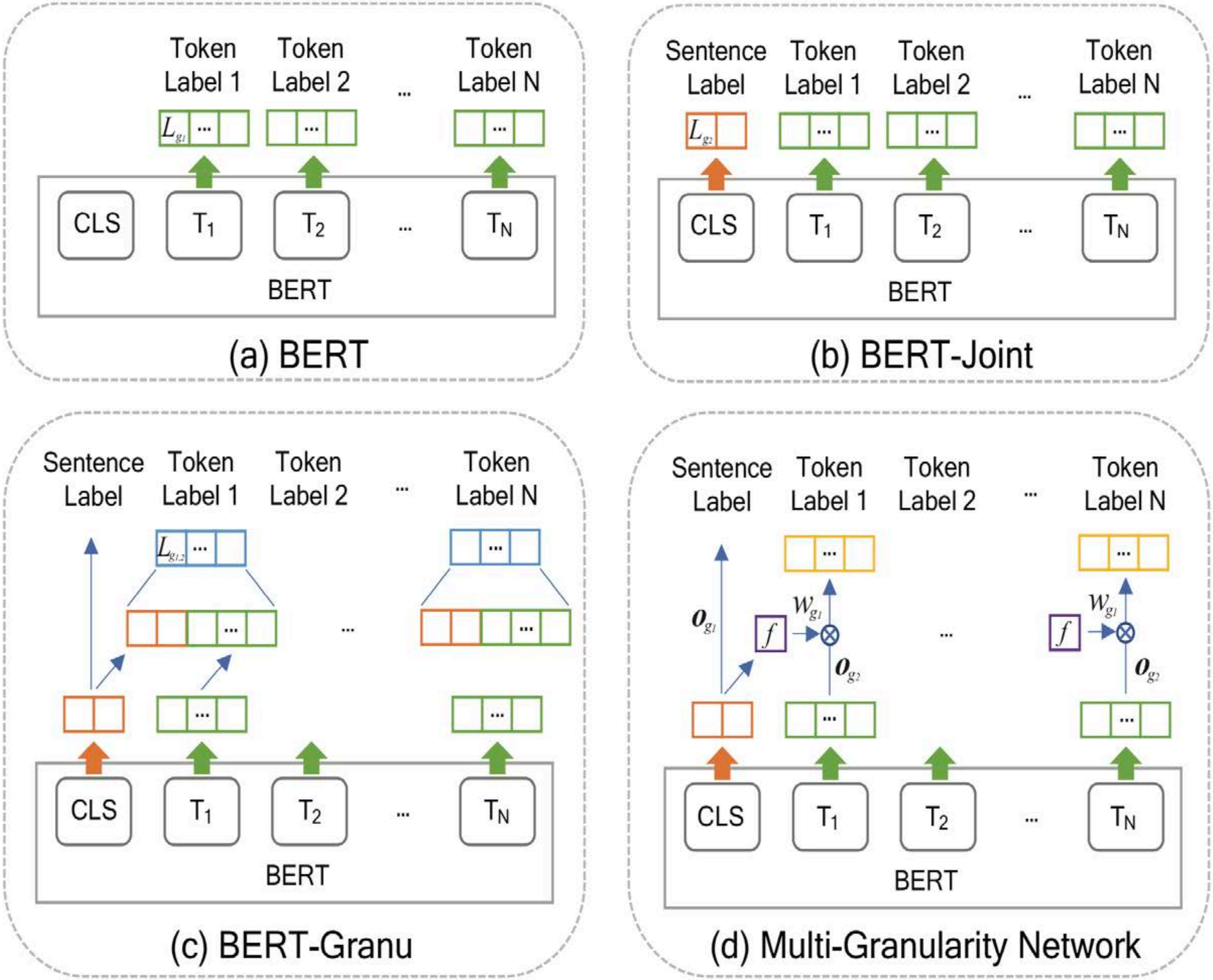
Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, Preslav Nakov
Fine-Grained Analysis of Propaganda in News Articles. EMNLP 2019

