Measuring and Characterizing Hate Speech on News Websites

Savvas Zannettou
Max Planck Institute for Informatics
szannett@mpi-inf.mpg.de

Mai ElSherief
Georgia Institute of Technology
melsherief@gatech.edu

Elizabeth Belding
University of California, Santa Barbara
ebelding@cs.ucsb.edu

Shirin Nilizadeh
University of Texas at Arlington
shirin.nilizadeh@uta.edu

Gianluca Stringhini
Boston University
gian@bu.edu

ABSTRACT

The Web has become the main source for news acquisition. At the same time, news discussion has become more social: users can post comments on news articles or discuss news articles on other platforms like Reddit. These features empower and enable discussions among the users; however, they also act as the medium for the dissemination of toxic discourse and hate speech. The research community lacks a general understanding on what type of content attracts hateful discourse and the possible effects of social networks on the commenting activity on news articles.

In this work, we perform a large-scale quantitative analysis of 125M comments posted on 412K news articles over the course of 19 months. We analyze the content of the collected articles and their comments using temporal analysis, user-based analysis, and linguistic analysis, to shed light on what elements attract hateful comments on news articles. We also investigate commenting activity when an article is posted on either 4chan's Politically Incorrect board (/pol/) or six selected subreddits. We find statistically significant increases in hateful commenting activity around real-world divisive events like the "Unite the Right" rally in Charlottesville and political events like the second and third 2016 US presidential debates. Also, we find that articles that attract a substantial number of hateful comments have different linguistic characteristics when compared to articles that do not attract hateful comments. Furthermore, we observe that the post of a news articles on either /pol/ or the six subreddits is correlated with an increase of (hateful) commenting activity on the news articles.

ACM Reference Format:
Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites . In *12th ACM Conference on Web Science (WebSci '20), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394231.3397902

WebSci '20, July 6–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7989-2/20/07…$15.00
https://doi.org/10.1145/3394231.3397902

1 INTRODUCTION

As the Web becomes more social, so becomes the discourse around news events. People share news articles on social media and discuss them with their friends [40, 68]. At the same time, news websites have become "social," allowing users to post comments and discuss stories among themselves [20, 61]. While the ability to post comments empowers users to discuss news stories in a constructive fashion, discussion can also become toxic, leading to racist remarks and hate speech [24, 33, 37]. In particular, recent research showed that polarized Web communities such as 4chan's Politically Incorrect Board (/pol/) and Reddit's The_Donald board often organize coordinated campaigns in which users are instructed to "attack" a target by using hate speech [25, 35, 44]. In some cases, these "*raids*" can be directed towards news stories from sites that advocate policies that these users do not agree with. Despite the problem that hate speech in news comments poses to news platforms and users, comment moderation remains an open problem [51].

While hate speech and toxic discourse on social media has been the subject of study by a number of researchers [17, 19, 23], as a research community we still lack understanding on the characteristics and the dynamics of hateful comments on news articles. In this paper, we perform a large-scale quantitative study of hateful news comments. We analyze 125M comments from 412K news articles posted between July, 2016 and February, 2018. To select the articles, we use all the news articles that are posted by popular news sites and for which links to them appear on 4chan's /pol/ and six selected subreddits from Reddit.

Research Questions. We aim to answer the following research questions: 1) Is hateful commenting activity correlated with real-world events? 2) Can we find important differences between the users that are posting on news sites according to their partisanship? 3) Can we find linguistic differences in articles that attract substantial numbers of hateful comments when compared to articles that do not? and 4) Do news articles attract more hate comments after they are posted on other Web communities like 4chan and Reddit?

To shed light on these research questions, we present a temporal and content analysis. We leverage changepoint analysis [39] to find significant changes in the time series of (hateful) commenting activity.We also use linguistic analysis that reveals the writing and linguistic peculiarities of news articles and whether articles that attract hate comments have differences to articles that do not attract hate. Overall, this paper provides an unprecedented view on hateful

commenting activity on news websites and on the characteristics of news articles that attract significant hate from users.

Findings. Among others, we make the following findings:

- We find a substantial increase in (hate) comments in close temporal proximity with important real-world events; e.g., we find statistically significant changes in hateful comments in news articles in close temporal proximity with the "Unite the Right" rally in Charlottesville during August, 2017, as well as the second and third US Presidential debates in 2016.
- We find differences between the users that are commenting on news articles according to the site's partisanship. Users that post on extreme-right sites tend to be more active overall by posting more comments and they tend to post more hateful content compared to users that are active on sites with other partisanships. Also, we find a higher percentage of hateful comments from users that choose to remain anonymous.
- Our linguistic analysis reveals that there is a correlation between articles using the highest number of Clout words (probably for influencing the readers) and attracting more hate comments. We also find that the articles that had more than 10% hateful comments, use more social references and include negative emotions, such as, *anxiety* and *anger* emotions, compared to those articles that receive no hate comment.
- We find a correlation between a link being posted on Reddit or /pol/, and receiving more (hateful) comments on that article. In particular, we find that the posting of news articles from domains with specific partisanships (i.e., Left, Center, Center-Right) to /pol/ or the six selected subreddits is correlated with an increase in hateful commenting activity in close temporal proximity with the posting of the news article on /pol/ or Reddit. We also discover that once a news article receives a substantial amount of hateful comments, it continues to receive a high fraction of such comments for a long period of time.

2 RELATED WORK

Hate Speech Detection. A large body of work focuses on detecting hate speech. HateSonar is a classifier [19] that uses Logistic Regression to classify text into: offensive language, or hate speech. Recently, Google has released a state of the art hate speech detection tool, called Perspective API [53], that detects textual toxic content, including hate speech. This tool uses machine learning techniques and a manually curated dataset of texts, to identify the rudeness, disrespect, or toxicity of any comment. Most previous work [30, 41, 59, 63, 64] proposes the use of supervised machine learning approaches, such as Support Vector Machines, Naive Bayes, and Logistic Regression, as well as Natural Language Processing techniques. Others [21, 26, 29, 55] propose the use of neural network-based classifiers.Another work [31] uses a semi-supervised approach to detect different forms of hate speech like implicit and explicit hate content. Chandrasekharan et al. [16] propose Bag of Communities: an approach that uses data from 4chan, Voat, Reddit, and Metafilter, and aims to detect abusive content. Finally, Saleem et al. [54] focus on multiple networks like Reddit and Voat, and propose the use of a community-driven detection approach.

