

Marie Skłodowska Curie,

Research and Innovation Staff
Exchange (RISE)



European
Commission

Ref. Ares(2018)3481899 - 30/06/2018

Horizon 2020
European Union funding
for Research & Innovation

ENhancing seCurity and privAcy in the Social wEb: a user-centered approach for the protection of minors



WP5– Fake activity detection and suppression Deliverable D5.1 “Software libraries built on Graphos.ml for detection of fake activity in large scale OSNs”

Editor(s): Ilias Leontiadis (TID)

Author(s): Ilias Leontiadis, Nicolas Kourtelis, Ioannis Arapakis (TID), Emiliano De Cristofaro, Gullermo Tangil, Jeremiah Onalapo, Lucky Onwuzurike, Gianluca Stringhini, Enrico Mariconti, Tristan Caufield (UCL), Savvas Zannetou, Michael Sirivianos, Costas Tziouvas, Antonis Papasavva, Herodotos Herodotou (CUT), Dimitrianos Savva (LST)

Dissemination Level: Public

Nature: Report

Version: 0.6










PROPRIETARY RIGHTS STATEMENT

This document contains information, which is proprietary to the ENCASE Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to any third party, in whole or in parts, except with prior written consent of the ENCASE consortium.

ENCASE Project Profile

Contract Number	691025
Acronym	ENCASE
Title	ENhancing seCurity and privacy in the Social wEb: a user-centered approach for the protection of minors
Start Date	Jan 1 st , 2016
Duration	48 Months

Partners

	Cyprus University of Technology	Cyprus
	Telefonica Investigacion Y Desarrollo SA	Spain
	University College London	United Kingdom
	Cyprus Research and Innovation Center, Ltd	Cyprus
	SignalGenerix Ltd	Cyprus
	Aristotle University	Greece
	Innovators, AE	Greece
	Universita Degli Studi, Roma Tre	Italy
	LSTech	Spain

Document History

AUTHORS

- (CUT) Savvas Zannettou, Costas Tziouvas, Michael Sirivianos, Antonis Papasavva, Herodotos Herodotou
- (UCL) Emiliano De Cristofaro, Gullermo Tangil, Jeremiah Onalapo, Lucky Onwuzurike, Gianluca Stringhini, Enrico Mariconti, Tristan Caufield
- (TID) Ilias Leontiadis, Nicolas Kourtelis, Ioannis Arapakis
- (LST) Dimitrianos Savva

VERSIONS

Version	Date	Author	Remarks
0.1	2.05.2018	TID	Initial Table of Contents (TOC)
0.2	15.05.2018	TID	Complete TOC
0.3	10.6.2018	All authors	Contributions from partners received
0.4	15.6.2018	TID	Revision and restructure
0.5	27.6.2018	TID	Final editing and restructure
0.6	29.6.2018	TID, CUT	Final version

Executive Summary

Online Social Networks (OSNs) play an important role in the way that people communicate and consume information. This is mainly because OSNs provide an ideal environment for communication and information acquisition, as users have access to a staggering amount of posts and articles that can share with others in real-time. Unfortunately, OSNs have also become the mechanism for massive campaigns to diffuse false information. In particular, recent reporting has highlighted how OSNs are exploited by powerful actors, potentially even state level, in order to manipulate individuals via targeted disinformation campaigns.

The extensive dissemination of false information in OSNs can pose a major problem, affecting society in extremely worrying ways. For example, false information can hurt the image of a candidate, potentially altering the outcome of an election. During crisis situations (e.g., terrorist attacks, earthquakes, etc.), false information can cause result in wide spread panic and general chaos. False information diffusion in OSNs is achieved via diverse types of users, which typically have various motives. A diverse set of users are involved in the diffusion of false information, some unwittingly, and some with particular motives. For example, terrorist organizations exploit OSNs to deliberately diffuse false information for propaganda purposes. Malicious users might utilize sophisticated automation tools (i.e., bots) or fake accounts that target specific benign users with the goal of influencing ideology. No matter the motivation, however, the effects false information has on society clearly indicate the need for better understanding, measurement, and mitigation of false information in OSNs.

In this deliverable, we provide a taxonomy of the false information ecosystem, we attempt to understand how misinformation is spread and we describe our methodology towards detecting such attempts.

Furthermore, all detection mechanisms have been implemented using state-of-the-art machine and deep learning methods (using keras, tensorflow, Theano, sk-learn)

Table of Contents

Executive Summary.....	4
1. Introduction	8
2 The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans	9
2.1. Project description and motivation	9
2.2. Taxonomy.....	10
2.2.1 Types of False Information.....	10
2.2.2 False Information Actors.....	11
2.2.3 Motives behind false information propagation	12
2.3. Line of Works and Future Directions	13
2.3.1 User Perception	13
2.3.2 Propagation of False Information	13
2.3.3 Detection and Containment of False Information	13
2.4. False Information In the political stage	14
2.5. Conclusion.....	14
2.6. References	15
3. The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube	18
3.1. Project description and motivation	18
3.2. Methodology.....	19
3.3. Data Acquisition	19
3.4. Clickbait Detection Model.....	20
3.4.1 Processed Modalities	20
3.4.2 Model Formulation	20
3.5. Results.....	23
3.5.1 Ground truth Analysis	23
3.5.2 Clickbait Detection Model Evaluation.....	27
3.6. Conclusion.....	28
3.7. References	29
4. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?	31
4.1. Project description and motivation	31
4.2. Background	32
4.3. Analysis	33

4.3.1	Ranking of users.....	33
4.3.2	Posts Analysis.....	34
4.4.	Conclusion.....	38
4.5.	References	38
5.	Understanding Web Archiving Services and Their (Mis)Use on Social Media.....	40
5.1.	Project description and motivation	40
5.2.	Background	41
5.3.	Datasets	42
5.4.	Results.....	43
5.4.1	URL characterization	43
5.4.2	Original Content Availability	45
5.4.3	User Base.....	46
5.4.4	Ad Revenue Deprivation	48
5.5.	Conclusion.....	50
5.6.	References	51
6.	Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web	53
6.1.	Project description and motivation	53
6.2.	Dataset	54
6.3.	Analysis	54
6.3.1	General Characterization	55
6.3.2	Account Evolution	58
6.3.3	Influence Estimation	60
6.4.	Conclusion.....	65
6.5.	References	66
7.	The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources	67
7.1.	Project description and motivation	67
7.2.	Methodology.....	68
7.3.	Results.....	70
7.3.1	Temporal Analysis.....	70
7.3.2	Influence Estimation	72
7.3.3	Hawkes Processes	73

7.3.4	Methodology.....	75
7.3.5	Influence Estimation Results.....	76
7.4.	Conclusion.....	80
7.5.	References	81
8.	Youtube Raids	83
8.1.	Project description and motivation	83
8.2.	Methodology.....	84
8.2.1	General Overview	84
8.2.2	Feature Engineering.....	86
8.2.3	Prediction Models.....	87
8.3.	Results.....	90
8.3.1	Experimental Setup.....	90
8.3.2	Experimental Results.....	91
8.3.3	Choosing an Ensemble	93
8.4.	Conclusion.....	94
8.5.	References	95
10.	Conclusions	99
11.	Copyright and Intellectual Property.....	99

1. Introduction

Over the past few years, a number of high-profile conspiracy theories and false stories have originated and spread on the Web. Due to the negligible cost of distributing information over social media, fringe sites can quickly gain traction with large audiences. At the same time, the explosion of information sources also hinders the effective regulation of the sector, while further muddying the water when it comes to the evaluation of news information by readers.

While there are many plausible motives for the rise in alternative narratives the manner in which they proliferate throughout the Web is still unknown. In this work, we address this gap by providing the first largescale measurement of how mainstream and alternative news flows through multiple social media platforms.

Furthermore, we implement state-of-the-art machine and deep learning models in order to detect such posts.

Our work is summarized in the following parts:

- In section 2, we provide a taxonomy of the false information ecosystem.
- In section 3, we describe our methodology to detect and characterize clickbait on YouTube.
- In section 4, we present the first work that conducts a characterization of the Gab social network.
- Section 5, we demonstrate how Web Archiving Services and are misused to propagate false information.
- Section 6 helps us to understand state-sponsored trolls on Twitter and their influence.
- Section 7 also help us to understand misinformation but studying how Web communities influence each other.
- Finally, section 8 sheds some light on YouTube raids and introduces a state-of-the-art methodology that aims to put an end to them.

2 The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans

2.1. Project description and motivation

In this work, we provide a taxonomy of the false information ecosystem that sheds light on the following questions: 1) What are the various types and instances of false information on the Web? 2) Who are the different actors that diffuse false information on the Web? and 3) What are the motives behind the spread of false information? Our taxonomy is built after an extensive study of the existing literature, where we identify the following lines of work regarding the false information ecosystem:

- **User perception of false information.** This refers to how users perceive and interact with false information that is disseminated in OSNs. For example, can users distinguish real stories from fake? If yes, what cues do they use to make this distinction?
- **Propagation dynamics of false information in OSNs.** Understanding the underlying propagation dynamics provides useful insights regarding the problem that can assist in detecting and mitigating false information.
- **Detection and containment of false information.** Detecting and containing the propagation of false information is a desired outcome. However, no robust platform, service or system is in-place that can effectively and efficiently mitigate the problem in a timely manner.
- **False information in politics.** We distinguish works that refer to politics for several reasons. First, anecdotal evidence and existing literature suggest that false information is often disseminated for politics-related reasons. Second, it can affect the course of history, especially in election periods. For instance, during the 2016 US Election, as well as the 2017 French Elections, there were examples where trolls from 4chan tried to change the election outcome by disseminating false information about the candidate they opposed [7, 8]. On top of this, extensive anecdotal evidence suggests the involvement of state-sponsored troll accounts that actively try to mislead and mutate the opinions of users on OSNs [9, 10].

For each of these lines of work, we provide an overview of the most relevant research papers (available at <https://arxiv.org/abs/1804.03461>), as well as possible future research directions that will address existing gaps and will better help the community to alleviate the emerging problem of false information on the Web.

Contributions. In summary, with the present study we make the following contributions: First, we provide a holistic view of the Web’s information ecosystem by analyzing the types of false information as well as the various actors that are involved and their motives. Second, we provide a comprehensive overview of the existing works focusing on the false information ecosystem organized into four high level categories (available at <https://arxiv.org/abs/1804.03461>). This overview can assist the research community in gaining a quick understanding of existing work and identify possible gaps. Finally, for each

category of literature, we provide some future research directions that the community can follow to extend the state-of-art on the Web’s information ecosystem.

2.2. Taxonomy

This section is a synthesis of the literature review in the following sections. We believe this taxonomy will provide a succinct roadmap for future work. The taxonomy is based on [11] and extended to build upon the existing literature. Specifically, we describe the various types of false information that can be found in OSNs, the various types of actors that contribute in the distribution of false information, as well as their motives.

2.2.1 Types of False Information

False information on the Web can be found in various forms, with varying severity metrics. Due to this, we propose the categorization of false information into ten types, broken into three groups based on their severity. Namely, we distinguish the types of false information according to how misleading and hurtful they are for the receiving users.

- (1) **Fake News.** The fake news group is the most severe group of false information that is found on the Web. We identify the following types of false information that fall within this group:
 - **Fabricated [12].** Completely fictional stories disconnected entirely from real facts. This is the most severe type of false information, as their only goal is to mislead their readers.
 - **Propaganda [13].** Consists mostly of fabricated stories that aim to hurt the interest of a particular party. This kind of false news is not new, as it was greatly used also shortly after World War II and the Cold War. Propaganda stories are profoundly utilized in political contexts to mislead people with the overarching goal of inflicting damage to a particular political party or nation-state. Due to this, propaganda is a consequential type of false information as it can change the course of human history (e.g., by changing the outcome of an election).
 - **Imposter [11].** Refer to news stories whose author/source is impersonated. This is particularly misleading type of false information, as one of the main feature that is used to verify the credibility of a news story is its source/author.
 - **Conspiracy Theories [14].** Refer to stories that try to explain a situation or an event by invoking a conspiracy without proof. Usually, such stories are about illegal acts that are carried out by governments or powerful individuals. They also typically present unsourced information as fact or dispense entirely with an “evidence” based approach, relying on leaps of faith instead.
- (2) **Biased/Inaccurate News.** This group refers to news stories that are misleading but incorporate the truth to some extent. Within this group, we identify the following types:
 - **Hoaxes [15].** News stories that contain facts that are either false or inaccurate and are presented as legitimate facts. This category is also known in the research community either as half-truth or factoid stories.
 - **Hyperpartisan [16].** Used primarily in political contexts, hyperpartism refers to stories that are extremely one-sided (i.e., left-wing or right wing). In the entire context of the false information ecosystem, we refer to hyperpartisan as stories that are extremely biased towards a person/party/situation/event. Some examples include the wide spread diffusion of false information to the alt-right community from small fringe Web communities like 4chan’s /pol/ board [29] and Gab, an alt-right echo chamber.

- **Fallacy [17]**. Stories that make use of invalid reasoning in the construction of an argument. Such arguments are deceptive by appearing to be better than they really are.
- (3) **Misleading/Ambiguous News**. This group is the mildest. Usually, news stories that fall within this group do not have serious consequences for the community. For instance, clickbait news articles primarily inflict only frustration on readers.
- **Rumors [18]**. Refers to stories whose truthfulness is ambiguous or never confirmed. This kind of false information is widely propagated on OSNs; hence several works have analyzed this type of false information.
 - **Clickbait [19]**. Refers to the deliberate use of misleading headlines and thumbnails of content on the Web. This type is not new as it appeared years before, during the "newspaper era," a phenomenon known as yellow journalism [20]. However, with the proliferation of OSNs, this problem is rapidly growing, as many users add misleading descriptors to their content with the goal of increasing their traffic for profit or popularity [21]. This is one of the least severe types of false information because if a user reads/views the whole content then he can distinguish if the headline and/or the thumbnail was misleading.
 - **Satire News [22]**. Stories that contain a lot of irony and humor. This kind of news is getting considerable attention on the Web in the past few years. Some popular examples of sites that post satire news are TheOnion and SatireWire. Usually, these sites disclose their satire nature in one of their pages (i.e., About page). However, as their articles are usually disseminated via social networks, this fact is obfuscated, overlooked, or ignored by users who often take them at face value with no additional verification.

2.2.2 False Information Actors

In this section, we describe the different types of actors that constitute the false information propagation ecosystem. We identified a handful of different actors that we describe below.

- **Bots [23]**. In the context of false information, bots are programs that are part of a bot network (Botnet) and are responsible for controlling the online activity of several fake accounts with the aim of disseminating false information. Botnets are usually tied to a large number of fake accounts that are used to propagate false information in the wild. A Botnet is usually employed for profit by 3rd party organizations to diffuse false information for various motives (see Section 2.3 for more information on their possible motives).
- **Criminal/Terrorist Organizations**. Criminal gangs and terrorist organizations are exploiting OSNs as the means to diffuse false information to achieve their goals. A recent example is the ISIS terrorist organization that diffuses false information in OSNs for propaganda purposes. Specifically, they widely diffuse ideologically passionate messages for recruitment purposes. This creates an extremely dangerous situation for the community as there are several examples of individuals from European countries recruited by ISIS that ended-up perpetrating terrorist acts.
- **Political Persons [24]**. Such persons or individuals use OSNs to disseminate false information in order to empower their public image or hurt the public image of other political persons. This problem received extensive attention related to the 2016 US election, where Facebook was accused of allowing the dissemination of false information, perhaps even affecting the outcome of the election.
- **Hidden Paid Posters [25] and State-sponsored Trolls [10]**. They are a special group of users that are paid in order to disseminate false information on a particular content or targeting a specific demographic. Usually, they are employed for pushing an agenda; e.g., to influence people to adopt certain social or business trends. Similar to bots, these

actors disseminate false information for profit. However, this type is substantially harder to distinguish than bots because they exhibit characteristics similar to regular users.

- **Journalists [26].** Individuals that are the primary entities responsible for disseminating information both to the online and to the offline world. However, in many cases, journalists are found in the center of controversy as they post false information for various reasons. For example, they might change some stories so that they are more appealing, in order to increase the popularity of their platform, site or newspaper.
- **Useful Idiots [27].** The term originates from the early 1950s in the USA as a reference to a particular political party’s members that were manipulated by Russia in order to weaken the USA. Useful idiots are users that share false information mainly because they are manipulated by the leaders of some organization or because they are naive. Usually, useful idiots are normal users that are not fully aware of the goals of the organization, hence it is extremely difficult to identify them. Like hidden paid posters, useful idiots are hard to distinguish and there is no study that focuses on this task.
- **Trolls [28].** The term troll is used in great extent by the Web community and refers to users that aim to do things to annoy or disrupt other users, usually for their own personal amusement. An example of their arsenal is posting provocative or off-topic messages in order to disrupt the normal operation or flow of discussion of a website and its users. In the context of false information propagation, we define trolls as users that post controversial information in order to provoke other users or inflict emotional pressure. Traditionally, these actors use fringe communities like Reddit and 4chan to orchestrate organized operations for disseminating false information to mainstream communities like Twitter, Facebook, and YouTube [29, 30].

2.2.3 Motives behind false information propagation

False information actors and types have different motives behind them. Below, we describe the categorization of motives ranked according to maliciousness (from the most malicious to the least).

- **Malicious Intent.** Refers to a wide spectrum of intent that drives actors that want to hurt others in various ways. Some examples include inflicting damage to the public image of a specific person, organization, or entity. This is the most severe type of intent, as the primary goal is to inflict damage to a particular organization, entity or person.
- **Political Influence.** This motive refers to the intent of misleading other people in order to influence their political decisions or manipulate public opinion with respect to specific topics [31]. It includes the enhancement of an individuals’ public image to seek a better result during election period. That said, as noted by [32], political influence usually tries to hurt the public images of an opposing politician or party, rather than promoting the politician or party they support.
- **Profit.** Many actors in the false information ecosystem seek popularity and monetary profit for their organization or website. To achieve this, they usually disseminate false information that increases the traffic on their website. This leads to increased ad revenue that results in monetary profit for the organization or website, at the expense of manipulated users. Some examples include the use of clickbait techniques, as well as fabricated news to increase views of articles from fake news sites that are disseminated via OSNs [33]
- **Passion.** A considerable number of users are passionate about a specific idea, organization, or entity. This affects their judgment and can contribute to the dissemination of false information. Specifically, passionate users are blinded by their ideology and perceive the false information as correct and contribute in its overall propagation [34].
- **Fun.** As discussed in the previous section, online trolls are usually diffusing false information for their amusement. This is the least severe motive, as their intentions are

usually not bad; however, their actions can sometimes inflict considerable damage to other individuals, and thus should not be dismissed.

2.3. Line of Works and Future Directions

2.3.1 User Perception

After reviewing the literature on how users perceive false information on the Web, we identify a few gaps. First, there is a lack of rigorous temporal analysis of user perception around the dissemination of false information and/or conspiracy theories. For example, perceptions might differ during the course of evolution of any particular conspiracy theory. Next, none of the studied reviewed take into consideration the interplay between multiple OSNs. Users on one platform might perceive the same false information differently depending on a variety of factors. For example, certain communities might be focused around a particular topic, affecting their susceptibility to false information on that topic, or the way the platform calls home presents information (e.g., news feed) can potentially influence how it users perceive false information. This issue becomes further muddled when considering users that are active on multiple platforms. In particular, we note that there is a substantial gap with respect to which OSNs have been studied with respect to false information; e.g., YouTube, which has become a key player in information consumption.

2.3.2 Propagation of False Information

Here we provide identified research gaps with respect to the propagation of false information on the Web. First, most of the studies focus on specific communities or events. Therefore, as a future research direction we propose studying the problem from a holistic point-of-view. That is, study how information propagates across multiple communities and fusing information that exists in multiple formats (e.g., images, textual claims, URLs, video, etc.). Furthermore, systems or tools that visualize the propagation of information across OSNs do not exist. These types of tools will enable a better understanding of false information propagation, as well as finding the source of information. Finally, to the best of our knowledge, the propagation of information via orchestrated campaigns has not been rigorously studied by the research community. An example of such a campaign is the posting comments in YouTube video by users of 4chan [29].

2.3.3 Detection and Containment of False Information

Despite the fact that several studies exist attempting to detect and contain false information on the Web, the problem is still emerging and prevalent. This is mainly because the problem requires higher cognitive and context awareness that current systems do not have. To assist in achieving a better detection of false information on the Web, we foresee the following tangible research directions.

First, information on the Web exists in multiple formats, and thus false information is disseminated via textual claims, screenshots, videos, etc. Most studies, however, take into consideration only one format. To achieve a multi-format false information detection system requires correlating information in multiple formats, which in turn requires understanding the similarities and differences in the content each format delivers. To the best of our

knowledge, a system that can meaningfully assist in detecting false information across multiple formats does not exist.

Next, to the best of our knowledge, no prior work has rigorously assessed credibility based on user profiling. For example, a post from an expert on a particular subject should not be treated with the same weight as a post by a typical user. We foresee studying false information detection from a users' perspective is a way forward in effectively detecting and containing the spread of false information on the Web.

Finally, most previous studies focus on detection and containment of false information on a single OSN, but the Web is much larger than any single platform or community. Therefore, future work should address the problem with a holistic view of the information ecosystem. This requires an understanding of how information jumps from one Web community to another, how Web communities influence each other, and how to correlate accounts that exist in multiple Web communities (e.g., how to find that a particular Facebook and Twitter account belong to the same user). Such an understanding will be particularly useful for containing the spread of false information from one Web community to another.

2.4. False Information In the political stage

As future directions for understanding and mitigating the effects of false information on the political stage, we propose the following. First, there is extensive anecdotal evidence highlighting that Web communities are used by state-sponsored troll factories, e.g., the recent news regarding Russian troll factories deployed to influence the outcomes of the 2016 Brexit [35] referendum and the 2016 US presidential election [36]. We thus propose investigating this phenomenon both from the perspective of user analytics as well as societal impact. Additionally, there is a lack of studies providing insight on how politics-related false information is disseminated across the Web; most studies focus on a single Web community or to specific events or do not examine politics. Understanding how political information propagates across the Web will help society identify the source of false information and lead to successful containment efforts.

2.5. Conclusion

In this work, we have presented an overview of the false information ecosystem. Specifically, we have presented the various types of false information that can be found online, the different actors of the false information ecosystem as well as their motives for diffusing controversial information. Through the identification of several line of works, we have presented the existing works on the false information ecosystem. Namely, we have presented works on user perception, propagation dynamics, detection and containment of false information as well as the dynamics of false information on the political stage (only available at <https://arxiv.org/abs/1804.03461>). To conclude, we present some gaps of the existing literature that can be exploited by researchers in order to further study the increasing problem of false information on the Web.

2.6. References

- [1] Facebook's failure: did fake news and polarized politics get Trump elected? <https://www.theguardian.com/technology/2016/nov/10/Facebook-fake-news-election-conspiracy-theories>
- [2] People shared nearly as much fake news as real news on Twitter during the election. <https://qz.com/1090903/people-shared-nearly-as-much-fake-news-as-real-news-on-twitter-during-the-election/>
- [3] How Russian trolls support of third parties could have cost Hillary Clinton the election. <https://qz.com/1210369/russia-donald-trump-2016-how-russian-trolls-support-of-us-third-parties-may-have-cost-hillary-clinton-the-election/>
- [4] How Trump Consultants Exploited the Facebook Data of Millions. <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>
- [5] False smartphone alert of huge earthquake triggers panic in Japan. <https://www.theguardian.com/world/2016/aug/01/false-alert-of-huge-earthquake-triggers-panic-in-japan>
- [6] Samer Al-khateeb and Nitin Agarwal, 2015 Examining botnet behaviors for propaganda dissemination: A case study of isil's beheading videos-based propaganda. In ICDMW.
- [7] French fear Putin and Trump followers are using 4chan to disrupt presidential election. <https://venturebeat.com/2017/05/05/french-fear-putin-and-trump-followers-are-using-4chan-to-disrupt-presidential-election/>
- [8] "We actually elected a meme as president": How 4chan celebrated Trump's victory", https://www.washingtonpost.com/news/the-intersect/wp/2016/11/09/we-actually-elected-a-meme-as-president-how-4chan-celebrated-trumps-victory/?utm_term=.ddf3a0ea9905
- [9] The Independent 2017 StPetersburg troll farm had 90 dedicated staff working to influence US election campaign <https://ind.pn/2yuCQdy>. (2017).
- [10] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn 2018 Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. arXiv preprint arXiv:1801.09288 (2018).
- [11] Fake news. it's complicated. <https://firstdraftnews.com/fake-news-complicated/>.
- [12] Victoria L Rubin, Niall J Conroy, and Yimin Chen. Towards news verification: Deception detection methods for news discourse. 2015.
- [13] Garth S Jowett and Victoria O'donnell 2014 Propaganda & persuasion Sage.

- [14] Mark Fenster 1999 Conspiracy theories: Secrecy and power in American culture U of Minnesota Press.
- [15] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes.
- [16] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein 2017 A Stylometric Inquiry into Hyperpartisan and Fake News. In arXiv preprint arXiv:1702.05638.
- [17] Frans H Van Eemeren, Bart Garssen, and Bert Meuffels 2009 Fallacies and judgments of reasonableness: Empirical research concerning the pragma-dialectical discussion rules. Springer Science & Business Media.
- [18] Warren A Peterson and Noel P Gist 1951 Rumor and public opinion In American Journal of Sociology.
- [19] Yimin Chen, Niall J Conroy, and Victoria L Rubin 2015 Misleading Online Content: Recognizing Clickbait as False News In MDD.
- [20] W Joseph Campbell 2001 Yellow journalism: Puncturing the myths, defining the legacies.
- [21] Politifact. 2017. The more outrageous, the better: How clickbait ads make money for fake news sites. <http://www.politifact.com/punditfact/article/2017/oct/04/more-outrageous-better-how-clickbait-ads-make-mone/>. (2017).
- [22] Clint Burfoot and Timothy Baldwin 2009 Automatic satire detection: Are you having a laugh? In ACL-IJCNLP.
- [23] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripean .2011 The socialbot network: when bots socialize for fame and money. In ACSAC.
- [24] 2016 Lie of the Year: Fakenews <http://www.politifact.com/truth-o-meter/article/2016/dec/13/2016-lie-year-fake-news/>
- [25] Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. 2013. Battling the internet water army: Detection of hidden paid posters. In ASONAM.
- [26] Seow Ting Lee 2004 Lying to tell the truth: Journalists and the social context of deception In Mass Communication & Society.
- [27] Useful Idiot Wiki. http://rationalwiki.org/wiki/Useful_idiot
- [28] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov 2015 Finding Opinion Manipulation Trolls in News Community Forums.In CoNLL.
- [29] Hine, Gabriel Emile, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web." *arXiv preprint arXiv:1610.03452* (2016).

