بسم الله الرحمن الرحیم

عصر فضای مجازی

گزارش شماره ۶۳
خرداد ۱۴۰۰

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

گفتار نفرت افکن آنلاین
(الکساندرا. سیگل)

محتوای انتشار یافته در این اثر
الزاماً بیانگر دیدگاه مرکز ملی فضای مجازی نیست

تهیه شده در پژوهشگاه فضای مجازی
(گروه مطالات بنیادین)

مترجم: دکتر علیرضا کاظمی (دکتری فلسفه علم و فناوری دانشگاه صنعتی شریف و همکار پژوهشی گروه مطالعات بنیادین)

ناظر علمی: دکتر حسین مطلبی کربکندی

نشانی: تهران، میدان آرژانتین، خیابان بیهقی، نبش خیابان ۱۶ غربی، پلاک ۲۰
تلفن: ۰۲۱-۸۶۱۵۱۰۶۱
کد پستی: ۱۵۱۵۶۷۴۳۱۱

فهرست

سخن نخست

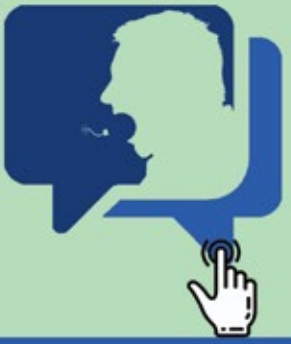
سخن نخست

فضای مجازی با شتاب شگرف و رو به تزایدی که در حال بسط و گسترش است تمام ساحات اجتماعی، اقتصادی، سیاسی و فرهنگی زندگی بشر را درنوردیده و هر روز بخش بزرگی از زندگی واقعی را در خود فرو برده و حیات متفاوت و جدیدی به آن می‌دهد. لذا به نظر می‌رسد دو نگاه کلان به فضای مجازی وجود دارد: نگاه اول که بالاخص در ابتدای رشد و تکوین فضای مجازی مسلط شده بود، آن را همچون ابزاری کنار سایر ابزارهای بشری تصویر می‌کرد که تنها طریقیت داشت. اما نگاه دوم، در نتیجه رشد تحولات خیره‌کننده فضای مجازی و سایه گستری آن در حوزه‌ها و شئون بشر در یک دهه اخیر آن را چون سکویی می‌داند که بسیار فراتر از شأن ابزاری حیات انسان‌ها را سامان جدیدی داده و ادعای تمدن نوینی را دارد. رویکردی که از قضا از چشمان بصیر رهبر انقلاب نیز دور نمانده و انتظاری تمدنی از فضای مجازی در ایران را مطالبه داشته‌اند.

در همین راستا گزارش‌های عصرفضای مجازی تلاش می‌کند تا فهم سازمان‌ها و دستگاه‌های مرتبط با حوزه فضای مجازی را ارتقاء بخشیده و آن‌ها را برای مواجهه فعال و خردمندانه با تحولات این عرصه مهیا سازد.

سید ابوالحسن فیروزآبادی
دبیر شورای عالی و رئیس مرکز ملی فضای مجازی

چکیده



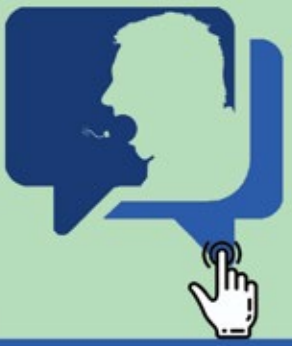
چکیده

مقالـه «گفتـار نفرت‌افکـن آنلاین» نوشـته الکسـاندرا ا. سـیگل، اسـتادیار علـوم سیاسـی دانشـگاه کلـورادو بولـدر اسـت کـه در سـال ۲۰۲۰ در کتـاب «رسـانه‌های اجتماعـی و مردم‌سـالاری» چـاپ انتشـارات دانشـگاه کمبریج به چاپ رسـیده اسـت. محـور این مقالـه گفتارهـای نفرت‌افکن آنلاین اسـت که در سـطح اینترنت و علی‌الخصوص شـبکه‌های اجتماعی توسـط افراد مسـتقل یـا گروه‌های سـازمان‌یافته منتشـر می‌شـود. گفتـار نفرت‌افکن آنلایـن افراد و گروه‌هـا را بـر اسـاس امـوری چون جنسـیت، نـژاد، مذهب و گرایش سیاسـی هـدف می‌گیـرد. ایـن سـنخ گفتـار ماننـد سـمی مهلـک بـرای همبسـتگی ملـی و نظام‌هـای مردم‌سـالار اسـت و موجـب تنش‌هـای اجتماعـی و تشـدید تضادهـای سیاسـی می‌شـود. از این روسـت کـه گفتـار نفرت‌افکـن آنلاین در میـان دانشـگاهیان و سیاسـت‌گذاران بـه یـک دغدغـه جـدی تبدیـل شـده اسـت. در ایـن مقالـه مـروری بـر آخریـن تکنیک‌های پیشـرفته کشـف گفتار نفرت‌افکـن آنلاین به بحث گذاشـته شـده اسـت. پس از آن بـه ادبیات موجود روی شناسـایی افـراد تولیدکننـده و هدف‌قرارگرفتـه گفتـار نفرت‌افکن آنلاین پرداخته شـده اسـت. سـپس میزان شـایع بـودن این نـوع ادبیات آسـیب‌زا در فضـای مجـازی مـورد تحلیـل قـرار گرفتـه اسـت و علی‌رغم رسـیدن بـه این

نتیجه که درصد این ادبیات بسیار کم است، با نشان‌دادن عواقب احتمالی این گفتار در فضای حقیقی که شامل آسیب‌های روانی، خشونت‌ها و حملات تروریستی است، بر جدی بودن این چالش تأکید شده است. در نهایت روش‌های مبارزه با گفتار نفرت‌افکن آنلاین که شامل ایجاد ممنوعیت‌ها توسط سکوها تا فراهم‌کردن ضدگفتار در برابر آن‌هاست ارائه شده است و میزان کارآمدی آن‌ها به بحث گذاشته شده است. باتوجه‌به اهمیت چالش گفتار نفرت‌افکن آنلاین در ایجاد آشوب و ازهم‌گسیختگی اجتماعی و ادبیات گسترده‌ای بررسی شده در این مقاله (بالغبر ۱۷۰ اثر به‌روز) و بررسی تمامی جوانب این پدیده، این مقاله می‌تواند اطلاعات ارزشمندی جهت مقابله با این آسیب فضای مجازی برای سیاستگذاران و فعالان این حوزه داشته باشد.

واژگان کلیدی: گفتار نفرت‌افکن آنلاین، کشف گفتار نفرت‌افکن، عواقب گفتار نفرت‌افکن

مقدمه



مقدمه

گفتار نفرت‌افکن آنلاین [1]

الکساندرا ا. سیگل [2]

گفتار نفرت‌افکن آنلاین که زمانی در کنج تاریک اینترنت جای داشت، به طرز روزافزونی در سکوهای رسانه‌های اجتماعی مطرح در حال مشاهده است. عواقب شوم گفتار نفرت‌افکن آنلاین خارج از فضای مجازی هر روز نمود بیشتری پیدا می‌کند؛ از حملات هدفمند یهودی‌ستیزانه [3] به خبرنگاران یهود تا گزارش‌های نقش رسانه‌های جمعی در دامن‌زدن به خشونت قومیتی در میانمار و سریلانکا. حکومت‌ها در سرتاسر جهان با ترس از این‌که این شیوه گفتار مضر در حال برانگیختن خشونت و ایجاد افراطی‌گرایی است، در حال تصویب مقررات و فشار آوردن به شرکت‌های رسانه‌های اجتماعی هستند تا سیاست‌هایی را برای توقف انتشار گفتار نفرت‌افکن آنلاین پیاده‌سازی کنند. [4]

بااین‌وجود، این فراخوان‌ها به در عمل به ندرت شواهد تجربی جامعی را پشت سر خود دارد. به علاوه، علی‌رغم توجه روزافزون به گفتار نفرت‌افکن آنلاین در ادبیات علمی، تعجب‌آور است که دانش کمی در مورد گستره، علل و عواقب انواع مختلف زبان مضر در سکوهای گوناگون در دسترس است. علاوه بر این، پژوهشگران تنها اخیراً شروع به بررسی کارآمدی رویکردهای مقابله با نفرت آنلاین کرده‌اند و فهم ما از هزینه‌های

1. Online Hate Speech
2. Alexandra A. Siegel
3. Anti-Semitic
4. Gagliardone et al. 2016

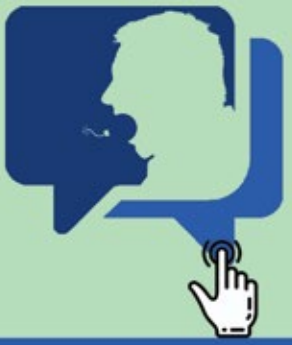
جانبی چنین مداخلاتی به طور خاص محدود است.

این فصل به بررسی ادبیات روز – شامل پژوهش‌های علمی، تحقیقات حقوقی و گزارش‌های سیاست‌گذارانه – در مورد گفتار نفرت‌افکن آنلاین می‌پردازد. به طور خاص، این فصل بحث‌های روز و محدودیت‌ها در رویکردهای کنونی به تعریف و کشف گفتار نفرت‌افکن آنلاین را بارز می‌کند؛ طرحی کلی از این‌که داده‌ها و نظرسنجی‌های رسانه‌های اجتماعی چه چیزی را می‌توانند در مورد تولیدکنندگان، هدف‌قرارگیرندگان[1] و شیوع کلی زبان مضر به ما بدهند، فراهم می‌کند؛ شواهد تجربی از عواقب گفتار نفرت‌افکن آنلاین خارج از فضای مجازی را مرور می‌کند؛ و بصیرت‌های کیفی در مورد این‌که مؤثرتری مداخلات در مبارزه با گفتار مضر آنلاین چیست ارائه می‌کند.

1. Targets

بخش اول

تعریف گفتار نفرت‌افکن آنلاین



بخش اول
تعریف گفتار نفرت‌افکن آنلاین

تعریف واحد مورد توافقی از گفتار نفرت‌افکن - چه آنلاین و چه غیر آنلاین- وجود ندارد و این موضوع به گرمی توسط دانشگاهیان، متخصصان حقوقی و سیاست‌گذاران در حال بحث است. رایج‌تر از همه، گفتار نفرت‌افکن این‌گونه فهمیده می‌شود؛ زبان سوگیرانه، تهاجمی و دارای سوءنیت که یک شخص یا گروه خاصی را به‌خاطر ویژگی‌های درونی[1] واقعی یا درک شده‌شان، هدف می‌گیرد.[2] بااین‌وجود، همان‌طور که سلارز (۲۰۱۶) استدلال می‌کند، «علی‌رغم ادبیات وسیع در مورد علل، مضرات و پاسخ‌ها به گفتار نفرت‌افکن، محققان کمی تلاش کرده‌اند تا این عبارت را به صورت نظام‌مند تعریف کنند.»

بسته به زمینه، تنوعی وسیعی از محتوا می‌تواند یا نمی‌تواند برای یک تعریف از گفتار نفرت‌افکن مناسب باشد.[3] برای مثال، در حالی که اهانت و توهین به سادگی به عنوان گفتار نفرت‌افکن قابل تشخیص هستند، زبانی که شامل القاب[4] است ضرورتاً توسط گوینده یا مخاطب، به عنوان گفتار نفرت‌افکن در نظر گرفته نمی‌شود.[5] در طرف مقابل، زبان لطیف‌تری که شخصی خارج از گروه را هدف می‌گیرد و ممکن است برای ناظران عادی به سختی به عنوان گفتار نفرت‌افکن تلقی شود، می‌تواند به صورت خاص

1. Innate
2. Cohen-Almagor 2011; Faris et al. 2016
3. Parekh et al. 2012; Sellars 2016
4. Epithet
5. Delgado 1982

صدماتی را به افراد و روابط گروهی وارد آورد.[1] این مسئله به طور خاص در فضای آنلاین صادق است چرا که گفتار به سرعت تکامل می‌یابد و می‌تواند بسیار تخصصی شود.[2] استفاده از واژگان رمزی به عنوان جایگزین اهانت‌های نژادی نیز در جوامع آنلاین رایج است که این امر تعریف گفتار نفرت‌افکن را پیچیده‌تر می‌کند.[3] برای مثال، در میان اعضاء آلت‌رایت،[4] خبرنگاران استفاده از واژه «گوگلز»[5] برای ارجاع به کاکاسیاه[6]؛ «اسکایپز»[7] به عنوان یک اهانت یهودی‌ستیزانه؛ «یاهوز»[8] به عنوان عبارتی تحقیرآمیز برای اسپانیایی‌تبارها و «اسکیتلز»[9] به عنوان یک عبارت مسلمان‌ستیزانه را ثبت کرده‌اند.[10] اجتماعات آلت‌رایت همچنین از استگانوگرافی،[11] مانند براکت‌های سه‌گانه استفاده کرده‌اند تا یهودیان را در فضای آنلاین شناسایی کنند و مورد اذیت قرار دهند.[12] به این صورت، در هنگام تعریف گفتار نفرت‌افکن – و به طور خاص گفتار نفرت‌افکن آنلاین – عبارت مشهور «وقتی آن را ببینم می‌شناسمش»، که برای محتوای مستهجن به کار می‌رود، کارایی ندارد.

در نتیجه، تعاریف موجود از گفتار نفرت‌افکن می‌توانند خیلی گسترده یا نسبتاً محدود باشند. در یک طرف طیف، تعاریفی وجود دارند که انواع گسترده‌ای از گفتار را که علیه یک فرد یا گروهی مشخص (یا به‌راحتی قابل تشخیص)، بر اساس ویژگی‌های دلبخواهی یا از نظر هنجاری بی‌ربط، نشانه می‌رود، به شمار می‌آورند.[13]

1. Parekh et al. 2012
2. Gagliardone et al. 2015
3. Duarte et al. 2018
4. Alt Right.

یک جنبش راست در ایالات متحده که سیاست رایج را رد می‌کند و به استفاده از فضای مجازی برای انتشار محتوای تحریک‌آمیز، علی‌الخصوص مقابله با برابری جنسیتی، مذهبی و نژادی مشهور است. م.

5. Googles

۶. واژه‌ای اهانت‌آمیز و نژادپرستانه برای ارجاع به سیاه‌پوستان

7. Skypes
8. Yahoos
9. Skittles
10. Sonnad 2016

۱۱. استگانوگرافی تکنیک مخفی کردن اطلاعات محرمانه در میان اطلاعات معمولی و غیرمحرمانه است.