Hate Speech on the Web. Some recent work studies the prevalence and characteristics of hate speech on specific web communities, such as Gab [66], 4chan's Politically Incorrect board (/pol/) [35], Twitter and Whisper [58]. Some works [47] study the effects of anonymity and forms of hate speech. Others [22, 23] perform an analysis on the personality of the targets and instigators of hate speech on Twitter. Another study by Zannettou et al. [69] shows the rise of racial slurs and in particular anti-semitism on 4chan and Gab. Chandrasekharan et al. [15] study the degree of hate speech on the platform after the bans of some prominent hateful subreddits like r/fatpeople and r/CoonTown, finding that these bans helped decrease the site's hate speech usage. This is because a lot of accounts that were active on these subreddits stopped using the site and others that migrated to other subreddits did not post hateful content. Olteanu et al. [50] focus on understanding the effect that real-world extremist attacks, involving Arabs and Muslims, have on hateful speech on the Web. Among other things, they observe an increase in the use of hate speech after such attacks and in particular increase in posts that advocate violence. Jhaver et al. [38] study the effects of blocklists (i.e., blocking users) on online harassment, finding that users are not adequately protected online, while others feel that they are blocked unfairly. Finally, a recent work by Zannettou et al. [67] studies the dissemination of hateful memes across the Web.

Hate Speech on News Comments. Some studies analyze aspects of hate speech on comments posted on news articles. Erjavec and Kovacic [24] undertake interviews with posters of hate speech on news sites to uncover their motives and strategies to share hateful content, finding that posters are driven by thrill and fun, while others are organized. Hughey and Daniels [37] analyze the methodological pitfalls for studying racist comments posted on news articles. Specifically, they analyze various strategies employed by news platforms, such as extreme moderation policies, not storing comments or disabling comments, and their implications on the Web. Harlow [33] analyzes comments posted on US news sites to understand racist discourse. They find that the comments included racial slurs despite the fact that the article did not; Latinos were the most targeted ethnicity.

3 METHODOLOGY

Dataset. Our dataset includes news articles and the comments posted on them between July 2016 and February 2018, on 4chan's Politically Incorrect board (/pol/) and six subreddits from Reddit, namely AskReddit, politics, conspiracy, The_Donald, news, and worldnews. We select these subreddits because they are among the most important subreddits when it comes to sharing news articles on Reddit [68]. These subreddits attract both a general audience (i.e., news, politics, worldnews, AskReddit subreddits), as well as users that are more into conspiracy theories and the alt-right (i.e., conspiracy, The_Donald, and /pol/). Due to this diversity in the Web communities where we collect news articles from, we expect that the collected articles will include a mixture of both mainstream, and possibly unbiased articles, as well as biased articles likely towards the alt-right community.

First, we extract all URLs that are posted on /pol/ and the six selected subreddits between July 2016 and February 2018. For obtaining the datasets for /pol/ we use the methodology presented by [35], while for Reddit we use publicly available data from Pushshift [12]. Then, we select the top 100 domains according to their popularity

Table 1: Top news sources that support comments as of June, 2018, that appear on /pol/ and the six selected subreddits.

News site	Com. platform (as of June 2018)	# of articles on /pol/	# articles on 6 subreddits	# collected articles	# collected comments
dailymail.co.uk	Custom	14,124	31,861	38,463	14,287,096
theguardian.com	Custom	10,430	49,318	42,137	11,090,592
nytimes.com	Custom	9,288	89,359	54,107	4,995,119
washingtonpost.com	Custom	9,213	136,120	-	-
breitbart.com	Disqus	7,698	39,793	41,918	46,684,682
independent.co.uk	Custom	6,232	28,971	-	-
rt.com	Spot.IM	5,980	13,913	17,075	2,707,512
thehill.com	Disqus	3,610	46,957	47,226	28,862,389
almasdarnews.com	Oneall	3,589	477	-	-
express.co.uk	Spot.IM	3,344	6,353	8,609	99,569
huffingtonpost.com	Facebook	3,009	34,999	27,092	1,089,113
cbc.ca	Custom	2,743	11,127	-	-
dailycaller.com	Disqus	2,727	18,516	19,457	5,326,962
politico.com	Facebook	2,684	26,247	19,916	626,386
latimes.com	Custom	2,091	15,902	-	-
thesun.co.uk	Custom	1,848	3,822	-	-
washingtontimes.com	Spot.IM	1,793	12,531	13,236	1,745,613
mirror.co.uk	Custom	1,734	5,001	-	-
infowars.com	Disqus	1,533	8,682	8,789	3,799,653
newsweek.com	Facebook	1,481	11,110	9,336	66,380
sputniknews.com	Facebook+Custom	1,380	3,808	4,343	29,368
timesofisrael.com	Facebook	1,301	4,367	4,588	110,466
dailywire.com	Disqus	1,173	6,892	7,343	603,208
welt.de	Custom	1,139	504	-	-
jpost.com	Spot.IM	1,080	4,037	4,707	294,250
slate.com	Custom	916	9,049	-	-
salon.com	Spot.IM	794	9,673	9,792	292,370
huffpost.com	Facebook	583	7,106	5,996	1,711,612
townhall.com	Disqus	548	7,015	7,235	693,372
firstpost.com	Facebook	76	23,310	20,759	555
Total		**104,141**	**666,820**	**412,124**	**125,116,267**

in each online service. However, not every popular domain in these communities is actually a news site. For example, the most popular domain on /pol/ is YouTube [35]. Therefore, to identify domains that refer to *news* sites, we used the Virus Total URL categorization API [10], which provides categories given a domain. After obtaining the set of categories for each domain, we select the domains that have the "news" term in either of the returned categories, thereby obtaining a set of 64 news sites. Then, during June 2018, we manually inspected these news sites to identify whether they allowed users to post comments, and if so what technology they used. We found that 34 (53.1%) sites do not support comments on their platform, six (9.3%) sites use *Disqus* [1], five (7.8%) sites use *Spot.IM* [9], seven (10.9%) sites use *Facebook* [2], while twelve (18.7%) sites use custom solutions. The full list with all the sites is available at [4].

Next, we aimed to implement tools to collect comments from the articles. Initially, we looked at multiple domains that use the same commenting platforms; e.g., Disqus, Spot.IM, and Facebook. For each of these, we built a crawler that uses the platform's API to get all the comments on articles posted on /pol/ or the six subreddits. For news sites that use custom solutions as their commenting platforms, we had to implement a separate crawler for each domain, which is not efficient. Therefore, we focused on the domains for which we have the most articles; we implemented custom crawlers for dailymail.co.uk, theguardian.com, and nytimes.com. Note that we initially aimed to also implement a crawler for washingtonpost.com but we were unable due to implementation issues. Table 1 summarizes the number of the collected articles and comments for each news site that supports comments as of June 2018. Note that since we collect the data well after their publication date (collection period between June and November 2018), there is a small percentage of articles that are not available either because they were removed or because the URL was not available. In total, we obtained 125M comments posted on 412K news articles. Finally, for each article, we collected its content and associated article metadata using Newspaper3k [7].

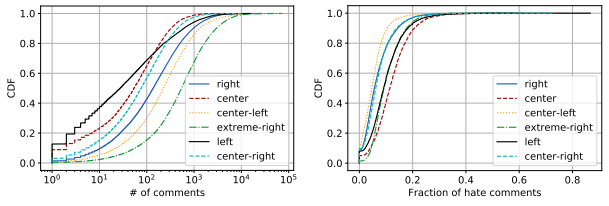
Table 2: News sites in our dataset and their partisanship.