- [30] Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. "The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources." In *Proceedings of the 2017 Internet Measurement Conference*, pp. 405-417. ACM, 2017.
- [31] Kingsley Napley. 2017. The Impact of Fake News: Politics. <https://www.lexology.com/library/detail.aspx?g=6c63091c-e81f-4512-8c47-521eadce65ff>. (2017).
- [32] Budak, Ceren, Sharad Goel, and Justin M. Rao. "Fair and balanced? quantifying media bias through crowdsourced content analysis." *Public Opinion Quarterly* 80, no. S1 (2016): 250-271.
- [33] Adperfect 2017 How fake news is creating profits.<http://www.adperfect.com/how-fake-news-is-creating-profits/>.(2017).
- [34] Eliot Higgins. 2016. Fake news is spiraling out of control - and it is up to all of us to stop it. <https://www.ibtimes.co.uk/fake-news-spiralling-out-control-it-all-us-stop-it-1596911>. (2016).
- [35] Alex Hern Robert Booth, Matthew Weaver and Shaun Walker 2017 Russia used hundreds of fake accounts to tweet about Brexit, data shows. <https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets>. (2017).
- [36] Scott Shane 2017 The Fake Americans Russia Created to Influence the Election <https://www.nytimes.com/2017>

3. The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube

3.1. Project description and motivation

Recently, YouTube surpassed cable TV in terms of popularity within teenagers [1]. This is because YouTube offers a vast number of videos, which are always available on demand. However, because videos are generated by the users of the platform, known as *YouTubers*, a plethora of them are of vague quality.

The ultimate goal of YouTubers is to increase their ad revenue by ensuring that their content will get viewed by millions of users. Several YouTubers deliberately employ techniques that aim to deceive viewers into clicking their videos. These techniques include: (i) use of eye-catching thumbnails, such as depictions of abnormal stuff or attractive adults, which are often irrelevant with the video’s content; (ii) use of headlines that aim to intrigue the viewers; and (iii) encapsulate false information to either the headline, the thumbnail or the video’s content. We refer to videos that employ such techniques as *clickbaits*. The continuous exposure of users to clickbaits cause frustration and degraded user experience.

The clickbait problem is essentially a peculiar form of the well-known spam problem [2], [3], [4]. In spam, malicious users try to deceive legitimate users by sending them misleading messages in order to: (i) propagate false information; (ii) increase their website’s views; (iii) advertise various websites; and (iv) perform attacks (e.g., phishing) by redirecting benign users to malicious websites. Nowadays, the spam problem is not as prevalent as a few years ago due to the deployment of systems that diminish it. Furthermore, users have an increased awareness of typical spam content (e.g., emails, etc.) and they can effortlessly discern it. However, this is not the case for clickbait. Clickbaits usually contain hidden false or highly ambiguous information that users or systems might not be able to perceive. Moreover, the users may lack context awareness that is crucial for identifying clickbait videos.

Recently, the aggravation of the fake news problem has induced broader public attention to the clickbait problem. For instance, Facebook aims at removing clickbaits from its newsfeed [5], [6]. In this work, we focus on YouTube for various reasons: i) we can anecdotally confirm that the problem exists in great extent on YouTube [7] and ii) to the best of our knowledge, YouTube relies on users to flag suspicious videos and then manually review them. To this extent, this approach is deemed to be inefficient. The need for an automated approach that minimizes human intervention is indisputable.

To attain this goal, we leverage some recent advances in the field of Deep Learning [8]. Specifically, we devise a novel formulation of variational autoencoders (VAEs) [9], [10] that fuses different modalities pertaining to a YouTube video, and infers latent correlations and dynamics between them. The proposed model infers for each video a latent variable vector that encodes a high-level representation of the content and the correlations between the various modalities. The significance of learning to compute such concise representations is that: (i) this learning procedure can be robustly performed by leveraging large unlabeled data corpora; and (ii) the obtained representations can be subsequently utilized to drive the classification process, with very limited requirements in labeled data.

To this end, we formulate the encoder part of the devised VAE model as a (2-component) finite mixture model [11]. That is, we consider a set of alternative encoder models that may generate the data pertaining to each video; the decision of which specific encoder (corresponding to one possible class) generates each is obtained via a trainable probabilistic gating network [12], which constitutes an integral part of the developed autoencoder. The whole model is trained in an end-to-end fashion, using the available training data, both the unlabeled and the few labeled ones. The latter are specifically useful for appropriately fitting the postulated gating network that infers the posterior distribution of mixture component (and corresponding class) allocation.

Contributions. We propose a deep generative model that allows for combining data from as *diverse modalities* as video headline text, thumbnail image and tags text, as well as various numerical statistics, including statistics from comments. Most importantly, the proposed model allows for successfully addressing the problem of learning from limited labeled samples and numerous unlabeled ones (semi-supervised learning). This is achieved by postulating a deep variational autoencoder that *employs a finite mixture model as its encoder*. In this context, mixture component assignment is regulated via an appropriate gating network; this also constitutes the eventually obtained classification mechanism of our deep learning system. We provide a large-scale analysis on YouTube, showing that with respect to the collected data its recommendation engine does not consider how misleading a recommended video is; hence recommending clickbait videos to its users.

3.2. Methodology

In this section we describe our data acquisition methodology and the formulation of the proposed clickbait detection model.

3.3. Data Acquisition

By leveraging YouTube’s Data API, between August and November of 2016, we collect metadata of videos published between 2005 and 2016. Specifically, we collected the following data descriptors for 206k videos: (i) basic details like headline, tags, etc.; (ii) thumbnail; (iii) comments from users; (iv) statistics (e.g., views, likes, etc.); and (v) related videos based on YouTube’s recommendation system. Note that for the data collection infrastructure we use a clean slate with respect to user history, hence the recommendations are not affected by any viewing history. We started our retrieval from a popular (108M views) clickbait video [13]. Subsequently, we collected all the related videos as were recommended by YouTube. We iterated the same procedure for all the captured videos, until we reached 206k examples. Note that this approach also enables us to study interesting aspects of the problem, by constructing a graph that captures the relations (recommendations) between videos.

To get labeled data, we opted for two different approaches. First, we manually reviewed a small subset of the collected data by taking into consideration the headline, the thumbnail, comments from users, and video content. Specifically, we watched the whole video and compared it with the thumbnail and headline. A video is considered clickbait only if the thumbnail and headline deviate substantially from its content.

However, this task is both cumbersome and time consuming; thus, we elected to retrieve more data that are labeled. To this end, we compiled a list of channels that habitually employ clickbait techniques and channels that do not (The list of channels is available on <https://goo.gl/ZSabkn>). To obtain the list of channels, we used a pragmatic approach; we found channels that are outed by other users as clickbait channels. For each channel, we retrieved up-to 500 videos, hence creating a larger labeled dataset. The overall labeled dataset consists of (i) 1,071 clickbaits and 955 non-clickbaits obtained from the manual review process and (ii) 8,999 clickbaits and 8,470 non-clickbaits obtained from the distinguished clickbait and non-clickbait channels. The importance of this dataset is two-fold, as it enables us to get insights regarding the problem and is instrumental for training our devised deep learning model.

3.4. Clickbait Detection Model

3.4.1 Processed Modalities

Our model processes the following modalities: 1) **Headline**: For the headline, we consider both the content and the style of the text. For the content of the headline, we use sent2vec embeddings [15] trained on Twitter data. For the style of the text, we use the features proposed in [16]; 2) **Thumbnail**: We scale down the images to 28x28 and convert them to grayscale. This way, we decrease the number of trainable parameters for our developed deep network, thus speeding training time up without compromising achievable performance; 3) **Comments**: We preprocess user comments to find the number of occurrences of words used for flagging videos. We consider the following words: “misleading, bait, thumbnail, clickbait, deceptive, deceiving, clicked, flagged, title”; 4) **Tags**: We encode the tags’ text as a binary representation of the top 100 words that are found in the whole corpus; 5) **Statistics**: We consider various statistics (e.g., views, likes, etc.).

3.4.2 Model Formulation

In Figure 1, we provide a high-level overview of the proposed model. The thumbnail modality is initially processed, at the encoding part of the proposed model, by a CNN [17]. We use a CNN that comprises four consecutive convolutional layers, with 64 filters each, and ReLU activations. The first three of these layers are followed by max-pooling layers. The fourth is followed by a simple densely connected layer, which comprises 32 units with ReLU activations. This initial processing stage allows for learning to extract a high-level, 32-dimensional vector of the raw image thumbnail, which contains the most useful bits of information for driving the classification decision. The fact that we utilize a CNN to extract

these representations offers rotation and translation invariance, and a reduced number of trainable parameters compared to alternatives, like dense-layer deep networks [17].

The overarching goal of the devised model is to limit the required availability of labeled data, while making the most out of large available corpora of (unlabeled) examples. To this end, after this first processing stage, we split the encoding part into two distinct subnetworks (subencoders) that work in tandem. Both these subencoders are presented with the aforementioned 32-dimensional thumbnail representation, fused with the data stemming from all the other available modalities. This results in an 855-dimensional input vector, first processed by one dense layer network (Fusing Network) that comprises 300 ReLU units. Due to its large number of parameters, this dense layer network may become prone to overfitting; hence we regularize using the prominent Dropout technique [18]. We use a Dropout level of $d = 0.5$; this means that, at each iteration of the training algorithm, 50% of the units are randomly omitted from updating their associated parameters. The so-obtained 300-dimensional vector, say $h()$, is the one eventually presented to both the postulated subencoders.

The rationale behind this novel configuration is motivated by a key observation; the two modeled classes of clickbaits and non-clickbaits are well-expected to entail significantly different patterns of correlations and latent underlying dynamics between the processed modalities. Hence, it is plausible that each class can be adequately and effectively modeled by means of distinct, and quite different, encoder distributions, inferred by means of the two encoder subnetworks. Each of these subnetworks are dense-layer networks comprising a hidden layer with 20 ReLU units, and an output layer with 10 units.

Since the devised model constitutes a VAE, the output units of the postulated subencoders are of a stochastic nature; specifically, we consider stochastic outputs, say \tilde{z} and \hat{z} , with Gaussian (posterior) densities, as usual in the literature of VAEs [9], [10]. Hence, what the postulated subencoder networks actually compute are the means, $\tilde{\mu}$ and $\hat{\mu}$, as well as the (diagonal) covariance matrices, $\tilde{\sigma}^2$ and $\hat{\sigma}^2$, of these Gaussian posteriors. On this basis, the actual subencoder output vectors, \tilde{z} and \hat{z} , are sampled each time from the corresponding (inferred) Gaussian posteriors, in a straightforward fashion. Note that our modeling selection of sharing the initial CNN-based processing part between the two subencoders allows for significantly reducing the number of trainable parameters, without limiting the eventually obtained modeling power.

Under this mixture model formulation, we need to establish an effective mechanism for inferring which observations (YouTube videos) are more likely to match the learned distribution of each component subencoder. This is crucial for effectively selecting between the samples of \tilde{z} or \hat{z} at the output of the encoding stage of the devised model. In layman terms, this can be considered to be analogous to a (soft) classification mechanism. This

mechanism can be obtained by computation of the posterior distribution of mixture component membership of each video (also known as "responsibility" in the literature of finite mixture models [19]). To allow for effectively inferring this posterior distribution, in this work we postulate an appropriate gating network. This is a dense-layer network, which comprises one hidden layer with 100 ReLU units, and is presented with the same vector, $h()$, as the two postulated subencoders. It is trained alongside the rest of the devised model, and it essentially constitutes the only part of the model that requires availability of labeled data for its training.

Note that this gating network entails only a modest number of trainable parameters, since both the size of its input as well as of its single hidden layer are rather small. As such, it can be effectively trained even with limited availability of (labeled) data. This is a key merit of our approach, which fully differentiates it from conventional classifiers that are presented with the extremely high-dimensional sets of raw observed data.

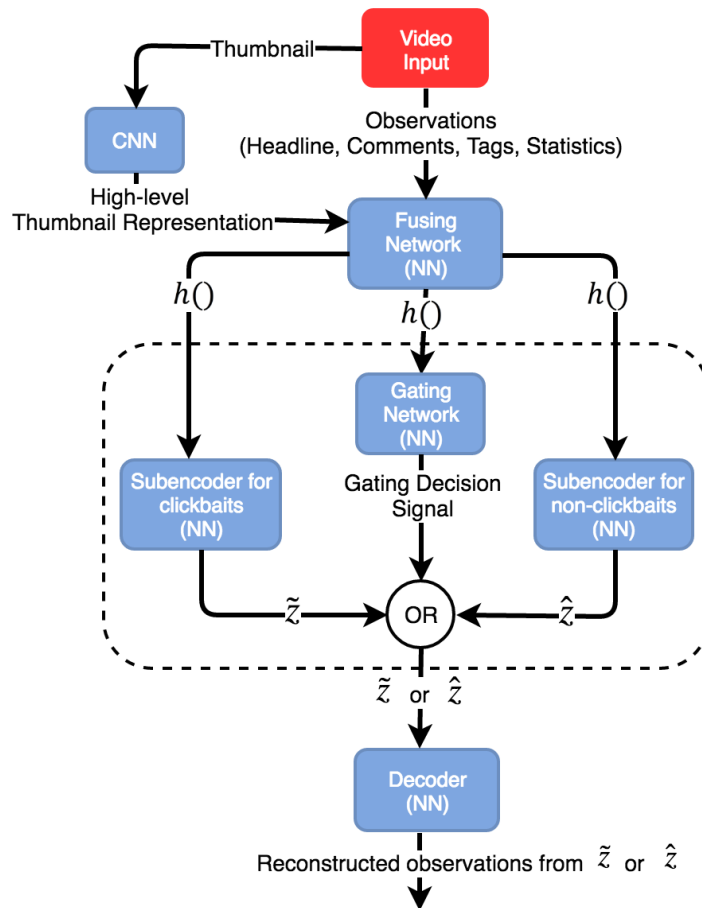


Figure 1 Overview of the proposed clickbait detection model. The dotted rectangle represents the encoding component of our model.

To conclude the formulation of the proposed VAE, we need to postulate an appropriate decoder distribution, and a corresponding network that infers it. In this work, we opt for a simple dense-layer neural network, which is fed with the (sampled) output of the postulated finite mixture model encoder and attempts to reconstruct the original modalities. Specifically, we postulate a network comprising one hidden layer with 300 ReLU units, and a set of 823 output units, that attempt to reconstruct the original modalities, with the exception of the thumbnail. The reason why we ignore the thumbnail modality from the decoding process is the need of utilizing deconvolutional network layers to appropriately handle it, which is quite complicated. Hence, we essentially treat the thumbnail modality as side-information, that regulates the inferred posterior distribution of the latent variables z (encoder) in the sense of a covariate, instead of an observed modality in the conventional sense. It is empirically known that such a setup, if appropriately implemented, does not undermine modeling effectiveness [20].

3.5. Results

In this section we provide our results from the analysis of the manually reviewed dataset as well as from the experimental evaluation of the devised clickbait detection model.

3.5.1 Ground truth Analysis

In order to better grasp the problem, we perform a comparative analysis of the manually reviewed ground truth dataset. Category. Table 1 reports the distribution of the videos in various categories. In total, we capture videos from 15 categories but, we only show the top five in terms of counts for brevity. We observe that most clickbaits exist in the Entertainment and Comedy categories, whereas non-clickbaits are prevalent in the Sports category. This indicates that, within this dataset, usually YouTubers employ clickbait techniques on “soft” videos used for entertainment.

Table 1 Top five categories (and their respective percentages) in our ground truth dataset

Category	Clickbaits (%)	Non-clickabaits (%)
Entertainment	406 (38%)	308 (32%)
Comedy	318 (29%)	228 (24%)
People & Blogs	155 (14%)	115 (12%)
Autos & Vehicles	33 (3%)	49 (5%)
Sports	29 (3%)	114(12%)

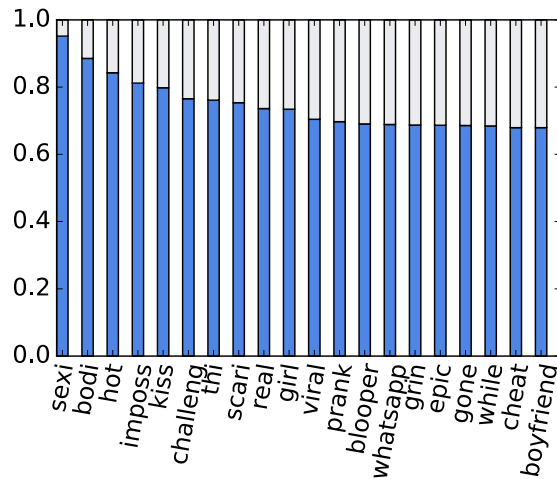


Figure 2 Top 20 stems found in the headline of clickbait videos

Headline: YouTubers normally employ deceptive techniques on the headline, such as the use of captivating and/or exaggerating phrases. To verify that this applies to our ground truth dataset, we perform stemming to the words that are found in clickbait and non-clickbait

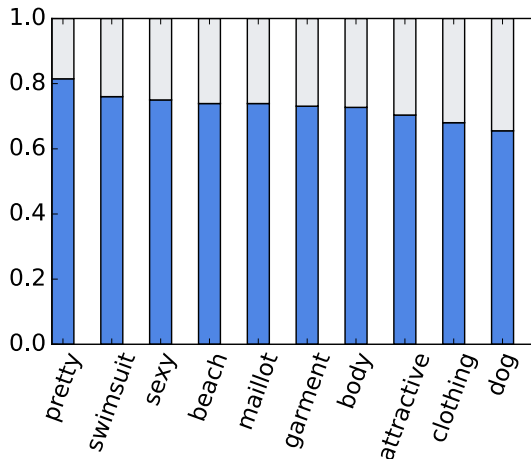


Figure 3 Top 10 words obtained from thumbnails of clickbait videos

headlines. Figure 2 depicts the ratio of the top 20 stems that are found in our ground truth clickbait videos (i.e., 95% of the videos that contain the stem “sexi” are clickbait). In essence, we observe that magnetizing stems like “sexi” and “hot” are frequently used in clickbait videos, whereas their use in non-clickbaits is low. The same applies to words used for exaggeration, like “viral” and “epic”.

Thumbnail: To study the videos’ thumbnails, we make use of the Imagga service [14], which offers descriptive tags for an image. Specifically, we perform tagging of all the thumbnails in our ground truth dataset. Figure 3 demonstrates the ratio of the top 10 Imagga tags that are found in the manually reviewed ground truth videos. We observe that clickbait videos typically use sexually appealing thumbnails in their videos in order to attract viewers. For instance, 81% of the videos the thumbnail of which contains the “pretty” tag are clickbaits.

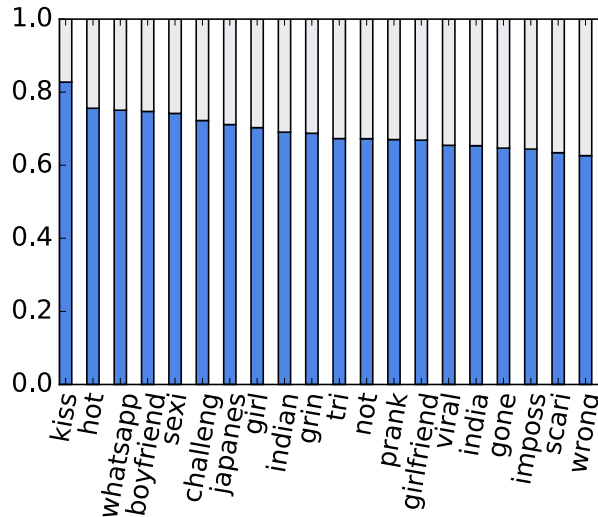


Figure 4 Top 20 tags found in clickbait videos

Tags: Tags are words that are defined by YouTubers before publishing and can dictate whether a video will emerge on users’ search queries. We notice that clickbaits use specific words on tags, whereas non-clickbaits do not. Figure 4 depicts the ratio of the top 20 stems that are found in clickbaits. We observe that many clickbait videos use tags such as “gone wrong”, “try not to laugh”, “viral”, “hot” and “impossible”. Such phrases are usually used for exaggeration.

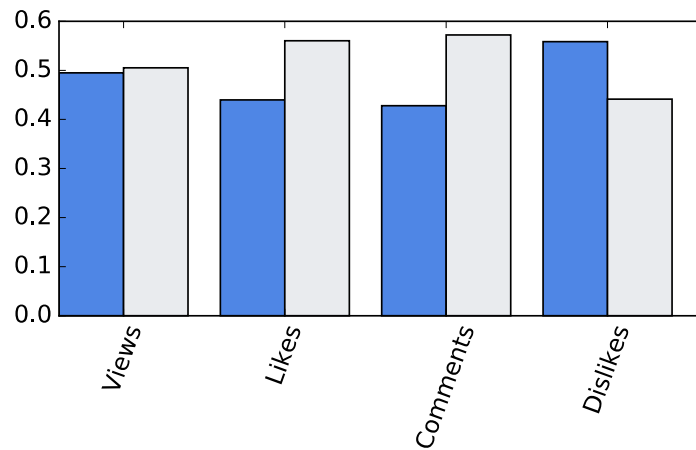


Figure 5 Statistics for clickbaits and non-clickbaits

Statistics: Figure 5 shows the normalized score of the number of views, likes, comments and dislikes in clickbait and non-clickbait videos. Interestingly, clickbaits and non-clickbaits videos have similar view counts. This indicates the severity of the problem on YouTube and suggests that viewers are not able to easily discern clickbait videos, hence clicking on them. Furthermore, we observe that non-clickbait videos have more likes and less dislikes than clickbaits. This is reasonable as many users feel frustrated after watching clickbaits.

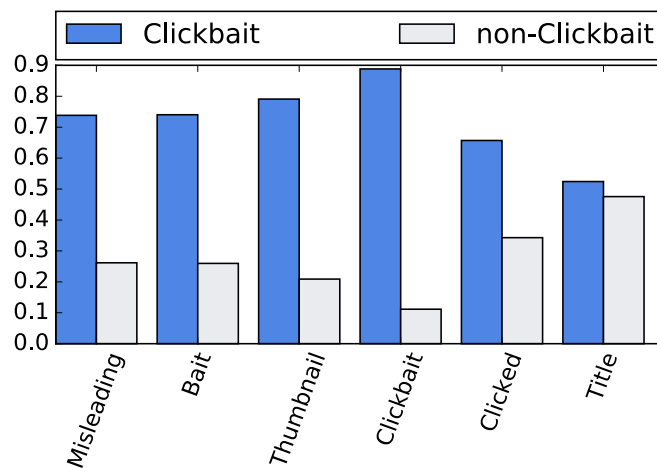


Figure 6 Words that were used for flagging videos and their respective ratio in clickbait videos

Comments: We notice that users on YouTube implicitly flag suspicious videos by commenting on them. For instance, we note several comments like the following: “the title is misleading”, “i clicked because of the thumbnail”, “where is the thumbnail?” and “clickbait”. Hence, we argue that comments from viewers is a valuable resource for assessing videos. To this end, we analyze the ground truth dataset to extract the mean number of occurrences

of words widely used for flagging clickbait videos. Figure 6 depicts the normalized mean scores for the identified words. We observe that these words were greatly used in clickbait comments but not in non-clickbaits. Also, it is particularly interesting that comments referring to video thumbnails were found 2.5 times more often in clickbait than in non-clickbaits.

Graph Analysis: Users often watch videos according to YouTube’s recommendations. From manual inspections, we have noted that when watching a clickbait video, YouTube is more likely to recommend another clickbait video. To confirm this against our ground truth, we create a directed graph $G = (V, E)$, where V are the videos and E are the connections between videos pointing to one another via a recommendation. Then, for all the videos, we select their immediate neighbors in the graph, and count the videos that are clickbaits and non-clickbaits. Table 2 reports the normalized mean of the number of connected videos for each class. We apply a normalization factor to mitigate the bias towards clickbaits, which have a slightly greater number in our ground truth. We observe that, when a user watches a clickbait video, they are recommended 4.1 clickbait videos on average, as opposed to 2.73 non-clickbait recommendations. A similar pattern holds for the non-clickbaits; a user is less likely to be served a clickbait when watching a non-clickbait.

Table 2 Normalized mean of related videos for clickbait and non-clickbait videos in the ground truth dataset

Source	Destination	Normalized Mean
Clickbait	Clickbait	4.1
Clickbait	Non-clickbait	2.73
Non-clickbait	Clickbait	2.75
Non-clickbait	Non-clickbait	3.57

YouTube’s countermeasures: To get an insight on whether YouTube employs any countermeasures, we calculate the number of offline (either deleted by YouTube or removed by the uploader) videos in our manually reviewed ground truth, as of January 10, 2017 and April 30, 2017. We found that only 3% (as of January 10th) and 10% (as of April 30th) of the clickbaits are offline. Similarly, only 1% (as of January 10th) and 5% (as of April 30th) of the non-clickbaits are offline. To verify that the ground truth dataset does not consist of only recent videos (thus, just published and not yet detected) we calculate the mean number of days that passed from the publication date up to January 10, 2017. We find that the mean number of days for the clickbaits is 700, while it is 917 days for the non-clickbaits. The very low offline ratio, as well as the high mean number of days, indicate that YouTube is not able to effectively tackle the problem in a timely manner.

3.5.2 Clickbait Detection Model Evaluation

Baselines: To evaluate our model, we compare it against two baseline models. First, a simple Support Vector Machine (SVM) with parameters $\gamma = 0.001$ and $C = 100$. Second, a

supervised deep network (SDN) that comprises (i) the same CNN as the proposed model; and (ii) a 2-layer fully-connected neural network with Dropout level of $d = 0.5$.

We train our proposed model using the entirety of the available unlabeled dataset, as well as a randomly selected 80% of the available labeled (ground truth) dataset, comprising an equal number of clickbait and non-clickbait examples. Subsequently, the resulting trained model is used to perform out-of-sample evaluation; that is, we compute the classification performance of our approach on the fraction of the available labeled examples that were not used for model training.

Table 3 Performance metrics for the evaluated methods. We also report the performance of our model when using only 25% or 50% of the available unlabeled data

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.882	0.909	0.884	0.896
SDN	0.908	0.920	0.907	0.917
Proposed Model (U = 25%)	0.915	0.918	0.926	0.923
Proposed Model (U = 50%)	0.918	0.918	0.934	0.926
Proposed Model (U = 100%)	0.924	0.921	0.942	0.931

Table 3 reports the performance of the proposed model as well as the two considered baselines. Our evaluation metrics are the accuracy, precision, recall and F1 scores. We observe that neural network-based approaches, such as a simple neural network and the proposed model, outperform SVMs in terms of all the considered metrics. Specifically, the best performance is obtained by the proposed model, which outperforms SVMs by 3.8%, 1.2%, 5.8% and 3.5% on the accuracy, precision, recall, and F1 score metrics, respectively.

Further, to assess the importance of using unlabeled data for the performance of our model, we also report results with reduced unlabeled data. We observe that, using only 25% of the available unlabeled data, the proposed model undergoes a substantial performance decrease, as measured by all the employed performance metrics. This performance deterioration only slightly improves when we elect to retain 50% of the available unlabeled data.

3.6. Conclusion

In this work, we have explored the use of variational autoencoders for tackling the clickbait problem on YouTube. Our approach constitutes the first proposed semi-supervised deep learning technique in the field of clickbait detection. This way, it enables more effective automated detection of clickbait videos in the absence of large-scale labeled data. Our analysis indicates that, according to our data, YouTube’s recommendation engine does not take into account the clickbait problem in its recommendations.