Technique	Snippet
`loaded language`	until forced to act by **a worldwide storm of outrage**.
`name calling, labeling`	dismissing the protesters as **lefties** and hugging Barros publicly
`repetition`	Farrakhan repeatedly refers to Jews as **Satan**. He states to his audience [...] call them by their real name, '**Satan**.'
`exaggeration, minimization`	heal the situation of **extremely grave** immoral behavior
`doubt`	**Can the same be said for the Obama Administration**?
`appeal to fear/prejudice`	**A dark, impenetrable and irreversible winter of persecution of the faithful by their own shepherds will fall**.
`flag-waving`	conflicted, and **his 17 Angry Democrats that are doing his dirty work are a disgrace to USA**! —Donald J. Trump
`flag-waving`	attempt (Mueller) **to stop the will of We the People**!!! It's time to jail Mueller
`causal oversimplification`	he said **The people who talk about the "Jewish question" are generally anti-Semites**. Somehow I don't think
`causal oversimplification`	will not be reversed, **which leaves no alternative as to why God judges and is judging America today**
`slogans`	**BUILD THE WALL!"** Trump tweeted.
`appeal to authority`	**Monsignor Jean-Franois Lantheaume, who served as first Counsellor of the Nunciature in Washington, confirmed that "Vigan said the truth. Thats all"**
`black-and-white fallacy`	Francis said these words: **Everyone is guilty for the good he could have done and did not do ...If we do not oppose evil, we tacitly feed it**.
`thought-terminating cliches`	**I do not really see any problems there.** Marx is the President
`whataboutism`	President Trump —**who himself avoided national military service** in the 1960's— keeps beating the war drums over North Korea
`reductio ad hitlerum`	"Vichy journalism," a term which now fits so much of the mainstream media. **It collaborates in the same way that the Vichy government in France collaborated with the Nazis.**
`red herring`	It describes the tsunami of vindictive personal abuse that has been heaped upon Julian from well-known journalists, many claiming liberal credentials. The Guardian, **which used to consider itself the most enlightened newspaper in the country**, has probably been the worst.
`bandwagon`	He tweeted, "**EU no longer considers #Hamas a terrorist group. Time for US to do same.**"
obfusc., int. vagueness, confusion	**The cardinal's office maintains that rather than saying "yes," there is a possibility of liturgical "blessing" of gay unions, he answered the question in a more subtle way without giving an explicit "yes."**
`straw man`	"Take it seriously, but with a large grain of salt." **Which is just Allen's more nuanced way of saying: "Don't believe it**."

Propagandistic News Outlets and Number of Articles

News Outlet	**#**	**News Outlet**	**#**
Freedom Outpost	133	The Remnant Magazine	14
Frontpage Magazine	56	Breaking911	11
shtfplan.com	55	truthuncensored.net	8
Lew Rockwell	26	The Washington Standard	6
vdare.com	20	www.unz.com	5
remnantnewspaper.com	19	www.clashdaily.com	1
Personal Liberty	18		

Propaganda Technique	**inst**	**avg. length**
loaded language	2,547	23.70 ± 25.30
name calling, labeling	1,294	26.10 ± 19.88
repetition	767	16.90 ± 18.92
exaggeration, minimization	571	45.36 ± 35.55
doubt	562	123.21 ± 97.65
appeal to fear/prejudice	367	93.56 ± 74.59
flag-waving	330	61.88 ± 68.61
causal oversimplification	233	121.03 ± 71.66
slogans	172	25.30 ± 13.49
appeal to authority	169	131.23 ± 123.2
black-and-white fallacy	134	98.42 ± 73.66
thought-terminating cliches	95	34.85 ± 29.28
whataboutism	76	120.93 ± 69.62
reductio ad hitlerum	66	94.58 ± 64.16
red herring	48	63.79 ± 61.63
bandwagon	17	100.29 ± 97.05
obfusc., int. vagueness, confusion	17	107.88 ± 86.74
straw man	15	79.13 ± 50.72
all	7,485	46.99 ± 61.45



Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, Preslav Nakov
Fine-Grained Analysis of Propaganda in News Articles. EMNLP 2019