Partisanship	News sites
Left	salon.com, huffpost.com, huffingtonpost.com, newsweek.com, firstpost.com
Center-Left	nytimes.com, theguardian.com, thehill.com, timesofisrael.com
Center	jpost.com, politico.com
Center-Right	rt.com, washingtontimes.com, sputniknews.com
Right	dailymail.co.uk, express.co.uk, dailycaller.com, dailywire.com, townhall.com
Extreme-Right	breitbart.com, infowars.com

Identifying partisanship. To identify the partisanship of news sites, we use information about news media listed on the Media Bias/Fact Check (MBFC) website [6], which contains annotations and analysis of the factual reporting and/or bias for news sites. MBFC has been used to annotate data in prior work for analyzing the factuality of reports and bias of news media [11]. Table 2 shows the partisanship/bias of each news site in our dataset.

Identifying hate comments. To identify comments that are hateful, we explore the use of two popular hate speech classifiers: Hatesonar [19] and the Perspective API [53]. The former is a classifier that uses Logistic Regression to classify comments as hateful, offensive, or neither. The classifier is trained on a corpus of 24K tweets annotated as either "Hate Speech," "Offensive Language," or "Neither" by workers on CrowdFlower. Similarly, the Perspective API leverages crowdsourced annotations of text to train machine learning models that predict the degree of rudeness, disrespect, or unreasonableness of a comment. In particular it offer two distinct models: the "Toxicity" and "Severe Toxicity" models. The difference between the two models is that the latter is more robust to the use of swear words. To assess the performance of these classifiers in our dataset, we extract a set of 100 random comments. Then, three of the authors of this study independently marked each comment as hateful or not, and we treat the majority agreement of these annotations as groundtruth. Then, all comments in our random sample were evaluated both with HateSonar and the Perspective API. We find that HateSonar performs poorly on our random sample (precision 0.5 and recall 0.31), while the Severe Toxicity model of Perspective API performs substantially better (precision 0.71 and recall 0.52). Interestingly, the Toxicity model of Perspective API performs better with respect to recall but is subpar in terms of precision (precision 0.53 and recall 0.84). Based on these results, we elect to use the Severe Toxicity model available from Perspective API, mainly because we favor precision over recall and we aim to be more robust to the use of swear words (i.e., not everything that includes a swear word is hateful).

Note that hate speech detection is an open research problem and, to the best of our knowledge, there is no classifier that can detect all kinds and forms of hate speech. This task is even difficult for humans as there are no clear definitions of what constitutes hate speech. For instance, in our random sample the three human annotators had a Fleiss Inter-Annotator agreement score of 0.39 that can be regarded as "fair agreement" [3]. Due to this, in this work, we follow a best effort approach to study the prevalence and spread of hate speech using Perspective API that outperforms other readily available alternatives, such as the HateSonar classifier.



(a) (b)

Figure 1: CDF of the number of (a) comments per article and (b) fraction of hate comments over all the comments per article.

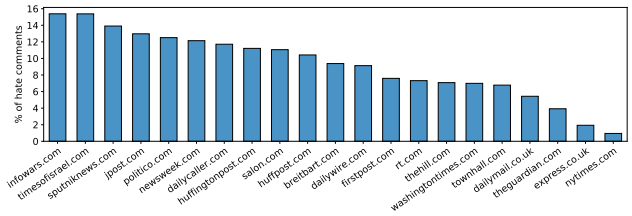
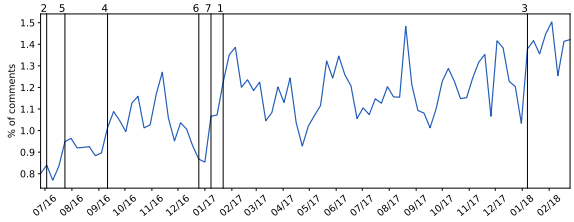
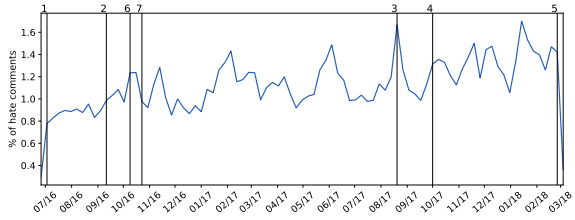


Figure 2: Percentage of hate comments.



(a) All comments



(b) Hate comments

Figure 3: Temporal overview of the collected comments. The figures are annotated with significant changes in the time series using changepoint analysis. See Tables 3 and 4, respectively, for real-world events that coincide with each changepoint.

4 RESULTS

In this section, we first provide a general characterization of the collected data with a focus on hateful content. Next, we provide a user-based analysis to understand user activity on news article comments and then we investigate whether news articles with specific linguistic features attract more hateful content. Finally, we examine whether there is any correlation between posting an article on 4chan's /pol/ and six subreddits and receiving hateful comments on those articles.

4.1 General Characterization

Prevalence of Hate Comments. We present statistics of the comments that are posted for news articles and the prevalence of hate speech in these comments. Fig. 1 shows the cumulative distribution function (CDF) of the number of comments and the fraction of hate comments over all comments per news article, grouped by the partisanship of the news sites (see Table 2). Note that for readability purposes we only show the distributions for articles that have at least one comment. When looking at the distribution of all the comments (Fig. 1(a)), we observe that extreme-right sites attract more comments, while left and center sites have a substantially lower commenting activity. To assess whether these results are affected by the different size of audiences for each news site, we use SimilarWeb [8] to obtain the number of monthly views per news site (as of December 2018). The full list of these views are publicly available via [5]. Interestingly, we find that the most visited partisanship of news sites in our dataset is center-left (669M visits), followed by right (491M visits), center-right (286M visits), left (251M visits), extreme-right (77M visits), and last center (65M visits). These findings indicate that the audience of left and extreme-right news sites are more active in posting comments despite the fact that center-left, right, and center-right news sites have a larger number of visits.

For hate comments (Fig. 1(b)), we plot the fraction of hate comments over the overall number of comments per article. We find that center and left-leaning sites attract more hate speech, while center-left sites have the lowest rate of hate comments. To assess whether the distributions shown in Fig. 1 have statistically significant differences, we perform a two-sample Kolmogorov-Smirnov (KS) test for each pair of distributions; in all cases we find statistically significant differences with $p < 0.01$.

Fig. 2 shows the percentage of hate comments over all the comments posted in news articles, grouped by news site. We find that infowars.com, a popular alt-right conspiracy-oriented news site, and timesofisrael.com are the sites with the highest percentage of hate comments (15.3%), followed by sputniknews.com (13.9%), jpost.com (12.9%), and politico.com (12.5%). When looking at the news sites with the least hateful commenting activity we find nytimes.com (0.9%), followed by express.co.uk (1.9%), and theguardian.com (3.9%). These results highlight the audience and comment moderation for each site: *i.e.,* infowars.com is likely to attract users that post hate comments and the site might not apply strict moderation policies, while nytimes.com might not attract hate comments or it might enforce strict moderation policies.