3.7. References

- [1] Piperjaffray, "Survey," 2016, <http://archive.is/AA34y>.
- [2] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in ACM CSA, 2010.
- [3] H. Gao, Y. Chen, K. Lee, D. Palsetia et al., "Towards online spam filtering in social networks." In NDSS, 2012.
- [4] N. Jindal and B. Liu, "Review spam detection," in WWW, 2007.
- [5] A. Peysakhovich, "Reducing clickbait in facebook feed," 2016, <http://newsroom.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed>.
- [6] J. Constine, "Facebook feed change fights clickbait post by post in 9 more languages," 2017, <http://archive.is/l8hlt>.
- [7] R. Campbell, "You Won't Believe How Clickbait is Destroying YouTube!" 2017, <https://tritontimes.com/11564/columns/you-wont-believe-how-clickbait-is-destroying-youtube/>.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, 2015.
- [9] D. Kingma and M. Welling, "Auto-encoding variational Bayes," in ICLR, 2014.
- [10] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," in ICML, 2016.
- [11] S. P. Chatzis, D. I. Kosmopoulos, and T. A. Varvarigou, "Signal modeling and classification using a robust latent space model based on t distributions," TOSP, 2008.
- [12] E. A. Platanios and S. P. Chatzis, "Gaussian process-mixture conditional heteroscedasticity," TPAMI, 2014.
- [13] "Clickbait video" <https://www.youtube.com/watch?v=W2WgTE9OKyg>.
- [14] Imagga, "Tagging Service," 2016, <https://imagga.com/>.
- [15] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," arXiv, 2017.
- [16] P. Biyani, K. Tsioutsoulouklis, and J. Blackmer, "'8 amazing secrets for getting more clicks': Detecting clickbaits in news streams using article informality," 2016.
- [17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in ECCV, 2014.
- [18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," JMLR, 2014.
- [19] G. McLachlan and D. Peel, Finite Mixture Models, 2000.

[20] I.Porteous, A.Asuncion and M Welling,“Bayesian matrix factorization with side information,” in AAAI, 2010.

4. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?

4.1. Project description and motivation

The Web’s information ecosystem is composed of multiple communities with varying influence [24]. As mainstream online social networks become less novel, users have begun to join smaller, more focused platforms. In particular, as the former have begun to reject fringe communities identified with racist and aggressive behavior, a number of alt-right focused services have been created. Among these emerging communities, the Gab social network has attracted the interest of a large number of users since its creation in 2016 [8], a few months before the US Presidential Election. Gab was created, ostensibly as a censorship-free platform, aiming to protect free speech above anything else. From the very beginning, site operators have welcomed users banned or suspended from platforms like Twitter for violating terms of service, often for abusive and/or hateful behavior. In fact, there is extensive anecdotal evidence that the platform has become the alt-right’s new hub [23] and that it exhibits a high volume of hate speech [13] and racism [5]. As a result, in 2017, both Google and Apple rejected Gab’s mobile apps from their stores because of hate speech [13] and non-compliance to pornographic content guidelines [1].

In this work, we provide, to the best of our knowledge, the first characterization of the Gab social network. We crawl the Gab platform and acquire 22M posts by 336K users over a 1.5 year period (August 2016 to January 2018).

Overall, the main findings of our analysis include:

- Gab attracts a wide variety of users, ranging from well-known alt-right personalities like Milo Yiannopoulos to conspiracy theorists like Alex Jones. We also find a number of “troll” accounts that have migrated over from other platforms like 4chan, or that have been heavily inspired by them.
- Gab is predominantly used for the dissemination and discussion of world events, news, as well as conspiracy theories. Interestingly, we note that Gab reacts strongly to events related to white nationalism and Donald Trump.
- Hate speech is extensively present on the platform, as we find that 5.4% of the posts include hate words. This is 2.4 times higher than on Twitter, but 2.2 times lower than on 4chan’s Politically Incorrect board (/pol/) [9].
- There are several accounts making coordinated efforts towards recruiting millennials to the alt-right. In summary, our analysis highlights that Gab appears to be positioned at the border of mainstream social networks like Twitter and “fringe” Web communities like 4chan’s /pol/. We find that, while Gab claims to be all about free speech, this seems to be merely a shield behind which its alt-right users hide.

4.2. Background

Gab is a new social network, launched in August 2016, that “champions free speech, individual liberty, and the free flow of information online.” It combines social networking features that exist in popular social platforms like Reddit and Twitter. A user can broadcast 300-character messages, called “gabs,” to their followers (akin to Twitter). From Reddit, Gab takes a modified voting system (which we discuss later). Gab allows the posting of pornographic and obscene content, as long as users label it as Not-Safe-For-Work (NSFW). Posts can be reposted, quoted, and used as replies to other gabs. Similar to Twitter, Gab supports hashtags, which allow indexing and querying for gabs, as well as mentions, which allow users to refer to other users in their gabs. Topics and Categories. Gab posts can be assigned to a specific topic or category. Topics focus on a particular event or timely topic of discussion and can be created by Gab users themselves; all topics are publicly available and other users.

Topics and Categories. Gab posts can be assigned to a specific topic or category. Topics focus on a particular event or timely topic of discussion and can be created by Gab users themselves; all topics are publicly available and other users can post gabs related to topics. Categories on the other hand, are defined by Gab itself, with 15 categories defined at the time of this writing. Note that assigning a gab to a category and/or topic is optional, and Gab moderates topics, removing any that do not comply with the platform’s guidelines.

Voting system. Gab posts can get up- and down-voted; a feature that determines the popularity of the content in the platform (akin to Reddit). Additionally, each user has its own score, which is the sum of up-votes minus the sum of downvotes that it received to all his posts (similar to Reddit’s user karma score [3]). This user-level score determines the popularity of the user and is used in a way unique to Gab: a user must have a score of at least 250 points to be able to down-vote other users’ content, and every time a user down-votes a post a point from his user-level score is deducted. In other words, a user’s score is used as a form of currency expended to downvote content. Moderation. Gab has a lax moderation policy that allows most things to be posted, with a few exceptions. Specifically, it only forbids posts that contain “illegal pornography” (legal pornography is permitted), posts that promote terrorist acts, threats to other users, and doxing other users’ personal information [18].

Monetization. Gab is ad-free and relies on direct user support. On October 4, 2016 Gab’s CEO Andrew Torba announced that users were able to donate to Gab [19]. Later, Gab added “pro” accounts as well. “Pro” users pay a per-month fee granting additional features like live-stream broadcasts, account verification, extended character count (up to 3K

characters per gab), special formatting in posts (e.g., italics, bold, etc.), as well as premium content creation. The latter allows users to create “premium” content that can only be seen by subscribers of the user, which are users that pay a monthly fee to the content creator to be able to view his posts. The premium content model allows for crowdfunding particular Gab users, similar to the way that Twitch and Patreon work. Finally, Gab is in the process of raising money through an Initial Coin Offering (ICO) with the goal to offer a “censorship-proof” peer-to-peer social network that developers can build application on top [2].

Dataset. Using Gab’s API, we crawl the social network using a snowball methodology. Specifically, we obtain data for the most popular users as returned by Gab’s API and iteratively collect data from all their followers as well as their followings. We collect three types of information: 1) basic details about Gab accounts, including username, score, date of account creation; 2) all the posts for each Gab user in our dataset; and 3) all the followers and followings of each user that allow us to build the following/followers network. Overall, we collect 22,112,812 posts from 336,752 users, between August 2016 and January 2018.

4.3. Analysis

In this section, we provide our analysis on the Gab platform. Specifically, we analyze Gab’s user base and posts that get shared across several axes.

4.3.1 Ranking of users.

To get a better handle on the interests of Gab users, we first examine the most popular users using three metrics: 1) the number of followers; 2) user account score; and 3) user PageRank. These three metrics provide us a good overview of things in terms of “reach,” appreciation of content production, and importance in terms of position within the social network. We report the top 20 users for each metric in Table 4. Although we believe that their existence in Table 4 is arguably indicative of their public figure status, for ethical reasons, we omit the “screen names” for accounts in cases where a potential link between the screen name and the user’s real-life names existed and it was unclear to us whether or not the user is a public figure. While Twitter has many celebrities in the most popular users [11], Gab seems to have what can at best be described as alt-right celebrities like Milo Yiannopoulos and Mike Cernovich.

Number of followers. The number of followers that each account has can be regarded as a metric of impact on the platform, as a user with many followers can share its posts to a large number of other users. We observe a wide variety of different users; 1) popular alt-right users like Milo Yiannopoulos, Mike Cernovich, Stefan Molyneux, and Brittany Pettibone; 2) Gab’s founder Andrew Torba; and 3) popular conspiracy theorists like Alex Jones. Notably lacking are users we might consider as counter-points to the alt-right right, an indication of Gab’s heavily right-skewed user-base.

Score. The score of each account is a metric of content popularity, as it determines the number of up-votes and down-votes that they receive from other users. In other words, is the degree of appreciation from other users. By looking at the ranking using the score, we observe two new additional categories of users: 1) users purporting to be news outlets, likely pushing false or controversial information on the network like PrisonPlanet and USSANews; and 2) troll users that seem to have migrated from or been inspired by other platforms (e.g., 4chan) like Kek Magician and CuckShamer.

PageRank. We also compute PageRank on the followers/followings network and we rank the users according to the obtained score. We use this metric as it quantifies the structural importance of nodes within a network according to its connections. Here, we observe some interesting differences from the other two rankings. For example, the account with username “realdonaldtrump,” an account reserved for Donald Trump, appears in the top users mainly because of the extremely high number of users that follow this account, despite the fact that it has no posts or score.

Table 4 Top 20 popular users on Gab according to the number of followers, their score, and their ranking based on PageRank in the followers/followings network. We omit the “screen names” of certain accounts for ethical reasons.

Followers			Scores			PageRank		
Name	Username	#	Name	Username	#	Name	Username	PR score
Milo Yiannopoulos	m	45,060	Andrew Torba	a	819,363	Milo Yiannopoulos	m	0.013655
PrisonPlanet	PrisonPlanet	45,059	John Rivers	JohnRivers	606,623	Andrew Torba	a	0.012818
Andrew Torba	a	38,101	Ricky Vaughn	Ricky_Vaughn99	496,962	PrisonPlanet	PrisonPlanet	0.011762
Ricky Vaughn	Ricky_Vaughn99	30,870	Don	Don	368,698	Mike Cernovich	Cernovich	0.006549
Mike Cernovich	Cernovich	29,081	Jared Wyand	JaredWyand	281,798	Ricky Vaughn	Ricky_Vaughn99	0.006143
Stefan Molyneux	stefanmolyneux	26,337	[omitted]	TukkRivers	253,781	Sargon of Akkad	Sargonofakkad100	0.005823
Brittany Pettibone	BrittPettibone	24,799	Brittany Pettibone	BrittPettibone	244,025	[omitted]	d_seaman	0.005104
Jebs	DeadNotSleeping	22,659	Tony Jackson	USMC-Devildog	228,370	Stefan Molyneux	stefanmolyneux	0.004830
[omitted]	TexasYankee4	20,079	[omitted]	causticbob	228,316	Brittany Pettibone	BrittPettibone	0.004218
[omitted]	RightSmarts	20,042	Constitutional Drunk	USSANews	224,261	Vox Day	voxday	0.003972
Vox Day	voxday	19,454	Truth Whisper	truthwhisper	206,516	Alex Jones	RealAlexJones	0.003345
[omitted]	d_seaman	18,080	Andrew Anglin	AndrewAnglin	203,437	Lauren Southern	LaurenSouthern	0.002984
Alex Jones	RealAlexJones	17,613	Kek_Magician	Kek_Magician	193,819	Donald J Trump	realdonaldtrump	0.002895
Jared Wyand	JaredWyand	16,975	[omitted]	shorty	169,167	Dave Cullen	DaveCullen	0.002824
Ann Coulter	AnnCoulter	16,605	[omitted]	SergeiDimitrovicIvanov	169,091	[omitted]	e	0.002648
Lift	lift	16,544	Kolja Bonke	KoljaBonke	160,246	Chuck C Johnson	Chuckcjohnson	0.002599
Survivor Medic	SurvivorMed	16,382	Party On Weimerica	CuckShamer	155,021	Andrew Anglin	AndrewAnglin	0.002599
[omitted]	SalguodNos	16,124	PrisonPlanet	PrisonPlanet	154,829	Jared Wyand	JaredWyand	0.002504
Proud Deplorable	luther	15,036	Vox Day	voxday	150,930	Pax Dickinson	pax	0.002400
Lauren Southern	LaurenSouthern	14,827	W.O. Cassity	wocassity	144,875	Baked Alaska	apple	0.002292

4.3.2 Posts Analysis.

Basic Statistics. First, we note that 63% of the posts in our dataset are original posts while 37% are reposts. Interestingly, only 0.14% of the posts are marked as NSFW. This is surprising given the fact that one of the reasons that Apple rejected Gab’s mobile app is due to the share of NSFW content [1]. From browsing the Gab platform, we also can anecdotally

confirm the existence of NSFW posts that are not marked as such, raising questions about how Gab moderates and enforces the use of NSFW tags by users. When looking a bit closer at their policies, Gab notes that they use a 1964 United States Supreme Court Ruling [21] on pornography that provides the famous “I’ll know it when I see it” test. In any case, it would seem that Gab’s social norms are relatively lenient with respect to what is considered NSFW. We also look into the languages of the posts, as returned by Gab’s API. We find that Gab’s API does not return a language code for 56% of posts. By looking at the dataset, we find that all posts before June 2016 do not have an associated language; possibly indicating that Gab added the language field afterwards. Nevertheless, we find that the most popular languages are English (40%), Deutsch (3.3%), and French (0.14%); possibly shedding light to Gab users’ locations which are mainly the US, the UK, and Germany.

Hashtags & Mentions. Gab supports the use of hashtags and mentions similar to Twitter. Table 5 reports the top 20 hashtags/mentions that we find in our dataset. We observe that the majority of the hashtags are used in posts about Trump, news, and politics. We note that among the top hashtags are “AltRight”, indicating that Gab users are followers of the alt-right movement or they discuss topics related to the alt-right; “Pizzagate”, which denotes discussions around the notorious conspiracy theory [20]; and “BanIslam”, which indicate that Gab users are sharing their islamophobic views. It is also worth noting the use of hashtags for the dissemination of popular memes, like the Drain the Swamp meme that is popular among Trump’s supporters [14]. When looking at the most popular users that get mentioned, we find popular users related to the Gab platform like Andrew Torba (Gab’s CEO with username @a). We also note users that are popular with respect to mentions, but do not appear in Table 4’s lists of popular users. For example, Amy is an account purporting to be Andrew Torba’s mother. The user Stargirlx, who we note changed usernames three times during our collection period, appears to be an account presenting itself as a millennial “GenZ” young woman. Interestingly, it seems that Amy and Stargirlx have been organizing Gab “chats,” which are private groups of users, for 18 to 29-year olds to discuss politics; possibly indicating efforts to recruit millennials to the alt-right community.

Table 5 Top 20 hashtags and mentions found in Gab. We report their percentage over all posts

Hashtag	(%)	Mention	(%)
MAGA	6.06%	a	0.69%
GabFam	4.22%	TexasYankee4	0.31%
Trump	3.01%	Stargirlx	0.26%
SpeakFreely	2.28%	YouTube	0.24%
News	2.00%	support	0.23%
Gab	0.88%	Amy	0.22%
DrainTheSwamp	0.71%	RaviCrux	0.20%
AltRight	0.61%	u	0.19%
Pizzagate	0.57%	BlueGood	0.18%
Politics	0.53%	HorrorQueen	0.17%
PresidentTrump	0.47%	Sockalexix	0.17%
FakeNews	0.41%	Don	0.17%
BritFam	0.37%	BrittPettibone	0.16%
2A	0.35%	TukkRivers	0.15%
maga	0.32%	CurryPanda	0.15%
NewGabber	0.28%	Gee	0.15%
CanFam	0.27%	e	0.14%
BanIslam	0.25%	careyetta	0.14%
MSM	0.22%	PrisonPlanet	0.14%
1A	0.21%	JoshC	0.12%

Hate Speech Assessment. As previously discussed, Gab was openly accused of allowing the dissemination of hate speech. In fact, Google removed Gab’s mobile app from its Play Store because it violates their hate speech policy [13]. Due to this, we aim to assess the extent of hate speech in our dataset. Using the modified Hatebase [4] dictionary used by the authors of [9], we find that 5.4% of all Gab posts include a hate word. In comparison, Gab has 2.4 times the rate of hate words when compared to Twitter, but less than halve the rate of hate words compared to 4chan’s Politically Incorrect board (/pol/) [9]. These findings indicate that Gab resides on the border of mainstream social networks like Twitter and fringe Web communities like 4chan’s Politically Incorrect (/pol/) board.

Temporal Analysis. Finally, we study the posting behavior of Gab users from a temporal point of view. Figure 7 shows the distribution of the Gab posts in our dataset according to each day of our dataset. We observe that the general trend is that the number of Gab’s posts increase over time.

To isolate significant days in the time series in Figure 7, we perform a changepoint analysis using the Pruned Exact Linear Time (PELT) method [10]. First, we use our knowledge of the weekly variation in average post numbers to subtract from our timeseries the mean number of posts for each day. This leaves us with a mean-zero timeseries of the deviation of the number of posts per day from the daily average. We assume that this timeseries is drawn from a normal distribution, with mean and

variance that can change at a discrete number of changepoints. We then use the PELT algorithm to maximize the log-likelihood function for the mean(s) and variance(s) of this distribution, with a penalty for the number of changepoints. By ramping down the penalty function, we produce a ranking of the changepoints.

Examining current events around these changepoints provides insight into the dynamics that drive Gab behavior. First, we note that there is a general increase in activity up to the Trump inauguration, at which point activity begins to decline. When looking later down the timeline, we see an increase in activity after the changepoint marked 1 in Figure 7. Changepoint 1 coincides with James Comey’s firing from the FBI, and the relative acceleration of the Trump-Russian collusion probe [16].

The next changepoint (2) coincides with the so-called “March Against Sharia” [12] organized by the alt-right, with the event marked 4 corresponding to Trump’s “blame on both sides” response to violence at the Unite the Right Rally in Charlottesville [15]. Similarly, we see a meaningful response to Twitter’s banning of abusive users [7] marked as changepoint 5.

Changepoint 3, occurring on July 12, 2017 is of particular interest, since it is the most extreme reduction in activity recognized as a changepoint. From what we can tell, this is a reaction to Donald Trump Jr. releasing emails that seemingly evidenced his meeting with a Russian lawyer to receive compromising intelligence on Hillary Clinton’s campaign [17]. I.e., the disclosure of evidence of collusion with Russia corresponded to the single largest drop in posting activity on Gab.

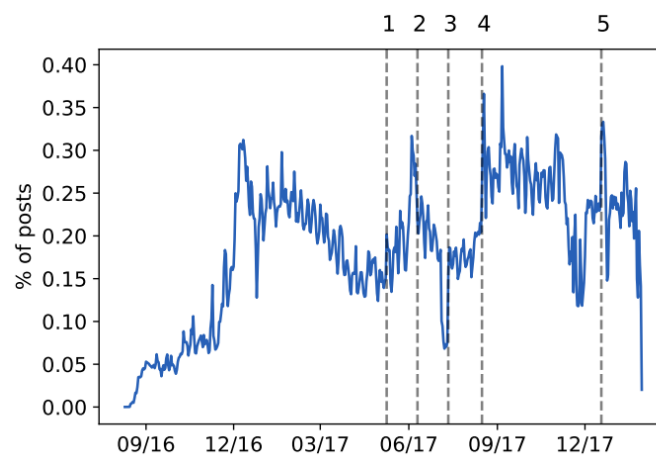


Figure 7 Temporal Analysis of Gab posts

4.4. Conclusion

In this work, we have provided the first characterization of a new social network called Gab. We analyzed 22M posts from 336K users, finding that Gab attracts the interest of users ranging from alt-right supporters and conspiracy theorists to trolls. We showed that Gab is extensively used for the discussion of news, world events, and politics-related topics, further motivating the need to take it into account when studying information cascades on the Web. By looking at the posts for hate words, we also found that 5.4% of the posts include hate words. Finally, using changepoint analysis, we highlighted how Gab reacts very strongly to real-world events focused around white nationalism and support of Donald Trump.

4.5. References

- [1] Apple’s Double Standards Against Gab. <https://medium.com/@getongab/apples-double-standardsagainst-gab-1bffa2c09115>, 2016.
- [2] A Censorship-Proof P2P Social Media Protocol. <https://www.startengine.com/gab-select>, 2017.
- [3] Reddit FAQ - Karma. <https://www.reddit.com/wiki/faq>, 2017.
- [4] Hatebase API. <https://www.hatebase.org/>, 2018.
- [5] T. Benson. Inside the “Twitter for racists”: Gab – the site where Milo Yiannopoulos goes to troll now. <https://goo.gl/Yqv4Ue>, 2016.
- [6] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In CHI, 2017.
- [7] K. Donaldson and J. Brustein. Twitter Bans Some White Supremacists and Other Extremists. <https://www.bloomberg.com/news/articles/2017-12-19/twitter-bans-some-white-supremacists-and-otherextremists>, 2017.
- [8] Gab. Gab site. <https://gab.ai>, 2017.
- [9] G. E. Hine, J. Onalapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web. In ICWSM, 2017.
- [10] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

- [11] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is Twitter, a social network or a news media? In WWW, 2010.
- [12] C. Mathias. The 'March Against Sharia' Protests Are Really Marches Against Muslims. <https://www.huffingtonpost.com/entry/march-against-sharia-anti-muslim-act-foramerica-us-5939576ee4b0b13f2c67d50c>, 2017.
- [13] R. Price. Google's app store has banned Gab, a social network popular with the far-right, for 'hate speech'. <http://uk.businessinsider.com/google-appstore-gab-ban-hate-speech-2017-8>, 2017.
- [14] A. Romano. The 2016 culture war, as illustrated by the alt-right. <https://www.vox.com/culture/2016/12/30/13572256/2016-trump-culture-war-alt-right-meme>, 2016.
- [15] M. Shear and M. Haberman. Trump Defends Initial 7 Remarks on Charlottesville; Again Blames 'Both Sides'. <https://www.nytimes.com/2017/08/15/us/politics/trumppress-conference-charlottesville.html>, 2017.
- [16] D. Smith. Trump fires FBI director Comey, raising questions over Russia investigation. <https://www.theguardian.com/us-news/2017/may/09/jamescomey-fbi-fired-donald-trump>, 2017.
- [17] D. Smith and S. Siddiqui. 'I love it': Donald Trump Jr posts emails from Russia offering material on Clinton. <https://www.theguardian.com/us-news/2017/jul/11/donald-trump-jr-email-chain-russia-hillary-clinton>, 2017.
- [18] P. Snyder, P. Doerfler, C. Kanich, and D. McCoy. Fifteen minutes of unwanted fame: detecting and characterizing doxing. In IMC, 2017.
- [19] A. Torba. How You Can Help Support Gab. <https://medium.com/@Torbahax/gab-donations9ca2a5c0557e>, 2016.
- [20] M. Wendling. The saga of 'Pizzagate': The fake story that shows how conspiracy theories spread. <http://www.bbc.com/news/blogs-trending-38156985>, 2016.
- [21] Wikipedia. Jacobellis v. Ohio. https://en.wikipedia.org/wiki/Jacobellis_v._Ohio, 1964.
- [22] Wikipedia. Unite the Right rally. https://en.wikipedia.org/wiki/Unite_the_Right_rally, 2017.
- [23] J. Wilson. Gab: alt-right's social media alternative attracts users banned from Twitter. <https://www.theguardian.com/media/2016/nov/17/gabalt-right-social-media-twitter>, 2016.
- [24] S. Zannettou, T. Caulfield, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In IMC, 2017.

5. Understanding Web Archiving Services and Their (Mis)Use on Social Media

5.1. Project description and motivation

In today’s digital society, the availability and persistence of Web resources are very relevant issues. A substantial number of URLs shared on the Web becomes unavailable after some time as websites are shutdown or redesigned in a way that does not preserve old URLs – a phenomenon known as “link rot” [9]. Moreover, content might be taken down by authorities on a legal basis, deleted by users who have shared it on social media, removed as per the “right to be forgotten”, etc [3]. Overall, the ephemerality of Web content often prompts debate with respect to its impact on the availability of information, accountability, or even censorship.

In this context, an important role is played by services like the Wayback Machine (archive.org), which proactively archives large portions of the Web, allowing users to search and retrieve the history of more than 300 billion pages. At the same time, on-demand archiving services like archive.is have also become popular: users can take a snapshot of a Web page by entering its URL, which the system crawls and archives, returning a permanent short URL serving as a time capsule that can be shared across the Web.

Archiving services serve a variety of purposes beyond addressing link rot. Platforms like archive.is are reportedly used to preserve controversial blogs and tweets that the author may later opt to delete [11]. Moreover, they also reduce Web traffic toward “source URLs” when the original content is still accessible, thus depriving them of potential ad revenue streams (users do not visit the original site, but just the archived copy). In fact, anecdotal evidence has emerged that alt-right communities target outlets they disagree with by nudging their users to share archive URLs instead [8], or discrediting them by pointing at earlier versions of articles [12].

Given the role in helping content persist, their use on social networks, as well as anecdotal evidence of their misuse in contexts where information could be weaponized [10], archiving services are arguably impactful actors that should be thoroughly analyzed. To this end, this work aims to shed light on the Web archiving ecosystem, aiming to answer the following research questions: How are archive URLs disseminated across popular social networks? What kind of content gets archived, by whom and why? Are archiving services misused in any way? To answer these questions, we perform a large-scale quantitative analysis of Web archives, based on two data sources: 1) 21M URLs collected from the archive.is live feed, and 2) 356K archive.is plus 391K Wayback Machine URLs that were shared on four social networks: Reddit, Twitter, Gab, and 4chan’s Politically Incorrect board (/pol/).

Our main findings include:

- News and social media posts are the most common types of content archived, likely due to their (perceived) ephemeral and/or controversial nature.
- URLs of archiving services are extensively shared on “fringe” communities within Reddit and 4chan to preserve possibly contentious content, or to refer to it without increasing the Web traffic to the source. We also find that /pol/ and Gab users favor archive.is over Wayback Machine (respectively, 15x and 16x), highlighting a particular use case in “controversial” online communities.
- Reddit bots are responsible for posting a very large portion of archive URLs in the subreddits we study (respectively, 44% and 85% of archive.is and Wayback Machine URLs). This is due to moderators aiming to alleviate the effects of link rot on the platform; however, this pro-active archival of content also impacts traffic to archived sites originating from Reddit.
- The Donald subreddit systematically targets ad revenue of news sources with conflicting ideologies: moderation bots block URLs from those sites and prompt users to post archive URLs instead (e.g., nydailynews.com have 46% of their content censored). According to our conservative estimates, popular news site like the Washington Post lose yearly approx. \$70K from their ad revenue because of the use of archiving services on Reddit.

5.2. Background

archive.is offers a free, on-demand archival service of Web pages: a user visits the service and enters a URL to be archived. It also acts as a link shortener which obfuscates the source URL, by generating a 5-character URL. For instance, <http://archive.is/HVbU> shows the snapshot of Google’s homepage, archived on July, 03, 2012 at 07:03:24.