12. Fleishman and Smith 2016
13. Parekh et al. 2012

در طرف دیگر طیف، تعاریفی هستند که قصد آسیب رساندن را لازم می‌دانند. محدودترین تعریف معتقد است که گفتار نفرت‌افکن بایستی «گفتار خطرناک» باشد – کلامی که مستقیماً به دنبال برانگیختن خشونت جمعی یا آسیب فیزیکی علیه فرد یا گروهی خارج از یک گروه دیگر است.[1] این تنش میان تعاریف نمایانگر دشواری ارائه تعریفی است که با کفایت به گستره پدیده‌هایی که ممکن است به عنوان گفتار نفرت‌افکن قلمداد شوند، بپردازد بدون این که تمایزهای ارزشمند را از دست بدهد. گفتار نفرت‌افکن آنلاین می‌تواند شامل تحریک‌کنندگان، جوامع هدف، انگیزه‌ها و تاکتیک‌های مختلف باشد. گاهی اوقات، مرتکب شوندگان این گفتار افرادی را که به ایشان حمله می‌کنند می‌شناسند، در حالی که در موارد دیگر ممکن است ایشان دنبال‌کنندگان آنلاین را تحریک کنند تا افراد خاصی را هدف بگیرند. گفتاری که خشونت را تحریک می‌کند با گفتاری که «صرفاً» توهین‌آمیز هستند تفاوت می‌کند و استفاده از زبان آسیب‌زا توسط یک حمله‌کننده واحد با پویش‌های نفرت‌افکن هماهنگ‌شده که توسط یک گروه دیجیتال انجام می‌شود فرق دارد.[2] کارهای اخیر به دنبال ارائه تعاریف جامع‌تر و شاکله‌هایی برای تشخیص گفتار نفرت‌افکن هستند که زمینه[3] و تبیینی برای تفاوت‌ها در شدت و قصد فراهم می‌کند.[4] ولی علی‌رغم این پیشرفت‌ها، هنوز وفاقی در ادبیات علمی روی این که چگونه بایستی گفتار نفرت‌افکن آنلاین را تعریف کرد وجود ندارد.

تعاریف حقوقی گفتار نفرت‌افکن به صورت مشابه مبهم است. حکومت‌ها در تلاش برای تنظیم‌گری مستقیم گفتار زیان‌آور - چه آنلاین و چه آفلاین- به طور روزافزونی گفتار نفرت‌افکن را در کدهای کیفری خود تعریف

1. Benesch 2013
2. Sellars 2016
3. Context
4. Waseem and Hovy 2016; Kennedy et al. 2018; Olteanu et al. 2018 Gagliardone et al. 2016;

می‌کنند.[1] مانند تعاریف دانشگاهی، این شامل طیفی از تعاریف نسبتاً گسترده، مانند تعریف کانادا از گفتار نفرت‌افکن به عنوان کلامی که «عامدانه نفرت را میان هر گروه قابل شناسایی اشاعه دهد»، تا تعاریف محدودتر می‌شود؛ مانند چهارچوب اتحادیه اروپا که گفتار نفرت‌افکن را این گونه تعریف می‌کند: «تحریک عمومی به خشونت یا نفرت علیه یک گروه از اشخاص یا عضوی از یک گروه که بر اساس نژاد، رنگ، تبار، دین یا باور، یا منشأ ملی یا قومیتی» باشد و «چشم‌پوشی، انکار یا بدیهی جلوه‌دادن فاحش نسل‌کشی، جنایات علیه بشریت، و جنایات جنگی (آن‌طور که در قوانین اتحادیه اروپا تعریف شده است) در سطح عمومی، وقتی که این عمل به نحوی انجام شود که احتمال برانگیختن خشونت یا نفرت علیه چنین گروه یا عضوی از چنین گروه را در پی داشته باشد».[2] در انگلستان، برانگیختن نفرت مذهبی یا نژادی، یک جرم کیفری است و نسخه‌های متفاوتی از این قانون- اگرچه خلاف قانون اساسی ایالات متحده است - در اکثریت دموکراسی‌های پیشرفته، مانند استرالیا، دانمارک، فرانسه، آلمان، هند، آفریقای جنوبی، سوئد و نیوزلند دیده می‌شود. همچنین در رژیم‌های اقتدارگرا، به طور خاص در جهان عرب نیز شاهد قوانینی هستیم که گفتار نفرت‌افکن آنلاین را در کنار قوانینی که با افراط‌گرایی مقابله می‌کنند، منع می‌کنند.[3] علی‌رغم وجود قوانینی که صراحتاً گفتار نفرت‌افکن را ممنوع می‌کند، این که این قوانین چگونه بایستی در عمل، علی‌الخصوص در عصر دیجیتال، به اجرا در آید هم اکنون محل بحث است.

اخیراً خود سکوهای آنلاین تعاریفی را از گفتار نفرت‌افکن به منظور تعدیل محتوای تولیدشده توسط کاربران ارائه کرده‌اند. برای مثال، بخش

1. Haraszti 2012
2. Sellars 2016
3. Chetty and Alathur 2018

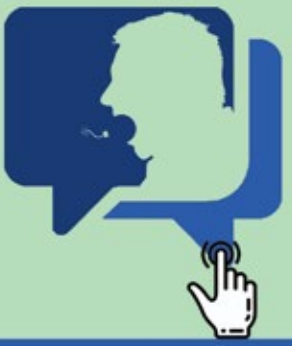
«محتوای نفرت‌افکن» راهنمای جامعه یوتیوب[1] می‌گوید «ما محتوایی را که خشونت علیه افراد یا گروه‌ها را بر اساس نژاد و منشأ قومی، دین، معلولیت، جنسیت، سن، ملیت، وضعیت خدمت نظامی،[2] یا گرایش جنسی/هویت جنسیتی، اشاعه دهد یا از آن چشم‌پوشی کند حمایت نمی‌کنیم و همچنین محتوایی را که هدف اصلی‌اش برانگیختن نفرت بر اساس این ویژگی‌های مرکزی باشد».[3] به طور مشابه، شرایط استفاده از خدمات[4] توییتر بیان می‌دارد که این شرکت «رفتار نفرت‌افکن» را که شامل «اشاعه خشونت یا حمله و تهدید مستقیم علیه دیگر انسان‌ها بر اساس نژاد، قومیت، منشأ ملی، گرایش جنسی، جنسیت، هویت جنسی، وابستگی دینی، سن، معلولیت یا بیماری» است، ممنوع می‌کند. این شرکت همچنین تأکید می‌کند که به حساب‌های کاربری‌ای که «هدف اصلی‌شان برانگیختن آسیب به دیگران بر اساس این ویژگی‌هاست» اجازه فعالیت نخواهد داد.[5] تعریف فیسبوک از گفتار نفرت‌افکن شامل عبارت تحریک به خشونت استفاده شده توسط توییتر و یوتیوب نیست، و به جای آن گفتار نفرت‌افکن را به عنوان «محتوایی که مستقیماً به مردم بر اساس نژاد؛ قومیت؛ منشأ ملی؛ وابستگی دینی؛ گرایش جنسی؛ جنس، جنسیت یا هویت جنسی؛ یا معلولیت‌ها یا بیماری‌های جدی، حمله می‌کند» تعریف می‌نماید.[6]

درمجموع، فقدان تعاریف سازگار و روشن از گفتار نفرت‌افکن در پژوهش‌های دانشگاهی، تحقیقات حقوقی و میان بازیگرانی که در تلاش برای حکمرانی فضای مجازی هستند به این معناست که علی‌رغم پژوهش‌های گسترده و مداخلات سیاست‌گذارانه‌ای که به خوبی ثبت شده‌اند، دانش ما از علل، عواقب و راه‌های مؤثر مبارزه با گفتار نفرت‌افکن آنلاین تا حدودی به خاطر ابهام تعریفی، غیرقطعی خواهد بود.

1. Youtube's Community Guideline
2. Veteran Status
3. YouTube 2018
4. Terms of Service
5. Twitter 2018
6. Facebook 2018

بخش دوم

کشف گفتار نفرت‌افکن آنلاین



بخش دوم
کشف گفتار نفرت‌افکن آنلاین

همان‌طور که وفاق روشنی روی تعریف گفتار نفرت‌افکن وجود ندارد، وفاقی هم در مورد مؤثرترین روش کشف آن در سکوهای مختلف در دست نیست. اکثر رویکردهای خودکار[1] به کشف گفتار نفرت‌افکن با یک طبقه‌بندی دوتایی شروع می‌شوند که در آن پژوهشگران به دنبال کدگذاری یک سند به عنوان «گفتار نفرت‌افکن یا خیر» هستند، هرچند رویکردهای چندکلاسه[2] نیز به کار گرفته شده‌اند.[3]

کشف خودکار گفتار نفرت‌افکن عموماً متکی بر پردازش زبان طبیعی یا استراتژی‌های متن‌کاوی[4] است.[5] آسان‌ترین این رویکردهای روش‌های لغت‌نامه-محور است که شامل توسعه فهرستی از واژگان است که درون یک متن جستجو و شمرده می‌شوند. رویکردهای لغت‌نامه-محور عموماً از واژگان محتوایی – شامل توهین و اهانت – استفاده می‌کنند تا گفتار نفرت‌افکن را تشخیص دهند.[6] این روش‌ها همچنین شامل طبیعی‌سازی[7] یا در نظر گرفتن کل واژگان در هر متن نیز هستند. با علم به این که گفتار نفرت‌افکن آنلاین می‌تواند واژگان اهانت‌آمیز را با اشتباه‌های املایی عمدی و سهوی مخفی کند، برخی پژوهشگران پیشنهاد استفاده از سنجه‌های

1. Automated
2. Multiclass
3. Davidson et al. 2017
4. Text mining
5. Facebook 2018
6. Dinakar et al. 2011; Dadvar et al. 2012; Liu and Forss 2015; Isbister et al. 2018
7. Normalizing

فاصله[1]، مانند کمترین تعداد ویرایش مورد نیاز برای تبدیل یک واژه به واژه‌ای دیگر را داده‌اند تا روش‌های لغت‌نامه‌محورشان را بهبود ببخشند.[2] به علاوه، با در نظر داشتن این‌که واژگان رمزی ممکن است برای اجتناب از کشف عبارت‌های نفرت‌افکن به کار رود، دیگر پژوهشگران واژگان رمزی شناخته شده ضد افراد خارج از گروه[3] را در لغت‌نامه‌هایشان گنجانده‌اند.[4]

فراتر از روش‌های صرفاً مبتنی بر لغت‌نامه، اکثر تکنیک‌های روزِ کشف گفتار نفرت‌افکن شامل وظایف طبقه‌بندی نظارت‌شده متن[5] هستند. این رویکردها، مانند استفاده از طبقه‌بندی‌کننده‌های[6] نائیو بیز،[7] ماشین‌های برداری حمایت خطی،[8] درخت‌های تصمیم‌گیری،[9] یا مدل‌های جنگل تصادفی،[10] غالباً متکی بر تکنیک‌های «کیسه واژگان»[11] و «n-گرام»[12] هستند. در روش کیسه واژگان، به جای این که از یک لغت‌نامه از پیش‌تعریف‌شده استفاده شود، مجموعه‌ای از واژگان[13] بر اساس واژگانی که در یک مجموعه داده آموزش دهنده پدیدار می‌شود، ایجاد می‌شود. سپس بسامدهای[14] واژگانی که در متن ظاهر می‌شوند، که به صورت دستی به عنوان «گفتار نفرت‌افکن یا خیر» حاشیه خورده‌اند، به عنوان ویژگی‌هایی برای آموزش یک طبقه‌بندی‌کننده استفاده می‌شوند.[15] برای اجتناب از طبقه‌بندی نادرست، زمانی که واژه‌ای در زمینه‌های دیگر استفاده شده یا به صورت نادرستی هجی شده باشد، برخی پژوهشگران از n-گرام‌ها استفاده می‌کنند که مشابه روش کیسه واژگان است و واژگان ترتیبی را به بایگرام‌ها، تریگرام‌ها یا فهرست‌های با طول «n» ترکیب می‌کند.[16] کارهای متأخرتر این رویکردها را برای بهبود دقت روش‌های لغت‌نامه-

1. Distance Metrics
2. Warner and Hirschberg 2012
3. Out-group
4. Magu et al. 2017
5. Supervised Text Classification Tasks
6. Classifiers
7. Naïve Bayes
8. Linear Support Vector Machines (SVM)
9. Decision Trees
10. Random Forest Models
11. Bag-of-Words
12. n-gram
13. Corpus
14. Frequency
15. Greevy and Smeaton 2004; Kwok and Wang 2013; Burnap and Williams 2016
16. Burnap and Williams 2016; Waseem and Hovy 2016; Badjatiya et al. 2017; Davidson et al. 2017

محـور- بـا حـذف مثبت‌های کاذب با شناسـایی این که چه توییت‌های شـامل اهانتـی بایسـتی در واقـع بـه عنـوان گفتـار نفرت‌افکـن شناسـایی شـوند - به کار گرفته‌انـد.[1] رویکردهـای قاعده-محور[2] و الگوهای گرامـری موضوع-محور،[3] کـه سـاختارهای عبـارت را وارد می‌کننـد نیز مورد اسـتفاده قـرار گرفته‌اند.[4]

دیگـر پژوهشـگران، گفتـار نفرت‌افکـن را بـا اسـتفاده از مدل‌سـازی موضـوع[5] تشـخیص می‌دهنـد که ایـن روش بـه دنبال شناسـایی پسـت‌هایی اسـت کـه متعلـق بـه یـک موضـوع تعریف‌شـده مثـل نـژاد و مذهـب اسـت.[6] دیگـران نیـز احساسـات را وارد تحلیلشـان کرده‌انـد، بـا ایـن فرض کـه گفتار نفرت‌افکـن احتمـالاً لحـن منفی دارد.[7] جایگـذاری واژه[8] یـا بازنمایی‌های بردار تکنیک‌هـای متـن[9] ماننـد doc2vec، paragraph2vec و FastText هـم مورد اسـتفاده قـرار گرفتـه انـد[10] و تکنیک‌هـای یادگیری عمیـق[11] که از شـبکه‌های عصبـی اسـتفاده می‌کنند هم بـرای طبقه‌بندی متـن و هم تحلیل احساسـات مرتبـط بـا کشـف گفتـار نفرت‌افکـن، رواج بیشـتری یافته‌اند.[12]

بـا شـناخت ایـن امـر کـه ایـن تکنیک‌هـا ممکـن اسـت بـرای شناسـایی صورت‌هـای ظریـف و غیرمسـتقیم نفرت‌افکنـی آنلایـن مناسـب نباشـند، محققـان همچنیـن رویکردهـای تئوری-محورتـری را بـه کار برده‌انـد. بـرای مثـال برناپ و ویلیامـز[13] (۲۰۱۶) و الشـریف، کولکارنی و همـکاران[14] (۲۰۱۸) مفهـوم دیگرسـازی[15] یـا زبانِ «مـا در مقابـل آن‌هـا»[16] را در اندازه‌گیری گفتار

1. Siegel et al. 2020
2. Rule-based approaches
3. Theme-based Grammatical Patterns
4. Fortuna and Nunes 2018
5. Topic-Modelling
6. Agarwal and Sureka 2017
7. Liu and Forss 2014; Gitari et al. 2015; Davidson et al. 2017; Del Vigna et al. 2017
8. Word Embedding
9. Vector Representations of Text Techniques
10. Djuric et al. 2015; Schmidt and Wiegand 2017; Siegel et al. 2020
11. Deep earning techniques
12. Yuan et al. 2016; Zhang et al. 2018, Al-Makhadmeh and Tolba 2020
13. Burnap and Williams
14. ElSherief, Kulkarni et al.
15. Othering
16. us vs. them