Temporal Analysis. Here, we examine the temporal aspect of the collected comments to understand how (hateful) commenting activity changes over time. This is a particularly interesting and important analysis since it will allow us to understand whether hateful commenting activity is correlated with real-world events and whether hateful commenting activity is increasing or decreasing over time. Fig. 3 shows the weekly percentage of comments and hateful comments for the whole dataset. We focus on the time period after July, 2016, as the vast majority of the collected comments are within the depicted time period. We find that the overall commenting activity started increasing during the months leading to the 2016 US

Table 3: Statistically significant changepoints and coinciding real-world events in the time series of all the comments.

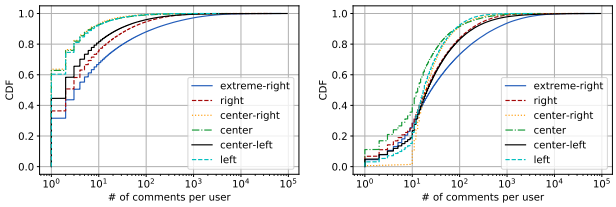
Changepoint	Events
1 - 2017/01/22	2017/01/20: Presidential Inauguration of Donald Trump [36].
2 - 2016/07/03	2016/07/02: Thousands of people protest in London against Brexit [43].
3 - 2018/01/07	2018/01/02: Donald Trump responds to Kim Jong-Un stating that his nuclear missile launch button is larger and more powerful [65].
4 - 2016/09/11	2016/09/09: US congress passes a law to allow families of 9/11 victims to sue Saudi Arabia [27]. 2016/09/11: Hillary Clinton is treated for pneumonia after leaving a ceremony honoring the anniversary of 9/11 attacks [45].
5 - 2016/07/24	2016/07/19: Donald Trump is nominated as the Republican's candidate for the 2016 US election [18].
6 - 2016/12/25	2016/12/22: Donald Trump names Kellyanne Conway as Counselor to the President and Sean Spicer as White House Press Secretary [13, 57].
7 - 2017/01/08	2017/01/06: A US intelligence document reports that Vladimir Putin ordered a campaign to influence the 2016 US election [28].

Table 4: Statistically significant changepoints and coinciding real-world events in the time series of hateful comments.

Changepoint	Events
1 - 2016/07/03	2016/07/02: Thousands of people protest in London against Brexit [43].
2 - 2016/09/11	2016/09/09: US congress passes a law to allow families of 9/11 victims to sue Saudi Arabia [27]. 2016/09/11: Hillary Clinton is treated for pneumonia after leaving a ceremony honoring the anniversary of 9/11 attacks [45].
3 - 2017/08/13	2017/08/11: Unite the Right rally begins in Charlottesville, Virginia [60].
4 - 2017/10/01	2017/10/02: Shooting in Las Vegas leads to the death of 59 people [49].
5 - 2018/02/18	2018/02/14: Shooting at Stoneman Douglas High School with 17 people dead [32].
6 - 2016/10/09	2016/10/09: Second presidential debate of the 2016 US election take place [56]
7 - 2016/10/23	2016/10/19: Third presidential debate of the 2016 US election take place at Las Vegas [34].

election (between September and November 2016), decreased after the election, while again started increasing after Trump's Inauguration (January 2017). Furthermore, we note that the biggest peak in commenting activity coincides with the "Unite the Right" rally in Charlottesville [60], during August 2017, which lead to the death of one woman [14]. When looking at the hate comments (Fig. 3(b)), we find a somewhat similar activity with all the comments (Fig. 3(a)). Some peaks in hateful commenting activity coincide with the 2016 US election period, with Trump's Inauguration in January 2017, with the Charlottesville rally in August 2017. Since our dataset is based on articles posted on 4chan's /pol/ and the six subreddits, these findings indicate that their users are particularly interested in discussing these political events and that they likely comment on them both on their platform as well as in the comments section of each article.

We further investigate whether the peaks in overall and hate commenting activity are statistically significant with respect to the time series of the comments. We run changepoint analysis that provides points in time where statistically significant changes occur on a time series. Specifically, we run the Pruned Exact Linear Time (PELT) algorithm [39] on the weekly time series of both all comments and hate comments. This algorithm maximizes the log-likelihood of the means and variances of the time series with a penalty function that enables us to rank the changepoint according to their statistical significance. Fig. 3 is annotated with the obtained changepoints for both all comments and hate comments, while Tables 3 and 4 report each changepoint and real-world events that coincide with each changepoint. For the overall commenting activity we find statistically significant changepoints that coincide with the Presidential Inauguration of Donald Trump (changepoint 1 in Table 3), Brexit protests (changepoint 2 in Table 3), and developments on the USA-North Korea relations (changepoint 3 in Table 3). For hateful commenting activity we find statistically significant changepoints that coincide with Brexit developments (changepoint 1 in Table 4),



(a) All users (b) Users with at least ten comments

Figure 4: CDF of the number of comments per user.

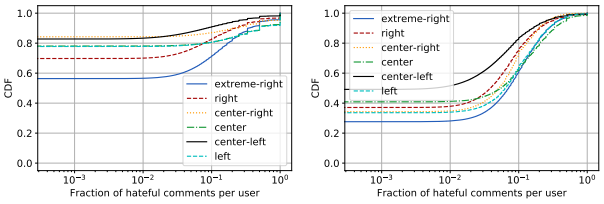
the Las Vegas shooting during October 2017 (changepoint 4 in Table 4), developments in US politics (changepoint 2 in Table 4), and the presidential debates during the 2016 US election (changepoints 6 and 7 in Table 4). Finally, we find a changepoint coinciding with the Charlottesville protest (changepoint 3 in Table 4).

4.2 User Analysis

In this section, we analyze the users that comment on news articles. We are particularly interested in understanding how these users interact in the comments of news articles, how persistent users are in disseminating hateful comments, and whether users that post on news sites with specific partisanship are more hateful. Furthermore, since some commenting platforms (e.g., Disqus) allow users to post comments anonymously, we investigate the effect of anonymity with respect to the dissemination of hateful comments on news articles. Note that due to ethical reasons, we do not make any attempt to link users across the multiple commenting platforms we study, while at the same time we make no attempt to de-anonymize users.