Wayback Machine. Launched in 2001, the Wayback Machine archives a large portion of Web content, storing periodic snapshots of various pages. It mainly works through a proactive crawler, which visits various sites and captures a snapshot of the content.¹ However, users can also trigger information archival on demand. When a page is archived, an archive URL is created in the following format: [https://web.archive.org/web/\[time of archival\]/\[source URL\]](https://web.archive.org/web/[time of archival]/[source URL]). For example, the archive URL <https://web.archive.org/web/20100205062719/http://www.google.com/> returns the version of Google’s home page on February 5, 2010, at 06:27:19 (UTC).

We opt to study the Wayback Machine and archive.is for a few reasons. First, they are popular services: as of January 2018, their Alexa Global Rank is, resp., 300 and 2,920. We also choose these two because of some important differences between them. The Wayback Machine is run by a 501(c)(3) non-profit organization, while archive.is is hosted by Russian provider Hostkey (only accessible via HTTP in Russia). Moreover, the former respects

robots exclusion standards (even retroactively) and gives website owners the right to request removal of pages from the archive, while the latter only complies (albeit inconsistently) with DMCA take-down requests. Finally, archive.is is reportedly used in “fringe” Web communities within 4chan and Reddit, which are known for generating [13] and incubating [7] fake news, and for their influence on the information ecosystem [15].

5.3. Datasets

We now present our datasets as well as our data collection methodology. We perform two crawls: 1) archive.is URLs obtained from the live feed page and 2) Wayback Machine and archive.is URLs posted on four social networks, namely, Twitter, Reddit, Gab, and 4chan’s /pol/.

archive.is live feed. To gather a large dataset of archive.is generated URLs, we use the live feed page (<http://archive.is/livefeed/>), which provides a view of the archive based on archival time (e.g., the first page lists URLs archived in the previous 10 minutes). In Aug 2017, we crawl the first 100K pages of the live feed, acquiring 45.2M URLs, archived between Oct 7, 2015 and Aug 26, 2017. Next, we visit the archive.is URLs, and scrape the content to get the archival time and the source URL. To avoid issues for the site operators, we throttle our crawler and do not visit all 45.2M URLs. Instead, we randomly sample them while ensuring temporal coverage, visiting 21.5M (48%) archive URLs, corresponding to 20.6M unique source URLs from 5.3M unique domains. Note that given the substantial size of our sample, which guarantees temporal coverage over almost two years, the resulting dataset is representative of the archive. In other words, our sampling strategy does not likely introduce substantial biases affecting our results.

Archive URLs posted on social networks. We search for archive.is and Wayback Machine URLs on Twitter, Reddit, and /pol/, between Jul 1, 2016 and Aug 31, 2017, and on Gab between Aug 1, 2016–Aug 31, 2017. We obtain the 4chan dataset from [6], the Gab one from [14], the Reddit one from pushshift.io, while, for Twitter, we rely on the 1% Streaming API.

Overall, the resulting dataset includes 50K posts from /pol/, 528K posts from Reddit, 7K posts from Gab, and about 9K tweets. Note that we have some gaps due to failure of our data collection infrastructure, specifically, there are 70 and 13 days missing for Twitter and /pol/, respectively

Basic Statistics. In Table 6, we report statistics from our archive.is live feed crawl as well as the crawl of archive.is and Wayback Machine URLs shared on Twitter, Reddit, /pol/, and Gab. We report the number of posts with archive URLs, along with the percentage over the

total number of posts, as well as the number of unique archive URLs, unique source URLs, unique source domains, and the percentage of URLs that are filtered out. Specifically, besides malformed URLs, we exclude, for archive.is, URLs unreachable between Aug 29 and Oct 7, 2017, while for Wayback Machine those pointing to types of information other than Web pages (e.g., images, videos, software, etc.). Overall, /pol/ and Gab users often share Wayback Machine URLs that point to non-Web pages: around 83% and 61% of the total, respectively, suggesting that archive.is is used mostly for the dissemination of Web pages, while Wayback Machine is preferred for other content. Also, a high percentage of malformed archive.is URLs are shared on Reddit (35%), due to bots trying to pro-actively archive resources but failing. From the normalized percentages, we observe that Twitter users rarely share URLs from archiving services, while Reddit users do so from both archiving services. On /pol/ and Gab, we find 15 and 16 times, respectively, more archive.is URLs than Wayback Machine ones.

Table 6 Overview of our datasets: number and percentage of posts that include archive URLs, unique number of archive URLs, source URLs, and source domains. We also filter URLs that are malformed, unreachable, or point to resources other than Web pages.

Platform	Archive	#Posts with Archive URLs (%all posts)	Archive URLs	Source URLs	Source Domains	Filtered
Live Feed	archive.is		21,537,554	20,608,834	5,388,112	-
Reddit	archive.is	327,050 (2.9 · 10 ⁻⁴ %)	310,392	291,382	15,994	35.70%
	Wayback	320,379 (2.8 · 10 ⁻⁴ %)	387,081	343,851	21,124	17.20%
/pol/	archive.is	46,912 (1.1 · 10 ⁻³ %)	36,277	33,824	3,970	4.67%
	Wayback	3,848 (9.7 · 10 ⁻⁵ %)	2,325	2,207	976	83.12%
Gab	archive.is	6,602 (3.4 · 10 ⁻⁴ %)	5,943	5,773	1,300	5.54%
	Wayback	478 (5.1 · 10 ⁻⁵ %)	361	349	240	61.18%
Twitter	archive.is	6,750 (3.1 · 10 ⁻⁶ %)	3,772	3,669	845	8.23%
	Wayback	1,905 (9.0 · 10 ⁻⁷ %)	1,290	1,257	846	7.49%

5.4. Results

5.4.1 URL characterization

First, we aim to characterize the type of archived content. To this end, we extract the domain categories of source URLs using the free Virus Total API (virustotal.com), which we choose since it consolidates categories from multiple services (e.g., Bit Defender and Alexa). Although categorization is done at domain-level, results are presented at a per-URL level (a URL is assigned the same category as its domain) to capture the popularity of each domain.

Live Feed. Due to throttling enforced by the API, we are not able to categorize all the 20.6M source URLs in our archive.is live feed dataset. Therefore, we first aggregate URLs into their domain, then, we follow a sampling approach using: 1) the top 100K most popular domains in our dataset, which correspond to 15M source URLs, and 2) a sample of 121K domains drawn according to their empirical distribution in our archive datasets, resulting in 1.4M (7%) archive URLs. In Figure 8 we report the top 15 categories obtained from Virus Total for both

samples. Note that Virus Total is unable to provide a category for 1% and 7% of the URLs for the two sets of domains that we checked, respectively. From Figure 8(a), we observe that the most popular category is Reference Materials (23%), which is due to the fact that, as discussed earlier, many archive.is URLs archive Wayback Machine URLs. Other popular categories include Social Networks (15%), News Sources (14%), Education (13%), and Business (12%). Adult Content accounts for 4% of source URLs. Figure 8(b) shows that, for the empirically distributed sample, the top 15 categories are slightly different, including Business (21%), News (13%), and Adult Content (12%).

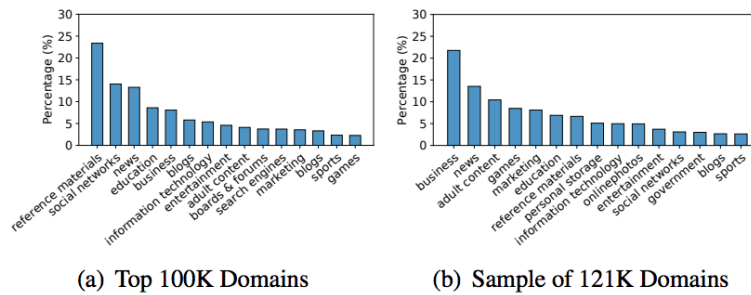


Figure 8 Top 15 domain categories for the archive.is live feed.

Social Networks. Unlike the live feed dataset, we perform URL characterization for all source URLs (aggregated by domain) found on Reddit, /pol/, Gab, and Twitter, again using the Virus Total API. In Figure 9, we report the top categories and their corresponding percentages for both archiving services (specifically, the union of categories that appear in the top 10 categories for each service). The Virus Total API is unable to provide a category for, on average, 1.5% and 9% of the archive.is and Wayback Machine URLs found on Reddit, /pol/, Gab, and Twitter, respectively. Overall, both archiving services are often used to disseminate URLs from news sources, social networks, and marketing sites on all social networks. However, there are interesting differences for the two archiving services: Education and Government URLs appear as top categories for the Wayback Machine (see Figure 9(b), Figure 9(c), and Figure 9(d)), while sites that contain obscene language appear only for archive.is (see Figure 9(c)). This suggests that the latter is used more extensively for “questionable” content. Moreover, we observe that Adult Content is among the top categories for all social networks except Twitter, while Gab and Reddit users often share archive URLs for domains related to Boards and Forums. Also, on /pol/, archive.is is used to archive and disseminate pages with obscene language, which is somewhat in line with previous observations [6] showing that /pol/ conversations often include hate speech.

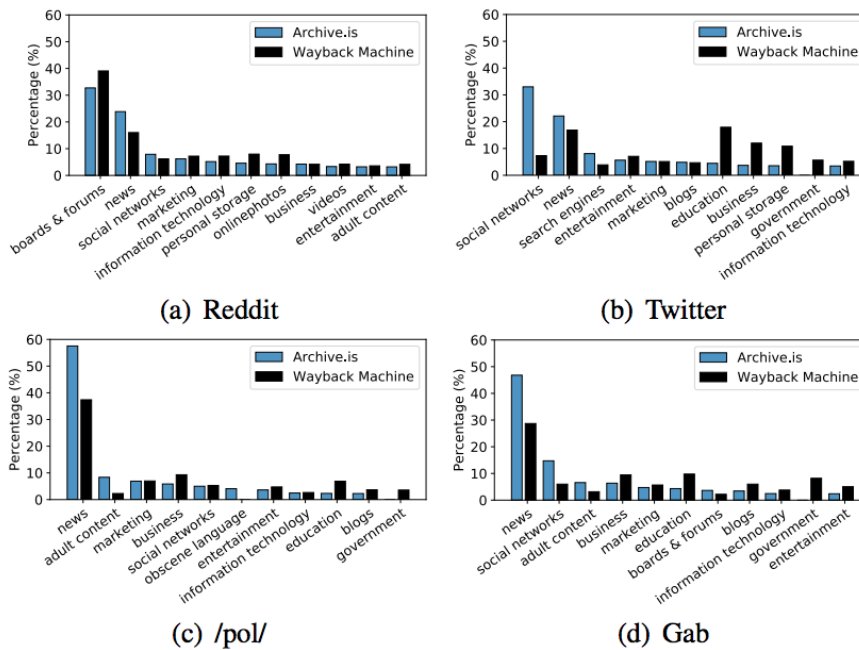


Figure 9 Top domain categories for archive URLs appearing on the four social networks.

5.4.2 Original Content Availability

We then assess the availability of the original archived content; this allow us to determine whether users are archiving URLs that are subsequently deleted. To this end, we make an HTTP request for each source URL in our datasets, on Oct 14– 21, 2017 for the live feed dataset, on Oct 4–5, 2017 for Reddit, Twitter, /pol/ datasets and on Jan 3, 2018 for Gab dataset.

We treat each URL as unavailable if we receive HTTP codes 404/410/451/5xx, or if the request times out.

Live Feed. We find that 12% of the source URLs corresponding to archive URLs on archive.is live feed are no longer available. Domains with most unavailable content include twitter.com (6%), nhk.or.jp (6%), googleusercontent.com (3%), aaaarg.fail (3%), and 4chan.org (3%).

Social Networks. In Reddit, source URLs corresponding to both archive.is and Wayback Machine are still available to a large degree (93% and 89% of them, respectively). This can be explained by the fact that Reddit bots archive URLs without considering the content. In /pol/, 82% and 66% of the original content is available for archive.is and Wayback Machine URLs, while on Gab it is 87% and 48%. Percentages decrease further for Twitter, with 76% and 49% for archive.is and Wayback Machine URLs, respectively. We also find that the top

domains for which content is no longer available differ across platforms. Except for Gab, the top unavailable domain are the social networks themselves: 10%, 54%, and 28%, for Reddit, /pol/, and Twitter, respectively. URLs from cache servers (i.e., googleusercontent.com) and Twitter are also frequently unavailable; 9% and 10% in Reddit, 5% and 4% in /pol/, 8% and 28% in Twitter, and 12% and 19% in Gab, for googleusercontent.com and Twitter, respectively. We also note the presence of unavailable 8ch.net URLs (another ephemeral imageboard) with 5% and 4% on /pol/ and Gab, respectively.

5.4.3 User Base

Reddit. Our analysis shows that archiving services are extensively used by Reddit bots. In fact, 31% of all archive.is URLs and 82% of Wayback Machine URLs in our Reddit dataset are posted by a specific bot, namely, SnapshillBot (which is used by subreddit moderators to preserve “drama-related” happenings discussed earlier or just as a subreddit specific policy to preserve every submission). Other bots include AutoModerator, 2016VoteBot, yankbot, and autotldr. We also attempt to quantify the percentage of archive URLs posted from bots, assuming that, if a username includes “bot” or “auto,” it is likely a bot. This is a reasonable strategy since Reddit bots are extensively used for moderation purposes, and do not usually try to obfuscate the fact that they are bots. Using this heuristic, we find that bots are responsible for disseminating 44% of all the archive.is and 85% of all the Wayback Machine URLs that appear on Reddit between Jul 1, 2016 and Aug 31, 2017. We also use the score of each Reddit post to get an intuition of users’ appreciation for posts that include archive URLs. In Figure 10(a), we plot the CDF of the scores of posts with archive.is and Wayback Machine URLs, as well as all posts that contain URLs as a baseline, differentiating between bots and non-bots. For both archiving services, posts by bots have a substantially smaller score: 80% of them have score of at most one, as opposed to 37% for non-bots and 59% of the baseline.

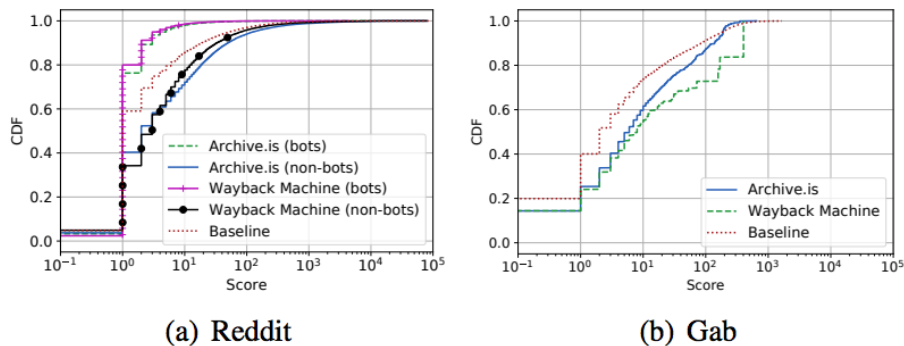


Figure 10 CDF of the scores of posts that include archive.is and Wayback Machine URLs.

Reddit Sub-Communities. We then study how specific subreddits share URLs from archiving services. In Table 7, we report the top subreddits that share the most archive URLs from archive.is and the Wayback Machine. Among these, we find a variety of subreddits ranging from politics (e.g., EnoughTrumpSpam, The Donald) to gaming (e.g., Gamingcirclejerk) and “drama-related” communities (e.g., SubredditDrama and Drama). Several subreddits prefer to use archive.is rather than the Wayback Machine, e.g., KotakuInAction, which historically covers the GamerGate controversy [2], The Donald, which discusses politics with a focus on Donald Trump, and Conspiracy, which focuses on various conspiracy theories.

Table 7 Top 15 subreddits sharing archive.is and Wayback Machine URLs.

Subreddit (archive.is)	(%)	Subreddit (Wayback)	(%)
The_Donald	24.48%	EnoughTrumpSpam	31.82%
KotakuInAction	15.83%	MGTOW	7.38%
EnoughTrumpSpam	12.06%	SnapshillBotEx	7.19%
MGTOW	3.48%	undelete	5.90%
undelete	2.74%	SubredditDrama	5.50%
SubredditDrama	2.61%	Drama	5.03%
Drama	2.33%	Gamingcirclejerk	3.47%
Gamingcirclejerk	1.57%	ShitAmericansSay	1.63%
conspiracy	1.44%	TopMindsOfReddit	1.51%
MensRights	1.12%	TheBluePill	1.25%
savedyouaclick	1.00%	Buttcoin_1000	1.15%
politics	0.98%	AgainstHateSubreddits	1.06%
DerekSmart	0.76%	subredditcancer	0.99%
ShitAmericansSay	0.75%	The_Donald	0.95%
PoliticsAll	0.72%	badeconomics	0.75%

Gab. On Gab, each post has a score that determines the popularity of the content. In Figure 10 (b), we report the CDF of the scores in posts that contain archive.is and Wayback Machine URLs, between August 2016 and August 2017. Once again, we also include a baseline, which is the scores for all the posts with URLs. We find that posts with Wayback Machine URLs have higher scores than those with archive.is URLs, and the baseline. Specifically, the mean score for Wayback Machine is 90, while for archive.is and the baseline the mean score is 35 and 30, respectively. This trend mirrors the one observed on Reddit for posts not authored by bots.

/pol/. As mentioned earlier, 4chan is an anonymous imageboard, which prevents us from performing user-level analysis. However, we can use the flag attribute to provide a country-level estimation. The top country sharing archive URLs is the USA, which is in line with previous characterizations of the board [6]. We also find a substantial percentage of “troll” flags: 9% and 5% for archive.is and Wayback Machine, respectively (see Sec. Background for a description of “troll” flags). This is somewhat surprising, since troll flags were reintroduced to /pol/ on June 13, 2017, thus they were only available for about 3 months of our 14-month dataset.

5.4.4 Ad Revenue Deprivation

During our experiments, we find evidence that at least one Reddit bot, AutoModerator, is used to remove links to unwanted domains and nudge users to share archive.is instead. In particular, it posts: *“Your submission was removed because it is from cnn.com, which has been identified as a severely anti-Trump domain. Please submit a cached link or screenshot when submitting content from this domain. We recommend using www.archive.is for this purpose.”*

This kind of notification appears in five different subreddits that discuss mainly politics and news, specifically, The Donald, Mr Trump, TheNewRight, Vote Trump, and Republicans. In particular, in The Donald, there are 13K such comments. AutoModerator blocks URLs from 23 news sources likely to be considered as anti-Trump by that community. In Table 8 we report the number of submissions deleted for each of the sources, along with the percentage over all submissions that include that source. Mainstream news outlets like Washington Post and CNN are the top domains that get removed from The Donald (3.8K and 3.3K submissions, respectively), and this happens slightly less than half the times (44% and 39% of the submissions, respectively). Interestingly, only URLs posted via the URL submission field are censored by AutoModerator, but not URLs that are inserted as part of the title field.

We attempt to estimate possible ad revenue deprivation due to the practice of forcing users to share archive URLs instead of source URLs on Reddit. We do so by providing a conservative approximation of the ad revenue loss. Since we do not have knowledge of how many times a particular URL is clicked, we use the up- and down-votes of a post. That is, we assume that when a user up-votes or down-votes a post, he also clicks on the URL included on the post. This constitutes a best-effort technique as prior work shows that a substantial portion of users on Reddit do not vote [4], while, at the same time, users that do vote do not necessarily read or click on the articles [5].

Table 8 Number and percentage of submissions deleted from The Donald with links to different news sources.

News Source	Count	(%)	News Source	Count	(%)
washingtonpost.com	3,814	44.13%	change.org	96	7.52%
cnn.com	3,354	39.39%	huffpost.com	62	13.39%
nydailynews.com	1,070	46.32%	fusion.net	60	44.77%
huffingtonpost.com	978	43.77%	cnn.it	58	44.61%
nationalreview.com	774	45.58%	altnet.org	26	20.01%
theblaze.com	704	46.74%	infostormer.com	16	27.11%
buzzfeed.com	588	45.97%	dailynewsbin.com	4	26.67%
salon.com	373	44.88%	todayvibes.com	4	7.27%
vice.com	372	45.14%	usanewsjets.ga	4	10.52%
vox.com	323	45.23%	fullycucked.com	1	1.78%
weeklstandard.com	253	46.25%	northcrane.com	1	0.13%
politifact.com	185	33.09%			

That said, this approach is reasonably conservative considering the influence that Reddit has with respect to news dissemination [15]. We then calculate the potential revenue loss using only ad impressions, i.e., we conservatively estimate the revenue generated when a user visits the website without taking into account any potential further action (e.g., clicking on the actual ad). To this end, we use an average Cost per 1,000 impressions (CPM) of \$24.74, as reported by Statista (<https://www.statista.com/statistics/308015>), while we assume an average of 3 ads per page [1]. In other words, we calculate the monthly revenue loss, for each domain, based on the average CPM value as well as the conservative estimate of the visits using the up- and down-votes. Overall, replacing URLs with archive URLs, as done, e.g., by the AutoModerator bot, yields an estimate of \$30K per month in revenue loss (for the top 20 domains in terms of views). This is detailed in Table 9, where we break down the estimate for each of the top 20 revenue deprived domains. On a purely pragmatic level, consider that our conservative estimate of ad revenue deprivation is around \$70K per year for the Washington Post alone. Although a more detailed impact analysis is out of the scope of this work, we suspect that even \$70K could have a real-world effect, e.g., on intern budgets or even early career hires.

Table 9 Top 20 domains with the largest ad revenue losses because of the use of archiving services on Reddit. We report an estimate of the average monthly visits from Reddit and the monthly ad loss.

Domain	Visits	Loss (\$)	Domain	Visits	Loss (\$)
washingtonpost.com	79,880	5,928	wsj.com	11,389	845
cnn.com	70,483	5,231	breitbart.com	11,357	842
nytimes.com	46,442	3,446	bbc.com	10,708	794
huffingtonpost.com	27,125	2,013	salon.com	10,364	769
thehill.com	18,643	1,383	buzzfeed.com	10,359	768
theguardian.com	16,376	1,215	foxnews.com	9,638	715
politico.com	15,774	1,170	yahoo.com	9,497	704
dailymail.co.uk	14,442	1,071	latimes.com	9,277	688
dailycaller.com	12,735	945	vox.com	8,976	667
google.com	11,576	859	washingtontimes.com	8,862	657

5.5. Conclusion

This work presented a large-scale analysis of the use of Web archiving services, such as archive.is and the Wayback Machine, on social media. Our study is based on two data crawls: 1) 21M URLs obtained from the archive.is live feed, covering almost two years, and 2) 356K archive.is plus 391K Wayback Machine URLs that were shared, over 14 months, on Reddit, Twitter, Gab, and 4chan’s Politically Incorrect board (/pol/). We showed that these services are extensively used to archive and disseminate news, social network posts, and controversial content—in particular by users of fringe communities within Reddit and 4chan. We also found that users not only rely on them to ensure persistence of Web content, but also to bypass certain censorship policies of some social networks. Some subreddits, as well as 4chan’s /pol/, actually nudge or force users to share archive URLs instead of direct link to news sources they perceive as having contrasting ideologies, taking away potentially hundreds of thousands of dollars in ad revenue. Overall, our measurements highlight the importance of archiving services in the Web’s information and ad ecosystems, and the need to carefully consider them when studying social media.

5.6. References

- [1] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan. Adscope: harvesting and analyzing online display ads. In WWW, 2014.
- [2] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter. In HyperText, 2017.
- [3] European Commission. General Data Protection Regulation (GDPR), Art. 17. <https://gdpr-info.eu/art-17-gdpr/>, 2017.
- [4] E. Gilbert. Widespread underprovision on Reddit. In CSCW, 2013.
- [5] M. Glenski, C. Pennycuff, and T. Weninger. Consumers and Curators: Browsing and Voting Patterns on Reddit. IEEE Transactions on Computational Social Systems, 2017.
- [6] G. E. Hine, J. Onalapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In ICWSM, 2017.
- [7] J. Jackson. Moderators of pro-Trump Reddit group linked to fake news crackdown on posts. <https://www.theguardian.com/technology/2016/nov/22/moderatorstrump-reddit-group-fake-news-crackdown>, 2016.
- [8] J. Koebler. Dear GamerGate: Please Stop Stealing Our Shit. https://motherboard.vice.com/en_us/article/ypw5mj/deargamergate-please-stop-stealing-our-shit, 2014.
- [9] W. Koehler. A longitudinal study of Web pages continued: A consideration of document persistence. Information Research, 9(2), 2004.
- [10] N. MacFarquhar. A Powerful Russian Weapon: The Spread of False Stories. <https://nyti.ms/2k6880n>, 2016.
- [11] M. Mondal, J. Messias, S. Ghosh, K. P. Gummadi, and A. Kate. Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data. In SOUPS, 2016.
- [12] N. Ralph. VICE Has Disabled Archiving Sites To Stop People Using Their Own Words Against Them. <http://theralphretort.com/vice-disabled-archiving-sites-against-them/>, 2017.
- [13] M. Wendling. The saga of 'Pizzagate': The fake story that shows how conspiracy theories spread. <http://www.bbc.com/news/blogs-trending-38156985>, 2016.
- [14] S. Zannettou, B. Bradlyn, E. De Cristofaro, M. Sirivianos, G. Stringhini, H. Kwak, and J. Blackburn. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? In WWW Companion, 2018.
- [15] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In ACM IMC, 2017.

6. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web

6.1. Project description and motivation

Recent political events and elections have been increasingly accompanied by reports of disinformation campaigns attributed to state-sponsored actors [3]. In particular, “troll farms,” allegedly employed by Russian state agencies, have been actively commenting and posting content on social media to further the Kremlin’s political agenda [10]. In late 2017, the US Congress started an investigation on Russian interference in the 2016 US Presidential Election, releasing the IDs of 2.7K Twitter accounts identified as Russian trolls.

Despite the growing relevance of state-sponsored disinformation, the activity of accounts linked to such efforts has not been thoroughly studied. Previous work has mostly looked at campaigns run by bots [3, 4, 8]; however, automated content diffusion is only a part of the issue, and in fact recent research has shown that human actors are actually key in spreading false information on Twitter [9]. Overall, many aspects of state-sponsored disinformation remain unclear, e.g., how do state-sponsored trolls operate? What kind of content do they disseminate? And, perhaps more importantly, is it possible to quantify the influence they have on the overall information ecosystem on the Web?

In this work, we aim to address these questions, by relying on the set of 2.7K accounts released by the US Congress as ground truth for Russian state-sponsored trolls. From a dataset containing all tweets released by the 1% Twitter Streaming API, we search and retrieve 27K tweets posted by 1K Russian trolls between January 2016 and September 2017. We characterize their activity by comparing to a random sample of Twitter users. Then, we quantify the influence of these trolls on the greater Web, looking at occurrences of URLs posted by them on Twitter, Reddit, and 4chan. Finally, we use Hawkes Processes [5] to model the influence of each Web community (i.e., Russian trolls on Twitter, overall Twitter, Reddit, and 4chan) on each other.

Main findings. Our study leads to several key observations:

- (1) Trolls actually bear very small influence in making news go viral on Twitter and other social platforms alike. A noteworthy exception are links to news originating from RT (Russia Today), a state-funded news outlet: indeed, Russian trolls are quite effective in “pushing” these URLs on Twitter and other social networks.
- (2) The main topics discussed by Russian trolls target very specific world events (e.g., Charlottesville protests) and organizations (such as ISIS), and political threads related to Donald Trump and Hillary Clinton.