نفرت‌افکن وارد کرده‌اند. ایشان به این یافته رسیده‌اند که گفتار نفرت‌افکن غالباً از ضمایر سوم‌شخص استفاده می‌کند که شامل عبارت‌هایی چون «همه‌شان را به خانه بفرستید» می‌شود. مطالعات دیگر اظهارات برتری درون‌گروهی را – علاوه بر حملات معطوف به خارج از گروه – در اندازه‌گیری‌هایشان وارد کرده‌اند.[1] یک رویکرد دیگر شامل به‌حساب آوردن کلیشه‌های[2] ضد-خارج-از-گروه[3] می‌شود. برای مثال گفتار ضد اسپانیایی‌تباری ممکن است به عبور از مرز ارجاع دهد، یا زبان یهودی‌ستیزانه ممکن است به بانکداری، پول یا رسانه ارجاع داشته باشد.[4] کارهای بیشتر میان گفتار نفرت‌افکنِ معطوف به یک گروه (گفتار نفرت‌افکن عمومیت‌یافته) و گفتار نفرت‌افکن معطوف به افراد (گفتار نفرت‌افکن جهت‌دار) تمایز گذاشته‌اند تا ظرافت‌های مهمی را در اهداف گفتار نفرت‌افکن آنلاین فرا چنگ بیاورند.[5] علاوه بر اتکا روی ویژگی‌های متنی، پژوهشگران مشخصات کاربران را نیز دخیل کرده‌اند که شامل ویژگی‌های شبکه و شمارش‌های دوست/دنبال‌کننده به منظر بهبود دقت کشف گفتار نفرت‌افکن است.[6]

یک دسته متأخرتر از رویکردهای از مجموعه داده‌های از پیش طبقه‌بندی شده از سکوهای آنلاین برای کشف محتوای نفرت‌افکن آنلاین بهره می‌گیرند. این‌ها می‌توانند شامل تکنیک‌ها کیسه اجتماعات[7] باشند[8] که مشابهت یک پست را به زبان استفاده شده در ۹ جامعه نفرت‌افکن شناخته‌شده دیگر از Voat، Reddit، 4chan و MetaFilter محاسبه می‌کند. تکنیک‌های مشابه توسط سلیم و همکاران[9] (۲۰۱۷) و سیگل و

1. Warner and Hirschberg 2012
2. Stereotypes
3. Anti-out-group
4. Alorainy et al. 2018
5. ElSherief, Kulkarni et al. 2018
6. Unsvåg and Gambäck 2018
7. Bag-of-Communities
8. Chandrasekharan, Samory et al. 2017
9. Saleem et al.

همکاران[1] (۲۰۲۰) به کار گرفته شده‌اند که از داده‌های زیر صفحه‌های ردیت[2] که مشهور به مشهور به نفرت‌افکنی هستند، برای طبقه‌بندی گفتار نفرت‌افکن در توییتر استفاده می‌کنند. یک امتیاز این روش‌ها این است که وثاقت‌پذیری میان‌کُدری[3] اندکی که میان مجموعه داده‌ها قابل کشف است یا این امر که تکامل سریع الگوهای گفتار آنلاین می‌تواند استفاده از یک داده آموزش‌دهنده واحد در طول زمان را دشوار کند، مانعی برایشان نیستند.[4]

علی‌رغم این پیشرفت‌های مهم در کشف خودکار گفتار نفرت‌افکن آنلاین، روش‌های موجود عمدتاً در سکوهای متعدد یا انواع مختلف گفتار نفرت‌افکن آزمون نشده‌اند. به لطف راحتی جمع‌آوری داده‌ها، اکثر مطالعات موجود به داده‌های توییتر اتکا کرده‌اند. هرچند کارهای دیگری داده‌های ردیت،[5] یوتیوب، فیسبوک، ویسپر،[6] تامبلر،[7] مای‌اسپیس،[8] گب[9] و بخش‌های نظرات سایت‌ها و وبلاگ‌ها را وارد کرده‌اند، اما این موارد نسبتاً نادر بوده‌است.[10] به علاوه، اکثریت غالب مطالعات، محتوای زبان انگلیسی را بررسی می‌کنند، هرچند برخی پژوهشگران روش‌هایی را برای کشف گفتار نفرت‌افکن در دیگر زبان‌ها توسعه داده‌اند. این شامل بررسی‌های تجربی گفتار نفرت‌افکن در امهری،[11] عربی،[12] هلندی،[13] آلمانی،[14] هندی،[12] اندونزیایی،[16] ایتالیایی،[17] کره‌ای،[18] لهستانی،[19] رومانیایی[20] و اسپانیایی[21] می‌شود. لغت‌نامه‌های چندزبانه جمع‌سپاری شده[22] گفتار نفرت‌افکن آنلاین شامل

1. Siegel et al.
2. Subreddits
3. Intercoder Reliability
4. Waseem 2016
5. Reddit
6. Whisper
7. Thumblr
8. Myspace
9. Gab
10. Fortuna and Nunes 2018; Mathew, Dutt et al. 2019
۱۱. زبان رسمی اتیوپی. م. 2018 Mossie and Wang
12. Siegel 2015; De Smedt et al. 2018; Siegel et al. 2018, Albadi et al. 2019, Chowdhury et al. 2019
13. Van Hee et al. 2015
14. Ross et al. 2017
15. Santosh and Aravind 2019
16. Aulia and Budi 2019
17. Lingiardi et al. 2019
18. Kang et al. 2018
19. Czapla et al. 2019
20. Meza 2016
21. Basile et al. 2019
22. Crowd-sourced Multilingual Dictionaries

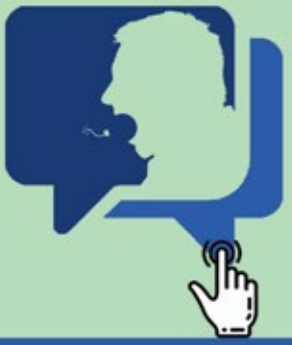
هیت‌بیس،[1] پایگاه داده اهانت نژادی[2] و هیت‌ترک[3] نیز توسعه داده شده‌اند که نشانگر راه‌های نویدبخش برای کارهای آینده است.[4] هنوز توسعه رویکردهای کشف گفتار نفرت‌افکن خودکاری که طراحی شده‌اند تا زبان‌های مختلفی را پوشش دهند دشوار است و کارهای بیشتری در این حوزه مورد نیاز است.

علاوه بر این، اکثر مطالعات گفتار نفرت‌افکن آنلاین به دنبال کشف همزمان همه انواع گفتار نفرت‌افکن یا «گفتار نفرت‌افکن عمومی»[5] هستند. با این وجود، آثار دیگر انواع مشخص زبان آسیب‌زا، شامل گفتار نفرت‌افکن جهادی،[6] گفتار نفرت‌افکن فرقه‌ای،[7] گفتار نفرت‌افکن مسلمان‌ستیز،[8] گفتار نفرت‌افکن سیاه‌پوست‌ستیز،[9] گفتار نفرت‌افکن زن‌ستیز،[10] و گفتار نفرت‌افکن مهاجرستیز[11] را مورد بررسی قرار داده‌اند. پژوهش‌های اخیر همچنین به کاوش روی تفاوت‌های میان انواع گفتار نفرت‌افکن، شامل مقایسه گفتار نفرت‌افکنی که خارج-از-گروه‌های[12] متنوعی را هدف قرار می‌دهند و تمییز میان انواع با شدت کم و زیاد گفتار نفرت‌افکن پرداخته‌اند.[13]

1. Hatebase
2. Racial Slur Database
3. HateTrack
4. ElSherief, Kulkarni et al. 2018, Siapera et al. 2018
5. Fortuna and Nunes 2018
6. Jihadist hate seech, De Smedt et al. 2018
7. Secratarian, Siegel 2015; Siegel et al. 2018
8. Olteanu et al. 2018
9. Kwok and Wang 2013
10. Mysogynistic, Citron 2011
11. Ross et al. 2017
12. out-groups
13. Beauchamp et al. 2018; Saha et al. 2019; Siegel et al. 2019

بخش سوم

تولیدکنندگان گفتار نفرت‌افکن آنلاین



بخش سوم
تولیدکنندگان گفتار نفرت افکن آنلاین

اگرچه پژوهش‌های گسترده‌ای به کاوش روی استفاده گروه‌های نفرت‌افکن سازمان‌یافته از گفتار نفرت‌افکن آنلاین پرداخته‌اند، اطلاعات کمتری در مورد بازیگران در اجتماعات غیررسمی مختص تولید محتوای آسیب‌زا، یا حساب‌هایی که گفتار نفرت‌افکن را در سکوهای اصلی تولید می‌کنند در دست است، به علاوه، هیچ پژوهش تجربی‌ای به صورت نظام‌مند به این مسئله نپرداخته است که این بازیگران چگونه درون و میان سکوها با هم تعامل می‌کنند.

گروه‌های نفرت‌افکن سازمان‌یافته خیلی زود پس از ابداع اینترنت، حضور آنلاین خود را تثبیت کردند[1] و در طول زمان تکثیر شده‌اند. بیش از یک دهه پژوهش عمدتاً کیفی نشان داده است که گروه‌های نفرت‌افکن سازمان‌یافته از اینترنت استفاده می‌کنند تا گفتار نفرت‌افکن را در سایت‌های رسمی‌شان منتشر کنند.[2] این شامل استفاده از انجمن‌های تعاملی[3] مانند صفحات چت و بازی‌های کامپیوتری می‌شود.[4] گروه‌های نفرت‌افکن از این کانال‌ها هم برای وسیع‌تر کردن گستره کارشان و هم برای هدف‌گیری مخاطبان خاص استفاده می‌کنند. برای مثال، بازی‌های ویدیویی صریحاً نژادپرستانه که از وبسایت‌های افراط‌گرایان

1. Bowman-Grieve 2009
2. Adams and Roscigno 2005; Chau and Xu 2007; Douglas 2007; Flores Yeffal et al. 2011; Castle 2012; Parenti 2013
3. Interactive Forums, Holtz and Wagner 2009
4. Selepak 2010

راست[1] نشئت می‌گیرند طراحی شده‌اند تا به حامیان تندوتیز و اعضاء بالقوه، بالاخص مخاطبان جوان، متوسل شوند.[2] در همین راستا، گروه‌های نفرت‌افکن از اینترنت برای جذب اعضاء جدید و تقویت هویت گروهی استفاده کرده‌اند.[3] سکوهای آنلاین همچنین به طور خاص برای سفارشی‌سازی پیام‌ها برای گروه‌ها یا افراد خاص مناسب هستند.[4] با فراهم کردن روش‌های کارآمد برای دسترسی به مخاطبان جدید و انتشار زبان نفرت‌افکن، اینترنت گروه‌های نفرت‌افکن را قادر می‌سازد تا به خوبی در قلمروی دیجیتال مطرح شوند، یک حس اجتماع را میان اعضایشان اشاعه می‌دهد و توجه خبرنگاران و شهروندان عادی را جلب می‌کند.[5] علاوه بر وب‌سایت‌های رسمی گروه‌های نفرت‌افکن سازمان‌یافته، تعداد سایت‌ها مختص تولید محتوای نفرت‌افکن که توسط گروه‌های غیررسمی و افراد هدایت می‌شوند نیز در طول زمان افزایش یافته است.[6] این شامل صفحات کانال‌ها یا اجتماعات نژادپرستی فاحش، زن‌ستیز، یا دیگر انواع تبعیض‌گرایی روی سکوهای رایج شبکه‌های اجتماعی مانند فیسبوک، توییتر، یوتیوب و همچنین انجمن‌هایی مانند ردیت، 4chan، 8chan، listserves، اجتماعات چت اینترنتی، انجمن‌های گفتگو و وبلاگ‌های طراحی‌شده برای انتشار گفتار نفرت‌افکن هستند.[7] این‌ها گستره‌ای را شامل پروفایل‌های فیسبوک جعلی طراحی شده برای تحریک خشونت علیه اقلیت‌ها[8] تا انجمن‌های ردیت بدنام (و هم اکنون ممنوع شده) مانند /CoonTown و /fatpeoplehate در بر می‌گیرند.[9] ملی‌گرایان سفیدپوست شناخته شده و حساب‌های کاربری نفرت‌افکن همچنین به

1. Far-right extremist
2. Selepak 2010
3. Chau and Xu 2007; Parenti 2013; Weaver 2013
4. Castle 2012
5. Bowman-Grieve 2009; McNamee et al. 2010
6. Potok 2015
7. Douglas 2007; Marwick 2017
8. Farkas and Neumayer 2017
9. Chandrasekharan, Pavalanathan et al. 2017

صورت آشکارا در سکوهای رسانه‌های اجتماعی رایج به فعالیت پرداخته‌اند. برای مثال، ریچارد اسپنسر[1] که تظاهرات آلت‌رایت شارلوتسویل[2] «راست را متحد کن» سازمان‌دهی کرد، بیش از ۷۵ هزار دنبال‌کننده داشت و تا نوامبر ۲۰۱۷ که رد صلاحیت شد، توسط توییتر تأییدشده بود. حساب‌های کاربری مانند @SageGang و @WhiteGenocide مکرراً توییت‌های مشتمل بر ادبیات خشونت‌بار نژادپرستانه و یهودی‌ستیزانه را منتشر می‌کنند.[3]

با این وجود، چنین فعالیت‌های متمرکزی حول محتوای نفرت‌افکن گاهی اوقات در سکوهای مشخص ممنوع و حذف می‌شوند. در نتیجه، این اجتماعات غالباً ناپدید می‌شوند و به صورت‌های جدید سر بر می‌آورند. برای مثال، در ۲۰۱۱، بنیان‌گذار 4chan، تابلو[4] خبری (n/) را به خاطر نظرات[5] نژادپرستانه حذف کرد و /pol/ را به عنوان یک انجمن جایگزین برای بحث سیاسی ایجاد نمود.

تابلوی /pol/ خیلی زود حتی با استانداردهای 4chan خانه‌ای برای گفتارهای به طورخاص نفرت‌افکن شد.[6]

به طور مشابه، زیرصفحات ردیت مانند Coontown به Voat نقل مکان کردند، سکویی که مقرراتی درمورد گفتار نفرت‌افکن نداشت.[7] هرچند داده‌های نظرسنجی و کارهای قوم‌نگاری[8] دلالت بر این دارند که کاربران 4chan و ردیت به طور عمده جوان، سفیدپوست و مذکر هستند،[9] با این وجود به خاطر ماهیت گمنام این سایت‌ها چیز زیادی در مورد کاربرانی که بیشترین گفتار نفرت‌افکن را تولید می‌کنند نمی‌دانیم. به

1. Richard Spencer
2. Charlottesville
3. Daniels 2017
4. Board
5. Comments
6. Hine et al. 2016
7. Chandrasekharan, Pavalanathan et al. 2017
8. Ethnographic
9. Daniels 2017; Costello and Hawdon 2018

طـور خـاص، ما میزانـی را که بیان ایشـان نمایانگر باورهای واقعی‌شـان اسـت یـا ایـن کـه صرفـاً رفتـاری جلب‌توجه‌طلبانـه یا سـوء رفتـار اینترنتی[1] اسـت، چیـزی که عمومـاً در ایـن اجتماعات رایـج اسـت،[2] نمی‌دانیم.