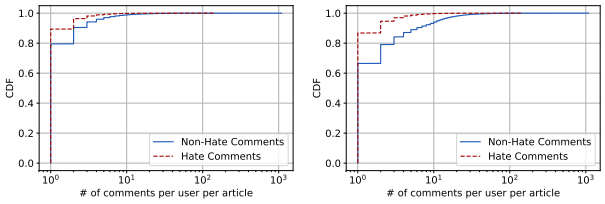
Effect of anonymity. We investigate the prevalence of posting comments anonymously. We find that in our dataset 6.5M (5.2%) comments are posted by anonymous users, while the rest of the comments are posted by users that have accounts with the various commenting platforms we study. Next, we look into the prevalence of hateful comments in each of these subsets: we find that in the anonymous subset there are relatively more hateful comments (10.7% of them), while for the subset where users had accounts we find a lower percentage of hateful comments (7.6%), which is inline with previous work focusing on hate speech on anonymous and non-anonymous platforms [48]. We also assess the statistical significance of these results with a Chi-square test on the number of hateful and non-hateful comments for anonymous and non-anonymous users, finding statistically significant differences ($p < 0.01$). Overall, these findings indicate that most users do not mind creating an account on these commenting platforms and that users that choose to remain anonymous are more likely to share hateful comments.

Overall User Activity. Since we want to analyze the dataset in the granularity of specific users, we therefore next focus on the subset of the dataset where users posted comments by creating accounts on the commenting platforms. Overall, we find 3.1M accounts across all the commenting platforms. To get a better understanding of how users interact with news comments, we plot the CDF of the number of comments per user in our dataset in Fig. 4. Since a substantial percentage of users had only posted one comment, we show the results for users that posted at least ten comments through all the articles in Fig. 4(b). Specifically, we find that from the users that



(a) All users (b) Users with at least ten comments

Figure 5: CDF of the fraction of hateful comments per user.



(a) All users (b) Users with at least ten comments

Figure 6: CDF of the number of comments per user per article.

are active on extreme-right news articles comments, 31% of them posted only once across all news articles, while the same percentage increases for other partisanships: 36% for right, 44% for center-left, 60% for left, and 63% for center and center-right. Furthermore, we note that users that post on extreme-right news articles comments are more active (mean number of comments 134.32), followed by users on center-left (mean number of comments 38.6) and right (mean 29.9).

Fig. 5 shows the fraction of hateful comments over all the comments that a user made per partisanship. We make several observations. First, a large percentage of users across all partisanships post only non-hateful comments: e.g., for extreme-right 56% of the users post only non-hateful comments, while for other partisanships like center-right and center-left the percentage is much higher reaching 84%. When we look at the results for the users with at least ten comments (see Fig. 5(b)), however, we note that these percentages are substantially lower compared to all users. This indicates that "power-users" are more likely to share hateful comments, while users that are posting only a few times are less likely to post hateful comments. Second, we note that users that post on extreme-right and right news articles are more likely to post hateful comments compared to users active on center- or left- leaning news articles.

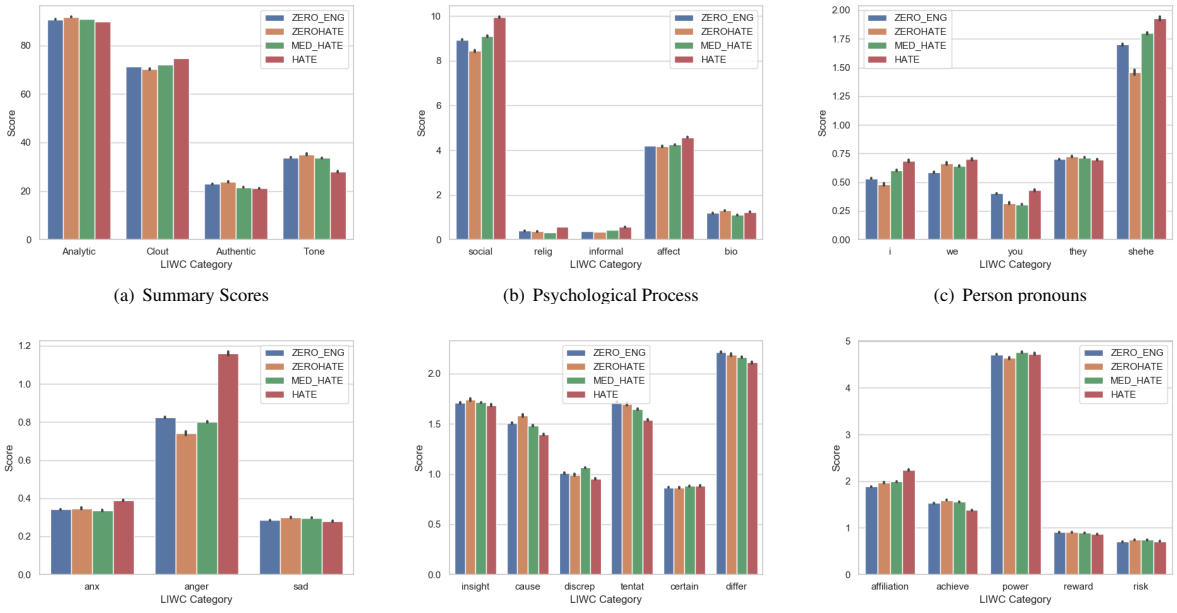
User Activity per Article. Finally, we analyze the user commenting activity in the granularity of specific articles. This analysis allows us to understand the discussion on specific news articles and whether users that post hateful comments are persistent (i.e., posting multiple hateful comments) or whether they are "one-off." We plot the CDF of the number of comments per user for each article by distinguishing between hateful and non-hateful comments in Fig. 6. We observe that for both hateful and non-hateful comments, a large percentage of users post only once on the news article. This happens for 79% for non-hateful comments and 89% for hateful comments, while by only considering users that posted over ten times (see Fig. 6(b)) the percentages decline to 66% for non-hateful and 86% for hateful comments. Also, we run a KS test on the distributions in Fig. 6, finding that the distributions exhibits statistically significant differences ($p < 0.01$). These results indicate that it is more likely that users that post non-hateful comments to hold a lengthy discussion on news articles, while users that post hateful comments are more likely to just post a single hateful comment once and then do not post other hateful comments. Note that we performed the same analysis by dividing the users according to their activity in news articles per partisanship finding no substantial differences between the results across partisanships (we omit these results from the manuscript).

4.3 Content Linguistic Analysis

In this section, we assess whether specific linguistics used in news articles have any correlation with hate intensity. This analysis is important as it sheds light into the linguistics that drive hateful activity in news article comments. These cues can later be used to predict whether an article is likely to attract hate based on linguistics.

In our analysis, we divide the collection of news articles into four types of articles based on their comment engagement and hate intensity in their associated comments: first, those that do not receive any engagement in terms of number of comments (ZERO_ENG); second, those that receive no hate comments (ZERO_HATE); third, those for which the number of hateful comments exceeds a pre-defined threshold k (HATE); and finally, the rest of the articles, which are the ones that receive at least one hate comment but less than the pre-defined threshold k (MED_HATE). By checking the CDF of the hate fraction in different articles (see Fig. 1(b)), we observe that a threshold of 10% over all comments represents a substantial number of articles; hence we set $k = 10\%$. Using this threshold, we find that 52.4% of the articles are ZERO_ENG, 7.3% are ZERO_HATE, 33.2% are MED_HATE, 7.1% are HATE articles.