Results: Fragment-Level

Model	Spans			Full Task		
	P	R	F_1	P	R	F_1
BERT	39.57	36.42	37.90	21.48	**21.39**	21.39
Joint	39.26	35.48	37.25	20.11	19.74	19.92
Granu	43.08	33.98	37.93	23.85	20.14	21.80
Multi-Granularity						
ReLU	43.29	34.74	38.28	23.98	20.33	21.82
Sigmoid	**44.12**	**35.01**	**38.98**	**24.42**	21.05	**22.58**

Results: Sentence-Level

Model	Precision	Recall	F1
All-Propaganda	23.92	1.00	38.61
BERT	**63.20**	53.16	57.74
BERT-Granu	62.80	55.24	58.76
BERT-Joint	62.84	55.46	58.91
MGN Sigmoid	62.27	59.56	60.71
MGN ReLU	60.41	**61.58**	**60.98**

- Long press release: https://www.datasciencesociety.net/global-datathon-aims-at-detecting-the-use-of-propaganda-in-the-news/
- Summary: https://www.datasciencesociety.net/hack-news-datathon/
- Detailed description: https://www.datasciencesociety.net/hack-news-datathon-case-propaganda-detection/
- Leaderboard: https://www.datasciencesociety.net/events/hack-the-news-datathon-2019/leaderboard/