خـارج از صفحـات رسـمی و غیررسـمی و انجمن‌های مختـص به محتوای آسـیب‌زا، گفتار نفرت‌افکن همچنین در بحث‌های آنلاین عمومی در سـکوهای محبـوب، مانند فیسـبوک، یوتیوب، مای‌اسـپیس، تامبلر، ویسـپر، و ییک‌یاک[3] شـایع هسـتد.[4] هرچنـد دانـش کمـی در مـورد افـراد مشـخصی کـه گفتـار نفرت‌افکـن را در ایـن سـکوهای اصلـی تولیـد می‌کننـد وجـود دارد، کارهـای اخیـر شـروع بـه اندازه‌گیـری و توصیـف رفتـار آن‌هـا کـرده اسـت. بیچامپ و همـکاران[5] (۲۰۱۸) با بررسـی سـیر تولیدکنندگان گفتـار نفرت‌افکن در طول زمـان یافته‌انـد کـه تولیدکننـدگان گفتـار نفرت‌افکنِ نژادپرسـتانه و زن‌سـتیز در توییتـر عمومـاً در آغـاز از زبـان نفرت‌آلود غیرمسـتقیم و «نرم‌تر» اسـتفاده می‌کننـد و بعـداً بـه تولیـد نفـرت زهرآگین‌تـر می‌رسـند. ایـن نویسـندگان پیشـنهاد می‌کننـد کـه ایـن امـر می‌توانـد ناشـی از کاهـش تدریجـی شـرم[6] اجتماعـی باشـد، چـرا کـه ایـن کاربـران خودشـان را در شـبکه‌های اجتماعیِ بـه طور فزاینـده‌ای افراطی می‌یابند. الشـریف، نیلی‌زاده و همـکاران[7] (۲۰۱۸) در توییتـر یافته‌انـد کـه حسـاب‌های کاربـری‌ای کـه گفتـار نفرت‌افکـن را بـر می‌انگیزنـد عمومـاً جدیـد و بسـیار فعـال هسـتند و توجـه عاطفـی کمتـر و خشـونت و عدم تعادل بیشـتری[8] را در محتوای توییت‌هایشـان در مقایسـه با دیگر کاربران توییتر دارا هسـتند. به طریق مشـابه، با اسـتفاده از مجموعه‌داده حاشـیه‌خورده به صورت دسـتی[9] از نزدیک ۵۰۰۰ «کاربر نفرت‌افکن»، ریبایرو و همـکاران[10] (۲۰۱۸) یافتنـد که کاربـران نفرت‌افکن بیشـتر توئیت می‌کنند، افـراد بیشـتری را در هـر روز دنبـال می‌کننـد، و حسـاب‌ها کاربری‌شـان

1. Trolling Behaviour
2. Phillips 2015
3. Yik Yak
4. Black et al. 2016; Fortuna and Nunes 2018
5. Beauchamp et al.
6. Stigma
7. ElSherief, Nilizadeh et al.
8. Immoderation
9. Manually Annotated
10. Ribeiro et al.

تازه‌تر و کم‌عمرتر هستند. ایشان همچنین یافتند که هرچند کاربران نفرت‌افکن عموماً دنبال‌کنندگان کمتری دارند، ایشان به شدت به شبکه‌های بازتوییت متصل هستند. کاربران نفرت‌افکن ۷۱ برابر بیشتر محتمل است که کاربران نفرت‌افکن دیگر را بازتوییت کنند و کاربران تعلیق شده ۱۱ برابر محتمل‌تر است که دیگر کاربران تعلیق شده را در نسبت با کاربران غیرنفرت‌افکن بازتوییت کنند. متیو، دوت و همکاران[1] (۲۰۱۹) نیز با مقایسه کاربرانی که گفتار نفرت‌افکن تولید می‌کنند و آن‌های که چنین کاری نمی‌کنند به این نتیجه رسیدند که کاربران نفرت‌افکن به شدت به یکدیگر متصل هستند. ایشان در نتیجه استدلال می‌کنند که محتوای تولیدشده توسط کاربران نفرت‌افکن عموماً سریع‌تر و بیشتر انتشار می‌یابد و در مقایسه با محتوای تولیدشده توسط کاربرانی که گفتار نفرت‌افکن تولید نمی‌کنند، به گستره بیشتری از مخاطبان می‌رسد.

چنین رفتاری می‌تواند منجر به رؤیت‌پذیری کلی گفتار نفرت‌افکن در سکوهای آنلاین اصلی شود. برای مثال، در توییتر هرچند توییت‌های شامل گفتار نفرت‌افکن پاسخ‌ها و لایک‌های کمتری نسبت به توییت‌های غیرنفرت‌افکن دارند، با این وجود آنها شامل تعداد مشابهی از بازتوییت‌ها هستند.[2] ساختار به شدت شبکه‌ای کاربران نفرت‌افکنِ توییتر همچنین با شواهد کیفی‌ای سازگار است که دال بر این هستند که افراد در زیر صفحات ردیت با محتوای فاحش نفرت‌افکن یا اجتماعاتی مانند تابلو /pol/ در 4chan بسیج می‌شوند تا درگیر حملات هماهنگ شده نژادپرستانه یا جنسیت‌گرایانه در توییتر شوند.[3]

مگدی و همکاران[4] (۲۰۱۶) با مطالعه ساختار شبکه‌ای کاربرانی که گفتار نفرت‌افکن آنلاین تولید می‌کنند به این یافته رسیدند که می‌توانند

1. Mathew, Dutt et al.
2. Klubicka and Fernandez 2018
3. Daniels 2017
4. Magdy et al.

احتمال این که کاربران توییتر پیام‌های مسلمان‌ستیز را بعد از حملات ۲۰۱۵ پاریس توییت کنند با دقت بالا بر اساس شبکه‌های توییترشان پیش‌بینی کنند حتی اگر ایشان هیچ‌گاه کلمات مسلمانان یا اسلام را در توییت‌های قبلی‌شان ذکر نکرده باشند. کاربران توییت که مجاری رسانه‌ای محافظه‌کار را دنبال می‌کنند، نامزدهای اصلی جمهوری‌خواه، واعظان مسیحی انجیلی[1] و حساب‌های کاربری که مسائل مربوط به سیاست خارجی را بحث می‌کنند با احتمال بسیار بیشتری محتوای مسلمان‌ستیز را به دنبال حملات پاریس در نسبت با دیگران توییت می‌کردند.

در یکی از معدود نظرسنجی‌های موجود از کاربران رسانه‌های اجتماعی که به کاوشِ استفاده از گفتار نفرت‌افکن می‌پردازد، کاستلو و هاودون[2] (۲۰۱۸) به این یافته رسیدند که مردمی که زمان بیشتری را در ردیت و تامبلر صرف می‌کنند گفتار نفرت‌افکنی بیشتری را به صورت آنلاین منتشر می‌کنند. به علاوه، افرادی که نزدیک به یک اجتماع آنلاین هستند یا زمان بیشتری را در اجتماعاتی می‌گذرانند که گفتار نفرت‌افکن رایج است، بیشتر گرایش دارند تا محتوای نفرت‌افکن تولید کنند. ولی برخلاف انتظاراتشان این نویسندگان یافتند که صَرف زمان بیشتر آنلاین به طور کلی ارتباطی با تولید محتوای نفرت‌افکن ندارد و ارتباطی میان استفاده از بازی‌های تیراندازی اول‌شخص و تولید محتوای آنلاین نفرت‌افکن نیست.

مشابه صفحات و انجمن‌هایی که به طور صریح به گفتار نفرت‌افکن آنلاین اختصاص دارند، افراد تولیدکننده گفتار نفرت‌افکن آنلاین به طور فزاینده‌ای توسط توییتر و دیگر سکوهای اصلی طرد شده‌اند. هرچند بسیاری از این کاربران به سادگی بعد از تعلیقشان، حساب‌های کاربری جدیدی ایجاد می‌کنند، مابقی به سکوهای تخصصی‌تر نقل‌مکان می‌کنند

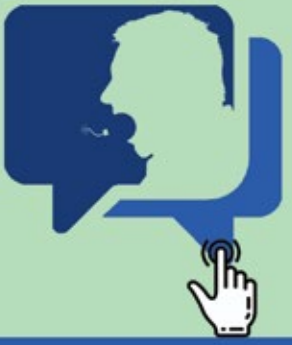
1. Evangelical Christian Preachers
2. Costello and Hawdon

کـه در آن‌هـا می‌تواننـد بـا آزادی بیشـتر محتـوای نفرت‌افکـن تولیـد کننـد. بـرای مثـال در اوت ۲۰۱۶، شـبکه اجتماعـی گـب بـه عنـوان جایگزینی برای توییتـر ایجـاد شـد. این سـکو بیـان می‌دارد کـه «اولاً به مـردم و گفتـار آزاد» اختصـاص دارد و بـا ایـن کار بـرای کاربـران طردشـده یـا تعلیـق شـده دیگـر شـبکه‌های اجتماعـی جذابیـت ایجـاد می‌کند.[1] مارویـک و لوئیـس[2] (۲۰۱۷) بـه ایـن نتیجـه رسـیدند کـه گـب عمدتـاً بـرای انتشـار و بحـث روی اخبار و رخدادهای جهان اسـتفاده می‌شـود و غالبـاً کاربران آلت-رایـت، نظریه‌پردازان توطئـه و ترول‌هـا را جـذب می‌کنـد. نویسـندگان یافتند که گفتـار نفرت‌افکن بـه مراتـب در گـب نسـبت بـه توییتـر رایج‌تـر اسـت ولـی در گـب نسـبت به تابلـو /pol/ در 4chan رواج کمتـری دارد. بـه طـور مشـابه، لیمـا و همکاران[3] (۲۰۱۸) بـه ایـن نتیجـه دسـت یافتند کـه گب به طـور کلی میزبـان کاربران طردشـده از دیگر شـبکه‌های اجتماعی اسـت که بسـیاری از ایشـان به خاطر اسـتفاده از گفتـار نفرت‌افکـن و افراط‌گرایانـه طـرد شـده‌اند.

1. Marwick and Lewis 2017
2. Marwick and Lewis
3. Lima et al.

بخش چهارم

گروه‌های هدف گفتار نفرت‌افکن آنلاین



بخش چهارم
گروه‌های هدف گفتار نفرت افکن آنلاین

یکی از معدود حوزه‌های مورد وفاق در تعریف گفتار نفرت‌افکن، که آن را از دیگر انواع گفتار آسیب‌زا متمایز می‌کند این است که گفتار نفرت‌افکن گروه‌ها یا افرادی را که مرتبط با یک گروه هستند هدف می‌گیرد.[1] بخش کوچکی از ادبیات به روشنی هدف‌قرارگرفتگان گفتار نفرت‌افکن آنلاین را تحلیل کرده است. سیلوا و همکاران[2] (۲۰۱۶) با مطالعه هدف‌قرارگرفتگان گفتار نفرت‌افکن آنلاین در ویسپر (یک سکو آنلاین گمنام) و توییتر با استفاده از الگوریتم عبارت-ساختار-محور دریافتند که افراد هدف‌گیری شده در هر دو سکو عمدتاً بر اساس قومیت، ویژگی‌های فیزیکی، گرایش‌های جنسی، طبقه یا جنسیت مورد حمله قرار گرفته‌اند. پژوهش‌های مبتنی بر نظرسنجی دلالت بر این دارند که قربانیان گفتار نفرت‌افکن آنلاین عموماً سطح بالایی از فعالیت آنلاین را دارند،[3] گمنانی آنلاین کمتری دارند و وارد مخالفت‌های بیشتری در فضای آنلاین می‌شوند.[4] با بررسی این گروه‌های هدف گفتار نفرت‌افکن آنلاین در توییتر، الشریف، نیلی‌زاده و دیگران (۲۰۱۸) یافتند که آن‌هایی که هدف حمله گفتار نفرت‌افکن قرار می‌گیرند ۶۰٪ نسبت به تحریک‌کنندگان و ۴۰٪ نسبت به سایر کاربران محتمل‌تر است که احراز کاربری شده باشند.

1. Sellars 2016
2. Silva et al.
3. Hawdon et al. 2014
4. Costello, Rukus, and Hawdon 2018

ایـن نشـان می‌دهـد کـه کاربـران مشـاهده‌پذیرترِ توییتـر (بـا دنبال‌کنندگان، بازتوییت‌هـا و فهرسـت‌های بیشـتر) بـا احتمـال بیشـتری هـدف نفرت‌افکنی قـرار می‌گیرند.

در همیـن راسـتا، پژوهش‌هـای کیفـی اخیـر پیشـنهاد می‌کننـد کـه خبرنـگاران، سیاسـتمداران، هنرمنـدان، بلاگرهـا و دیگر اشـخاص عمومی[1] به طـور نامتناسـبی هدف گفتـار نفرت‌افکن قـرار گرفته انـد.[2] برای مثـال، وقتی که نسـخه بازسازی‌شـده کاملاً زنانه فیلم روح‌شـکن[3] در جولای ۲۰۱۶ عرضه شـد، سـفیدبرترطلبی[4] بـه نـام میلـو ییانوپولـوس،[5] طوفـان توییتـری‌ای را به دنبال انتشـار یـک نقد فیلم منفـی در برایتبارت[6] برانگیخت. سـفیدبرترطلبان شـروع بـه حمله بـه خـط زمانـی[7] بازیگـر زن آمریکایی-آفریقایی تبار، لسـلی جونـز[8] بـا اهانت‌هـای جنسـیت‌زده[9] و نـژادی و رفتارهـای نفرت‌افکن، شـامل تجـاوز جنسـی و تهدیـد به مـرگ، کردند. وقتی که این سـوء اسـتفاده با اقدام یـیانوپولـوس بـه توییت مسـتقیم در مورد جونـز و گیردادن بـه دنبال‌کنندگان او تشـدید شـد، جونـز از توییتـر بیـرون رفـت. بعـد از ایـن کـه فشـار عمومی کمپانـی را متقاعـد بـه دخالـت کـرد، یـیانوپولـوس از توییتر طرد شـد و جونز بازگشـت.[10] بـه طور مشـابه، همـان طور که میکـی کندال[11] نویسـنده توصیف می‌کنـد: «مـن قصـد داشـتم زمانـی توییتـر را ترک کنـم. این برای مـن قابل اسـتفاده نبود. من وارد می شـدم و ۲۵۰۰ نظر منفی داشـتم. شـخصی که به نظـر انرژی تمام نشـدنی داشـت تصویر مـن را روی تصاویر افـراد در حال زجر کشـیدن فوتوشـاپ می‌کـرد و چیزهایـی مثل این که من باید توسـط سـگ‌ها مـورد تجـاوز قـرار بگیـرم می‌گفـت». اطلاعات کنـدال نیـز در فضـای آنلاین عیـان شـده بـود - آدرس او بـه صـورت آنلاین در دسـترس قرار داشـت - و او

1. Public Figures
2. Isbister et al. 2018
3. Ghostbuster
4. White Supremacist
5. Milo Yiannopoulos
6. Breitbart
7. Timeline
8. Leslie Jones
9. Sexist
10. Isaac 2016
11. Mikki Kendall

تصویری از خودش و خانواده‌اش دریافت کرد که «به نظر از طریق دوربین یک تک‌تیرانداز گرفت شده بود».[1]