Articles' Linguistic Styles and Hate Comments. The interplay of language use and journalism, media and society has been the focus of political science and journalism research [42, 62]. In particular, many principles of journalism are grounded in psycho-linguistic research, the study of how language is acquired, represented, and used [46]. To better understand the characteristics of the articles and their relation to receiving hate comments, we perform a psycho-linguistic analysis on the news articles. For a full psycho-linguistic analysis, we use a tool called Linguistic Inquiry and Word Count (LIWC) [52]. LIWC is a text analysis program that calculates the degree to which various categories of words are used in a text. LIWC has been widely adopted by researchers to study emotional, cognitive, and structural components present in individuals' verbal and written speech samples. We focus on the following dimensions provided by the tool: *summary scores*, *psychological processes*, and *linguistic dimensions*. *Summary scores* include general attributes derived from the text, like the authenticity of the text, and basic statistics, like words per sentence. *Psychological processes* describe the emotions that the text exposes, and *linguistic attributes* describe the linguistic style of the text. We perform the analysis on each article. Fig. 7 shows the mean scores for our key LIWC attributes. To assess the statistical significance of our results, we perform unpaired (two sample) t-tests with a 95% confidence interval for the difference between the means. Our analysis yields the following observations:



(a) Summary Scores (b) Psychological Process (c) Person pronouns

(d) Negative Emotions (e) Cognitive processes (f) Drives

Figure 7: Mean scores for LIWC categories across articles with different level of hate comment.

HATE articles include content with the highest *clout* scores and the least *tone* scores in comparison to all other articles. Fig. 7(a) shows the language values obtained from LIWC averaged over all content for ZERO_ENG, ZERO_HATE, MED_HATE, and HATE articles. We show that HATE articles have the highest mean (μ = 74.67, $p < 0.05$) for clout (influence and power) values and the lowest mean (μ = 28.06, $p < 0.05$) for tone. The high *clout* score suggests that the linguistic style of HATE articles is associated with high expertise and confident cues, which can be used to influence an audience. Also, the low *tone* scores suggest that the linguistic style of HATE articles is associated with the highest *negative* tone.

HATE articles include content with the highest *social*, *religion,* and *affect* references in comparison to all other articles. Fig. 7(b) shows that HATE articles have the highest mean for the social (μ = 9.94, $p < 0.05$), religion (μ = 0.57 , $p < 0.05$), and affect (μ = 0.56 , $p < 0.05$). *Social processes* include family, friends, female and male references. For example, an excerpt from a news article, belonging to the HATE category, that evokes the social category is "*Hillary Clinton has an explanation for why women white women in particular voted against her last November they caved in to pressure from their husbands fathers boyfriends and male bosses*." Our analysis also reveals that HATE articles reference religion-related entities and are on average more emotional than other types of articles.

On average, HATE articles include the highest first (*I*) and third person (*she/he*) singular pronouns in comparison to all other types of articles. Fig. 7(c) shows that HATE articles have the highest mean for scores associated with first (μ = 0.68, $p < 0.05$) and third singular pronouns (μ = 1.92 $p < 0.05$). These findings show that articles which are about individual people, or include and cite their opinions receive hate comments with higher probability.

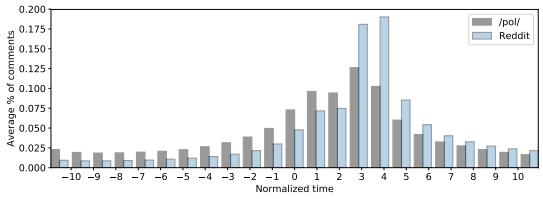
HATE articles include the highest *anger* and *anxiety* references. Fig. 7(d) shows that *anger* is the most prevalent negative emotion for all three types of articles. In particular, HATE articles on average have the highest level of anger (μ = 1.15, $p < 0.05$). Also, we find that HATE articles on average have the highest level of anxiety (μ = 0.39, $p < 0.05$).

HATE articles include the least number of words that suggest causation, discrepancy, tentative, and differentiation. Fig. 7(e) shows that HATE articles tend to have the lowest scores for causation (μ = 1.39, $p < 0.05$), discrepancy (words like "would" and "should," μ = 0.95, $p < 0.05$), tentative (words like "maybe" and "perhaps," μ = 1.53, $p < 0.05$), and differentiation (words like "hasn't," "but," and "else," μ = 2.1, $p < 0.05$). This can indicate that HATE articles tend to have less justification of arguments in terms of causes or effects.

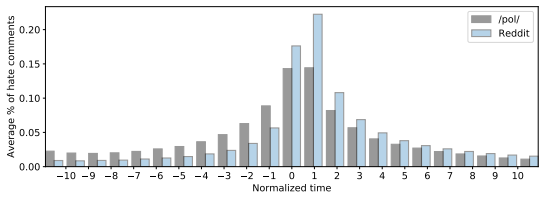
HATE articles include the highest references related to affiliation and the lowest references to achievement. Fig. 7(f) shows that HATE articles have the highest mean for words suggesting affiliation (μ = 2.23, $p < 0.05$) and the lowest achievement references (μ = 1.38 , $p < 0.05$). This likely suggests that HATE articles are motivated by the need to be affiliated to certain groups and because of their negative nature they might not mention achievements.

4.4 Activity after Social Network Posts

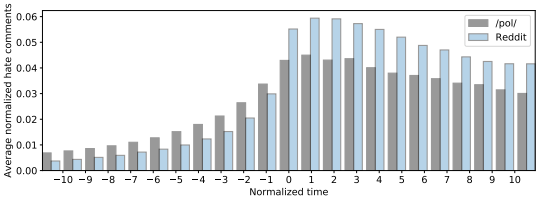
In this section, we study the commenting activity on news articles after they appear on social networks. We aim to provide answers to the following questions: 1) Is the appearance of news articles on social networks like 4chan and Reddit correlated with the (hateful) commenting activity on news articles? 2) How does the (hateful) commenting activity decay after the posting of news articles on



(a) All comments



(b) Hate comments



(c) Hate comments/All comments

Figure 8: Increase of comment activity over time after the post of news articles on six subreddits or /pol/.

4chan and Reddit? 3) What portion of news articles receive increased hateful activity shortly after appearing in other social networks?