Hack The News Datathon (January 2019): 250 participants from 50 countries

TASK RULES/DATES REGISTER RESULTS ORGANISERS

EMNLP-IJCNLP 2019

Passcode | Team Page

https://propaganda.qcri.org/nlp4if-shared-task/

Home | Call For Task Proposals | Tasks | CodaLab

SemEval-2020

International Workshop on Semantic Evaluation

Sponsored by SIGLEX

http://alt.qcri.org/semeval2020/index.php?id=tasks

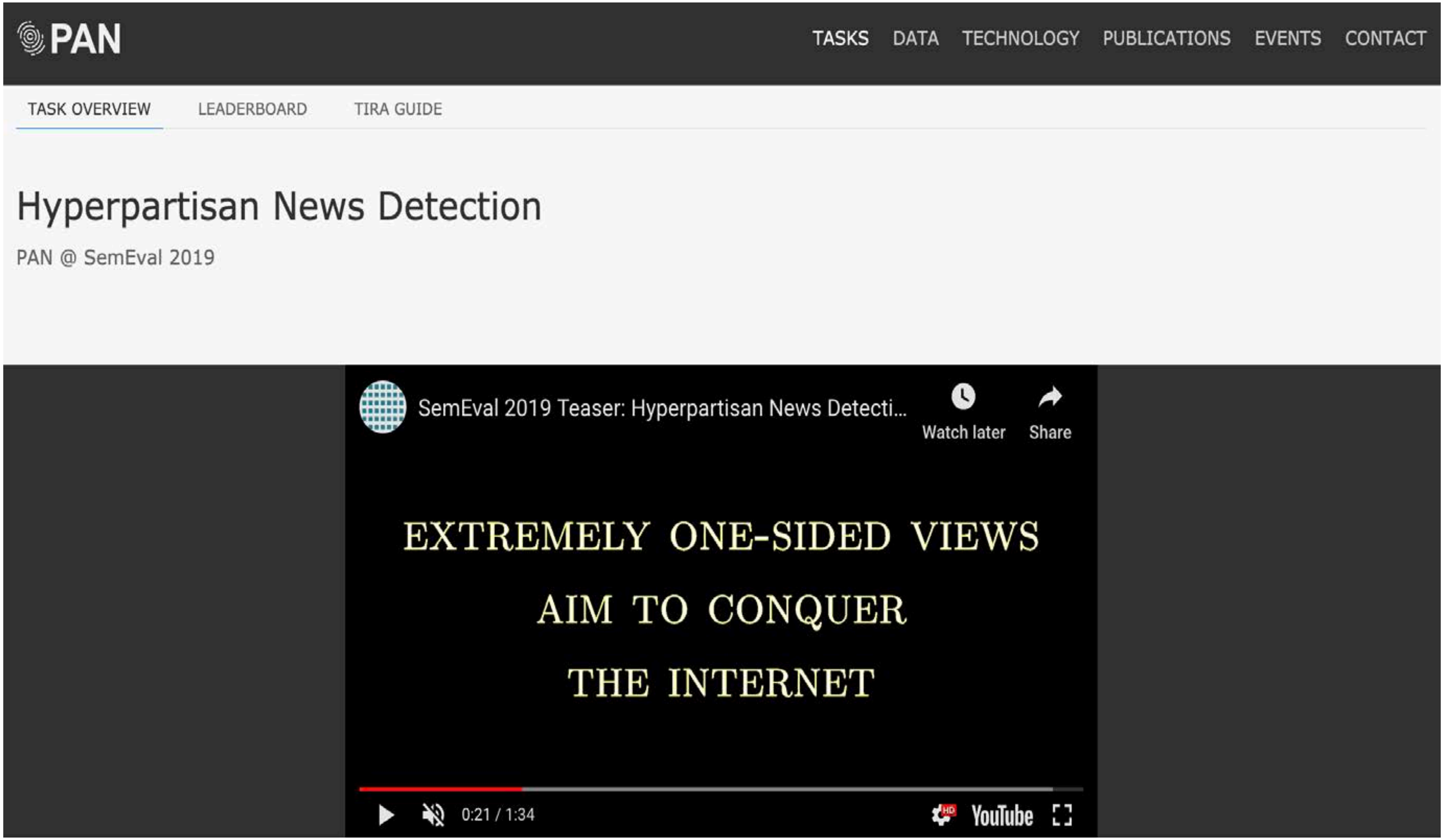
Tasks

We are pleased to announce the following tasks in SemEval-2020.