در ژوئن ۲۰۱۶، تعدادی از خبرنگاران یهودی کاملاً آشکار گزارش رگباری از نفرت‌افکنی آنلاین را که شامل استگانوگرافی – پرانتزهای سه گانه که در طرفین اسامی‌شان قرار داشت مانند (((این))) – بود دادند.[2] در نتیجه، اتحادیه ضد بی‌آبروسازی[3] پرانتزهای سه گانه را به پایگاه داده سمبل‌های نفرت‌افکنشان افزود. این «معادل دیجیتال یک ستاره زرد» به منظور شناسایی یهودیان به عنوان اهداف اذیت و آزار آنلاین بود.[4] برای مثال، جاناتان وایزمن[5] از نیویورک تایمز توییتر را بعد از قرار گرفتن در معرض اذیت و آزار یهودی‌ستیزانه، که با یک حساب کاربری توییر با عنوان @CyberTrump آغاز شد و با رگبار فعالیت توییتری نفرت‌افکن، نامه‌های صوتی[6] و رایانامه‌های شامل اهانت و تصاویر خشونت‌بار، تشدید شد، ترک کرد.[7]

همان‌طور که این مثال‌ها نشان می‌دهند، گفتار نفرت‌افکن آنلاین می‌تواند در حملات هماهنگ شده‌ای که این رفتار را کشف می‌کند بیشترین مشاهده‌پذیری را داشته باشد.[8] چنین حملاتی توجهاتی زیادی را هم به صورت آنلاین و هم از طریق مجاری رسانه‌ای سنتی جذب می‌کند، و این اهداف راهبردی را هم برای افراط‌گرایان و هم برای جستجوگران ترول‌ها[9] برای رسیدن به مخاطبان بیشتر و بالابردن پیام‌هایشان، مفید می‌کند. چنین پویش‌های آزار و اذیت هماهنگ شده‌ای، به گروه‌های افراد گمنام این اجازه را می‌دهد تا با یکدیگر همکاری نمایند تا کاربران مشخصی را با محتوای نفرت‌افکن به صورت مستمر بمباران کنند.[10]

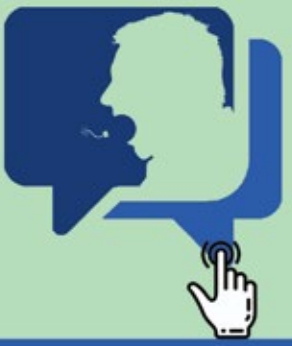
1. Isaac 2016
2. Fleishman and Smith 2016
3. Anti-Defamation League (ADL)
4. Gross 2017
5. Weisman
6. voicemails
7. Gross 2017
8. Mariconti et al. 2018
9. Trolls Seeking
10. Chess and Shaw 2015; Chatzakou et al. 2017

یکی از بروزات این رفتار به عنوان یورش[1] شناخته می‌شود، وقتی دار و دسته‌های دیجیتال تک‌کاره[2] حملاتی را به منظور اختلال دیگر سکوها و تضعیف کاربرانی که حامی مسائل و سیاست‌هایی هستند که ایشان با آن مخالف‌اند، سامان و ترتیب می‌دهند.[3] ولی هرچند یورش توجه رسانه‌ای زیادی را به خود جلب می‌کند، ما فهم کمی از حجم فراگیر بودن و رایج‌بودن این حملات یا سکوهایی که بیشتر در آن‌ها رخ می‌دهند داریم.

1. Raiding
2. Ad-hoc Digital Mobs
3. Hine et al. 2016; Kumar et al. 2018; Mariconti et al. 2018

بخش پنجم

شـــایع بودن گفتار نفرت افکن آنلاین



بخش پنجم
شـــایع بودن گفتـــار نفرت افکـــن آنلایـــن

در حالـی کـه بخـش عمـده پژوهش‌هـا مختـص تعریـف و کشـف گفتـار نفرت‌افکـن آنلایـن بـوده اسـت، مـا بـه طـرز عجیبـی دانـش کمـی در مـورد محبوبیـت گفتـار نفرت‌افکـن چـه در سـکوهای اصلـی و چـه حاشـیه‌ای، و یـا در مـورد ایـن کـه چگونـه حجـم گفتـار نفرت‌افکـن در واکنـش بـه رخدادهـای موجـود جابه‌جـا می‌شـود، داریم. سـکوهای رسـانه‌های اجتماعی مشـاهده‌پذیری گفتـار نفرت‌افکـن آنلایـن را افزایـش داده‌انـد کـه خبرنگاران و دانشـگاهیان را واداشـته اسـت تـا بگویند گفتـار نفرت‌افکن در حـال افزایش اسـت. در نتیجه، گرایشـی وجود دارد که کل سـکوهای رسـانه‌های اجتماعی اصلـی را بـه عنـوان سـنگرهای نفرت‌افکنـی آنلایـن توصیف کنند بـدون این کـه از شـواهد تجربـی بـرای ارزیابی این کـه این پدیـده واقعاً چه قـدر فراگیر اسـت اسـتفاده کنند. بـرای مثـال، سـردبیر Atlantic، جفری گلدبـرگ[1] بعد از ایـن کـه هـدف یـک حملـه نفرت‌آلـود آنلاین قـرار گرفـت، توییتـر را «چاه فاضـلاب یهودسـتیزان، ضدهمجنس‌گرایـان و نژادپرسـتان» نامیـد.[2] در حالی کـه هـر گفتـار نفرت‌افکـن آنلاینی یک مشـکل اسـت، گفتن این‌که سـکویی کـه توسـط بیـش از یک‌چهـارم آمریکایی‌هـا و میلیون‌هـا نفـر در سرتاسـر جهان اسـتفاده می‌شـود تحت سـیطره چنین گفتـاری قـرار دارد گمراه‌کننده

1. Jeffrey Goldberg
2. Lizza 2016

و بـه طـور بالقـوه‌ای مشـکل آفریـن اسـت؛ به طـور خـاص در کشـورهایی که آزادی سیاسـی و مدنـی هم‌اکنـون مـورد تهدید اسـت و رسـانه‌های اجتماعی مجـرای ارزشـمندی برای صداهای اپوزوسـیون اسـت.[1]

بـا توجـه بـه شـواهد تجربـی، تعـداد معـدودی مطالعـه شـروع بـه ارزیابی نظام‌منـد فراگیـریِ گفتـار نفرت‌افکن در سـکوهای اجتماعی کردنـد، هرچند کارهـای بیشـتری مـورد نیاز اسـت. با تحلیـل محبوبیت گفتـار نفرت‌افکن در بیـش از ۷۵۰ میلیـون توییـت سیاسـی و در ۴۰۰ میلیـون توییتی که توسـط نمونـه تصادفـی کاربـران توییتر آمریکایی بیـن ژوئن ۲۰۱۵ و ژوئـن ۲۰۱۷ به دسـت آمـده اسـت، سـیگل و همـکاران[2] (۲۰۲۰) بـه ایـن یافته رسـیدند که حتـی در فراوان‌تریـن روزهـا، تنهـا کمتر از یک درصـد از توییت‌هـا در فضای توییتـر آمریکایی شـامل گفتـار نفرت‌افکن هسـتند. به طور مشـابه، با مطالعه روی محبوبیـت گفتـار نفرت‌افکن در صفحات اتیوپیایی فیسـبوک، گگلیاردون و همـکاران[3] (۲۰۱۶) بـه ایـن نتیجـه رسـیدند کـه ۰.۴ درصـد عبارت‌هـا در نمونـه معـرف[4] ایشـان به عنـوان گفتـار نفرت‌افکن طبقه بندی شـده اسـت و ۰.۳ درصـد از توییت‌هـا به عنوان گفتار خطرناک که مسـتقیماً یا غیرمسـتقیم بـه خشـونت علیـه یـک گـروه خـاص فـرا می‌خوانـد، طبقه‌بنـدی شـده بـود.

هرچند این مطالعات گویای این هسـتند که گفتار نفرت‌افکن آنلاین یک پدیـده نسـبتاً نـادر اسـت، پژوهش‌های نظرسـنجی بین ملیتی[5] دلالـت بر این دارنـد کـه با این وجود، تعداد زیادی از افـراد به صورت اتفاقی در معرض گفتار نفرت‌افکـن آنلایـن قرار گرفته‌اند. در یک نظرسـنجی بین - ملیتـی از کاربران اینترنـت بیـن سـنین ۱۵ و ۳۰ سـال، ۵۳٪ پاسـخ‌دهندگان آمریکایی گزارش دادنـد کـه در معرض محتـوای آنلاین نفرت‌افکـن قرار گرفته‌انـد، در حالی که ۴۸٪ فنلاندی‌هـای، ۳۹٪ بریتانیایی‌هـا و ۳۱٪ آلمان‌هـا گـزارش قرارگیـری

1. Gagliardone et al. 2016
2. Siegel et al.
3. Gagliardone et al.
4. Representative Sample
5. Cross-National Survey Research

در معرض چنین محتوایی را دادند. استفاده زیاد از شبکه‌های اجتماعی آنلاین و بازدید از «سایت‌های خطرناک» دو تا از قوی‌ترین پیش‌بینی‌کننده‌های[1] قرارگرفتن در معرض این محتواهاست.[2] شاید برای توضیح شکاف میان یافته‌های تجربی که گفتار نفرت‌افکن در سکوهای اصلی نسبتاً نادر است و نرخ بالای در معرض قرار گرفتن بر اساس خود گزارشی، کاکینن و همکاران[3] (۲۰۱۸) یافتند که در حالی که محتوای نفرت‌افکن به ندرت تولید می‌شود، این محتوا نسبت به انواع دیگر محتوا مشاهده‌پذیری بالاتری دارد. گفتار نفرت‌افکن همچنین در برخی اجتماعات جمعیتی[4] نسبت به دیگران رایج‌تر است. برای مثال ساها و همکاران[5] (۲۰۱۹) یافتند که گفتار نفرت‌افکن در زیرصفحه‌های ردیتی که مرتبط با کالج‌ها و دانشگاه‌های خاصی هستند نسبت به زیرصفحات ردیتی که مرتبط با کالج‌ها و دانشگاه‌ها نیستند شایع‌تر هستند.

علاوه بر کاوش روی فراگیری گفتار نفرت‌افکن آنلاین، کارهای اخیر روی این که چگونه رخدادهای آفلاین می‌توانند منجر به افزایش محبوبیت چنین گفتاری شوند، تحقیق کرده‌اند. یک مسیر پژوهش به کاوش روی اثر رخدادهای آفلاین خشن روی انواع مختلف گفتار نفرت‌افکن می‌پردازد. برای مثال، با مطالعه روی اثر علی حملات تروریستی در کشورهای غربی روی استفاده از زبان نفرت‌افکن در ردیت و توییتر، اولتینو و همکاران[6] (۲۰۱۸) به این نتیجه رسیدند که رخدادهای خشونت افراط‌گرایانه منجر به افزایش گفتار نفرت‌افکن آنلاین، بالأخص پیام‌های مرتبط با حمایت از خشونت، در هر دو سکو می‌شود. نویسندگان استدلال می کنند که این شاهدی فراهم می‌کند که متأسفانه استدلال‌های نظری در مورد حلقه بازخورد میان خشونت آفلاین و گفتار نفرت‌افکن آنلاین درست هستند.

1. Predictors
2. Hawdon et al. 2017
3. Kaakinen et al.
4. Demographic Communities
5. Saha et al.
6. Olteanu et al.

این یافته پژوهش‌های دیگری را هم حمایت می‌کند که دلالت بر این دارند که گفتار نفرت‌افکن و جرایم نفرت‌افکن عموماً بعد از رخدادهای «رها کننده»[1] که می‌توانند محلی، ملی یا بین‌المللی باشند افزایش می‌یابند و غالباً احساسات منفی را به سمت گروه‌هایی که مرتبط با مظنونان به ارتکاب خشونت هستند می‌کشند.[2]

به صورت مشابه، سیگل و همکاران (۲۰۱۸) به منظور ارزیابی اثر رخدادهای خشونت روی محبوبیت گفتار نفرت‌افکن ضدشیعی در فضای توییتری عربستان، یافته‌اند که هم رخدادهای خشونت‌آمیز خارجی و هم حملات تروریستی داخلی روی مساجد شیعه، افزایش چشم‌گیری را در زبان ضد-شیعی در فضای توییتری عربستان سعودی ایجاد می‌کند. به منظور فراهم آوردن بصیرت‌های بیشتر در مورد مکانیزم‌هایی که از طریق آن‌ها رخدادهای خشونت‌آمیز آفلاین منجر به افزایش استفاده از گفتار نفرت‌افکن آنلاین می‌شود، نویسندگان نشان می‌دهند که هرچند روحانیون و دیگر بازیگران نخبه گفتار تحقیرآمیز را بعد از رخدادهای خارجی خشونت فرقه‌ای برمی انگیزانند و انتشار می‌دهند – که بیشترین افزایش را در زبان ضدشیعی ایجاد می‌کند - ایشان با احتمال کمتری این کار را به دنبال بمب‌گذاری‌های مساجد داخلی انجام می‌دهد.

سیگل و همکاران (۲۰۲۰) برای کاوش روی اثر رخدادهای سیاسی و نه خشونت‌آمیز روی محبوبیت گفتار نفرت‌افکن آنلاین، یافتند که برخلاف روایت ژورنالیستی محبوب، گفتار نفرت‌افکن آنلاین نه در طول پویش ۲۰۱۶ دونالد ترامپ و نه پس از پیروزی غیرمنتظره او افزایشی نیافت. با استفاده از مجموعه داده‌ی بیش از یک میلیارد توییت، نتایج ایشان مستحکم است؛ چه زمانی که گفتار نفرت‌افکن را با رویکرد لغت‌نامه محور

1. Trigger Events
2. Awan and Zempi 2015

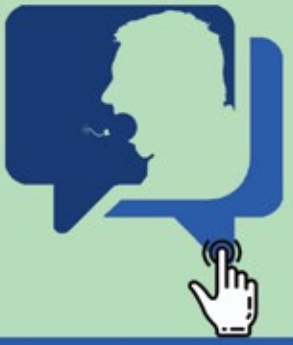
تقویت شده با یادگیری ماشین کشف می‌کنند، چه زمانی که از یک الگوریتم کشف اجتماع-محور که شباهت روزانه داده‌های توییتر را با محتوای تولیدشده در زیر صفحات نفرت‌افکن ردیت در طول زمان مقایسه می‌کند، استفاده می‌نمایند. به جای آن، گفتار نفرت‌افکن «مقطعی»[1] بود؛ جهش‌هایی که به دنبال رخدادهای مشخص زده می شوند و بعد از آن به سرعت به حالت تعادل باز می‌گردند. به طور مشابه، فاریس و همکاران[2] (۲۰۱۶) نشان می‌دهند که جهش‌ها در گفتار نفرت‌افکن آنلاین عموماً مرتبط با رخدادهای سیاسی هستند در حالی که سلیم و همکاران (۲۰۱۷) یافتند که گفتار نفرت‌افکن به دنبال رخدادهایی که واکنش‌های عاطفی شدید را تحریک می‌کند، مثال تظاهرات بالتیمور[3] یا تصمیم دیوان عالی آمریکا در مورد ازدواج هم‌جنسان، افزایش می‌یابد.