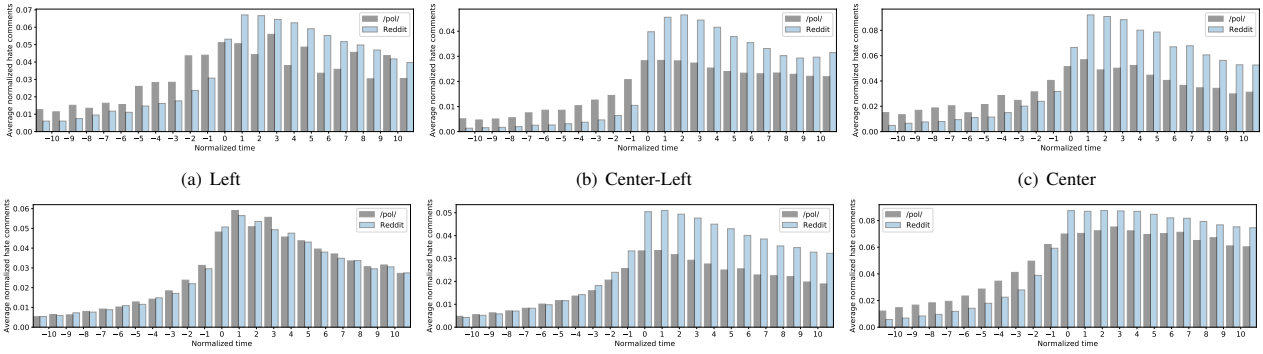
To provide answers to the above questions, we find the first occurrence of each news article on the six subreddits and on /pol/. Then, we normalize the occurrence of each comment in the news article, with respect to the first occurrence of the article in each platform, hence obtaining a view of whether comments, and in particular hate comments, increase after the appearance of articles on Reddit and 4chan. To do this, we subtract the timestamp of each comment in news articles with the timestamp of the first occurrence of the article on the six subreddits and /pol/, hence obtaining a normalized time for the comments. Fig. 8 shows the average percentage of comments that were posted in close proximity with the first occurrence of each article on the six subreddits and /pol/. Time zero corresponds to the first occurrence of the article on /pol/ or the six subreddits, while each bar corresponds to a time period of two hours. For instance, the bars that have the number zero correspond to the time interval between the first occurrence of the article and the next two hours. We report the results using three ways: Fig. 8(a) shows the occurrence of all comments per normalized time slot, Fig. 8(b) shows the occurrence of hateful comments per normalized time slot, while Fig. 8(c) shows the fraction of hateful comments over all comments per normalized time slot. The latter is useful as it captures the correlation between the hateful commenting activity and the overall activity.

We observe that for all comments (see Fig. 8(a)) the commenting activity increases after the first occurrence of the news articles in the six subreddits and /pol/ (normalized time 0) with a peak of activity at normalized time 3 and 4 for /pol/ and the six subreddits, respectively. Also, we find that the commenting activity close to the first occurrence (between 0 and 2 normalized time) is greater for /pol/ when compared to the six subreddits, while later on (after normalized time 2) the percentage activity is larger for the six subreddits. This is likely due to Reddit bots that post news articles without user interaction and likely because of 4chan's ephemeral nature: 4chan users are more likely to interact with the article closer to the article's post on the platform, as threads are short-lived. By only considering the hateful commenting activity (see Fig. 8(b)), we observe a similar pattern with the important difference that the peak in hateful activity is closer to the appearance of the articles on the six subreddits and /pol/, namely during normalized time 1. This indicates that hateful commenting activity increases substantially right after the appearance of news articles on the six subreddits and /pol/, in a far quicker pace when compared to the overall commenting activity.

To further study the interplay between the overall commenting activity and the hateful commenting activity, we plot the fraction of hate comments over all comments per normalized time in Fig. 8(c). We observe that despite the fact that the overall commenting activity and hateful activity decreases substantially after normalized time 4 (see Fig. 8(a) and Fig. 8(b)) the fraction of hateful comments over all comments decreases more gradually and it remains close to the peak (normalized time 4) even at normalized time 10. These results highlight that the hateful commenting activity remains high relative to the overall commenting activity in an article for a long period after the appearance of news articles on the six subreddits and/or 4chan's /pol/, hence indicating that once a news article receives substantial amount of hate it continues to receive a relatively high fraction of hateful comments for a long time period.

Next, we make the same analysis focusing on hate comments, by grouping the articles according to each news site's partisanship (see Table 2). Fig. 9 shows the fraction of hateful comments over all comments per normalized time period for each partisanship (we omit the figures for the overall commenting activity and overall hateful commenting activity due to space constraints). We find that extreme-right news sites are more persistent in hateful commenting activity as the fraction of hateful comments over all comments decays substantially slower compared to the other partisanships. On the other hand, news sites that are more on the center (*i.e.,* center, center-left, center-right) have the fastest decay of hateful comments over all comments. These findings indicate that extreme news sites are more likely to maintain a substantial percentage of hateful commenting activity after the appearance of news articles on the six subreddits and /pol/ when compared to other partisanships on the center.

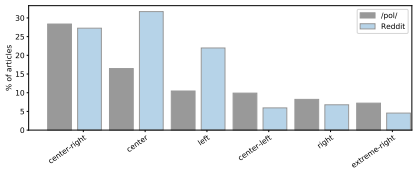
These results are based on all the articles in our dataset that have at least one comment. However, not all articles receive hate comments after their first occurrence in other platforms like /pol/ and the six subreddits. To understand this phenomenon and its prevalence on the Web, we filter the articles so that we select the ones that had the maximum (hateful) commenting activity during the normalized time zero: we find 39K articles for hateful commenting activity and 17K for all commenting activity. Fig. 10 reports the percentage of articles over all articles (with at least one comment) that have an increase in commenting activity, and in particular hate commenting activity, shortly after the first occurrence of the news articles on /pol/ or the six subreddits. We find that domains that are *center-right* have the most articles with commenting activity increase, while *extreme-right* domains have the least (see Fig. 10(a)). When considering only



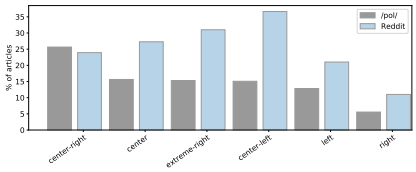
(a) Left (b) Center-Left (c) Center

(d) Center-Right (e) Right (f) Extreme Right

Figure 9: Fraction of hate comments over all comments for each normalized timeslot.



(a) All comments



(b) Hate comments

Figure 10: Percentage of news articles with increased commenting activity after appearing on /pol/ and/or the six subreddits.

hateful activity (see Fig. 10(b)), we find something similar: again, *center-right* domains have the most articles with activity increase and in this case it is hateful. A possible explanation is that users from the six subreddits or /pol/ disagree or have a different ideology with articles from center-right news sites, hence posting hateful comments in the comments section right after their appearance on their platform. Finally, we note that for hateful commenting activity the percentages are higher for Reddit across all partisanships with the exception of center-right, possibly indicating that Reddit users are more likely to post hateful comments on these news articles in close temporal proximity after their appearance on the six subreddits.

5 CONCLUSION

In this paper, we presented a large-scale quantitative analysis of the news commenting ecosystem. We analyzed 125M comments and 412K news articles across several axes: we performed a general characterization of hateful content in news comments, a temporal analysis, as well as a linguistics characterization. Overall, among other things, we found that (hateful) commenting activity increases with notable events that have a strong political nature, articles that attract varying hateful activity have significant linguistic differences, while our user-based analysis reveals that users that post comments in extreme-right sites tend to be more active and post more hateful comments compared to users that post on sites with other partisanships. Furthermore, we found a correlation between the posting of news articles on either /pol/ or the six selected subreddits and increased (hateful) commenting activity on the article.