Task 11: Detection of Propaganda Techniques in News Articles

Related PAN Task at SemEval

https://pan.webis.de/semeval19/semeval19-web/



Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, Martin Potthast: *SemEval-2019 Task 4: Hyperpartisan News Detection.* SemEval@NAACL-HLT 2019: 829-839

The Role of Education

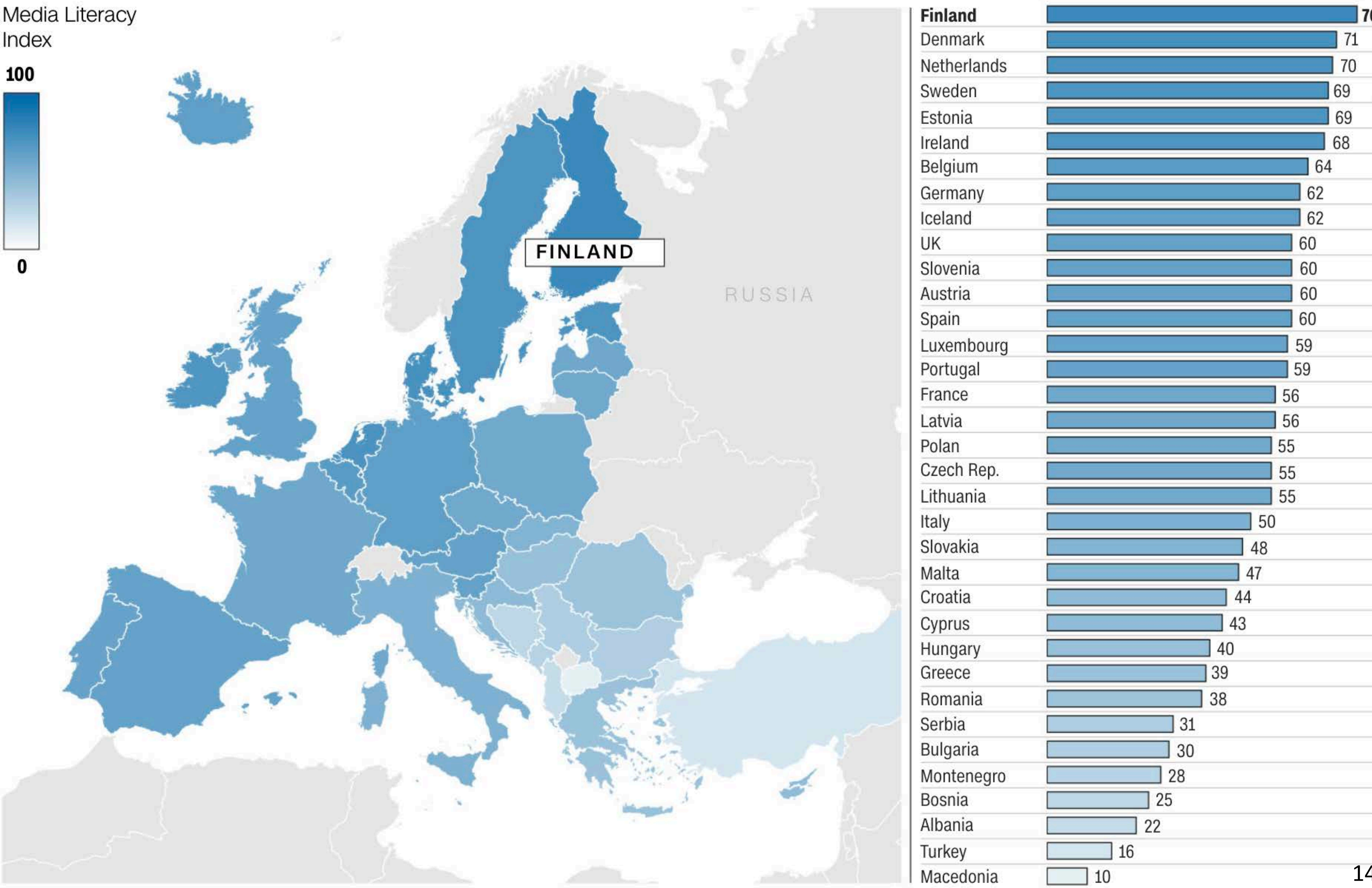
Finland is winning the war on fake news. What it's learned may be crucial to Western democracy