- (3) Trolls adopt different identities over time, i.e., they “reset” their profile by deleting their previous tweets and changing their profile name/information.
- (4) Trolls exhibit significantly different behaviors compared to other (random) Twitter accounts. For instance, the locations they report concentrate in a few countries like the USA, Germany, and Russia, perhaps in an attempt to appear “local” and more effectively manipulate opinions of users from those countries. Also, while random Twitter users mainly tweet from mobile versions of the platform, the majority of the Russian trolls do so via the Web Client.

6.2. Dataset

Russian trolls. We start from the 2.7K Twitter accounts suspended by Twitter because of connections to Russia’s Internet Research Agency troll farm. The list of these accounts was released by the US Congress as part of their investigation of the alleged Russian interference in the 2016 US presidential election and includes both Twitter’s user id (which is a numeric unique identifier associated to the account) and the handle. ¹ From a dataset storing all tweets released by the 1% Twitter Streaming API, we search for tweets posted between January 2016 and September 2017 by the user ids of the trolls. Overall, we obtain 27K tweets from 1K out of the 2.7K Russian troll accounts.

Note that the criteria used by Twitter to identify these troll accounts are not public. What we do know, is this is not the complete set of active Russian trolls, because 6 days prior to this writing Twitter announced they have discovered over 1K more troll accounts.² Nonetheless, it constitutes an invaluable “ground truth” dataset enabling efforts to shed light on the behavior of state-sponsored troll accounts.

Baseline dataset. We also compile a list of random Twitter users, while ensuring that the distribution of the average number of tweets per day posted by the random users is similar to the one by trolls. To calculate the average number of tweets posted by an account, we find the first tweet posted after January 1, 2016 and retrieve the overall tweet count. This number is then divided by the number of days since account creation. Having selected a set of 1K random users, we then collect all their tweets between January 2016 and September 2017, obtaining a total of 96K tweets.

We follow this approach as it gives a good approximation of posting behavior, even though it might not be perfect, since (1) Twitter accounts can become more or less active over time, and (2) our datasets are based on the 1% Streaming API, thus, we are unable to control the number of tweets we obtain for each account.

6.3. Analysis

6.3.1 General Characterization

Account creation. First, we examine the dates when the state-sponsored accounts infiltrated Twitter, by looking at the account creation dates. From Figure 11, we observe that 71% of them are actually created before 2016. There are some interesting peaks, during 2016 and 2017: for instance, 24 accounts are created on July 12, 2016, approximately a week before the Republican National Convention (when Donald Trump received the nomination), while 28 appear on August 8, 2017, a few days before the infamous Unite the Right rally in Charlottesville [11]. Taken together, this might be evidence of coordinated activities aimed at manipulating users’ opinions on Twitter with respect to specific events.

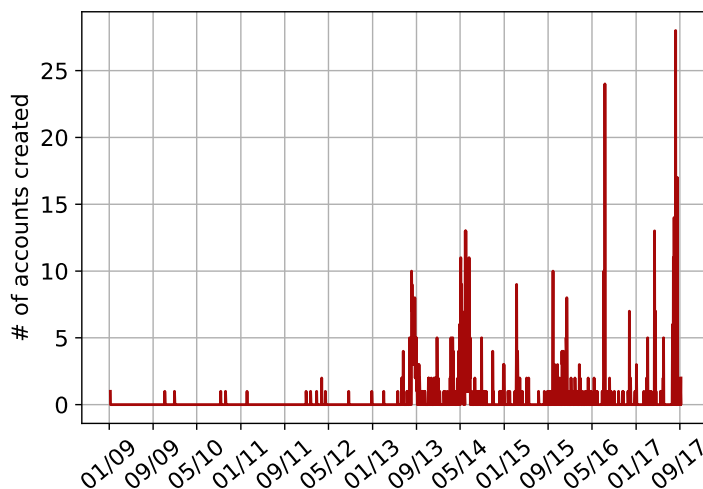


Figure 11 Number of Russian troll accounts created per day

Account Characteristics. We also shed light on the troll account profile information. In Table 10, we report the top ten words appearing in the names and the descriptions of Russian trolls, as well as character 4-grams for account names and word bigrams for profile descriptions. Interestingly, a substantial number of Russian trolls pose as news outlets, evident from the use of the term “news” in both the name (1.3%) and the description (10.7%). Also, it seems they attempt to increase the number of their followers, thus their reach of Twitter users, by nudging users to follow them (see, e.g., “follow me” appearing in almost 8% of the accounts). Finally, 10.3% of the Russian trolls describe themselves as Trump supporters: “trump” and “maga” (Make America Great Again, one of Trump campaign’s main slogans).

Table 10 Top 10 words found in Russian troll screen names and account descriptions. We also report character 4-grams for the usernames and word bigrams for the description.

Word	Name		Word	Description	
	(%)	4-gram (%)		(%)	Word bigram (%)
news	1.3%	news 1.5%	news	10.7%	follow me 7.8%
bote	1.2%	line 1.5%	follow	10.7%	breaking news 2.6%
online	1.1%	blac 1.3%	conservative	8.1%	news aus 2.1%
daily	0.8%	bote 1.3%	trump	7.8%	uns in 2.1%
today	0.6%	rist 1.1%	und	6.2%	deiner stdt 2.1%
ezeziel2517	0.6%	nlin 1.1%	maga	5.9%	die news 2.1%
maria	0.5%	onli 1.0%	love	5.8%	wichtige und 2.1%
black	0.5%	lack 1.0%	us	5.3%	nachrichten aus 2.1%
voice	0.4%	bert 1.0%	die	5.0%	aus deiner 2.1%
martin	0.4%	poli 1.0%	nachrichten	4.3%	die dn 2.1%

Client. We analyze the clients used to post tweets. We do so since previous work [2] shows that the client used by official or professional accounts are quite different that the ones used by regular users. Table 11 reports the top 10 clients for both Russian trolls and baseline users. We find the latter prefer to use Twitter clients for mobile devices (48%) and the TweetDeck dashboard (32%), whereas, the former mainly use the Web client (50%).

Table 11 Top 10 Twitter clients (as % of tweets)

Client (Trolls)	(%)	Client (Baseline)	(%)
Twitter Web Client	50.1%	TweetDeck	32.6%
twitterfeed	13.4%	Twitter for iPhone	26.2%
Twibble.io	9.0%	Twitter for Android	22.6%
IFTTT	8.6%	Twitter Web Client	6.1%
TweetDeck	8.3%	GrabInbox	2.0%
NovaPress	4.6%	Twitter for iPad	1.4%
dlvr.it	2.3%	IFTTT	1.0%
Twitter for iPhone	0.8%	twittbot.net	0.9%
Zapier.com	0.6%	Twitter for BlackBerry	0.6%
Twitter for Android	0.6%	Mobile Web (M2)	0.4%

Location. We then study users’ location, relying on the self-reported location field in their profiles. Note that users not only may leave it empty, but also change it any time they like, so we look at locations for each tweet. We retrieve it for 75% of the tweets by Russian trolls, gathering 261 different entries, which we convert to a physical location using the Google Maps Geocoding API (<https://developers.google.com/maps/documentation/geocoding/start>). The API does not return results for 11 queries, as they correspond to non-existing locations (e.g., “block corner street”).

In the end, we obtain 178 unique coordinate locations for the trolls, as depicted in Figure 12 Distribution of reported locations for tweets by Russian trolls (red circles) and baseline (green triangles).(red circles). The size of the circles on the map indicates the number of tweets that appear on each location. We do the same for the baseline, getting 2,037 different entries, converted by the API to 894 unique locations (950 queries do not return results). We observe that most of the tweets from Russian trolls come from locations within the USA and Russia, and some from European countries, like Germany, Belgium, and Italy. Whereas, tweets in our baseline are more uniformly distributed across the globe, with many tweets from North and South America, Europe, and Asia. This suggests that Russian trolls may be pretending to be from certain countries, e.g., USA or Germany, aiming to pose as locals and better manipulate opinions. This explanation becomes more plausible when we consider that a plurality of troll account tweets have their location set as a generic form of “US,” as opposed to a specific city, state, or even region. Interestingly, the 2nd, 3rd, and 4th most popular location for troll accounts to tweet from are Moscow, St. Petersburg, and a generic form of “Russia.” We also assess whether users change their country of origin based on the self-reported location. We find a negligible percentage (1%) of trolls that change their country, whereas for the baseline the corresponding percentage is 16%.

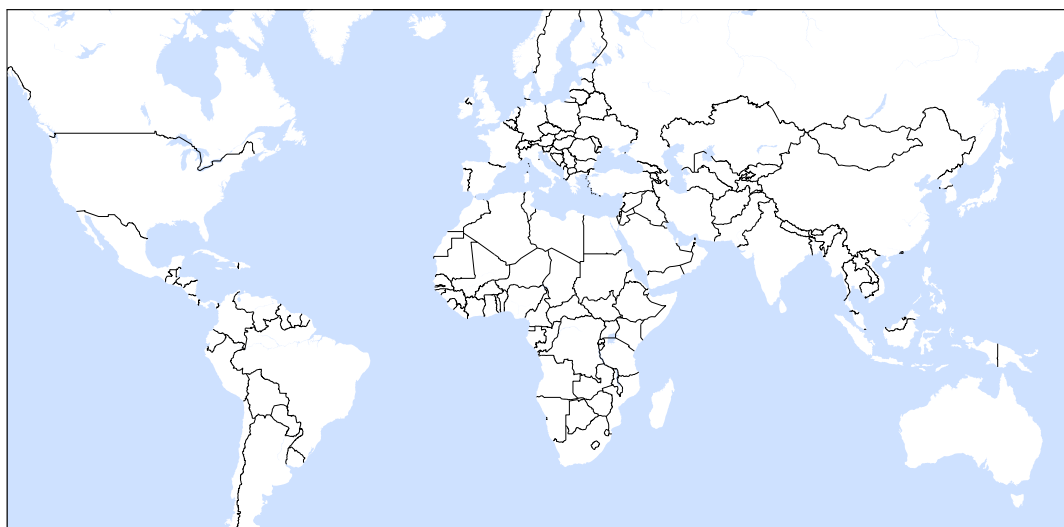


Figure 12 Distribution of reported locations for tweets by Russian trolls (red circles) and baseline (green triangles).

Hashtags. Our next step is to study the use of hashtags in tweets. Russian trolls use at least one hashtag in 32% of their tweets, compared to 10% for the baseline. Overall, we find 4.3K and 7.1K unique hashtags for trolls and random users, respectively, with 74% and 78% of them only appearing once. In Table 12, we report the top 20 hashtags for both datasets. State-sponsored trolls appear to use hashtags to disseminate news (7.2%) and politics (2.6%) related content, but also use several that might be indicators of propaganda and/or controversial topics, e.g., #ISIS, #IslamKills, and #BlackLivesMatter. For instance, we find among the tweets in our dataset some notable examples including: “We just have to close

the borders, ‘refugees’ are simple terrorists #IslamKills” on March 22, 2016, “#SyrianRefugees ARE TERRORISTS from #ISIS #IslamKills” on March 22, 2016, and “WATCH: Here is a typical #BlackLivesMatter protester: ‘I hope I kill all white babes!’ #BatonRouge” on July 17, 2016. Note that [that](#) denotes a link. We also study when these hashtags are used by the trolls, finding that most of them are well distributed over time. However, there are some interesting exceptions, e.g., with #Merkelmussbleiben (a hashtag seemingly supporting German Chancellor Angela Merkel) and #IslamKills. Specifically, tweets with the former appear exclusively on July 21, 2016, while the latter on March 22, 2016, when a terrorist attack took place at Brussels airport. These two examples illustrate how the trolls may be coordinating to push specific narratives on Twitter.

Table 12 Top 20 hashtags in tweets from Russian trolls and baseline users.

Hashtag	Trolls		Baseline				
	(%)	Hashtag	(%)	Hashtag	(%)		
news	7.2%	US	0.7%	iHeartAwards	1.8%	UrbanAttires	0.6%
politics	2.6%	tcot	0.6%	BestFanArmy	1.6%	Vacature	0.6%
sports	2.1%	PJNET	0.6%	Harmonizers	1.0%	mPlusPlaces	0.6%
business	1.4%	entertainment	0.5%	iOSApp	0.9%	job	0.5%
money	1.3%	top	0.5%	JouwBaan	0.9%	Directioners	0.5%
world	1.2%	topNews	0.5%	vacature	0.9%	JIMIN	0.5%
MAGA	0.8%	ISIS	0.4%	KCA	0.9%	PRODUCE101	0.5%
health	0.8%	Merkelmussbleiben	0.4%	Psychic	0.8%	VoteMainFPP	0.5%
local	0.7%	IslamKills	0.4%	RT	0.8%	Werk	0.4%
BlackLivesMatter	0.7%	breaking	0.4%	Libertad2016	0.6%	dts	0.4%

6.3.2 Account Evolution

Name Changes. Previous work [7] has shown that malicious accounts often change their profile name in order to assume different identifies. Therefore, we investigate whether state-sponsored trolls show a similar behavior, as they might change the narrative with which they are attempting to influence public opinion. Indeed, we find that 9% of the accounts operated by Russian trolls change their screen name, up to 4 times during the course of our dataset. Some examples include changing screen names from “OnlineHouston” to “HoustonTopNews”, or “Jesus Quintin Perez” to “WorldNewsPolitics,” in a clear attempt to pose as news-related accounts. In our baseline, we find that 19% of the accounts changed their Twitter screen names, up to 11 times during our dataset; highlighting that changing screen names is a common behavior of Twitter users in general.

Followers/Friends. Next, we look at the number of followers and friends (i.e., the accounts one follows) of the Russian trolls, as this is an indication of the possible overall impact of a tweet. In Figure 13, we plot the CDF of the number of followers per tweet measured at the time of that tweet. On average, Russian trolls have 7K followers and 3K friends, while our baseline has 25K followers and 6K friends. We also note that in both samples, tweets

reached a large number of Twitter users; at least 1K followers, with peaks up to 145K followers. These results highlight that Russian troll accounts have a non-negligible number of followers, which can assist in pushing specific narratives to a much greater number of Twitter users. We also assess the evolution of the Russian trolls in terms of the number of their followers and friends. To this end, we get the follower and friend count for each user on their first and last tweet and calculate the difference. Figure 14 plots the CDF of the increase/decrease of the followers and friends for each Russian troll as well as random user in our baseline. We observe that, on average, Russian trolls increase their number of followers and friends by 2,065 and 1,225, respectively, whereas for the baseline we observe an increase of 425 and 133 for followers and friends, respectively. This suggests that Russian trolls work hard to increase their reachability within Twitter.

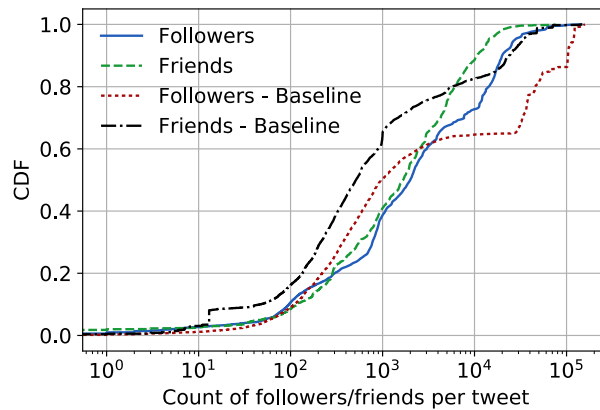


Figure 13 CDF of the number of followers/friends for each tweet

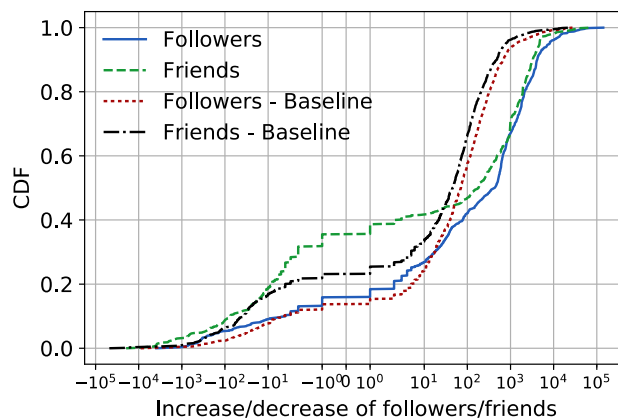


Figure 14 CDF of the number of increase in followers/friends for each user from the first to the last tweet

Tweet Deletion. Arguably, a reasonable strategy to avoid detection after posting tweets that aim to manipulate other users might be to delete them. This is particularly useful when troll accounts change their identity and need to modify the narrative that they use to influence public opinion. With each tweet, the Streaming API returns the total number of available tweets a user has up to that time. Retrieving this count allows us to observe if a user has deleted a tweet, and around what period; we call this an “observed deletion.” Recall that our dataset is based on the 1% sample of Twitter, thus, we can only estimate, in a conservative way, how many tweets are deleted; more specifically, in between subsequent tweets, a user may have deleted and posted tweets that we do not observe. In Figure 15, we plot the CDF of the number of deleted tweets per observed deletion. We observe that 13% of the Russian trolls delete some of their tweets, with a median percentage of tweet deletion equal to 9.7%. Whereas, for the baseline set, 27% of the accounts delete at least one tweet, but the median percentage is 0.1%. This means that the trolls delete their tweets in batches, possibly trying to cover their tracks or get a clean slate, while random users make a larger number of deletions but only a small percentage of their overall tweets, possibly because of typos.

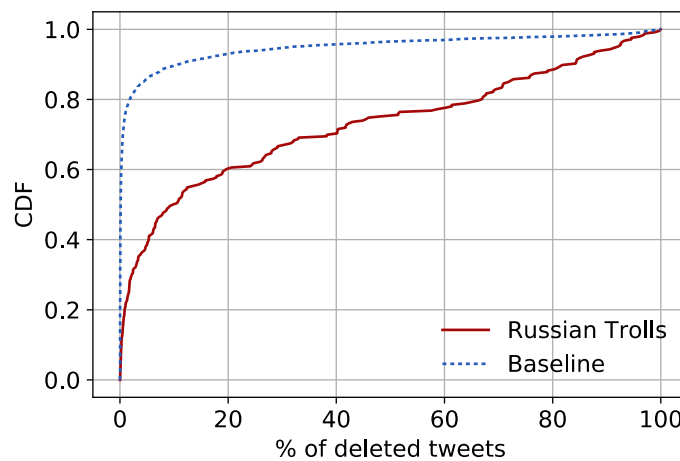


Figure 15 CDF of number of deleted tweets per observed deletion.

6.3.3 Influence Estimation

Thus far, we have analyzed the behavior of the Russian trolls on the Twitter platform, and how this differs from that of a baseline of random users. Allegedly, their main goal is to ultimately manipulate the opinion of other users and extend the cascade of disinformation they share (e.g., other users post similar content) [1]. Therefore, we now set out to shed light on their impact, in terms of the dissemination of disinformation, on Twitter and on the greater Web.

To assess their influence, we look at the URLs posted by four groups of users: Russian trolls on Twitter, “normal” accounts on Twitter, Reddit users, and 4chan users (/pol/ board). For each unique URL, we fit a statistical model known as Hawkes Processes [5, 6], which allows us to estimate the strength of connections between each of these four groups in terms of how likely an event – the URL being posted by either trolls or normal users to a particular platform – is to cause subsequent events in each of the groups. For example, a strong connection from Reddit to /pol/ would mean that a URL that appears on Reddit is likely to be seen and then re-posted on /pol/; whereas, a weak connection from Russian trolls to normal users on Twitter indicates that a URL posted by Russian trolls is less likely to be re-tweeted or re-posted by the latter. We fit the Hawkes Processes using the methodology presented by [13].

To study the dissemination of different types of content, we look at three different sets of URLs: 1) The complete set of all URLs posted by Russian troll accounts; 2) The subset of URLs for Russian state-sponsored news, namely, RT (Russia Today); and 3) The subset of URLs from other news sources, including both mainstream and alternative, using the list provided by [13]. Note that we initially planned to also include Sputnik news as a Russian state-sponsored outlet, but we did not find many instances of Sputnik URLs.

Table 8 summarizes the number of URLs, number of events (i.e., occurrences of a given URL) as well as the mean background rate for each category and social network. The background rate defines the rate at which events occur excluding the influence of the platforms included in the model; the background rate includes events created spontaneously on each network, such as by a user reading the original article and then posting a link to, or those generated by another platform not monitored by us, such as Facebook. The number of events for Russian state-sponsored news sources is substantially lower than the number of events from other mainstream and alternative news sources. This is expected since the former only includes one news source (RT), however, it is interesting that the background rates for these URLs are higher than for other mainstream and alternative news, meaning that events from Russian state-sponsored news are more likely to occur spontaneously, either from platforms we do not measure or posted by users directly.

		/pol/	Reddit	Twitter	Trolls
URLs	Russian state-sponsored	6	13	19	19
	Other news sources	47	168	159	192
	All	127	482	861	989
Events	Russian state-sponsored	19	42	118	19
	Other news sources	720	3,055	2,930	195
	All	1,685	9,531	1,537,612	1,461
Mean λ_0	Russian state-sponsored	0.0824	0.1865	0.2264	0.1228
	Other news sources	0.0421	0.1447	0.1544	0.0663
	All	0.0324	0.1557	0.1553	0.0753

Figure 16 Total URLs with at least one event in Twitter, /pol/, Reddit, and Russian trolls on Twitter; total events for Russian state-sponsored news URLs, other news URLs and all the URLs; and mean background rate (λ_0) for each platform.

Fitting a Hawkes model yields a weight matrix, which characterizes the strength of the connections between the groups we study. Each weight value, represents the connection strength from one group to another and can be interpreted as the expected number of subsequent events that will occur on the second group after each event on the first. The mean weight values over all URLs, as well as for the URLs from Russian state-sponsored outlets and other mainstream and alternative URLs are presented in Figure 17 and Figure 18.

In Figure 17, which shows the mean weights for all URLs, we observe that for /pol/, Reddit, and normal users on Twitter, the greatest weights are from each group to itself, meaning that reposts/retweets on the same site are more common than sharing the URL to the other platforms. For the Russian Trolls on Twitter, however, the weight is greater from the trolls to Twitter than from the trolls to themselves, perhaps reflecting their use as an avenue for disseminating information to normal Twitter users.

From Figure 18, we observe that, in most cases, the connections are stronger for non-Russia state-sponsored news, indicating that regular users are more inclined to share news articles from mainstream and alternative news sources. Looking at the Russian trolls and normal Twitter users, we see that the trolls are more likely to retweet or repost Russian state-sponsored URLs from normal Twitter users than other news sources; conversely, normal Twitter users are more likely to retweet or repost Russian state-sponsored URLs from the troll accounts.

	/pol/	Reddit	Twitter	Trolls
/pol/	0.127	0.125	0.124	0.120
Reddit	0.106	0.167	0.129	0.114
Twitter	0.064	0.083	0.263	0.075
Trolls	0.108	0.116	0.125	0.120

Destination

Figure 17 Mean weights for all URLs in our dataset

	/pol/	Reddit	Twitter	Trolls
/pol/	R: 0.1016 O: 0.1328 -23.5%*	R: 0.1286 O: 0.1224 5.1%	R: 0.1043 O: 0.1205 -13.4%	R: 0.0809 O: 0.1110 -27.1%
Reddit	R: 0.1040 O: 0.0992 4.8%	R: 0.1144 O: 0.1969 -41.9%	R: 0.1309 O: 0.1393 -6.0%	R: 0.1240 O: 0.1224 1.3%
Twitter	R: 0.0848 O: 0.0825 2.7%	R: 0.1086 O: 0.1252 -13.2%	R: 0.1783 O: 0.2465 -27.7%	R: 0.0969 O: 0.0815 18.9%
Trolls	R: 0.0658 O: 0.1144 -42.5%	R: 0.0995 O: 0.1130 -11.9%	R: 0.1668 O: 0.1113 49.8%**	R: 0.1150 O: 0.1157 -0.6%

Destination

Figure 18 Mean weights for news URLs categorized as Russian state-sponsored (R) and other mainstream and alternative news URLs (O). We also show the percent of increase/decrease between the two categories (also indicated by coloration). Note that * and ** refer to statistical significance with, resp., $p < 0.05$ and $p < 0.01$.

In order to assess the significance of our results, we perform two-sample Kolmogorov-Smirnov tests on the weight distributions for the Russian state-sponsored news URLs and the other news URLs for each source-destination platform pair (depicted as stars in the Figure 18). Small p value means there is a statistically significant difference in the way that Russian state-sponsored URLs propagate from the source to the destination platform. Most of the source-destination pairs have no statistical

significance, however for the Russian trolls–Twitter users pair, we find statistically significance difference with $p < 0.01$.

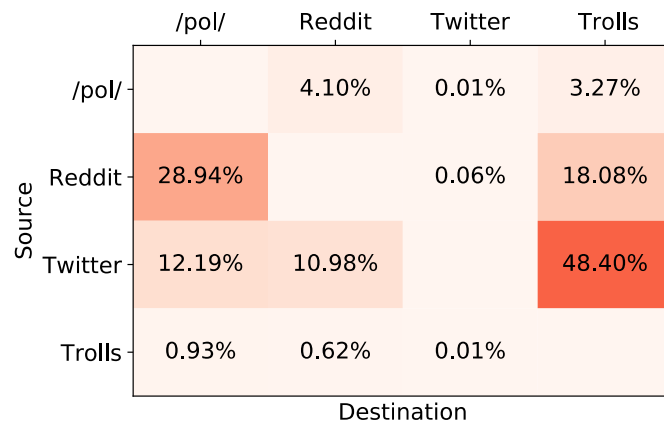


Figure 19 Estimated mean percentages of events created because of other events for all URLs

In Figure 19 and Figure 20, we report the estimated total impact for each pair of platforms, for both Russian state-sponsored news, other news sources as well as all the observed URLs. We determine the impact by calculating, based on the estimated weights and the number of events, the percentage of events on a destination platform caused by events on a source platform, following the methodology presented by [13]. (We omit the details due to space constraints.)

For all URLs (Figure 19), we find that the influence of Russian trolls is negligible on Twitter (0.01%), while for /pol/ and Reddit it is slightly higher (0.93% and 0.62%, respectively). For other pairs, the larger impacts are between Reddit–/pol/ and Twitter–Russian trolls, mainly due to the larger population of users. Looking at the estimated impact for Russian state-sponsored and other news sources (Figure 20), we note that the Russian trolls influenced the other platforms approximately the same for alternative and mainstream news sources (0.72%, 0.62%, and 0.61 for /pol/, Reddit, and Twitter, respectively). Interestingly, Russian trolls have a much larger impact on all the other platforms for the Russian state-sponsored news when compared to the other news sources: approximately 2 times more on /pol/, 5 times more on Reddit, and 4 times more on Twitter.

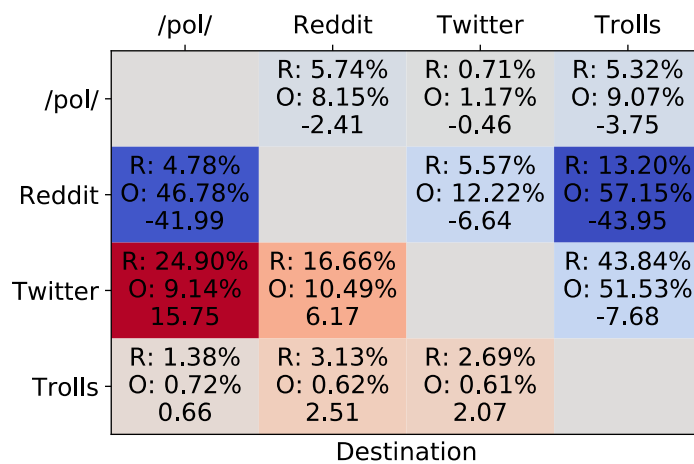


Figure 20 Estimated mean percentages of events created because of other events for Russian state-sponsored URLs (R) and other mainstream and alternative news URLs (O). We also show the difference between the two categories of news (also indicated by coloration).