Naturally our work has some limitations. First, our dataset was collected well after the publication of the articles and their comments, hence it is likely that some of the hateful content was moderated/deleted. Second, we relied on the Perspective API for detecting hate speech, which is expected to miss some hateful content (as mentioned in Section 3). This is because hate speech detection is an open research problem and available classifiers are unable to detect all possible types of hateful content.

To conclude, for our future work, we plan to work on pro-actively detecting organized campaigns, mainly from users of fringe Web communities, that aim to "raid" news articles with hate comments. Also, we aim to assess the effect that other mainstream social networks (e.g., Twitter) have on the commenting activity of news articles. Finally, we plan to build a classifier that will be able to detect whether news articles are likely to attract hateful comments.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (NSF) under Grant CNS-1942610. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] 2019. Disqus API. https://disqus.com/api/docs/.
[2] 2019. Facebook Graph API. https://developers.facebook.com/docs/graph-api/.
[3] 2019. Fleiss' Kappa. https://en.wikipedia.org/wiki/Fleiss_kappa.
[4] 2019. Full list of sites we use. https://bit.ly/2XZtwvA.
[5] 2019. List of monthly views. https://bit.ly/3bB9vzi.
[6] 2019. Media Bias/Fact Check Site. https://mediabiasfactcheck.com/.
[7] 2019. Newspaper3k. https://newspaper.readthedocs.io/en/latest/.
[8] 2019. SimilarWeb Site. https://www.similarweb.com/.
[9] 2019. Spot.Im API. https://developers.spot.im/.
[10] 2019. Virus Total API. https://www.virustotal.com/.
[11] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *EMNLP*.

[12] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *ICWSM*.

[13] Dylan Byers. 2016. Trump picks Sean Spicer as White House press secretary. http://cnnmon.ie/2hZDxUE.

[14] Christina Caron. 2017. Heather Heyer, Charlottesville Victim, Is Recalled as "a Strong Woman". https://nyti.ms/2vuxFZx.

[15] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *CSCW* (2017).

[16] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *CHI*.

[17] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *WebSci*.

[18] Stephen Collinson. 2016. It's official: Trump is Republican nominee. http://cnn.it/2a6ytZN.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

[20] Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In *CSCW*.

[21] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *WWW*.

[22] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding-Royer. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. In *ICWSM*.

[23] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth M. Belding-Royer. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. In *ICWSM*.

[24] Karmen Erjavec and Melita Poler Kovačič. 2012. "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society* (2012).

[25] Claudia Flores-Saviaga, Brian C Keegan, and Saiph Savage. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. In *ICWSM*.

[26] Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A Unified Deep Learning Architecture for Abuse Detection. *WebSci*.

[27] Fox News. 2016. Congress passes bill letting 9/11 victims sue Saudi Arabia, in face of veto threat. http://fxn.ws/2cKQFuW.

[28] Fox News. 2018. Intel report says Putin ordered campaign to influence US election. http://fxn.ws/2jjHnt0.

[29] Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *Workshop on Abusive Language Online*.

[30] Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *RANLP*.

[31] Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *IJCNLP*.

[32] Emanuella Gringberg and Eric Levenson. 2018. At least 17 dead in Florida school shooting, law enforcement says. https://edition.cnn.com/2018/02/14/us/florida-high-school-shooting/index.html.

[33] Summer Harlow. 2015. Story-chatterers stirring up hate: Racist discourse in reader comments on US newspaper websites. *Howard Journal of Communications* (2015).

[34] Barney Henderson. 2016. Donald Trump and Hillary Clinton to clash in Las Vegas "Fight Night" debate: US election briefing and polls. https://bit.ly/3cDVKQu.

[35] Gabriel Emile Hine, Jeremiah Onaolapo, 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. *ICWSM*.

[36] Steve Holland and Emily Stephenson. 2017. Trump, now president, pledges to put "America First" in nationalist speech. http://reut.rs/2iQMMmK.

[37] Matthew W Hughey and Jessie Daniels. 2013. Racist comments at online news sites: a methodological dilemma for discourse analysis. *Media, Culture & Society* (2013).

[38] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *TOCHI* (2018).

[39] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* (2012).

[40] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *WWW*.

[41] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. In *AAAI*.

[42] Annabelle Lukin. 2013. Journalism, ideology and linguistics: The paradox of Chomsky's linguistic legacy and his 'propaganda model'. *Journalism* (2013).

[43] Elaine Ly and Angela Dewan. 2016. Thousands say "No" to Brexit in colorful protest. http://cnn.it/29drqiT.

[44] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. In *CSCW*.

[45] Jonathan Martin and Amy Chozick. 2016. Hillary Clinton's Doctor Says Pneumonia Led to Abrupt Exit From 9/11 Event. https://nyti.ms/2cFiCkr.

[46] Katherine C McAdams. 1984. Psycholinguistics explains many journalism caveats. *The Journalism Educator* (1984).

[47] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *HT*.

[48] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *HT*.

[49] Nytimes. 2017. Multiple Weapons Found in Las Vegas Gunman's Hotel Room. https://nyti.ms/2fKkQ8p.

[50] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hateful speech online. In *ICWSM*.

[51] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *ACL*.

[52] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015.

[53] Perspective API. 2018. https://www.perspectiveapi.com/.

[54] Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *CoRR* (2017).

[55] Joan Serra, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. 2017. Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words.

[56] Rebecca Shabad. 2016. Second presidential debate 2016: What time, how to watch and live stream online. https://cbsn.ws/2S0a4eh.

[57] David Sherfinski. 2016. Kellyanne Conway selected as Donald Trump's counselor. https://go.shr.lc/2TOpkcv.

[58] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *ICWSM*.

[59] Tom De Smedt, Guy De Pauw, and Pieter Van Ostaeyen. 2018. Automatic Detection of Online Jihadist Hate Speech. *CoRR* (2018).

[60] Hawes Spencer. 2017. A Far-Right Gathering Bursts Into Brawls. https://nyti.ms/2uTmIgV.

[61] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2010. News comments: Exploring, modeling, and online prediction. In *ECIR*.

[62] Tom Van Hout. 2015. Between text and social practice: Balancing linguistics and ethnography in journalism studies. In *Linguistic Ethnography*.

[63] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *ITASEC*.

[64] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Workshop on Language in Social Media*.

[65] Eli Watkins. 2018. Trump taunts North Korea: My nuclear button is "much bigger," "more powerful". http://cnn.it/2A7Q4e5.

[66] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *WWW Companion*.

[67] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *IMC*.

[68] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *IMC*.

[69] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. A Quantitative Approach to Understanding Online Antisemitism. *ICWSM*.