در مجموع، این مطالعات گویای اهمیت بررسی فراگیری و پویایی گفتار نفرت‌افکن آنلاین به‌صورت نظام‌مند در طول زمان و با استفاده از نمونه‌های معرف بزرگ است. کارهای بیشتری برای فهم بهتر این‌که چگونه انواع گفتار نفرت‌افکن آنلاین در زمینه‌های متنوع جهانی به جنبش در می‌آیند و چگونه محبوبیت نسبی آن‌ها در سکوهای رسانه‌ها اجتماعی اصلی و تخصصی در طول زمان تغییر می‌کند، نیاز است.

1. Bursty
2. Faris et al.
3. Baltimore

بخش ششم

عواقب آفلاین گفتار نفرت افکن آنلاین



بخش ششم
عواقب آفلاین گفتار نفرت‌افکن آنلاین

اندازه‌گیری نظام‌مند اثر گفتار نفرت‌افکن آنلاین چالش‌برانگیز است[1] ولی حجم متنوعی از پژوهش‌ها دلالت بر این دارند که گفتار نفرت‌افکن آنلاین عواقب آفلاین جدی برای افراد و گروه‌ها دارد. نظرسنجی‌ها از کاربران اینترنت نشان می‌دهد که قرارگیری در معرض گفتار نفرت‌افکن آنلاین می‌تواند منجر به ترس شود،[2] بالأخص در جمعیت‌هایی که از نظر تاریخی به حاشیه رانده شده‌اند یا محروم هستند. کارهای دیگر پیشنهاد می‌کنند که این چنین در معرض قرار گرفتن می‌تواند مردم را به انزوا از بحث‌های عمومی آنلاین و آفلاین سوق دهد و بنابراین به مشارکت مدنی و گفتار آزاد آسیب وارد کند.[3] در واقع داده‌های تجربی نشان می‌دهند که قرارگیری در معرض گفتار نفرت‌افکن می‌تواند بسیاری از عواقبی را که هدف‌قرارگرفتن توسط جرایم نفرت‌افکن به دنبال دارد، داشته باشد که شامل آسیب‌های روان‌شناختی و ترس‌های اشتراکی[4] است.[5] در این راستا، گروه‌های حقوق بشر استدلال کرده‌اند که عدم موفقیت در پایش و شمارش گفتار نفرت‌افکن در فضای آنلاین می‌تواند انقیاد اقلیت‌های هدف‌قرارگرفته را تقویت کند و ایشان را نسبت به حملات آسیب‌پذیرتر نماید در حالی که جمعیت‌های اکثریت را نسبت به چنین تنفری بی‌تفاوت‌تر

1. Sellars 2016
2. Hinduia and Patchin 2007
3. Henson et al. 2013
4. Communal Fear
5. Gerstenfeld 2017

کنـد.[1] بـا ایـن وجـود، کارهـای اخیـر نشـان می‌دهـد کـه تفسـیرها از گفتـار نفرت‌افکـن – ایـن کـه چه چیزی محتـوای نفرت‌افکن در نظر گرفته می‌شـود و همچنیـن رتبه‌بندی‌هـا از شـدت و محتـوا- در کشـورها بـه طـور گسـترده تفـاوت می‌کنـد[2] و مردان و محافظه‌کاران سیاسـی عموماً محتـوای نفرت‌افکن را نسـبت بـه زنـان، معتدلین سیاسـی و لیبرال‌ها کمتـر آزاردهنـده می‌یابند.[3]

در سـطح فـردی، پژوهش‌هـای کیفـی گویـای آن هسـتند که مسـلمانان سـاکن غـرب کـه هـدف گفتـار نفرت‌افکـن آنلایـن قـرار گرفته‌انـد تـرس از ایـن دارنـد کـه تهدیدهـای آنلایـن در فضای حقیقی محقق شـود.[4] بـه علاوه، نظرسـنجی‌ها از کاربـران جـوان اینترنـت بـه ایـن یافتـه رسـیده اسـت کـه تعـداد زیـادی از پاسـخ‌دهندگان آمریکایی-آفریقایـی، تبعیـض شـخصی یـا فـردی را در فضـای آنلایـن تجربـه کرده‌انـد و قرارگیـری در معـرض چنیـن تبعیضـی بـا افسـردگی و نگرانـی، مرتبـط اسـت.[5] تینـز و مارکـو[6] (۲۰۱۰) در مطالعـه آثـار متغیـر قرارگیـری در معـرض گفتـار نفرت‌افکـن آنلایـن، از یک آزمایـش نظرسـنجی کـه روی کاربـران اینترنـت در مقطـع دانشـگاهی انجام شـد یافتنـد کـه شـرکت‌کنندگان آمریکایی-آفریقایـی بیشـترین اذیـت را از محتـوای نژادپرسـتانه (تصاویـر) در سـایت‌های شـبکه‌های اجتماعی متحمل شـدند، در حالـی کـه آمریکاهایی اروپایـی - بالأخص آن‌هایی کـه رویکردهای غیرنژادپرسـتانه[7] داشـتند - محتمل‌تـر بـود که بـا این تصاویر ناراحت نشـوند. بـه طریـق مشـابه، افـرادی کـه در معـرض گفتـار نفرت‌افکن در زیـر صفحات ردیت مرتبط با دانشـگاه بودند سـطوح بالاتری از اسـترس را نسـبت به کسانی کـه نبودنـد نشـان دادند.[8] داده‌های نظرسـنجی گویای این اسـت کـه جوانانی کـه در معـرض گفتـار نفرت‌افکـن آنلایـن قـرار گرفته‌اند تعلـق ضعیف‌تری به خانـواده دارنـد و سـطوح بالاتـری از ناراحتـی را گزارش می‌کننـد، هرچند این

1. Izsak 2015
2. Salminen et al. 2019
3. Costello et al. 2019
4. Awan and Zempi 2015
5. Tynes et al. 2008
6. Tyns and Markoe
7. Color-Blind Attitude
8. Saha et al. 2019

رابطـه ضرورتاً علی نیسـت.[1] قرارگیـری در معرض گفتـار نفرت‌افکن در فضای آنلایـن همچنیـن بـا اجتنـاب از بحث‌هـای سیاسـی در طـول زمـان مرتبـط اسـت.[2] در سـطح گروهـی، گفتار نفرت‌افکـن آنلاین تنش‌هـای میان‌گروهی را در زمینه‌هـای مختلـف دامن می‌زند کـه گاهی اوقات منجر به کشـمکش‌های خشـونت بار و تضعیف هم‌بسـتگی اجتماعی می‌شـود.[3] برای مثال، فیسـبوک بـرای نقشـش در بسـیج خشـونت‌های اوبـاش مسلمان‌سـتیز در سـریلانکا و بـرای برانگیختـن خشـونت علیه مـردم روهینگیـا در میانمار مـورد حمله قرار گرفته اسـت.[4] با روشـن کـردن مکانیزم‌هایی کـه از طریق آن‌هـا قرارگیری در معـرض گفتـار نفرت‌افکـن تنش‌هـای بیـن گروهـی ایجـاد می‌کنـد، داده‌های نظرسـنجی و شـواهد تجربـی از لهسـتان دلالـت بـر ایـن دارد کـه قرارگیری مکـرر و بسـیار در معـرض گفتار نفرت‌افکن منجر به حساسـیت‌زدایی نسـبت بـه محتـوای نفرت‌آلود، پایین‌تـر ارزیابی‌کـردن جمعیت‌هایـی هدف‌قرارگرفته توسـط گفتـار نفرت‌افکـن و فاصله‌گیـری بیشـتر از آن‌هـا می‌شـود کـه منجر به سـطوح بالاتـر تعصـب ضد-خارج-از-گـروه می‌گردد.[5]

حجـم متنوعـی از ادبیـات پیشـنهاد می‌کنـد کـه گفتـار نفرت‌افکـن می‌توانـد محیطـی را اشـاعه دهـد کـه در آن خشـونت تعصب-محـور یـا بـه صـورت ضمنـی یـا تصریحی تشـویق می‌شـود.[6] رخـداد و انتشـار تضـاد میان گروهـی زمانـی کـه افـراد و گروه‌هـا فرصت بیـان عمومـی شـکایت‌ها و افعال جمعـی هماهنگ‌شـده را دارند، با احتمال بیشـتری حاصل می‌شـود.[7] فناوری دیجیتـال بـه نظـر موانـع را برای فعـل جمعی میـان اعضاء یک گـروه مذهبی یـا قومیتـی بـا بهبـود دسترسـی بـه اطلاعـات در مـورد ترجیحـات یکدیگـر، کاهـش می‌دهـد. ایـن به نظر احتمـال تضاد میـان گروهـی را افزایش می‌دهد و انتشـار آن را در میـان مرزهـا شـتاب می‌بخشـد.[8]

1. Hawdon et al. 2014
2. Barnidge et al. 2019
3. Izsak 2015
4. Vindu, Kumar, and Frenkel 2018
5. Soral et al. 2018
6. Herek et al 1992; Greenawalt 1996; Calvert 1997; Tsesis 2002; Matsuda 2018
7. Weidmann 2009; Cederman et al. 2010
8. Pierskalla and Hollenbach 2013; Bailard 2015; Weidmann 2015

به علاوه، هرچند گفتار نفرت‌افکن تنها یکی از فاکتورهای بسیاری است که در تعامل با یکدیگر تضاد قومیتی را تحریک می‌کند، این فاکتور نقشی قدرتمند را در تشدید احساسات نفرت جمعی ایفا می‌کند.[1] این می‌تواند به طور خاص در فضای آنلاین صادق باشد که در آن گمنامی اجتماعات آنلاین می‌توانند مردم را به سمت اظهار نظرهای نفرت‌آلودتری نسبت به حالت عادی سوق دهد.[2] همان‌طور که افراد باور پیدا می‌کنند که قواعد «نرمال» رفتار اجتماعی حاکم نیست،[3] تنش‌های بین گروهی تشدید می‌شود. در همین راستا، گفتار نفرت‌افکن آنلاین فاصله‌ای فیزیکی میان گوینده و مخاطب قرار می‌دهد که افراد را جسور می‌کند تا بدون پیامد، خودشان را اظهار کنند.[4] شاید مهم‌تر این که شبکه‌های اجتماعی آنلاین این فرصت را برای افراد فراهم می‌کند تا با دیگر افراد همفکری که ممکن است در غیر این صورت هرگز با ایشان مرتبط نشوند یا حتی از وجودشان با خبر نشوند، ارتباط بگیرند.[5] با تشخیص اهمیت گفتار نفرت‌افکن آنلاین به عنوان یکی از نشانه‌های هشدار اولیه خشونت قومیتی، پایگاه‌های داده گفتار نفرت‌افکن چندزبانه به طور روزافزونی توسط حکومت‌ها، سیاست‌گذاران و سمن‌ها[6] برای کشف و پیش‌بینی بی‌ثباتی سیاسی، خشونت و حتی نسل‌کشی استفاده می‌شوند.[7]

بسیاری استدلال کرده‌اند که ارتباط مستقیمی میان نفرت‌افکنی آنلاین و جرایم نفرت‌افکنی وجود دارد و مرتکبین خشونت آفلاین غالباً نقشی را که اجتماعات آنلاین در سوق‌دادن ایشان به عمل بازی کرده‌اند، ذکر می‌کنند.[8] برای مثال، در ۱۷ ژوئن ۲۰۱۵؛ دیلان رووف[9] ۲۱ ساله وارد

1. Vollhardt et al. 2007; Gagliardone et al. 2014
2. Cohen-Almagor 2017
3. Citron 2014; Delgado and Stefancic 2014
4. Citron 2014
5. Posner 2001
6. NGOs
7. Gagliardone et al. 2014; Tuckwood 2014; Gitari et al. 2015
8. Citron 2014; CohenAlmagor 2017; Gerstenfeld 2017
9. Dylann Roof

کلیسای امانوئل آفریکن متودیست اپسیکوپال[1] شد و نه نفر را به قتل رساند. روف در بیانیه‌اش نوشت که وی اولین گرایش‌های نژادپرستانه‌اش را از وب‌سایت شورای شهروندان محافظه‌کار[2] دریافت کرده است.[3] به طور مشابه، گفته می‌شود که مرتکب‌شونده حمله به کنیسه پیتزبورگ در گب به موضع رادیکالش کشانده شده بود و مرتکب‌شونده تیراندازی‌های مسجد نیوزلند در ۲۰۱۹ مطابق گزارش‌ها در سکوهای آنلاین به نگاه رادیکالش کشانده شده بود و حمله‌اش را در یوتیوب پخش کرد.

هرچند خیلی دشوار است تا به صورت علی رابطه میان گفتار نفرت‌افکن آنلاین و جرایم نفرت‌افکن را بررسی کنیم، مطالعات تجربی اخیر در این راستا تلاش کرده‌اند. این آثار روی ادبیات وسیعی ساخته شده است که در جستجوی این هستند که چگونه استفاده از گفتار نفرت‌افکن از طریق سکوهای رسانه‌ای سنتی می‌توانند برای تحریک طغیان‌های خشونت‌آمیز یا نفرت قومیتی به کار روند. این شامل آثاری است که تأثیر رادیوی نفرت‌افکن را روی سطوح خشونت در طول نسل‌کشی رواندان[4] بررسی می‌کند،[5] پژوهش روی این که چگونه پروپاگاندای رادیویی، خشونت یهودی‌ستیزانه را در آلمان نازی برانگیخت،[6] و مطالعه‌ای روی این که چگونه رادیوی صرب‌های ملی‌گرا برای برانگیختن خشونت در کرواسی در دهه ۱۹۹۰ مورد استفاده قرار گرفت.[7]

با بررسی آثار نفرت‌افکنی آنلاین، چان و همکاران[8] (۲۰۱۵) یافتند که در دسترس بودن پهنای باند، جرایم نفرت‌افکن نژادپرستانه را در محیط‌هایی با سطوح بالای تبعیض و نسبت بالاتر واژگان نژادپرستانه جستجو شده در

1. Emanuel African Methodist Episcopal
2. Council of Conservative Citizens (CCC)
3. Cohen-Almagor 2018
4. Rwandan genocide
5. Yanagizawa-Drott 2014
6. Adena et al. 2015
7. DellaVigna et al. 2014
8. Chan et al.