By Eliza Mackintosh, CNN
Video by Edward Kiernan, CNN

https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

Media literacy across Europe

Finland ranked first out of 35 countries in a study measuring resilience to the post-truth phenomenon



Source: European Policies Initiative, Open Society Institute - Sofia, March 2018
Graphic: Henrik Pettersson, CNN

https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

"Propaganda becomes ineffective the moment we are aware of it."

Joseph Goebbels (1897-1945)

The Tanbih Project

Tanbih: News Aggregator

Show stance, bias, propaganda in the news.

Promote different viewpoints, engage users.

Limit the effect of disinformation.

Highlights:

- **Disinformation-aware news aggregator**
- **Media profiles:** can fact-check the news before they were even written
- **Fine-grained propaganda analysis:** adversarial attacks are very hard
- **Focus on MENA:** media, events, languages

Collaborations

MIT CSAIL

SOFIA UNIVERSITY
ST. KLIMENT OHRIDSKI
EST. 1888

Data Science Society

DATA PRO

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY
عضو في مؤسسة قطر
Member of Qatar Foundation

جامعة قطر
QATAR UNIVERSITY

Carnegie Mellon University Qatar

NORTHWESTERN UNIVERSITY IN QATAR

UNIVERSITY OF TEXAS ARLINGTON

ALJAZEERA

AP
Associated Press

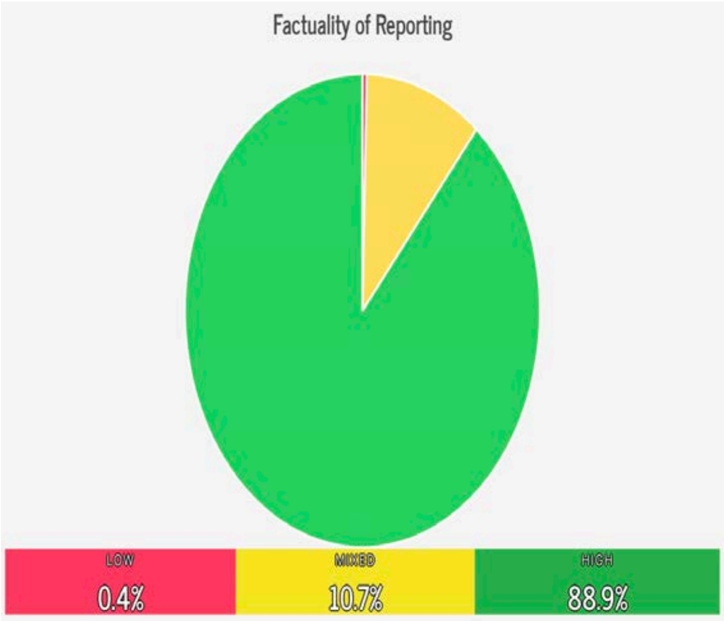
RTÉ

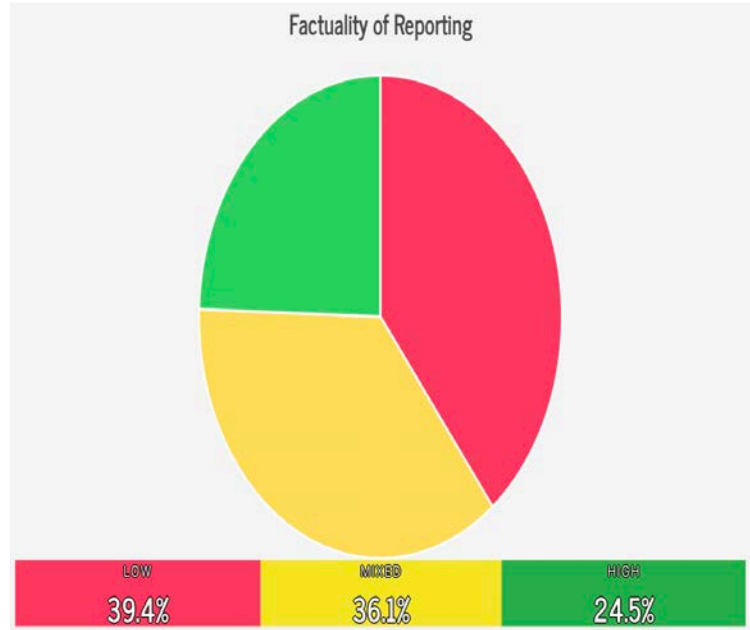
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