6.4. Conclusion

In this work, we analyzed the behavior and use of the Twitter platform by Russian state-sponsored trolls during the course of 21 months. We showed that Russian trolls exhibited interesting differences when compared with a set of random users, actively disseminate politics-related content, adopted multiple identities during their account’s lifespan, and that they aimed to increase their impact on Twitter by increasing their followers. Also, we quantified the influence that Russian trolls have on Twitter, Reddit, and /pol/ using a statistical model known as Hawkes Processes. Our findings show that trolls’ influence was not substantial with respect to the other platforms, with the significant exception of news published by the Russian state-sponsored news outlet RT (Russia Today). Our study also prompts some directions for future work. For instance, the consistent reinventing of troll accounts’ identities, batch message deletion, and aggressive collection of friends and followers could prove useful for designing detection and mitigation techniques. In particular, we believe our findings motivate the need for more sophisticated measurements of influence. Our analysis indicates that the troll accounts were not terribly effective in spreading disinformation, but it is likely that these curated troll accounts left simple information diffusion to automated bots, focusing, instead, on more nuanced methods of manipulation.

6.5. References

- [1] S. Earle. TROLLS, BOTS AND FAKE NEWS: THE MYSTERIOUS WORLD OF SOCIAL MEDIA MANIPULATION. <https://goo.gl/nz7E8r>, 2017.
- [2] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. Towards detecting compromised accounts on social networks. IEEE TDSC, 2017.
- [3] E. Ferrara. Disinformation and social bot operations in the run up to the 2017 French presidential election. ArXiv 1707.00086, 2017.
- [4] S. Hegelich and D. Janetzko. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In ICWSM, 2016.
- [5] S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. In ICML, 2014.
- [6] S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks. ArXiv 1507.03228, 2015.
- [7] E. Mariconti, J. Onaolapo, S. S. Ahmad, N. Nikiforou, M. Egele, N. Nikiforakis, and G. Stringhini. What's in a Name?: Understanding Profile Name Reuse on Twitter. In WWW, 2017.
- [8] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Goncalves, A. Flammini, and F. Menczer. Detecting and Tracking Political Abuse in Social Media. In ICWSM, 2011.
- [9] K. Starbird. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In ICWSM, 2017.
- [10] The Independent. St Petersburg 'troll farm' had 90 dedicated staff working to influence US election campaign. <https://ind.pn/2yuCQdy>, 2017.
- [11] Wikipedia. Unite the Right rally. https://en.wikipedia.org/wiki/Unite_the_Right_rally, 2017.
- [12] F. M. F. Wong, C.-W. Tan, S. Sen, and M. Chiang. Quantifying Political Leaning from Tweets and Retweets. In ICWSM, 2013.
- [13] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In ACM IMC, 2017.

7. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources

7.1. Project description and motivation

Over the past few years, a number of high-profile conspiracy theories and false stories have originated and spread on the Web. After the Boston Marathon bombings in 2013, a large number of tweets started to claim that the bombings were a “false flag” perpetrated by the United States government [17]. Also, the GamerGate controversy started as a blogpost by a jaded ex-boyfriend that turned into a pseudo-political campaign of targeted online harassment [2]. More recently, the Pizzagate conspiracy [19] – a debunked theory connecting a restaurant and members of the US Democratic Party to a child sex ring – led to a shooting in a Washington DC restaurant [8]. These stories were all propagated, in no small part, via the use of “alternative” news sites like Infowars and “fringe” Web communities like 4chan. Overall, the barrier of entry for such alternative news sources has been greatly reduced by the Web and large social networks. Due to the negligible cost of distributing information over social media, fringe sites can quickly gain traction with large audiences. At the same time, the explosion of information sources also hinders the effective regulation of the sector, while further muddying the water when it comes to the evaluation of news information by readers.

While there are many plausible motives for the rise in alternative narratives [16], ranging from libelous (e.g., to harm the image of a particular person or group), political (e.g., to influence voters), profit (e.g., to make money from advertising), or trolling [1], the manner in which they proliferate throughout the Web is still unknown. Although previous work has examined information cascades, rumors, and hoaxes [5, 11, 15], to the best of our knowledge, very little work provides a holistic view of the modern information ecosystem. This knowledge, however, is crucial for understanding the alternative news world and for designing appropriate detection/mitigation strategies. Anecdotal evidence and press coverage suggest that alternative news dissemination might start on fringe sites, eventually reaching mainstream online social networks and news outlets [14, 18]. Nevertheless, this phenomenon has not been measured and no thorough analysis has focused on how news moves from one online service to another.

In this work, we address this gap by providing the first largescale measurement of how mainstream and alternative news flows through multiple social media platforms. We focus on the relationship between three fundamentally different social media platforms, Reddit, Twitter, and 4chan, which we choose because of: 1) their fundamental differences as well as

their generally accepted “driving” of substantial portions of the online world; 2) anecdotal evidence that suggests that specific sub-communities within Reddit and 4chan act as generators [18] and incubators [10] of fake news stories; and 3) the substantial impact they have in forming and manipulating peoples’ opinions (and therefore actions), when they constantly disseminate false information [8].

Contributions. First, we undertake a large-scale measurement and comparison of the occurrence of mainstream and alternative news sources across three social media platforms (4chan, Reddit, and Twitter). Then, we provide an understanding of the temporal dynamics of how URLs from news sites are posted on the different social networks. Finally, we present a measurement of the influence between the platforms that provides insight into how information spreads throughout the greater Web. Overall, our findings indicate that Twitter, Reddit, and 4chan are used quite extensively for the dissemination of both alternative and mainstream news. Using a statistical model for influence – namely, Hawkes processes – we show that each of the platforms (and, in the case of Reddit, sub-communities) have varying degrees of influence on each other, and this influence differs with respect to mainstream and alternative news sources.

7.2. Methodology

News sites. Our analysis uses a set of news websites that can confidently be labeled as either “mainstream” or “alternative” news. More specifically, we create a list of 99 news sites including 45 mainstream and 54 alternative ones (list available at https://drive.google.com/open?id=0ByP5a__khV0dM1ZSY3YxQWF2N2c). For the former, we select 45 from the Alexa top 100 news sites, leaving out those based on user-generated content, those serving specialized content (e.g., finance news), as well as non-English sites. For the latter, we use Wikipedia (https://en.wikipedia.org/wiki/List_of_fake_news_websites) and FakeNewsWatch (<http://fakenewswatch.com/>). We also add two state-sponsored alternative news domains: sputniknews.com and rt.com, as they have recently attracted public attention due to their posting of controversial, and seemingly agenda-pushing stories [3].

We gather information from posts, threads, and comments on Twitter, Reddit, and 4chan that contain URLs from the 99 news sites. Our datasets cover activity on the three platforms between June 30, 2016 and February 28, 2017. Table 13 shows the total number of posts/comments crawled and the percentage of posts that contains links to URLs from the aforementioned news domains. We observe that mainstream news URLs are present in a greater percentage of posts on 4chan and Reddit than on Twitter, while alternative ones are about twice as likely to appear in posts on 4chan than on Twitter or Reddit. Table 14 provides a summary of our datasets, which we present in more detail below.

Table 13 : Total number of posts crawled and percentage of posts that contain URLs to our list of alternative and mainstream news sites.

Platform	Total Posts	% Alt.	% Main.
Twitter	587M	0.022%	0.070%
Reddit (posts + comments)	332M	0.023%	0.181%
4chan	42M	0.050%	0.197%

Table 14 : Number of posts/comments that contain a URL to one of our information sources, as well as the number of unique URLs linking to alternative and mainstream news sites in our list.

Platform	Posts/Comments	Alt. URLs	Main. URLs
Twitter	486,700	42,550	236,480
Reddit (six selected subreddits)	620,530	40,046	301,840
Reddit (all other subreddits)	1,228,105	24,027	726,948
4chan (/pol/)	90,537	8,963	40,164
4chan (/int/, /sci/, /sp/)	7,131	615	5,513

Twitter. We collect the 1% of all publicly available tweets with URLs from the aforementioned news domains between June 30, 2016 and February 28, 2017 using the Twitter Streaming API. In total, we gather 487k tweets containing 279k unique URLs pointing to mainstream or alternative news sites. Due to a failure in our collection infrastructure, we have some gaps in the Twitter dataset, specifically between Oct 28–Nov 2 and Nov 5–16, 2016, as well as Nov 22, 2016 – Jan 13, 2017, and Feb 24–28, 2017.

Reddit. We obtain all posts and comments on Reddit between June 30, 2016 and February 28, 2017, using data made available on Pushshift. We collect approximately 42M posts, 390M comments, and 300k subreddits. Once again, we filter posts and comments that contain URLs from one of the 99 news sites, which yields a dataset of 1.8M posts/comments and approximately 1.1M URLs.

4chan. For 4chan, we use all threads and posts made on the Politically Incorrect (/pol/) board, as well as /sp/ (Sports), /int/ (International), and /sci/ (Science) boards for comparison, using the same methodology as [9]. We opt to select both not safe for work boards (i.e., /pol/) and safe for work boards (i.e., /sp/, /int/, and /sci/) to observe how these compare to each other with respect to the dissemination of news. The resulting dataset includes 97k posts.

7.3. Results

7.3.1 Temporal Analysis

In this section, we present the results of our temporal analysis of the way news are posted on Twitter, Reddit, and 4chan.

URL Occurrence. In Figure 21, we measure the daily occurrence of news URLs over the three platforms normalized by the average daily number of URLs shared in each community.⁶ We find that /pol/ and the six selected subreddits exhibit a much higher percentage of occurrences of alternative news compared to the other communities (Figure 21(a)), whereas, for mainstream news, the sharing behavior is more similar across platforms (Figure 21(b)). There are also some interesting spikes, likely related to the 2016 US elections, on the date of the first presidential debate and election day itself. These findings indicate that the selected sub-communities are heavily utilized for the dissemination of alternative news. We also study the fraction of alternative news URLs with respect to overall news URLs (Figure 21(c)), highlighting that mainstream news URLs are overall more “popular” than the alternative news URLs. Note that the Twitter spike in Figure 21(c) appears to be an artifact of a failure in our collection infrastructure.

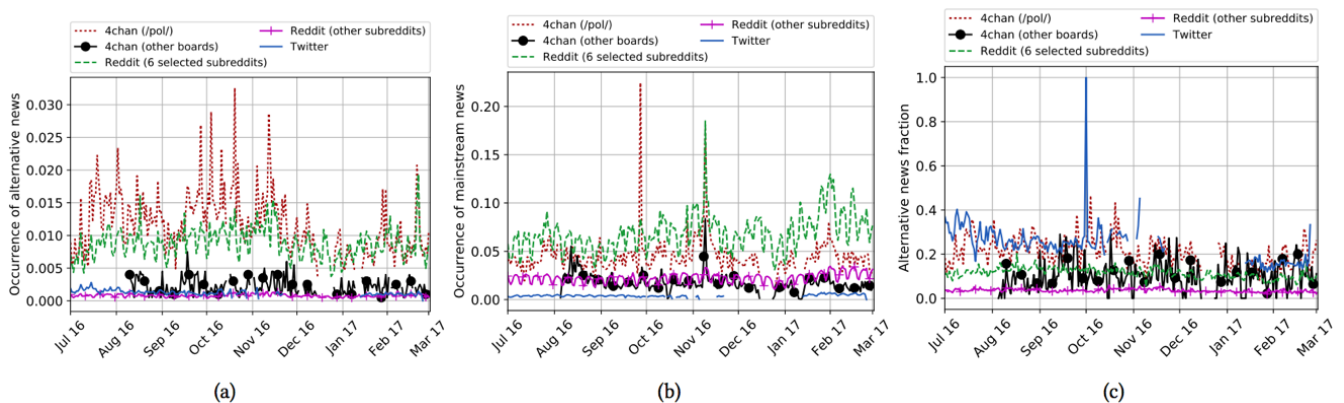


Figure 21 Normalized daily occurrence of URLs for (a) alternative news, (b) mainstream news, and (c) fraction of alternative news over all news.

Cross Platform Analysis. Next, given the set of unique URLs across all platforms and the time they appear for the first time, we analyze their appearance in one, two, or three platforms, and the order in which this happens. For each URL, we find the first occurrence on each platform and build corresponding “sequences,” e.g., if a URL first appears on the six subreddits (Reddit) and subsequently on /pol/ (4chan), the sequence is Reddit→4chan (R→4). Table 15 reports the distribution of the sequences of appearances considering only the first hop, i.e., up to the first two platforms in the sequence. The majority of URLs only appear on one platform: 82% of alternative URLs and 89% of mainstream news URLs. Also,

both alternative and mainstream news URLs tend to appear on the six subreddits first and later appear on either Twitter or /pol/, and on Twitter before /pol/.

Table 15 Distribution of URLs according to the sequence of first appearance within platforms for all URLs, considering only the first hop. “4” stands for /pol/ (4chan), “R” for the six selected subreddits (Reddit), and

Sequence	Alternative (%)		Mainstream (%)	
4 only	3,236	(4.4%)	18,654	(3.7%)
4→R	1,118	(1.5%)	4,606	(0.9%)
4→T	315	(0.5%)	861	(0.17%)
R only	24,292	(33.3%)	230,602	(46.1%)
R→4	2,181	(3.0%)	11,307	(2.3%)
R→T	4,769	(6.5%)	16,685	(3.35%)
T only	32,443	(44.5%)	204,836	(41%)
T→4	585	(0.8%)	1,345	(0.26%)
T→R	3,964	(5.5%)	10,640	(2.12%)

“T” for Twitter.

We also study the temporal dynamics of URLs that appear on all three platforms, with triplets of sequences. Table 16 reports the distribution of these sequences. The most common sequences are similar for both alternative and mainstream news URLs: R→T→4, R→4→T, and T→R→4 are the top three sequences. As already mentioned, the six selected subreddits “outperform” both other platforms in terms of the speed of sharing mainstream and alternative news URLs, as evidenced by the fact that it is at the head of the sequence for 51% and 59% of alternative and mainstream news URLs, respectively.

Table 16 Distribution of URLs according to the sequence of first appearance within a platform for URLs common to all platforms. “4” stands for /pol/ (4chan), “R” for the six selected subreddits (Reddit), and “T”

Sequence	Alternative (%)		Mainstream (%)	
4→R→T	128	(5.5%)	552	(8.9%)
4→T→R	145	(6.2%)	290	(4.7%)
R→4→T	335	(14.4%)	1,525	(24.5%)
R→T→4	841	(36.3%)	2,189	(35.3%)
T→4→R	192	(8.2%)	486	(7.8%)
T→R→4	673	(29%)	1,166	(18.8%)

for Twitter.

Finally, we analyze the source of the URLs for each of the three platforms, as follows. We create two directed graphs, one for each type of news, $G = (V, E)$, where V represents alternative or mainstream domains, as well as the three platforms, and E the set of sequences that consider only the first-hop of the platforms. For example, if a Breitbart.com URL appears first on Twitter and later on the six selected subreddits, we add an edge from

breitbart.com to Twitter, and from Twitter to the six selected subreddits. We also add weights on these edges based on the number of such unique URLs. By examining the paths, we can discern which domains’ URLs tend to appear first on each of the platforms. Figure 22 shows the graphs built for alternative and mainstream domains. Comparing the thickness of the outgoing edges, one can see that breitbart.com URLs appear first in the six selected subreddits more often than on Twitter and more frequently than on /pol/. However, for other popular alternative domains, such as infowars.com, rt.com, and sputniknews.com, URLs appear first on Twitter more often than the six selected subreddits and /pol/. Also, /pol/ is rarely the platform where a URL first shows up. For the mainstream news domains, we note that URLs from nytimes.com and cnn.com tend to appear first more often on the selected subreddits than Twitter and /pol/, however, URLs from other domains like bbc.com and theguardian.com tend to appear first more often on Twitter than the selected subreddits. Similar to the alternative domains graph, there is no domain where /pol/ dominates in terms of first URL appearance.

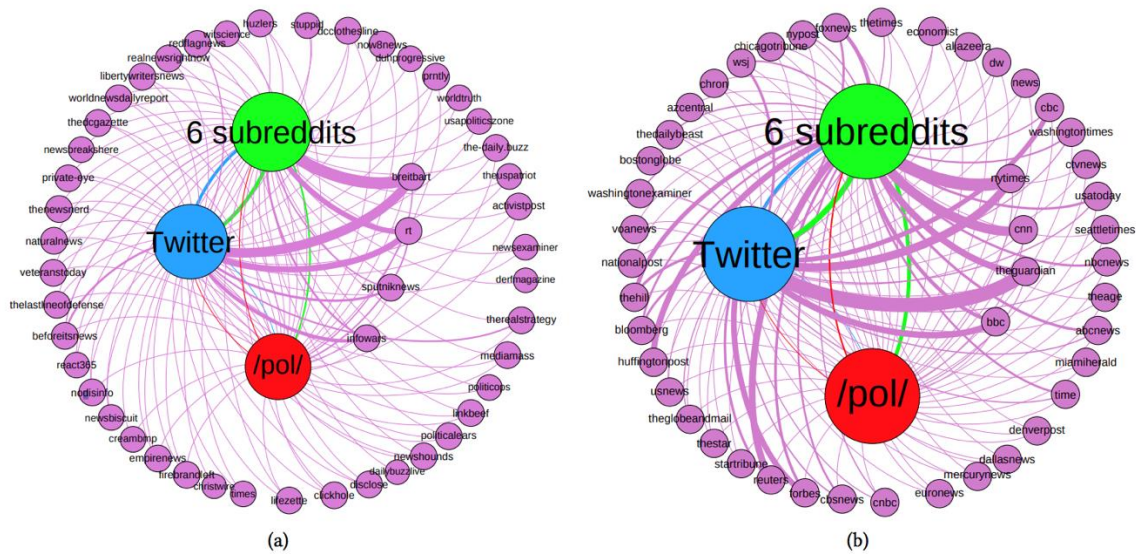


Figure 22 : Graph representation of news ecosystem (a) alternative news domains and (b) mainstream news domains. Edges are colored the same

7.3.2 Influence Estimation

Thus far, our measurements have shown relative differences in how news media is shared on Reddit, Twitter, and 4chan. In this section, we provide meaningful evidence of how the individual platforms influence the media shared on other platforms. We do so by using a mathematical technique known as Hawkes processes. These statistical models can be used

for modeling the dissemination of information in Web communities [4] as well as measuring social influence [6].

7.3.3 Hawkes Processes

Note that the three platforms we measure do not obviously exist in a vacuum, rather, they exist within the greater ecosystem of the Web. Imagine, however, that each of the platforms were entirely self-contained, with a completely disjoint set of users. In such a scenario, there would be a natural rate at which URLs will be posted, and it would be possible to model this using standard Poisson processes. However, our platforms are clearly not independent. While they do exhibit their own background URL posting rates and internal influence, they are also affected by each other, as well as by the greater Web.

A Hawkes model consists of a number, K , of point processes, each with a “background rate” of events $\lambda_{0,k}$. An event on one process can cause an impulse response on other processes, which increases the probability of an event occurring above the processes’ background rates. Figure 23 depicts an example of what a sequence of events on a Hawkes model with three processes might look like, using The_Donald, Twitter, and /pol/ communities for representative purposes.

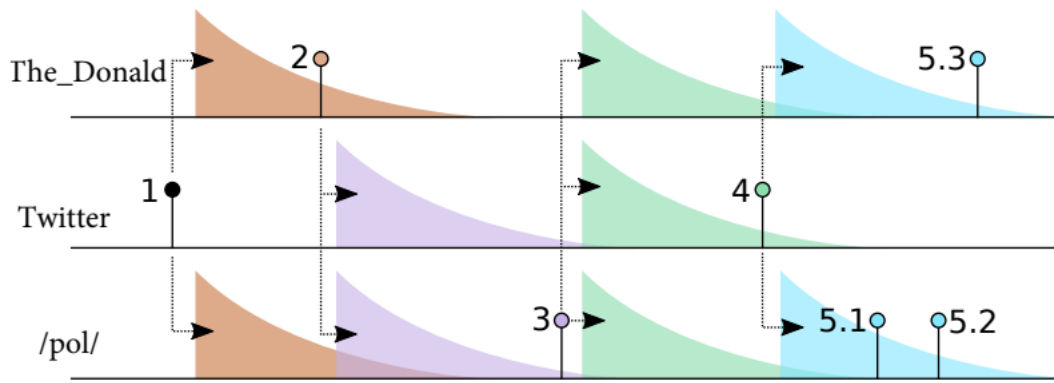


Figure 23: A depiction of a Hawkes model showing how the interaction between events on 3 processes might look like. Note: this example is inspired by [12].

First, an event 1 (a URL being posted) occurs on Twitter; this is caused by the background rate of the process, meaning that the URL was posted not because it was seen on any of the platforms in the model, but because it was seen elsewhere (including a user finding it organically). This initial event causes an impulse response on the rates of the other processes, The_Donald and /pol/, meaning that the URL is more likely to be posted on those platforms after having been seen on Twitter. Eventually, this causes another event on The_Donald (2), which in turn causes an event on /pol/. A process can cause an additional impulse response to itself, as seen with event 3, and multiple events can be caused in

response to a single event, as seen with event 4 causing events 5.1, 5.2, and 5.3. Naturally, the data we collect does not explicitly state which events are caused by other events, or which are caused by the background rate.

For our purposes, fitting a Hawkes model to a series of events on the different platforms gives us values for the background rates for each platform along with the probability of an event on one platform causing events on other platforms. We emphasize that the background rates of the Hawkes processes allow us to also account for the probability of an event caused by external sources of information. For example, it includes the probabilities of events caused by Facebook as well as other platforms. Thus, while we are only able to model the specific influences of the three platforms we study, the resulting probabilities are affirmatively attributable to each of them; the influence of the greater Web is captured by the background rates.

For a discrete-time Hawkes model, time is divided into a series of bins of duration Δt , and events occurring within the same time bin do not interact with each other. The rate of each k -th process, $\lambda_{t,k}$ is given by:

$$\lambda_{t,k} = \lambda_{0,k} + \sum_{k'=1}^K \sum_{t'=1}^{t-1} s_{t',k'} \cdot h_{k' \rightarrow k}[t - t']$$

where $s \in \mathbb{N} \times T \times K$ is the matrix of event counts (how many events occur for process k at time t) and $h_{k' \rightarrow k}[t - t']$ is an impulse response function that describes the amplitude of influence that events on process k' have on the rate of process k . Following [12], the impulse response function $h_{k' \rightarrow k}[t - t']$ can be decomposed into a scalar weight $W_{k \rightarrow k'}$ and a probability mass function $G_{k \rightarrow k'}[d]$. The weight specifies the strength of the interaction from process k to process k' and the probability mass function specifies how the interaction changes over time:

$$h_{k \rightarrow k'}[d] = W_{k \rightarrow k'} G_{k \rightarrow k'}[d]$$

The weight value $W_{k \rightarrow k'}$ can be interpreted as the expected number of child events that will be caused on process k' after an event on process k . The probability mass function $G_{k \rightarrow k'}$ specifies the probability that a child event will occur at each specific time lag $d\Delta t$, up to a maximum lag Δt_{\max} . This interpretation of $W_{k \rightarrow k'}$ is useful because it allows us to compare how much influence platforms have on each other. For instance, we can examine whether a URL posted on Twitter or on Reddit is more likely to cause the same URL to be posted on 4chan, or if there is a difference in influence from one platform to another between URLs for mainstream and alternative news sites.

7.3.4 Methodology

We now provide more details about our experiments, once again, considering 4chan (/pol/), Twitter, and the six subreddits. We study Hawkes processes at the subreddit granularity to get a better understanding of the various platforms and particular subreddits.

We aim to examine how these platforms and subreddits influence each other, so we model the arrival of URLs, in posts or tweets, with a Hawkes model with $K=8$ point processes—one for Twitter, one for /pol/, and one for each of the subreddits. The model is fully connected, i.e., it is possible for each process to influence all the others, as well as itself, which describes behavior where participants on a platform see a URL and re-post it on the same platform. For example, with Twitter, this value ($WTwitter \rightarrow Twitter$) would likely be quite high, given that tweets are commonly re-tweeted a number of times: the initial tweet containing a URL is likely to cause a number of re-tweets, also containing the URL, on the same platform.

We select URLs that have at least one event in Twitter, /pol/, and at least one of the subreddits, and we model each URL individually. The missing Twitter data affects 3,177 (37%) of the URLs. One way to mitigate the impact of this missing data is to remove events for which it has a larger impact. E.g., if an event spans 100 day, the missing Twitter data has less of an effect than if the event only spanned two days. Thus, we examine URLs from other platforms that overlap with any of the missing days and remove the 10% of URLs (895) with the shortest total duration from the first event recorded until the last event recorded. This results in the missing data making up a smaller portion of the overall duration of the events.

The number of remaining URLs and events included for each platform are shown in Table 17. For each URL, we create a matrix $s \in \mathbb{N}^{(T \times 8)}$ containing the number of events (URL posts) per minute for each of the platforms/subreddits. Here, T is the number of minutes from the first recorded post of the URL on any platform, to the last recorded post of a URL on any platform (NB: this value can be different for each URL). We select $\Delta t = 1$ minute as a reasonable compromise between accuracy and computational cost. Using this bin size, 92% of events are in a bin by themselves, and another 5.4% share a bin, but only with other events from the same platform or subreddit, meaning that timing interactions between the platforms are not lost.

Next, we fit a Hawkes model for each URL using the approach described in [12, 13], which uses Gibbs sampling to infer the parameters of the model from the data, including the weights, background rates, and shape of the impulse response functions between the different processes. By setting $\Delta t_{max} = 60 \cdot 12 = 720$ minutes, we say that a given event can cause other events within a 12-hour time window. Experiments with other values (6, 12, 24,

and 48 hours) gave similar results. After fitting the models, we have the values for the W matrix – i.e., the weights of the interactions between events on different processes for each URL. These weights can then be interpreted as the expected number of events. For example, $WTwitter \rightarrow /pol/ = 0.1$ would mean that an event on Twitter will cause n events on $/pol/$, where n is drawn from a Poisson distribution with rate parameter 0.1. Finally, for each URL, we also get the $\lambda_{0,k}$ values for each process, which are the background rates for event arrivals that are not caused by other events in the system we model. Again, these background rates capture events due to some other platform, e.g., someone posting the URL after reading it on the original site or seeing the URL on another site not included in the model, like Facebook.

Table 17 Total URLs with at least one event in Twitter, /pol/, and at least one of the subreddits; total events for mainstream and alternative URLs, and the mean background rate (λ_0) for each platform/subreddit.