گـوگل افزایـش می‌دهـد. کارهـای ایشـان دلالـت بـر ایـن دارد که دسترسـی آنلایـن بـروز جرایـم نفرت‌افکـن نژادپرسـتانه انجـام شـده توسـط مرتکبیـن مسـتقل را افزایـش می‌دهـد. به طور مشـابه، اسـتفانز-داویدوویتز[1] (۲۰۱۷) به ایـن نتیجه رسـیده اسـت کـه نرخ جسـتجو در گـوگل بـرای واژگان و عبارات مسلمان‌سـتیز، شـامل اصطلاحـات خشـونت‌باری ماننـد «همـه مسـلمانان را بکشـید» می‌توانـد بـرای پیش‌بینـی بروز جرایـم نفرت‌افکن مسلمان‌سـتیز در طـول زمان اسـتفاده شـود. مطالعـات دیگر ارتباطـی را میان گفتـار نفرت‌آلود در توییتـر و جرایـم نفرت‌افکن در بافت[2] آمریکا نشـان می‌دهـد، ولی ارتباطات علی بـه خوبی شناسـایی نشـده‌اند.[3]

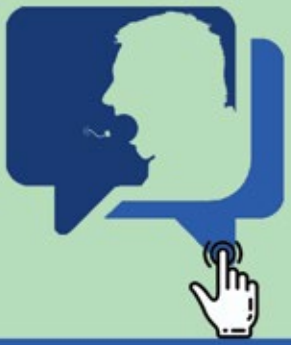
در یکـی از معـدود مطالعاتـی کـه صراحتاً رابطه علی میـان نفرت آنلاین و خشـونت آفلایـن را بررسـی می‌کند، مولر و شـوارز[4] (۲۰۱۷) تغییـرات بیرونی را در قطعی‌هـای گسـترده اینترنـت و فیسـبوک بررسـی مـی کننـد تا نشـان دهنـد کـه جرایـم نفرت‌افکـن مهاجرسـتیز در محیط‌هایـی بـا اسـتفاده بالاتر از فیسـبوک در بازه‌هـای زمانـیِ وجـود احساسـات زیـاد ضدمهاجـر آنلایـن، بـه طـور زیـادی افزایـش پیـدا می‌کنـد. ایشـان یافتنـد کـه ایـن اثـر بـه طور مشـخص بـرای رخدادهای خشـن علیـه مهاجرین، شـامل آتش‌زنـی و یورش آشـکار اسـت. بـه طـور مشـابه در مقالـه‌ای دیگـر، مولـر و شـوارز (۲۰۱۹) از تغییـرات در اسـتفاده‌های اولیـه از توییتر اسـتفاده می‌کنند تا نشـان دهند که اسـتفاده بیشـتر از توییتر با افزایش جرایم نفرت‌افکن مسلمان‌سـتیزانه از آغاز پویـش ترامـپ مرتبـط بـوده اسـت. نتایج ایشـان شـواهد اولیه‌ای را به دسـت می‌دهـد کـه رسـانه‌های اجتماعـی می‌تواننـد بـه عنوان یـک مکانیزم انتشـار بیـن گفتـار نفرت‌افکـن آنلایـن و جرایم خشـونت‌آمیز آفلاین عمـل کنند. در مجمـوع، ایـن آثـار نشـان می‌دهنـد کـه گفتـار نفرت‌افکـن آنلایـن می‌توانند

1. Stephens-Davidowitz
2. Context
3. Williams et al. 2019; Chyzh et al. 2019
4. Muller and Schwarz

عواقب قدرتمندی در جهان واقعی داشته باشد که در گستره‌ای از آثار روان‌شناختی منفی در سطح فردی تا حملات خشونت‌آمیز آفلاین قرار می‌گیرد.

بخش هفتم

مبارزه با گفتار نفرت‌افکن آنلاین



بخش هفتم
مبارزه با گفتار نفرت افکن آنلاین

نگرانی‌های روبه افزایش در مورد آثار گفتار نفرت‌افکن آنلاین در جهان واقعی، محققان، سیاست‌گذاران و سکوهای آنلاین را واداشته است تا راهبردهایی برای مبارزه با گفتار نفرت‌افکن آنلاین توسعه دهند. این رویکردها به طور کلی دو صورت دارد: تعدیل محتوا[1] و ضدگفتار.[2]

یکی از راهبردها برای مبارزه با گفتار نفرت‌افکن آنلاین تعدیل‌کردن محتوا بوده است که شامل ممنوع‌کردن حساب‌های کاربری یا اجتماعاتی است که از شرایط استفاده از خدمات[3] سکوها یا قواعد بیان‌شده تخطی می‌کنند.[4] در ۳۱ مه ۲۰۱۶، کمیسیون اروپا[5] به همراه فیسبوک، توییتر، یوتیوب و مایکروسافت، کد رفتار داوطلبانه‌ای را درمورد مقابله با گفتار نفرت‌افکن آنلاین غیرقانونی[6] صادر کردند که حذف هرگونه گفتار نفرت‌افکن، آن‌طور که توسط اتحادیه اروپا تعریف شده است را الزام می‌کند. انگیزه این اقدام ترس‌ها از افزایش گفتار متعصبانه[7] علیه مهاجرین و همچنین نگرانی‌ها در مورد این بود که گفتار نفرت‌افکن می‌تواند به حملات تروریستی دامن بزند.[8] به‌علاوه، از دسامبر ۲۰۱۷، در مواجهه با فشارها

1. Content Moderation
2. Counter-Speech
3. Terms of Service
4. Kiesler et al. 2012
5. European Commission
6. Code of Conduct on Countering Illegal Hate Speech Online
7. Intolerant Speech
8. Aswad 2016

به دنبال راه‌پیمایی مرگبار «راست را متحد کن» در اوت ۲۰۱۷ در شارلوتسویل ویرجینیا،[1] توییتر سیاستی جدید را برای ممنوع کردن حساب‌های کاربری که مرتبط با گروه‌هایی هستند که «از خشونت علیه غیرنظامیان برای جلو بردن اهدافشان استفاده می‌کنند یا آن را اشاعه می‌دهند»، اعلام کرد.[2] این سکو شروع به تعلیق تعدادی از حساب‌های کاربری با دنبال‌کنندگان زیاد کرد که درگیر ملی‌گرایی سفید[3] یا سامان‌دهی راه‌پیمایی شارلوتسویل بودند. در این بازه زمانی، توییتر همچنین یک فعال انگلیسی راست افراطی که توسط رئیس‌جمهور ترامپ بازتوییت شده بود و همچنین تعدادی دیگر حساب کاربری مرتبط با گروه ملی‌گرایان افراطی[4] او را تعلیق کرد.[5] این شرکت اعلام کرد که ممنوعیت ایشان روی تهدیدهای خشونت‌بار قابل گسترش است تا شامل هر محتوایی شود که خشونت را تکریم و ستایش می‌کند.[6] به طور مشابه در آوریل ۲۰۱۸ فیسبوک مجموعه ۲۵ صفحه‌ای از قواعدش را که محتوای مجاز در فیسبوک را دیکته می‌کند، اعلام کرد.[7] بخش در مورد گفتار نفرت‌افکن بیان می‌کند که «ما اجازه گفتار نفرت‌افکن در فیسبوک را نمی‌دهیم چرا که این امر محیطی از ارعاب و دفع دیگران را خلق می‌کند و در برخی موارد خشونت جهان واقعی را اشاعه می‌دهد». هدف از ممنوع کردن گفتار نفرت‌افکن توسط سکوهای آنلاین اصلی کاهش احتمال این بود که کاربران روزمره اینترنت به طور اتفاقی در معرض گفتار نفرت‌افکن آنلاین قرار گیرند.

با این وجود دانش کمی در دست است که این ممنوعیت‌ها در عمل چگونه پیاده سازی شده است یا این که آن‌ها به طور کلی چه قدر در

1. Charlottesville, Virginia
2. Twitter 2017
3. White Nationalism
4. Ultranationalist Group
5. Twitter 2017
6. Twitter 2017
7. Facebook 2018

کاهش گفتار نفرت‌افکن آنلاین در این سکوها یا قرارگیری در معرض چنین گفتاری مؤثر بوده‌اند. به علاوه، استفاده از کشف خودکار گفتار نفرت‌افکن، با بارز شدن محدودیت این روش‌ها در اشتباهات خجالت‌آور، تحت انتقاد قرار گرفته است؛ برای مثال وقتی که فیلترهای نزاکت[1] فیسبوک فقره‌ای از اعلامیه استقلال آمریکا[2] را به عنوان گفتار نفرت‌افکن مشخص کردند.[3] در حالی که یک مطالعه مروری در فوریه ۲۰۱۹ توسط کمیسیون اروپا دلالت بر این دارد که سکوهای رسانه‌های اجتماعی شامل فیسبوک و گوگل در حذف ۷۵٪ پست‌های پرچم‌خورده[4] توسط کاربران که از استانداردهای اتحادیه اروپا تخطی می‌کند ظرف ۲۴ ساعت موفق بوده‌اند، ما نمی‌دانیم که چه نسبتی از گفتار نفرت‌افکن پرچم خورده است یا چگونه این می‌تواند له یا علیه انواع خاصی از گفتار سیاسی سوگیری داشته باشد.[5]

کارهای تجربی روی کارآمدی ممنوع‌کردن محتوای نفرت‌افکن نتایج غیر روشنی را به دست داده است. چاندراسخاران، پاوالانثان و دیگران[6] (۲۰۱۷) با مطالعه اثر ممنوع کردن زیرصفحات ردیت /fatpeoplehate و /CoonTown در ۲۰۱۵، به این نتیجه رسیدند که این ممنوع کردن موفقیت‌آمیز بوده است. با تحلیل بیش از ۱۰۰ میلیون پست و نظر ردیت، نویسندگان یافتند که بسیاری از حساب‌های کاربری بعد از ممنوعیت، از ادامه فعالیت دست کشیدند و آن‌هایی که باقی ماندند استفاده از گفتار نفرت‌افکنشان را حداقل ۸۰٪ کاهش دادند. هرچند بسیاری از این کاربران به دیگر زیر صفحات ردیت مهاجرت کردند، این زیر صفحات جدید شاهد افزایشی در استفاده از گفتار نفرت‌افکن نبود که نشان می‌دهد ممنوعیت

1. proprietary filter
2. Declaration of Independence
3. Lapin 2018
4. Flagged
5. Laub 2019
6. Chandrasekharan, Pavalanathan et al.

در محدودکردن گفتار نفرت‌افکن آنلاین در ردیت موفقیت‌آمیز بوده است. سلیم و روثز[1] (۲۰۱۹) نیز در مطالعه ردیت یافتند که ممنوع کردن یک زیر صفحه ردیت نفرت‌آلود (r/fatpeoplehate) کاربران این زیر صفحه را واداشت که از پست محتوا در ردیت دست بکشند. به طور مشابه، دیگر پژوهش‌های دلالت بر این دارند که ممنوع کردن حساب‌های کاربری در توییتر شبکه‌های اجتماعی افراط‌گرا را با اخلال مواجه می‌کند، چرا که کاربرانی که مرتباً دچار ممنوعیت می‌شوند وقتی دوباره به یک سکوی خاص ملحق می‌شوند، افت جدی‌ای را در تعداد دنبال‌کنندگانشان تجربه می‌کنند.[2]

با این وجود، هرچند ممنوعیت‌ها حجم کلی گفتار نفرت‌افکن در ردیت را کاهش داده است و فعالیت‌های افراط‌گرایانه را در توییتر با اخلال مواجه کرده است، چنین کاربرانی به سادگی به دیگر سکوها مهاجرت کرده‌اند. نوول و همکاران[3] (۲۰۱۶) یافتند که در واکنش به ممنوعیت‌های سال ۲۰۱۵، کاربران خشمگین به دنبال سکوهای جایگزین چون Voat, Snapzu و Empeopled رفتند. کاربرانی که به این سکوهای حاشیه‌ای مهاجرت کردند غالباً نام کاربری‌شان را حفظ می‌کنند و تلاش می‌کنند اجتماعات ممنوع‌شده‌شان را در یک حوزه جدید و با تنظیم‌گری کمتر، مجدداً بسازند.[4] علاوه بر جابه‌جایی گفتار نفرت‌افکن از یک سکو به سکویی دیگر، آثار دیگر گویای این هستند که تولیدکنندگان محتوای نفرت‌افکن به راحتی در مورد این که چگونه به استفاده از گفتار نفرت‌افکن آنلاین در سکوهای مورد ترجیحشان ادامه دهند، خلاق‌تر می‌شوند. برای مثال، به دنبال اجتناب از تعدیل محتوا، همان‌طور که پیش‌تر توضیح داده شد، اعضاء اجتماعات آنلاین عموماً از واژگان رمزی برای فرار از کشف شدن استفاده می‌کنند.[5]

1. Saleem and Ruths
2. Berger and Perez 2016
3. Newell et al.
4. Chandrasekharan, Pavalanathan et al. 2017
5. Chancellor et al. 2016; Sonnad 2016

علاوه بر این، تلاش‌ها برای ممنوع کردن حساب‌های کاربری گاهی می‌تواند نتیجه معکوس داشته باشد و حمایت آن‌هایی را که نسبت به اجتماعاتِ نفرت‌افکن همدردی دارند، برانگیزد. وقتی کاربران شناخته شده مورد حمله قرار می‌گیرند، افرادی که باورهای مشابه دارند ممکن است انگیزه پیدا کنند تا به دفاع از ایشان بشتابند و/یا نگاه‌هایی را ابراز کنند که مورد مخالفت شرکت‌ها و مؤسسات قدرتمند قرار می‌گیرد. برای مثال، مطالعات تجربی رفتارهای افراط‌گرایانه آنلاین که حساب‌های طرفدار داعش[1] را بررسی می‌کنند نشان می‌دهند که افراط‌گرایانه آنلاین، مسدود شدن حساب‌های کاربری‌شان را یک نشان افتخار می‌دانند و افرادی که مسدود یا ممنوع شده‌اند غالباً قادرند حساب‌های کاربری‌شان را تحت اسامی جدید فعال سازی مجدد کنند.[2] به علاوه، ممنوع کردن کاربران عموماً ایشان را وا می‌دارد تا به سکوهای تخصصی‌تر مانند گب یا Voat نقل مکان کنند که می‌تواند افرادی را که محتوای نفرت‌افکن آنلاین تولید می‌کنند رادیکال‌تر کند. در واقع، ممنوع کردن کاربران نفرت‌آلود ایشان را از محیط متنوعی که در آن ایشان ممکن است در تماس با صداهای معتدل و مخالف قرار بگیرند جدا می‌کند و خشم و احساس اذیت شدن را در آن‌ها بالاتر می‌برد و ایشان را به اتاقک‌های پژواک[3] نفرت‌آلودی سوق می‌دهد که در آن افراط‌گرایی و فراخوان برای خشونت آفلاین تشویق می‌شود و طبیعی جلوه می‌کند.[4] هرچند این استدلال نظری قانع‌کننده‌ای علیه ممنوع کردن کاربران در سکوهای اصلی است، کارهای تجربی بیشتری برای یافتن میزانی که کاربران ممنوع شده به سکوهای افراطی‌تر مهاجرت می‌کنند و همچنین فهم این که آیا ایشان واقعاً در این سکوها رادیکال‌تر می‌شوند، مورد نیاز است.