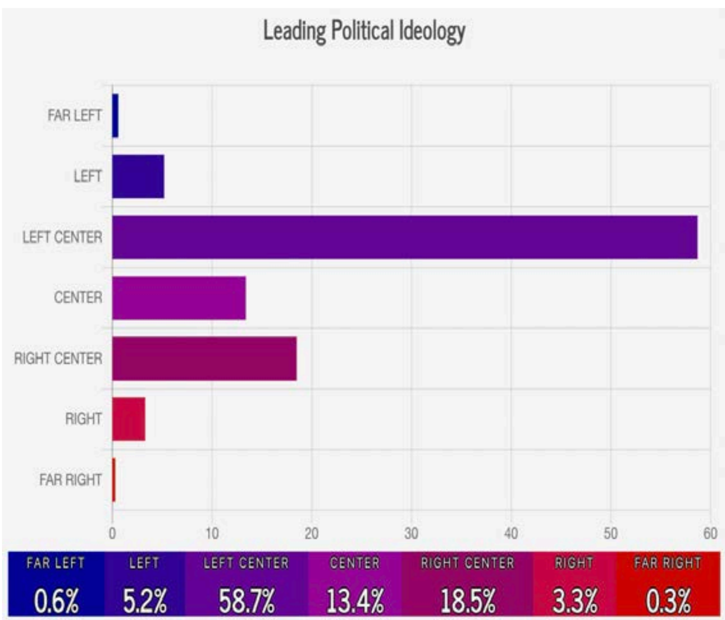
Tanbih: App

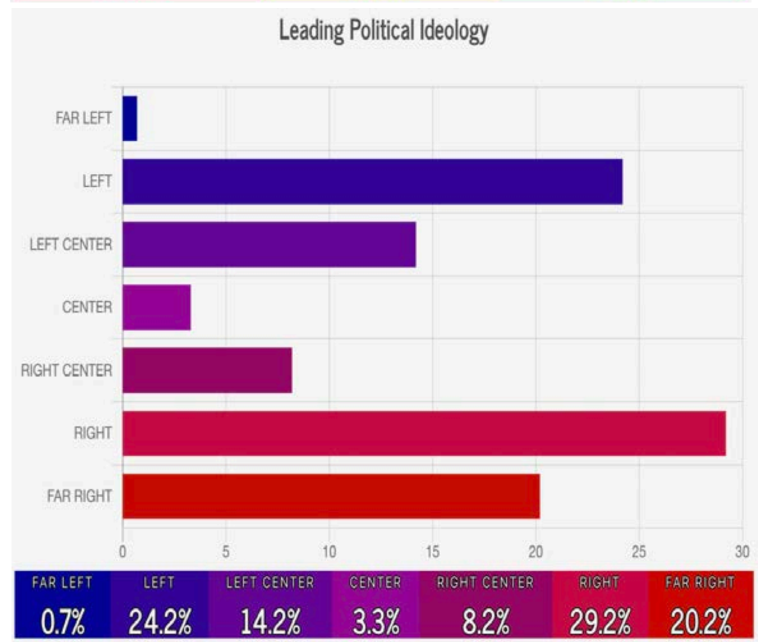
The New York Times

BREITBART









TANBIH
أخبار تنبيه

English | Arabic

http://www.tanbih.org

BUSINESS SPORTS POLITICS TECH & SCI

an hour ago

HEALTH

"We're OK" says British colleague of eight climbers feared dead in India

NODE

1 of 2

2 hours ago

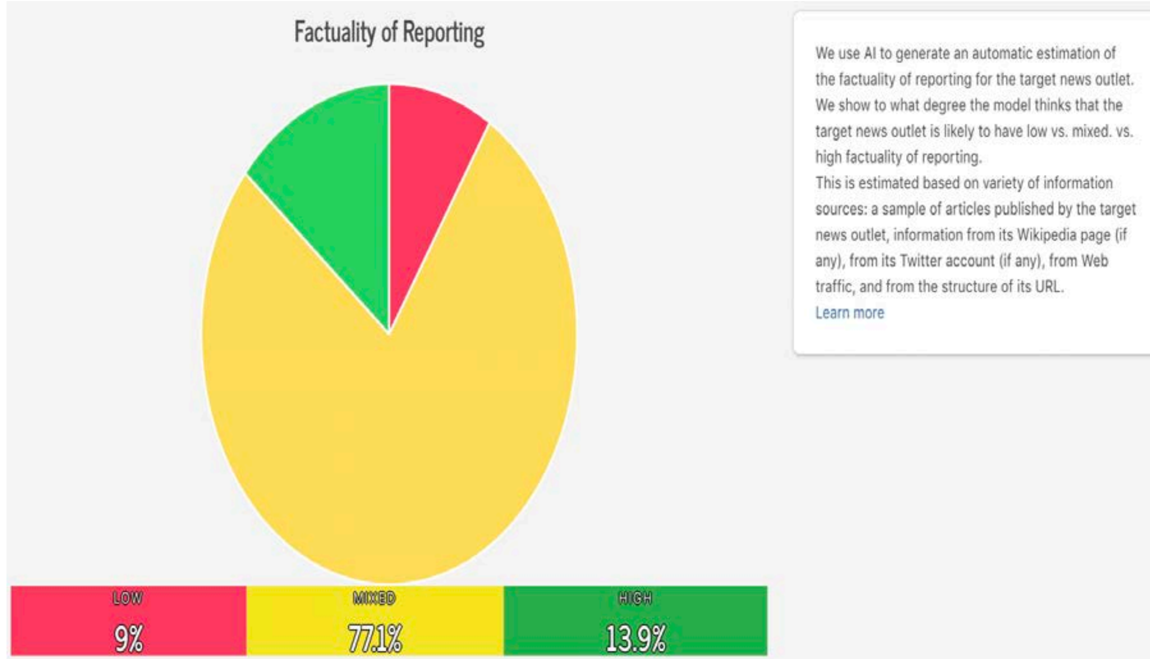
LEGAL

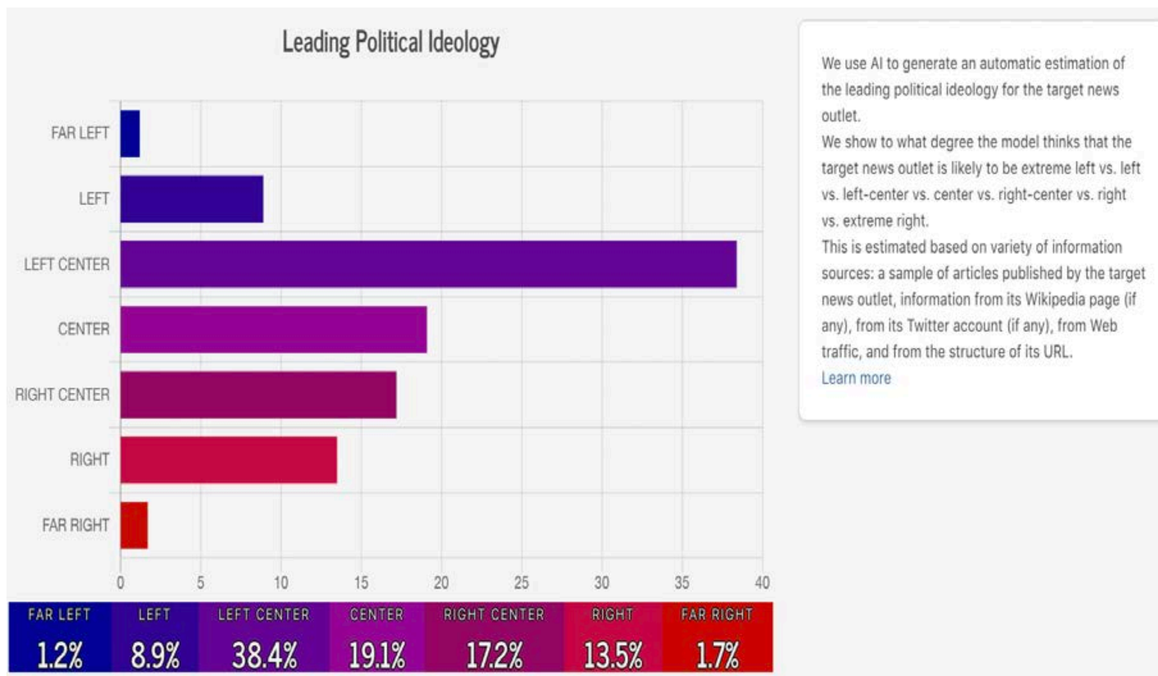
Propaganda? some

'Nothing Short of Madness': Netanyahu Accused of Trying to Turn Israel Into Iran

MIDDLEEAST

1 of 2





Try Tanbih: http://www.tanbih.org

TANBIH

أخبار تنبيه

Try Tanbih: http://www.tanbih.org

This work is part of the Tanbih project, developed in collaboration between the QCRI and MIT-CSAIL, with the aim to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading.

We are hiring postdocs and scientists!

References

- Atanas Atanasov, Gianmarco De Francisci Morales and Preslav Nakov: Understanding the Roles of Political Trolls in Social Media. CoNLL 2019
- Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, Preslav Nakov: Proppy: A System to Unmask Propaganda in Online News. AAAI 2019: 9847-9848
- Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, Preslav Nakov: Fine-Grained Analysis of Propaganda in News Articles. EMNLP 2019
- Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, Yelena Mejova: Seminar Users in the Arabic Twitter Sphere. SocInfo (1) 2017: 91-108
- Kareem Darwish, Michael Aupetit,Peter Stefanov, Preslav Nakov: Unsupervised User Stance Detection on Twitter. ICWSM 2020
- Todor Mihaylov, Georgi Georgiev, Preslav Nakov: Finding Opinion Manipulation Trolls in News Community Forums. CoNLL 2015: 310-314
- Todor Mihaylov, Ivan Koychev, Georgi Georgiev, Preslav Nakov: Exposing Paid Opinion Manipulation Trolls. RANLP 2015: 443-450
- Todor Mihaylov, Preslav Nakov: Hunting for Troll Comments in News Community Forums. ACL (2) 2016
- Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, Ivan Koychev: The dark side of news community forums: opinion manipulation trolls. Internet Research 28(5): 1292-1312 (2018)
- Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, Ivan Koychev: Do Not Trust the Trolls: Predicting Credibility in Community Question Answering Forums. RANLP 2017: 551-560