		The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter
URLs	Mainstream	3,097	2,523	3,578	2,584	907	841	5,589	5,589
	Alternative	2,008	252	813	362	321	100	2,136	2,136
	Total	5,105	2,775	4,391	2,946	1,228	941	7,725	7,725
Events	Mainstream	12,312	7,517	26,160	5,794	1,995	2,302	19,746	36,250
	Alternative	7,797	458	2,484	586	497	176	7,322	23,172
	Total	20,109	7,975	28,644	6,380	2,492	2,478	27,068	59,422
Mean λ_0	Mainstream	0.001502	0.001382	0.001265	0.001392	0.000501	0.000107	0.001564	0.002330
	Alternative	0.001627	0.000619	0.000696	0.000553	0.000423	0.000034	0.001525	0.002803

7.3.5 Influence Estimation Results

Looking at the number of URLs in Table 17, we note that there are substantially more events for mainstream than alternative news URLs. However, for Twitter, /pol/, and The_Donald, the ratios of events to URLs for alternative news URLs are similar to or greater than the ratios for mainstream ones. These high ratios explain the high background rates (cf. Table 17) for alternative news URLs for these platforms despite the lower number of events.

From the Hawkes models for each URL, we obtain the weight matrix W which specifies the strength of the connections between the different platforms and subreddits. The mean weight values over all URLs for alternative and mainstream news URLs, as well as the percentage difference between them are presented in Figure 24. First, we look at Twitter. Background rates are high for both mainstream and alternative news URLs, which is not surprising given the large number of users on the platform. The values for $WTwitter \rightarrow Twitter$ are also substantially higher than all other weights: 0.1096 for mainstream news URLs and 0.1554 for alternative news URLs. This reflects the ease and common practice of re-tweeting: a URL in a tweet is likely to generate other events as users re-tweet it. There are different possible explanations for why the Twitter to Twitter rate for alternative news URLs is much greater than the rate for mainstream news URLs. The first is bot activity—if automated

Twitter bots are used to spread alternative news URLs, it could result in a much higher rate of tweeting and re-tweeting. Another possible explanation is the behavior of users who read news stories from alternative sources; they might be more inclined to re-tweet the URL [7].

Mean Weights - Pct. Increase/Decrease of Alternative over Mainstream URLs

	The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter	
Source	The_Donald	A: 0.0741 M: 0.0720 2.8%	A: 0.0549 M: 0.0563 -2.5%	A: 0.0592 M: 0.0622 -4.8% **	A: 0.0562 M: 0.0556 1.2%	A: 0.0549 M: 0.0561 -2.2%	A: 0.0526 M: 0.0551 -4.6%	A: 0.0652 M: 0.0621 5.1%	A: 0.0797 M: 0.0700 13.8% **
	worldnews	A: 0.0624 M: 0.0569 9.7%	A: 0.0665 M: 0.0694 -4.2%	A: 0.0551 M: 0.0593 -7.0%	A: 0.0531 M: 0.0615 -13.5%	A: 0.0596 M: 0.0555 7.3%	A: 0.0606 M: 0.0551 10.0%	A: 0.0570 M: 0.0580 -1.8%	A: 0.0647 M: 0.0667 -3.0%
	politics	A: 0.0614 M: 0.0596 2.9%	A: 0.0539 M: 0.0522 3.3%	A: 0.0715 M: 0.0758 -5.7%	A: 0.0584 M: 0.0521 12.1% **	A: 0.0540 M: 0.0507 6.4%	A: 0.0549 M: 0.0505 8.8%	A: 0.0635 M: 0.0581 9.4%	A: 0.0677 M: 0.0655 3.4%
	news	A: 0.0652 M: 0.0640 1.8%	A: 0.0549 M: 0.0607 -9.6%	A: 0.0557 M: 0.0594 -6.2%	A: 0.0672 M: 0.0617 9.0%	A: 0.0579 M: 0.0571 1.4%	A: 0.0547 M: 0.0559 -2.1%	A: 0.0629 M: 0.0610 3.2%	A: 0.0664 M: 0.0673 -1.2%
	conspiracy	A: 0.0634 M: 0.0603 5.2%	A: 0.0570 M: 0.0588 -3.0%	A: 0.0566 M: 0.0600 -5.7%	A: 0.0558 M: 0.0555 0.7%	A: 0.0623 M: 0.0626 -0.4%	A: 0.0578 M: 0.0591 -2.3%	A: 0.0589 M: 0.0587 0.4%	A: 0.0675 M: 0.0625 8.1%
	AskReddit	A: 0.0680 M: 0.0550 23.5%	A: 0.0644 M: 0.0558 15.5%	A: 0.0624 M: 0.0585 6.7%	A: 0.0607 M: 0.0521 16.7%	A: 0.0546 M: 0.0563 -3.1%	A: 0.0534 M: 0.0637 -16.2%	A: 0.0623 M: 0.0573 8.8%	A: 0.0494 M: 0.0598 -17.4%
	/pol/	A: 0.0598 M: 0.0588 1.7%	A: 0.0554 M: 0.0576 -3.9% *	A: 0.0577 M: 0.0580 -0.6%	A: 0.0551 M: 0.0569 -3.2%	A: 0.0532 M: 0.0561 -5.2%	A: 0.0540 M: 0.0549 -1.6%	A: 0.0761 M: 0.0734 3.7%	A: 0.0639 M: 0.0634 0.6%
	Twitter	A: 0.0583 M: 0.0558 4.4% *	A: 0.0443 M: 0.0536 -17.5% **	A: 0.0471 M: 0.0575 -18.1% **	A: 0.0459 M: 0.0533 -13.8% **	A: 0.0454 M: 0.0501 -9.4% **	A: 0.0440 M: 0.0506 -12.9% **	A: 0.0579 M: 0.0606 -4.6%	A: 0.1554 M: 0.1096 41.9% **
	Destination								

Figure 24: Mean weights for alternative URLs (A), mainstream URLs (M), and the percent increase/decrease between mainstream and alternative (also indicated by the coloration). Stars indicate the level of statistical significance (p-value) between the weight distributions: no stars indicate no statistical significance, while * and ** indicate, resp., statistical significance with $p < 0.05$ and $p < 0.01$.

Looking at the weights for Twitter to the other platforms, except The_Donald, they are all greater for mainstream news URLs, meaning that the average tweet containing a mainstream URL is more likely to cause a subsequent post on the other platforms than the average tweet containing an alternative URL. The next communities most likely to cause events on others are The_Donald and /pol/. It is worth noting that The_Donald is the only platform/subreddit that has greater alternative URL weights for all of its inputs. Assuming that the population of The_Donald users that also read, say, worldnews is the same for both alternative and mainstream news URLs—which is reasonable—then the difference in weights implies that the users have a stronger preference for re-posting alternative news URLs back to The_Donald than for mainstream news URLs. The opposite can be seen for worldnews and politics, where most of the input weights are stronger for mainstream news. However, despite the higher weights for alternative news URLs, The_Donald is also, interestingly, influenced more strongly by mainstream news URLs than alternative news URLs on all platforms, with the exception of Twitter. This is in part because of the greater number of mainstream URL events, but The_Donald also has a higher background rate for

alternative news URLs than mainstream news URLs, which implies that a lot of the alternative news URLs on the platform are coming from other sources.

To assess the statistical significance of the results, we perform two sample Kolmogorov-Smirnov tests on the weight distributions of mainstream and alternative news URLs for each source destination pair (depicted as stars in Figure 24). Rejecting the null hypothesis here implies a difference in the way mainstream and alternative news URLs propagate from the source to the destination— either mainstream news URLs tend to cause more events on other platforms than alternative news URLs, or the opposite. Unsurprisingly, many of the source-destination pairs have no significant difference. However, in most cases where Twitter is the source community there is a significant statistical difference with $p < 0.01$. I.e., for some communities, Twitter is used not just to disseminate news, but to disseminate news from a specific type of source.

Figure 25 illustrates the estimated total impact of the different platforms on each other, for both mainstream and alternative news URLs. The impact is estimated using the weight values that are shown in Figure 24. Since the weight values can be interpreted as the expected number of additional events caused as a consequence of an event, we can estimate the percentage of events on each platform that were caused by each of the other platforms by multiplying the weight by the actual number of events that occurred on the source platform and dividing by the number of events that occurred on the destination platform:

$$\text{Pct}_{A \rightarrow B} = \frac{\sum_{u \in \text{urls}} \left(W_{A \rightarrow B} \cdot \sum_{t=1}^T s_{t,A} \right)}{\sum_{u \in \text{urls}} \sum_{t=1}^T s_{t,B}}$$

The percentages for mainstream and alternative news URLs as well as the difference between them are presented in Figure 25.

Twitter contributes heavily to both types of events on the other platforms—and is in fact the most influential single source for most of the other platforms. Despite Twitter’s lower weights for alternative news URLs, it actually has a greater influence on alternative than mainstream news URLs, in terms of percentage of events caused, on all the other platforms/subreddits. This is due to the fact that, even though it has lower weights, the largest proportion of alternative URL events are on Twitter. After Twitter, `The_Donald` and `/pol/` also have a strong influence on the alternative news URLs that get posted on other platforms. `The_Donald` has a stronger effect for alternative news URLs on all platforms except Twitter— although it still has the largest alternative influence on Twitter, causing an estimated 2.72% of alternative news URLs tweeted. Interestingly, `The_Donald` causes 8% of `/pol/`’s alternative news URLs, while `/pol/`’s influence on `The_Donald` is less, at 5.7%. For the mainstream news URLs the

strength of influence is reversed. Specifically, /pol/’s influence on The_Donald is 8.61% whereas The_Donald’s influence on /pol/ is 6.13%.

In descending order, the influences on Twitter for mainstream news URLs are politics (4.29%), /pol/ (3.01%), The_Donald (2.97%), worldnews (2.74%), news (1.81%), AskReddit (1.34%), and conspiracy (1.04%). The strongest influences for alternative news URLs are, unsurprisingly, The_Donald (2.72%) and /pol/ (1.96%), followed by politics (1.10%), worldnews (0.60%), AskReddit (0.55%), news (0.50%), and conspiracy (0.46%). Twitter influences the alternative news URLs on other platforms to a large degree—but the largest alternative URL inputs to Twitter are The_Donald and /pol/. While we are only looking at a closed system of 8 different platforms and subreddits, we note that Twitter is undoubtedly effective at propagating information. Thus, the influence these two communities have on Twitter is likely to have a disproportional impact on the greater Web compared to their relatively minuscule userbase.

Pct. of Alternative URLs - Pct. of Mainstream URLs

	The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter
The_Donald		A: 16.77% M: 5.68% 11.09	A: 11.25% M: 3.52% 7.74	A: 18.01% M: 7.69% 10.32	A: 20.68% M: 14.32% 6.36	A: 20.27% M: 8.01% 12.25	A: 8.00% M: 6.13% 1.87	A: 2.72% M: 2.97% -0.25
worldnews	A: 1.09% M: 3.75% -2.66		A: 1.37% M: 1.67% -0.30	A: 4.52% M: 7.86% -3.34	A: 5.96% M: 8.34% -2.39	A: 6.16% M: 7.44% -1.28	A: 1.63% M: 4.07% -2.43	A: 0.60% M: 2.74% -2.14
politics	A: 2.75% M: 9.16% -6.41	A: 11.13% M: 9.83% 1.30		A: 13.79% M: 12.57% 1.22	A: 12.12% M: 19.03% -6.91	A: 17.35% M: 17.17% 0.18	A: 3.50% M: 6.95% -3.45	A: 1.10% M: 4.29% -3.19
news	A: 1.30% M: 3.33% -2.04	A: 6.21% M: 4.21% 2.00	A: 1.86% M: 1.33% 0.54		A: 6.30% M: 6.30% -0.00	A: 4.99% M: 5.80% -0.81	A: 1.65% M: 3.14% -1.49	A: 0.50% M: 1.81% -1.31
conspiracy	A: 1.12% M: 1.58% -0.45	A: 5.86% M: 2.74% 3.13	A: 1.72% M: 0.80% 0.92	A: 3.79% M: 3.17% 0.61		A: 5.00% M: 3.81% 1.19	A: 1.62% M: 1.73% -0.10	A: 0.46% M: 1.04% -0.57
AskReddit	A: 0.66% M: 1.61% -0.95	A: 6.09% M: 2.94% 3.15	A: 0.92% M: 0.74% 0.19	A: 3.21% M: 3.30% -0.09	A: 4.24% M: 4.80% -0.56		A: 1.15% M: 2.00% -0.85	A: 0.55% M: 1.34% -0.79
/pol/	A: 5.70% M: 8.61% -2.91	A: 12.86% M: 6.31% 6.55	A: 7.80% M: 3.24% 4.56	A: 12.25% M: 8.31% 3.94	A: 15.42% M: 11.16% 4.26	A: 14.41% M: 9.02% 5.39		A: 1.96% M: 3.01% -1.05
Twitter	A: 14.32% M: 10.79% 3.53	A: 27.67% M: 9.28% 18.39	A: 18.95% M: 6.00% 12.94	A: 34.28% M: 15.15% 19.13	A: 37.07% M: 15.64% 21.43	A: 20.76% M: 11.63% 9.13	A: 16.54% M: 9.79% 6.75	

Destination

Figure 25 Mean estimated percentage of alternative URL events caused by alternative news URL events (A), mean estimated percentage of mainstream news URL events caused by mainstream news URL events (M), and the difference between alternative and mainstream news (also indicated by the coloration).

7.4. Conclusion

This work explored how mainstream and fringe Web communities share mainstream and alternative news sources with a particular focus on how communities influence each other. We collected millions of posts from Twitter, Reddit, and 4chan, and analyzed the occurrence and temporal dynamics of news shared from 45 mainstream and 54 alternative news sites. We found that users on the different platforms prefer different news sources, especially when it comes to alternative ones. We also explored complex temporal dynamics and we discovered, for example, that Twitter and Reddit users tend to post the same stories within a relatively short period of time, with 4chan posts lagging behind both of them. However, when a story becomes popular after a day or two, it is usually the case it was posted on 4chan first, lending some credence to 4chan’s supposed influence on the Web.

Using Hawkes processes, we also modeled the influence the individual platforms have on each other, while also taking into account influence that comes from external sources of information. We found that the interplay between platforms manifests in subtle, yet meaningful ways. For example, of all the platforms and subreddits, Twitter by far has the most influence in terms of the number of URLs it causes to be posted to other platforms and contributes to the share of alternative news URLs on the other platforms to a much greater degree than to the share of mainstream news URLs. After Twitter, `The_Donald` subreddit and `/pol/` are the next most influential when it comes to alternative news URLs. For such URLs, `The_Donald` is less influenced by the other platforms than `/pol/`, and has a higher background rate, i.e., more of the URLs posted there come from other sources.

To the best of our knowledge, our analysis constitutes the first attempt to characterize the dissemination of mainstream and alternative news across multiple social media platforms, and to estimate a quantifiable influence between them. Overall, our findings shed light on how Web communities influence each other and can be extremely useful to better understand and detect false information as well as informing the design of systems that aim to trace the origins of fake stories and mitigate their dissemination.

7.5. References

- [1] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31, 2017.
- [2] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Hate is not binary: Studying abusive behavior of #gamergate on twitter. In *WebSci*, 2017.
- [3] L. Dearden. Nato accuses Sputnik News of distributing misinformation as part of 'Kremlin propaganda machine'. <http://ind.pn/2luLjs0>, 2016.
- [4] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, and H. Zha. Fake news mitigation via point process based intervention. In *PMLR*, 2017.
- [5] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.
- [6] F. Guo, C. Blundell, H. Wallach, and K. Heller. The Bayesian Echo Chamber: Modeling social influence via linguistic accommodation. In *Artificial Intelligence and Statistics*, 2015.
- [7] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. In *WWW*, 2013.
- [8] M. Haag and M. Salam. Gunman in pizzagate shooting is sentenced to 4 years in prison. <https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html>, 2017.
- [9] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*, 2017.
- [10] J. Jackson. Moderators of pro-Trump Reddit group linked to fake news crackdown on posts. <https://www.theguardian.com/technology/2016/nov/22/moderators-trump-reddit-group-fake-news-crackdown>, 2016.
- [11] S. Kumar, R. West, and J. Leskovec. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *WWW*, 2016.
- [12] S. W. Linderman and R. P. Adams. Discovering Latent Network Structure in Point Process Data. In *ICML*, 2014.
- [13] S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv 1507.03228*, 2015.
- [14] M. Mohan. Macron Leaks: The anatomy of a hack. <http://www.bbc.co.uk/news/blogs-trending-39845105>, 2017.
- [15] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A Platform for Tracking Online Misinformation. In *WWW Companion*, 2016.

[16] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In ICWSM, 2017.

[17] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter After the 2013 Boston Marathon Bombing. In iConference, 2014.

[18] M. Wendling. The saga of 'Pizzagate': The fake story that shows how conspiracy theories spread. <http://www.bbc.com/news/blogs-trending-38156985>, 2016.

[19] Wikipedia. Pizzagate conspiracy theory. https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory, 2017

8. Youtube Raids

8.1. Project description and motivation

As people move significant portions of their lives to online interactions, online aggression phenomena have become a significant problem [27]. Online aggression can happen on a one-to-one basis, in which the victim is repeatedly abused [13], or take the form of a coordinated harassment campaign, where the perpetrators loosely coordinate to deliver harmful content in a repetitive fashion [12, 14]. Such coordinated aggression campaigns take the name of raids [28]. Recent research showed that raids are often organized by fringe Web communities such as 4chan with the goal of attacking and undermining users on other platforms (e.g., YouTube, Twitter) who advocate for issues or policies that they do not agree with [28].

Despite the increasing relevance of online aggression, social networks lack adequate countermeasures to mitigate this problem. In general, abusive activity is generated by humans and not by automated programs, therefore automated systems that have been proposed to detect bots and fake accounts [9, 44] are unsuitable in this scenario. For these reasons, most platforms make use of reactive systems: users can report abusive accounts to the social network, which will then verify the claim and potentially suspend offenders [34]. The efficacy of this modus operandi has been brought to question both by researchers [14], and by the social network operators themselves, e.g., Twitter's transparency in admitting that the company could do a much better job in dealing with abuse [46].

In this project, we focus on identifying YouTube videos that are likely to be raided. We focus on YouTube because it is one of the top visited video platforms, with more than 1 billion users and 1 billion hours of videos watched every day [4]. With such large amount of activity, it is only expected to attract a lot of hate speech, something that prompted YouTube to announce new measures to crackdown such abusive behavior [11].

Also, previous work identified YouTube as the most heavily targeted platform by fringe communities such as 4chan [28]. This work analyzed the behavior of 4chan users active on the /pol/ board. The authors observed a particular action, among others: some users were posting YouTube links with sentences like you know what to do; suddenly, some of those videos were commented on by people interested in spreading hateful comments. In some cases, the behavior was attacking the video and the people commenting it positively. In other cases, the 4chan users supported the message of the video and attacked other commenters detracting from it. In either case, this action has been defined by the authors as a raid. In fact, the variety of topics covered by YouTube uploaders, in conjunction with its comment-oriented interface, make it an attractive platform for abusers. We obtained a ground truth dataset of raided YouTube videos from the authors of [28], and compared such videos with random YouTube uploads that were not targeted by raiders.

Previous work [28] showed how to detect whether a raid was occurring by examining the degree of synchronous commenting behavior between 4chan and YouTube, validating it in terms of the rate of hate comments in YouTube, as well as commenter account overlap. In

this work package, instead, we take an orthogonal approach, looking raided videos alone, with the goal of gaining an understanding of which videos are likely to be raided in the future. To this end, we look at multiple features of YouTube videos, such as their title, category, what is displayed in the thumbnail image that is shown as a preview, and what the video is talking about (i.e., the audio transcript of the video). This information allows us to gain an understanding of what content attracts raids, in other words why these videos are raided.

Based on the insights gained from our analysis, we build a system that is able to tell at upload time whether a video is likely to be raided in the future. Our system leverages an ensemble of classifiers each looking at a different element of the video (metadata, thumbnails, and audio transcripts). Our approach allows video streaming platforms to assess whether a video that has been uploaded is at risk of being raided. Our ensemble detection algorithm performs very well, with 0:9 precision and 0:85 recall. Moreover, this approach could enable platforms to take countermeasures, by studying the comments posted on such videos and rate limit them, or temporarily blocking the comments to risky videos when a link to them is posted on fringe Web communities.

The contributions of the present work are as follows:

- We study the relatively new problem of raiding videos on YouTube from fringe communities like 4chan on a large set of 2.8k videos.
- We extract and use different properties of these videos in our analysis, including the transcripts of the videos, their metadata, as well as their thumbnails.
- We apply a novel, ensemble classifier based on deep-learning methods which performs better than state-of-art methods and baselines.
- We offer suggestions on how current video platforms can apply our methodology to detect such attacks and mitigate their impact.

8.2. Methodology

We now introduce our approach to provide a proactive detection tool of videos targeted by hate attacks on online services, and on YouTube in particular. Our goal is to systematize this task and, for this, we use supervised machine learning. In particular, we build a set of machine learning classifiers, each of which focuses on a different set of features extracted from online videos. We next describe our proposed set of features, which is strongly motivated by the findings reported above. As not all videos contain the same information, we then present three independent classifiers that are triggered accordingly, and we finally show how to ensemble and properly balance the different predictions to provide one single decision.

8.2.1 General Overview

A high-level description of our detection system is depicted in Figure 2. The system is first trained using a dataset of *raided* and *non-raided* videos from different sources. The purpose of this phase is to obtain the following key elements that will later be used to predict whether a video could be targeted by hate attacks or not:

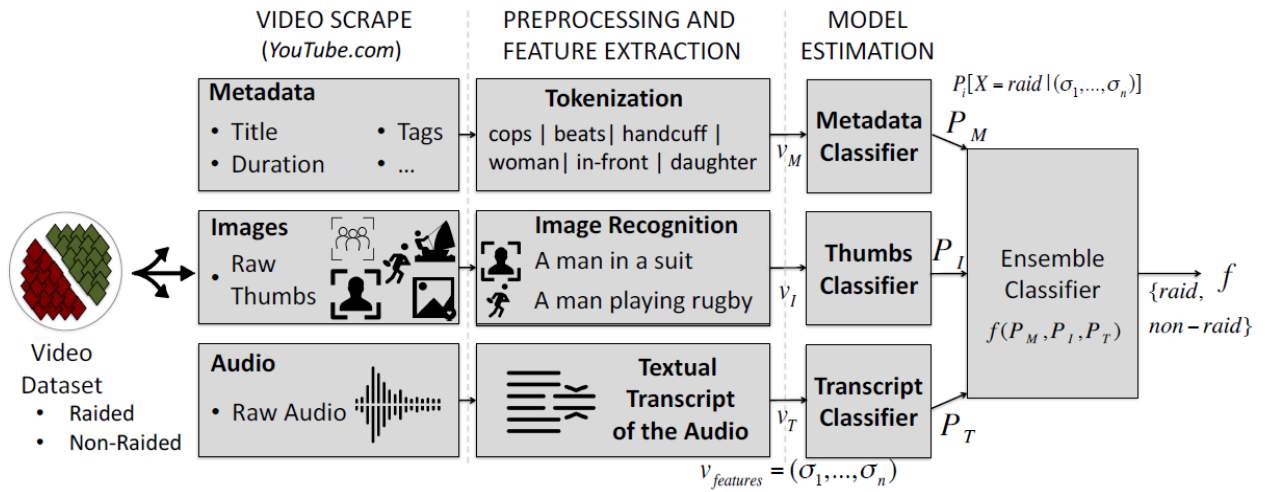


Figure 2: Architecture of our proactive detection system.

(i) A set of prediction models that output the probability $C_i(\sigma_1, \dots, \sigma_n) = P_i[Y = \text{raid} \mid (\sigma_1, \dots, \sigma_n)]$

of each video Y being raid given a feature vector obtained from different elements i of the video. Each of this prediction models are referred as individual classifiers.

(ii) A weighted model $f(C) = \sum w_i \cdot C_i$

that combines all predictions in C , where each of the classifiers is weighted by based on the performance obtained on a validation set. This set is different from the training set (used to build the individual probabilities), as well as different from the testing set (used to measure the efficiency of the classifiers). The process of weighting the individual predictors also serves as a way to calibrate the output of the probabilities. The final classifier will then output a decision based on a votation such that

$$f = \begin{cases} \text{raid} & \text{if } f(C) < \epsilon \\ \text{non-raid} & \text{otherwise,} \end{cases}$$

where ϵ is a threshold typically set to $\left\lfloor \frac{\sum w_i}{2} \right\rfloor$.

For the sake of simplicity, we refer to the model presented in (ii) as weighted-vote. One can simplify the model by giving equal weight to all (typically) and obtaining a nominal value for before voting. In other words, applying a threshold for each (e.g., 0.5) and creating an equal vote among participants. We refer to this non-weighted voting system as majority-vote. One can further simplify the scheme by combining each individual prediction using the arithmetic mean of the output the probabilities. We refer to this as average-prediction system. Note the parameters used in both majority-vote and average-prediction are fixed and they do not require calibration. Thus, the validation set is not used in these two modes.

8.2.2 Feature Engineering

One of the key elements to build a robust classification system is to consider a diverse set of features. Our work mainly extracts features from three different sources: (i) structured attributes of the metadata of the video, (ii) features extracted from raw audio from the video, and (iii) features extracted from raw images (thumbnails) from the video. Based on a preprocessing, we transform non-textual elements of a video (i.e., audio and images) into text representation. Other textual elements such as the title of the video and the tags are kept as text.

These textual representations are then transformed into a fixed-size input space vector of categorical features. This is done by tokenizing the input text to obtain a nominal discrete representation of the words described on it. Thus, feature vectors will have a limited number of possible values given by the bag of words representing the corpus in the training set. When extracting features from the text, we count the number of occurrences a word appears in the text.

As in large text corpus certain words—such as articles—can appear rather frequently (without carrying meaningful information), we transform occurrences into a score based on two relative infrequency known as term-frequency and inverse document-frequency (namely, tf-idfs). Intuitively, the term frequency represents how “popular” a word is in a text (in the feature vector), and the inverse document-frequency represents how “popular” a word appear in the text, provided that it does not appear very frequently in other in the corpus (the feature space). More formally, we compute as

where N is the total number of samples and $df(s,t)$ is the number of samples that contain term t .

As for the thumbnails, after extracting the most representative descriptions per image, we removed the least informative elements and retained only entities (nouns), actions (verbs), and modifiers (adverbs and adjectives). Each element in the caption is processed to a

common base to reduce inflectional forms and derived forms (known as stemming). We further abstracted the descriptions obtained from the images using topic-modeling.

In our current implementation, we extract features from only one image of the video (i.e., the thumbnail). As commented before, this is mainly because the thumbnails are typically purposely selected by the user as a video core and encapsulate semantically relevant context. However, we emphasize that our architecture supports the extraction of features from every frame in the video.

8.2.3 Prediction Models

We now present three independent classifiers to estimate the likelihood of a video being targeted by hate attacks. These classifiers are built to operate independently and at the moment a video is uploaded. In particular, each classifier is designed to model traits from different aspects of the video as described above. Available decisions are later combined to provide one unanimous output.

The motivation behind the use of three different classifiers that are used in an ensemble is as follows. Features obtained from different parts of a video are inherently incomplete since some fields are optional and others might not report meaningful features. For instance, a music video might not report a lengthy transcript or a thumbnail might not contain distinguishable context. Thus, any reliable decision system should be able to explicitly deal with incomplete video sections. Ensemble methods are suitable for this. Another reason behind it is that ensembles often perform better than a single classifier [18].

Metadata and thumbnail classifiers. We build a prediction model such that $P_i(X = \text{raid})$ based on the features extracted from the metadata (P_M) and from those extracted from the image thumbnails (P_I). The architecture of these two predictors is explicit and accepts a range of classifiers. Our current implementation supports Random Forests, Extra Randomized Trees, and Support Vector Machines (SVM), both radial and linear. For the purpose of this work, we selected Random Forest (RF) as the base classifier for P_T and SVM with linear kernel for P_M . Both SVM and RF have been successfully applied to different aspects of security in the past (e.g., fraud detection [10]) and have been shown to outperform other classifiers (when compared to 180 classifiers in real-world problems [22]).

Audio-transcript classifier Audio-transcript classifier Before feeding the transcripts into the classifier it's necessary to perform data cleaning. Firstly, we remove from the transcripts any words that have a transcription confidence $p_{\text{trans}} < 0.5$ as these are mostly wrong and unrelated to the video context. Including these terms will only confuse the classifier. This only removes 9.2% of the words (91% of words have a confidence of 0.5 or more). Furthermore, the transcripts contain a lot of repeated terms that are mostly exclamations such as uh uh, or hm hm, Finally, notice that the transcripts contain tags for non-verbal communication such as noise, laughter, etc. These were included in the text as they do carry predictive power.

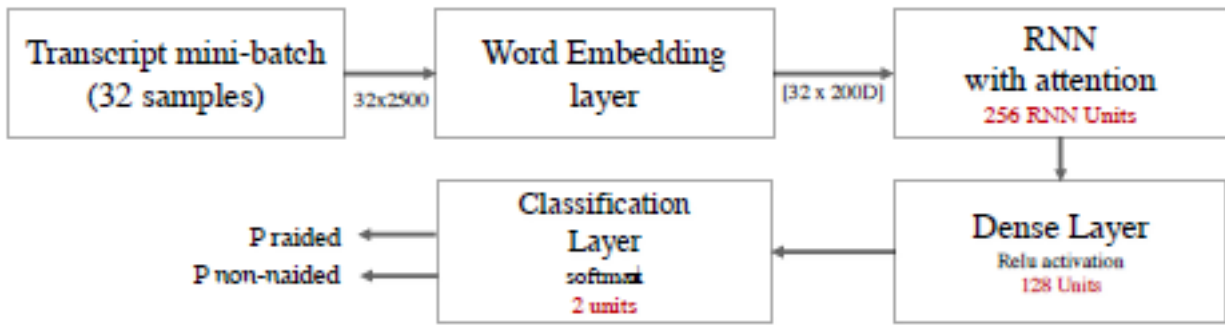


Figure 3: Architecture of the transcript classifier. An Recurrent neural network with attention and pre-trained word embeddings was used.

There are several choices in terms of selecting a classifier for long sequences of text. We experimented with transitional TF-IDF based approaches, with Convolutional Networks and with Recurrent Neural Networks. In the end we decided to use Recurrent Neural Networks (RNN) since they achieved the best performance as they are quite efficient at understanding sequences of words and interpret the overall context. Furthermore, we also use an attention mechanism [8] as they can really help the RNN to focus its attention to sequence of words that might indicate raidable videos. An illustration of the overall architecture is shown in Figure 3.

Before feeding any text to the network we need to transform each transcript to a sequence of words. Because neural networks process data in mini-batches, all transcripts within a mini-batch must have the same length (number of words). Transcripts with words that are larger than the sequence length will be trimmed whereas samples with less words are left-padded with zeros (the model will learn they carry no information). Ideally, we want to setup a sequence length that is large enough to contain the text from all samples in a mini-batch but not too long to waste resources (feeding zeros in the network). We thus take the 95th percentile of length (with respect to the number of words) in the input corpus as the optimal sequence length (i.e., 5% of samples will be truncated). This results in establishing a sequence length of 2500 words.

The first layer of the network performs a word embedding, which maps each word to a high-dimensional vector. Word embedding has proved to be a highly effective technique for text classification tasks especially for tasks for which we have few training samples. We use pre-trained word embeddings from GloVe, which was constructed on more than 2 billion tweets that map each word into a 200-D vector. If a word is not found in the GloVe dataset, we initialize a vector of random weights, which the word embedding layer eventually learns from the input data.

Next, after experimenting with several choices for the RNN architecture we use a layer of 256 GRU units. To reduce over-fitting we use a recurrent dropout with $p = 0.5$ as it empirically

provided the best results across our datasets. On top of the recurrent layer, we add an attention layer as we are working with large sequence lengths (2500 words). The network at this stage outputs one activation at the end of the whole sequence (the whole transcript).

Connected to the recurrent part, we add a fully-connected (Dense) layer to mix the recurrent layer outputs and to gradually bring the dimensionality down to 128 units. Finally, the output layer is another Dense layer with one neuron per class. We use softmax as the activation function to normalize output values between 0 and 1.

In terms of training, we use mini-batches of 32 transcripts (i.e., the input is of shape 32x2500). We use categorical cross-entropy as loss function and Adam as the optimization function. A maximum of 100 epochs was allowed but we also employed a separate validation set to perform early stopping: training was interrupted if the validation loss did not drop in 10 consecutive epochs and the weights of the best epoch were restored.

Finally, for our implementation we use Keras [2] with Theano [3] as a back-end.

Ensemble classifier The objective of this ensemble method is to aggregate the predictions of the different base estimators. Each classifier individually model the notion of videos that can potentially be targeted by hate attacks based on the set of features. The idea is that the decisions are then combined together so that the ensemble is able to make a more informed prediction. This not only allows to improve the robustness of the predictions (in terms of confidence), but also can result on a more accurate prediction. We design our ensemble method to take a weighted vote of the available predictions. In order to compute the best-performing set of weights, we estimate a function f that takes as input each of the individual probabilities and outputs the aggregated prediction. During training this function learns from a independent validation set, and it will be used during testing to weight each prediction model P_i . Formally,

For the decision function f we use a logistic distribution function that models how an expected probability in the validation set is affected by individual decisions P_i in a multiple regression. This function is approximated in the following form:

where p is the number of individual estimators and n is the number of observations in the validation set. This can be interpreted as the sum of the weights w times the probability score p_i given by the individual classifiers in the weighted voting system.

8.3. Results

In this section, we present the setup and the results of our experimental evaluation.

8.3.1 Experimental Setup

Our main objective is to show that we can distinguish between raided and non-raided videos. However, there are several subtasks we also want to evaluate, aiming to better characterize the problem and understand how our classifiers perform.

Experiments. We start by trying to distinguish between random YouTube videos and those that are linked from /pol/. Next, we distinguish between the videos that are raided and those that are not (whether posted on /pol/ or not). Finally, we predict whether videos posted on 4chan will be raided. More specifically, in Experiment 1, we set out to measure whether our classifiers are able to distinguish between videos linked from /pol/ and a random video uploaded to YouTube, aiming to gather insight into the ability to discriminate between videos potentially raided vs. those that are not at risk at all. Then, Experiment 2 evaluates whether or not the classifier can distinguish between any non-raided video (i.e., regardless of whether it is a random YouTube video or one posted on 4chan) and videos that will be raided. Finally, in Experiment 3, we focus on videos posted on 4chan, and determine which are going to be raided and which are not; this ensures that we can not only predict whether a video was posted on 4chan, but whether or not the video will be raided.

Train, Test, and Validate Splits. We split our datasets into three chunks: two for training and tuning parameters of the ensemble (training and testing) and one for validation, and report performance metrics on the latter. As we are dealing with highly unbalanced classes (there are multiple orders of magnitude more videos posted to YouTube than those posted to 4chan, let alone those that are raided), we balance the training and testing sets to model both classes properly, but leave the validation set unbalanced. Leaving the training split unbalanced would make it difficult for our models to properly learn the differences between the different classes. The validation set, however, remains unbalanced to more realistically model a real-world scenario.

The total number of videos in each split is proportionally sampled depending on the less populated class, assigning splits of 60%, 20%, and 20% to the training, testing, and validation sets. The more populated class uses the same amount of samples for training and test, while it will have all the remaining samples in the validation set. This procedure is repeated 10 times and the results are an average of the 10 different rounds. Table 3 summarizes all settings in our experiments, along with the number of samples used.

Evaluation Metrics. We evaluate our system using Accuracy, Precision, Recall, and F1-measure. Precision measures the performance of our algorithm only for the values of the class of interest, while Recall measures the proportion of positives that are correctly identified as such. The F1-measure is the harmonic mean of Precision and Recall. Finally, Accuracy quantifies the proportion of correct predictions made in both classes.

ID	Description	Training	Test	Validation
Exp. 1	Random YouTube vs. all on 4chan	731+731	243+243	13,470+244
Exp. 2	All non-raided vs. raided on 4chan	258+258	85+85	14,890+85
Exp. 3	Non-raided on 4chan vs. raided on 4chan	258+258	85+85	446+85

Table 2: Number of samples used in our experiments. The sets are balanced as there is the same amount of samples per each class (class 1 samples+class 2 samples).

Overall, these metrics are a good summary of the performance of an classifier in terms of True Negatives (TN), False Negatives (FN), False Positives (FP), and True Positives (TP); however, they are not ideal for comparing results across different experiments. Therefore, we will also plot the Area Under the Curve (AUC), which reports the TP-rate (Recall) against the FP-rate (1 - Recall).

8.3.2 Experimental Results

We now report the results of our experimental evaluations, as per the settings introduced above. To ease presentation, we only report metrics for the individual classifiers as well as two ensemble methods: weighted-vote and average-prediction. We do not report results for other ensemble classifiers (simple-voting and the other underlying algorithms for estimating the weights), since they underperform in our experiments. For weighted-vote, weights are fit using XTREE [23], as described in Section 4.3.3. Also note that, for average-prediction, we find that the thumbnails classifier tends to disagree with the metadata and the transcripts classifiers combined. Therefore, in this mode, we fix a weight of $w = 0$ for the thumbnails classifier (i.e., thumbnail = 0).

Experiment 1. As mentioned, in this experiment we study whether we can predict that a video is linked from 4chan. Results are reported in Table 4. Overall, we find that we can correctly identify 92% of the videos (see average-prediction ensemble), and maintain high Recall. Since we are dealing with a rather unbalanced validation set (in favor of the negative class), it is not surprising that Precision drops to values close to 0, even though we have high Accuracy. Looking at the results obtained by the individual classifiers, we note that metadata has the highest Accuracy (0.91), although audiotranscript scores highly as well (0.81), with the weighted-vote ensemble classifier matching the best Recall from metadata (0.91). The best AUC value is the same between the metadata classifier and the weighted-

vote ensemble (0.96). In Figure 5a, we also plot the ROC curve for all five classifiers. The individual AUC scores are 0.79, 0.96, 0.62 for the transcripts, metadata, and thumbnails, respectively, while the two ensembles

(weighted-vote and average-prediction) score, 0.96 and 0.95, respectively. The weighted-vote ensemble has the highest AUC throughout most of the x-axis, although, the ROC curve essentially overlaps with that of the the metadata classifier. The two ensemble have different strengths: the weighted-vote ensemble has the highest Recall and AUC values, but the average-prediction (with wthumbnail = 0) has the highest Accuracy, Precision, and F1-measure.

	Experiment 1					Experiment 2					Experiment 3				
Classifier	ACC	PRE	REC	F1	AUC	ACC	PRE	REC	F1	AUC	ACC	PRE	REC	F1	AUC
Transcripts	0.81	0.05	0.60	0.10	0.79	0.89	0.03	0.56	0.06	0.79	0.71	0.32	0.58	0.40	0.73
Metadata	0.91	0.13	0.89	0.23	0.96	0.87	0.03	0.85	0.06	0.94	0.73	0.32	0.71	0.44	0.79
Thumbnails	0.55	0.02	0.64	0.05	0.62	0.52	0.01	0.66	0.02	0.61	0.53	0.18	0.55	0.27	0.56
Weighted-vote ensemble	0.89	0.12	0.91	0.21	0.96	0.85	0.03	0.88	0.05	0.94	0.74	0.34	0.69	0.45	0.80
Average-prediction ensemble	0.92	0.15	0.85	0.26	0.95	0.90	0.04	0.82	0.07	0.92	0.75	0.35	0.69	0.46	0.80

Table 4: Results for Experiment 1 (videos posted on 4chan and random YouTube samples), Experiment 2 (raided videos posted on 4chan and all non raided videos), and for Experiment 3 (non-raided videos posted on 4chan and raided videos posted on 4chan). ACC stands for Accuracy, PRE for Precision, and REC for Recall. The ensemble classifiers have different inputs: the weighted-vote classifier receives inputs from all three the individual ones, while the average-prediction receives the inputs only from the metadata and the transcript classifier.

Experiment 2. In Figure 5b, we report the AUC when classifying raided and non-raided videos—regardless of whether the latter are random YouTube videos or non-raided ones

posted on 4chan. We find that the average-prediction ensemble classifier correctly labels 90% of the videos—as shown in Table 4). Unlike Experiment 1, among the individual classifiers, the best performance is achieved by the audio-transcript classifier, except for Recall, where the metadata classifier performs best. This setting also yields high Recall (0.88) when combining all classifiers into the weighted-vote ensemble. As in Experiment 1 the weighted-vote ensemble presents the highest Recall and AUC, but the average-prediction has higher Accuracy, Precision, and F1-measure. Figure 5b shows a similar situation as in the previous experiment: the ROC curve for the metadata classifier is really close to or overlapping with the ones for the two ensemble. AUC equals to 0.61 for thumbnails, 0.79 for transcripts, and 0.94 for metadata. Whereas, the weighted-vote ensemble achieves 0.94 AUC as the metadata individual classifier, and average-prediction 0.92.

Experiment 3. Finally, we evaluate how well our models discriminate between raided videos posted to 4chan and non-raided videos

also posted to 4chan. Our results confirm that this is indeed the most challenging task. Intuitively, these videos are much more similar to each other than those found randomly on YouTube. This is because /pol/ is interested in a particular type of content in general, regardless of whether or not the video ends up raided. Nonetheless, as shown in Table 4, around 75% of the videos are correctly classified by the best performing classifier, i.e., the average-prediction ensemble. This setting shows a clear case for the ensemble classification yielding appreciably better performance. Overall, the individual classifiers, i.e., transcripts, metadata, and thumbnails reach AUCs of 0.73, 0.79, and 0.56, respectively, whereas, both the ensemble classifiers reach 0.80. Nevertheless, the ROC curve in Figure 5c shows how the weighted-vote ensemble is sometimes penalized by the weakest performing classifier (i.e., thumbnails classifier). This is apparent by comparing the differences between weighted-vote and average-prediction (recall that $w_{\text{thumbnails}} = 0$ in the latter).

8.3.3 Choosing an Ensemble

The goal of our system is to make the best final “decision” possible given the choices made by the individual classifiers. In absolute terms, the weighted-vote (with XTREE as baseline estimator) yields the best performance in all three experiments in terms of Recall (and overall AUC). In particular, it outperforms the average-prediction ensemble in two of the tasks: modeling videos from /pol/ (Experiment 1), and detecting raided videos in general (Experiment 2). When restricting our attention to the detection of raids of videos posted on 4chan (Experiment 3), both ensemble methods are comparable in terms of Recall. However, when looking at Precision, we find that average-prediction outperforms weighted-vote. The trade-off between having good Precision and good Recall will have an important impact on the amount of vetting work required by providers deploying our system, as we discuss in more detail in Section 6. In the following, we provide an explanation as to why the two ensemble classifiers report similar results in some experiments and different ones in others. When using a base estimator to fit the best weights for the individual classifiers, we observe a bias towards the decisions made by the metadata classifier. This is expected, as this

classifier is the one that performs best among the individual classifiers (and substantially so in both Experiment 1 and Experiment 2). On the contrary, the thumbnails classifier performs worst, except for Recall in Experiment 2. As for the Precision, the thumbnails classifier always perform the worst.

One important thing to highlight is that our data includes videos with partially-available features. When this happens, the ensemble classifier is forced to make a decision based on the other inputs. This is precisely the case for the thumbnails, which are not always

available. This is why we evaluated the average-prediction ensemble method forcing a weight $w_{\text{thumbnails}} = 0$. In this setting, the weighted-vote method with XTREE provided similar results, since XTREE initially assigned a low weight (although not exactly 0) to the thumbnails. Overall, with the average-prediction method, Accuracy is always better than for both the metadata and the XTREE ensemble classifiers. This also applies for Precision and F1-measure. This means that this configuration reduces the number of false positives and, as a consequence, is slightly more accurate. In other words, this highlights how, when the individual classifiers have similar performance, the ensemble is better than the best options among the single classifiers. In fact, the metadata and transcripts classifiers have different performances, but the thumbnails classifier differs the most.

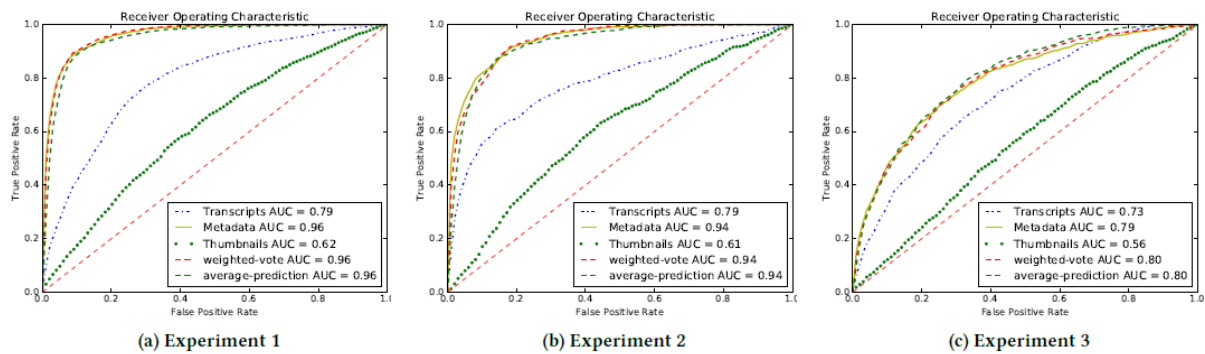


Figure 5: ROC curves for each experiment. AUC values for Thumbnails, Transcripts, Metadata, and Ensemble classifiers, XTREE and Average probabilities.

8.4. Conclusion

In this work, we introduce a system that aims to flag at upload-time whether a video is likely to be raided. Our results indicate that even single-input classifiers that use metadata, thumbnails or audio transcripts can be effective. However, an ensemble of these classifiers can significantly improve the detection performance and result in deployable early-warning systems. Furthermore, we demonstrate that our methodology can also associate videos with the fringe communities that are likely to raid them (e.g., videos that are typically attacked by the 4chan community) and, therefore, warn video providers about the possible source of a raid. In terms of future, work we plan to try more elaborate deep-learning methods that will

allow us to combine directly audio, video and metadata into a single classifier. Furthermore, we plan to look into raids from other communities such as reddit, twitter, etc.

8.5. References

- [1] [n. d.]. CMU Pronouncing Dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>. ([n. d.]). Accessed: 2017-10-30.
- [2] [n. d.]. Keras: The Python Deep Learning library. <https://keras.io/>. ([n. d.]).
- [3] [n. d.]. Theano. <http://deeplearning.net/software/theano/>. ([n. d.]).
- [4] 2017. YouTube for the press. <https://www.youtube.com/yt/about/press/>. (2017).
- [5] Swati Agarwal and Ashish Sureka. 2014. A Focused Crawler for Mining Hate and Extremism Promoting Videos on YouTube. In Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT '14). ACM, 294296.
- [6] Nisha Aggarwal, Swati Agrawal, and Ashish Sureka. 2014. Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos. In 12th Annual Conference on Privacy, Security and Trust, PST. 8493.
- [7] Saleem Alhabash, Jong hwan Baek, Carie Cunningham, and Amy Hagerstrom. 2015. To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in Human Behavior* 51, Part A (2015), 520-531.
- [8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).
- [9] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In CEAS, Vol. 6.
- [10] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems* 50, 3 (2011), 602-613.
- [11] Ali Breland. 2017. YouTube cracking down on hate speech. <http://thehill.com/policy/technology/347868-google-launches-initiative-to-reduce-hateful-content-on-youtube>. (2017).
- [12] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter. In ACM Hypertext.
- [13] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In International ACM Web Science Conference.
- [14] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying. In WWW CyberSafety Workshop.

- [15] Maura Conway and Lisa McInerney. 2008. Jihadi Video and Auto-radicalisation: Evidence from an Exploratory YouTube Study. Springer Berlin Heidelberg, 108118.
- [16] YouTube CreatorAcademy. [n. d.]. <https://creatoracademy.youtube.com/page/lesson/thumbnails/#yt-creators-strategies-5>. ([n. d.]). Online; accessed October 2017.
- [17] M. Dadvar, Rudolf Berend Trieschnigg, and Franciska M.G. de Jong. 2014. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. Springer Verlag, 275281.
- [18] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In Proceedings of the First International Workshop on Multiple Classifier Systems.
- [19] Ekaterina Egorova and Jordi Luque Serrano. 2016. Semi-Supervised Training of Language Model on Spanish Conversational Telephone Speech Data. *Procedia Computer Science* 81, Supplement C (2016), 114–120. <https://doi.org/10.1016/j.procs.2016.04.038> SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.
- [20] Daniel Ballcells Eichenberger. 2016. Speech activity detection: Application-specific tuning and context-based neural approaches. Bachelor Thesis. Universitat Politècnica de Catalunya, UPC.
- [21] Mattias Ekman. 2014. The dark side of online activism: Swedish right-wing extremist video activism on YouTube. *MedieKultur: Journal of media and communication research* 30, 56 (2014), 21.
- [22] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research (JMLR)* 15, 1 (Jan. 2014), 31333181.
- [23] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1 (ICASSP'92). IEEE Computer Society, Washington, DC, USA, 517520. <http://dl.acm.org/citation.cfm?id=1895550.1895693>
- [24] V. Goel, S. Kumar, and W. Byrne. 2004. Segmental Minimum Bayes-Risk Decoding for Automatic Speech Recognition. *IEEE Transactions on Speech and Audio Processing* 12, 3 (2004), 234249.
- [25] Ramesh A Gopinath. 1998. Maximum likelihood modeling with Gaussian distributions for classification. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, Vol. 2*. IEEE, 661664.
- [26] Michael Green, Ania Bobrowicz, and Chee Siang Ang. 2015. The lesbian, gay, bisexual and transgender community online: discussions of bullying and self-disclosure in YouTube videos. *Behaviour & Information Technology* (February 2015), 19.
- [27] Dorothy Wunmi Grigg. 2010. Cyber-aggression: Definition and concept of cyberbullying. *Australian Journal of Guidance and Counselling* 20, 02 (2010).

- [28] Gabriel Emile Hine, Jeremiah Onalapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In ICWSM.
- [29] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [30] Bo-June Paul Hsu and James R Glass. 2008. Iterative language model estimation: efficient data structure & algorithms.. In *Proc. Interspeech*. 841844.
- [31] Mallory Hussin, Savannah Frazier, and J. Kevin Thompson. 2011. Fat stigmatization on YouTube: A content analysis. *Body Image* 8, 1 (2011), 90-92.
- [32] L.A. Janson. 2013. Flaming motivation in YouTube users as a function of the traits Disinhibition seeking, Assertiveness and Anxiety? Technical Report.
- [33] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128-3137.
- [34] Imrul Kayes, Nicolas Kourtellis, Daniele Quercia, Adriana Iamnitchi, and Francesco Bonchi. 2015. The Social World of Content Abusers in Community Question Answering. In *WWW*.
- [35] Kyounghee Kwon and Anatoliy Gruzd. 2017. Is Aggression Contagious Online? A Case of Swearing on Donald Trump's Campaign Videos on YouTube. *50* (01 2017), 2165.
- [36] K. Hazel Kwon and Anatoliy Gruzd. 2017. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* 27, 4 (2017), 991-1010.
- [37] Patricia G. Lange. 2014. Commenting on YouTube rants: Perceptions of inappropriateness or civic engagement? *Journal of Pragmatics* 73, Supplement C (2014), 53-65. *The Pragmatics of Textual Participation in the Social Media*.
- [38] Jordi Luque, Carlos Segura, Ariadna Sanchez, Mart Umbert, and Luis Angel Galindo. 2017. The Role of Linguistic and Prosodic Cues on the Prediction of Self-Reported Satisfaction in Contact Centre Phone Calls. In *Proc. Interspeech 2017*. 2346-2350. <https://doi.org/10.21437/Interspeech.2017-424>
- [39] Shivraj Marathe and Kavita P. Shirsat. 2015. Approaches for Mining YouTube Videos Metadata in Cyber bullying Detection. *4* (05 2015), 680-684.
- [40] Peter J. Moor, Ard Heuvelman, and Ria Verleur. 2010. Flaming on YouTube. *Computers in Human Behavior* 26, 6 (November 2010), 1536-1546.
- [41] A. Oksanen, D. Garcia, A. Sirola, M. Nsi, M. Kaakinen, T. Keipi, and P. Rsnen. 2015. Pro-Anorexia and Anti-Pro-Anorexia Videos on YouTube: Sentiment Analysis of User Responses. *Journal of Medical Internet Research* 17, 11e256 (October 2015).

- [42] Jaimie Yejean Park, Jiyeon Jang, Alejandro Jaimes, Chin-Wan Chung, and Sung-Hyon Myaeng. 2014. Exploring the User-generated Content (UGC) Uploading Behavior on Youtube. In WWW Companion. 529534.
- [43] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
- [44] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In ACSAC.
- [45] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. 2010. Mining YouTube to Discover Extremist Videos, Users and Hidden Communities. Springer Berlin Heidelberg, Berlin, Heidelberg, 1324.
- [46] The Guardian. 2015. Twitter CEO: We suck at dealing with trolls and abuse. goo.gl/6CxnwP. (2015).
- [47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016).
- [48] Andrew Weaver, Asta Zelenkauskaitė, and Lelia Samson. 2012. The (Non)Violent World of Youtube: Content Trends in Web Video. Journal of Communication 62, 6 (December 2012), 1065-1083.

10. Conclusions

For this work, we explored how mainstream and fringe Web communities spread misinformation. Furthermore, we studied how these communities influence each other in their attempt to share mainstream and alternative news. To the best of our knowledge, our analysis constitutes the first attempt to characterize the dissemination of mainstream and alternative news across multiple social media platforms, and to estimate a quantifiable influence between them.

Moreover, we studied other types of activities related to fake information: clickbaits, raids, archiving, hoaxes, etc.

In addition, we demonstrate our ability to detect some of these attempts before they become a real problem. This work is an early step towards a system that can potentially monitor and advise users about the authenticity of the information that they are daily exposed to.

Last, within this work we implemented the Back-End and the Intelligent Web-Proxy environments of our infrastructure. Future work will focus on the deployment of the above research and implemented algorithms in the ENCASE aforementioned infrastructure. This deployment will support the browser add-on that warns users about false information dissemination and identity misrepresentation.

11. Copyright and Intellectual Property

The intellectual property will be jointly owned between the Institutions that each of the ENCASE partners. If a project partner decides to move institutions for the duration of the project the Institution to which they move would not become a join owner, and the ownership will remain with the institution at which partners are originally based.