1. pro-ISIS
2. Vidino and Hughes 2015; Berger and Perez 2016
3. Echo Chambers
4. Marwick and Lewis 2017; Lima et al. 2018; Zannettou et al. 2018; Jackson 2019

در این راه، کارهای تجربی موجود در مورد کارآمدی تعدیل محتوا گویای این است که هرچند این کار می‌تواند گفتار نفرت‌افکن را در سکوهایی مشخص کاهش دهد، چرا که کاربران خشمگین به دیگر گوشه‌های اینترنت مهاجرت می‌کنند، ولی روشن نیست که آیا چنین تلاش‌هایی گفتار نفرت‌افکن را در مجموع کاهش دهد. به علاوه، سؤالات حقوقی، اخلاقی، و فنی چالش‌برانگیزی در رابطه با مزایای ممنوع کردن گفتار نفرت‌افکن در سکوهای رسانه‌های اجتماعی جهانی، علی‌الخصوص خارج از دموکراسی‌های غربی وجود دارد. برای مثال، پژوهشی اخیر در ProPublica پی برده است که قواعد فیسبوک شفاف نیستند و به طور ناسازگاری توسط ده‌ها هزار پیمانکار جهانی که مسئول تعدیل محتوا هستند اعمال می‌شوند. در بسیاری از کشورها و قلمروهای تحت مناقشه، مانند سرزمین‌های فلسطینی، کشمیر و کریمه، فعالان و خبرنگاران برای گفتار نفرت‌افکن یا آسیب‌زا سانسور شده‌اند چون فیسبوک به نگرانی‌های حکومت‌ها پاسخ داده است و تلاش کرده است تا خودش را از مسئولیت حقوقی کنار بکشد. این گزارش نتیجه می‌گیرد که استانداردهای تعدیل محتوای گفتار نفرت‌افکن فیسبوک «عموماً از نخبگان و حکومت‌ها در مقابل فعالان مردمی و اقلیت‌های نژادی حمایت می‌کند». در همین راستا، حکومت‌ها ممکن است گفتار اپوزوسیون را گفتار افراط‌گرایانه یا نفرت‌افکن اعلام کنند تا از تعدیل محتوا به منظور ساکت کردن منتقدانشان استفاده کنند.[1] به علاوه، روش‌های کشف خودکار گفتار نفرت‌افکن به خوبی با زمینه‌های محلی تطبیق داده نشده است و تعدیل کننده‌های محتوای کمی استخدام شده‌اند که زبان‌های محلی حرف می‌زنند، شامل آن‌هایی که به کار گرفته می‌شوند تا به اقلیت‌های تحت

1. Laub 2019

خطـری را کـه عمومـاً هدف گفتار نفرت‌افکن هسـتند، بپردازنـد. در یک مثال مشـهور در ۲۰۱۵، علـی رغم خشـونت قومیتی و گزارش‌های متعـدد از گفتار نفرت‌افکن در فیسـبوک و دیگر سـکوهای رسـانه‌های اجتماعی که مسـلمانان را در میانمـار هـدف قـرار می‌دادند، گفته می‌شـود فیسـبوک صرفـاً دو تعدیل کننده محتـوای برمه‌زبان را اسـتخدام کرد.[1]

بـا شـناخت ایـن امر که سانسـور گفتار نفرت‌افکـن می‌تواند با حمایت‌های حقوقـی از آزادی بیـان در تضـاد قـرار گیـرد یا توسـط حکومت‌ها بـرای هدف قـرار دادن منتقـدان بـه کار گرفته شـود، آژانس‌های بین‌المللی مانند یونسـکو عمومـاً قائل‌انـد کـه «جریـان آزاد اطلاعـات بایسـتی همیشـه یـک هنجـار باشـد». در نتیجـه ایشـان غالبـاً اسـتدلال می‌کننـد کـه ضدگفتـار عمومـاً به سـرکوب گفتار نفرت‌افکن ترجیح دارد.[2] ضدگفتار پاسـخی مسـتقیم به گفتار نفرت‌افکـن اسـت که قصـد دارد روی گفتمـان و رفتار اثر بگـذارد.[3] پویش‌های ضدگفتـار دیرزمانی اسـت کـه برای مبارزه بـا اظهار عمومی گفتـار نفرت‌افکن و تبعیض‌آمیز توسـط کانال‌های رسـانه‌ای سـنتی اسـتفاده می‌شـوند.

مثال‌های آن در بافت آمریکا شـامل اسـتفاده از بیلبوردهای ضد-KKK[4] در دیپ سـاوث[5] و انتشـار اطلاعات در مورد گروه‌های نفرت‌افکن آمریکایی توسـط مرکـز قانـون فقـر جنوبـی[6] اسـت.[7] مداخلاتـی کـه طراحی شـده‌اند تـا جلوی تحریـک بـه خشـونت را بگیرنـد نیز مـورد اسـتفاده قـرار گرفته‌اند که شـامل اسـتفاده از درام‌هـا دربـاره رخدادهـای جـاری در زندگـی گروه‌هایـی خاصی[8]، بـرای مقابلـه با تنش‌های بین گروهی در رواندا، و اسـتفاده از کمدی تلویزیونی در کنیـا بـرای تضعیـف گفتـار نفرت‌افکـن اسـت.[9] ارزیابی‌هـای تجربـی بـه

1. Stecklow 2018
2. Gagliardone et al. 2015
3. Benesch 2014a, 2014b
4. KKK یک گروه نفرت‌افکن تروریست سفیدپوست آمریکایی است.
5. Deep South, Richards and Calvert 2000
6. Southern Poverty Law Center
7. McMillin 2014
8. Soap Operas
9. Staub et al. 2003; Paluck 2009; Kogen 2013

این نتیجه رسیده‌اند که این مداخلات می‌توانند شرکت‌کنندگان را برای شناسایی و مقاومت در برابر تحریک به نفرتِ ضد-خارج-از-گروه توانمندتر کنند.

کارهای متأخرتر به استفاده از ضدگفتار در فضای آنلاین پرداخته است. برای مثال، سمن‌های بین‌المللی، سلبریتی‌ها، و کسب‌وکارهای ملی با ترس از این که انتخابات پیش روی کنیا در ۲۰۱۳ می‌تواند به خشونت منجر شود، تلاش کردند تا پویش‌های «پروپاگاندای صلح»[1] را مورد حمایت مالی قرار دهند تا مانع از گسترش گفتار نفرت‌افکن آنلاین و خشونت آفلاین در کنیا شوند. برای مثال، یک شرکت، پول نقد و زمان مکالمه با تلفن همراه را به کنیایی‌هایی ارائه کرد که به یکدیگر پیام‌های صلح‌آمیز شامل عکس‌ها، اشعار و داستان‌ها، می‌فرستادند.[2] مگدی و همکاران (۲۰۱۶) با نشان دادن این که ضدگفتار به طور ارگانیک در سکوهای آنلاین رخ می‌دهد، به دنبال حملات پاریس در ۲۰۱۵ تخمین می‌زنند که اکثریت غالب توییت‌های پست شده پس از حملات در دفاع از مسلمانان بوده است در حالیکه توییت‌های نفرت‌افکن مسلمان‌ستیز نمایانگر بخش کوچکی از محتوا در فضای توییتر بوده است. به طور مشابه، با بررسی گفتار نفرت‌افکن در بحث‌های سیاسی نیجریه، بارلت و همکاران[3] (۲۰۱۵) یافتند که محتوای افراطی غالباً با مخالفت، تحقیر و ضدپیام‌ها[4] مواجه می‌شوند.

یک رشته نوظهور از ادبیات به ارزیابی تجربی این می پردازد که چه صورت‌هایی از پیام‌های ضدگفتار بیشترین کارآمدی را در کاهش گفتار نفرت‌افکن آنلاین دارند. مانگر[5] (۲۰۱۷) نشان می‌دهد که ضدگفتار با استفاده از بات‌های خودکار می‌تواند نمونه‌های گفتار نژادپرستانه را کاهش دهد اگر تحریک‌کنندگان توسط اعضاءعالی‌رتبه داخل گروه

1. Peace Propaganda
2. Benesch 2014a
3. Bartlett et al.
4. Counter-messages
5. Munger

تحریم[1] شوند، که در این مورد می‌تواند یک مرد سفیدپوست با تعداد زیادی دنبال‌کننده توییتر باشد. به طور مشابه، سیگل و بدان[2] (۲۰۲۰) از یک حساب جعلی[3] استفاده کردند تا با گفتار نفرت‌افکن فرقه‌ای در فضای توییتری عربی مقابله کنند. ایشان یافتند که صرف دریافت یک پیامِ تحریم کننده، استفاده از گفتار نفرت‌افکن را بالأخص برای کاربران در شبکه‌هایی که گفتار نفرت‌افکن نسبتاً غیر رایج است، کاهش می‌دهد. به علاوه، ایشان نشان دادند که پیام‌های نشان‌دهنده یک هویت مذهبی مسلمان رایج که تصدیقاتی را از بازیگران نخبه دارد به طور خاص در کاهش سطح گفتار نفرت‌افکن متعاقب کارآمد هستند. پژوهش‌های دیگری برای ارزیابی بیشتر این‌که چه انواعی از ضدگفتار از چه منابعی بیشتری تأثیر را در کاهش نفرت‌افکنی آنلاین در زمینه‌های مختلف دارند مورد نیاز است. لیتارو[4] (۲۰۱۷) با شناخت پتانسیل بات‌های ضدگفتار پیشنهاد استفاده از بات‌های هوش مصنوعی دست جمعی را برای مبارزه با گفتار نفرت‌افکن آنلاین می‌دهد، هرچند که امکان‌پذیری و عواقب چنین مداخلاتی به خوبی فهم نشده است. شیب و پروس[5] (۲۰۱۶) با شبیه‌سازی این که چگونه ضدگفتار می‌تواند برای در محاق قرار دادن گفتار نفرت‌افکن در فیسبوک به کار آید به این یافته رسیدند که ضدگفتار می‌تواند اثر قابل ملاحظه‌ای روی کاهش مشاهده‌پذیری گفتار نفرت‌افکن آنلاین، علی‌الخصوص زمانی که تولیدکنندگان گفتار نفرت‌افکن آنلاین در اقلیت یک اجتماع خاص هستند، داشته باشد. یکی از معدود مطالعاتی که به روشنی به کشف ضدگفتارهایی که به طور طبیعی در رسانه‌های اجتماعی رخ می‌دهد[6] می‌پردازد، یافته است که نظرات ضدگفتار، لایک‌ها و مشارکت‌های به مراتب بیشتری را نسبت به دیگر نظرات دریافت می‌کند

1. Sanction
2. Siegel and Badaan
3. Suckpuppet Account
4. Leetaru
5. Shieb and Preuss
6. Mathew et al. 2018; Mathew, Saha et al. 2019

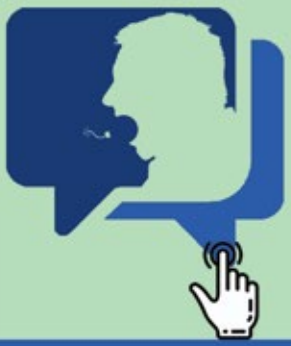
و می‌توانـد تولیدکننـدگان گفتـار نفرت‌افکـن را وادار بـه معذرت‌خواهی کنند یـا رفتارشـان را تغییـر دهند. با این وجـود کارهای تجربی بیشـتری مورد نیاز اسـت تـا ببینیـم چگونـه ایـن پویایی بـه صـورت نظام‌مندتـری در سـکوهای رسـانه‌های اجتماعـی واقعـی در طـول زمان توسـعه پیـدا می‌کند.

بـا مقایسـه صریح سانسـورکردن یا تعدیل محتوا بـا مداخلات ضدگفتاری، آلوارز-بنجومـی و وینتـر[1] (۲۰۱۸) بـه آزمـون ایـن می‌پردازنـد کـه آیا کاهش مقبولیـت اجتماعـی نظرات تهاجمی در یـک انجمن آنلاین، اسـتفاده از گفتار نفرت‌افکـن آنلایـن را کاهـش می‌دهـد؟ ایشـان در ابتـدا انجمـن آنلایـن را طراحـی کردنـد و از شـرکت‌کنندگان خواسـتند به بحثی در مـورد موضوعات اجتماعی روز ملحق شـوند و در آن مشـارکت کنند. ایشـان سـپس به صورت آزمایشـگاهی نظراتـی را کـه شـرکت‌کنندگان پیـش از پسـت نظر خودشـان مشـاهده می‌کردنـد دسـت‌کاری کردنـد. ایشـان یـک فرآینـد سانسـورکننده را گنجاندنـد کـه در آن شـرکت‌کنندگان هیـچ نظـر نفرت‌افکنـی را مشـاهده نمی‌کردنـد و یـک فرآینـد ضدگفتـار را تعبیـه کردنـد کـه در آن نظـرات ضدگفتـار سانسـور نمی‌شـد ولـی در کنار پسـت‌هایی قرار می‌گرفـت که روی ایـن حقیقـت تأکیـد می‌کردند کـه گفتار نفرت‌افکن در این سـکو مـورد قبول نیسـت. با مقایسـه سـطح خصومت نظـرات و نمونه‌های نفرت‌افکنـی در طول شـرایط این فرآیند، ایشـان یافتند که فرآیند سانسورکردن بیشترین کارآمدی را در کاهـش نظـرات خصومت‌آمیـز دارد. ولی نویسـندگان خاطرنشـان کردند کـه ایـن امر که ایشـان یک اثـر معنادار آمـاری از فرآیند ضدگفتار را مشـاهده نکرده‌انـد می‌توانـد ناشـی از اندازه‌هـای کوچـک نمونه‌هـا و ناتوانـی در پایـش تعامـلات تکرارشـونده در طـول زمـان در وضـع آزمایشگاهی‌شـان[2] باشـد. در مجمـوع، ایـن حجم رو به رشـد از ادبیـات روی آثار سانسـورکردن و ضدگفتار

1. Alvarez-Benjumea and Winter
2. Experimental Setup

روی گفتـار نفرت‌افکـن آنلایـن میزانـی خوش‌بینـی، بـه طور خـاص در مورد اثـر تعدیـل محتوا روی کاهش گفتار نفرت‌افکن در سـکوهای اصلـی و توانایی پویش‌هـای ضدگفتـار در کاهـش گسـتره، مشـاهده‌پذیری و آسـیب گفتـار نفرت‌افکـن آنلایـن، را فراهـم می‌کنـد. با این وجـود ما چیزهایی بسـیار کمی در مـورد آسـیب جانبی بالقوه ایـن مداخلات می‌دانیم. کارهای آینده بایسـتی نـه تنهـا مقیـاس بزرگ‌تـری از آزمون‌هـای تجربـی از انـواع ایـن مداخـلات را در زمینه‌هـای مختلـف فراهـم کننـد، بلکـه بایسـتی آثـار بلندمـدت چنین رویکردهایـی را نیز ارزیابـی نمایند.

جمع بندی



بخش اول
نتیجه‌گیری‌ها و گام‌هایی برای پژوهش آینده

همان‌طور که گفتار نفرت‌افکن آنلاین به طرز فزاینده‌ای در سکوهای رسانه‌های اجتماعی بیشتر مشاهده می‌شود، این مسئله در محور بحث‌های دانشگاهی، حقوقی و سیاست‌گذاری قرار گرفته است. علی‌رغم افزایش توجه به گفتار نفرت‌افکن آنلاین، همان‌طور که این فصل گویای آن بود، بحث روی این که چگونه گفتار آنلاین را تعریف کنیم به هیچ وجه حل نشده است. بخشی به خاطر این چالش‌های تعریفی و بخشی ناشی از ماهیت تکامل یابنده و به شدت مختص به زمینه[1] گفتار نفرت‌افکن آنلاین، کشف نظام‌مند محتوای نفرت‌آلود یک کار بسیار دشوار است.

در حالی که روش‌های روز که از یادگیری ماشین، شبکه‌های عصبی و واردکردن ویژگی‌های زمینه‌ای استفاده می‌کنند توانایی ما را برای اندازه‌گیری و پایش گفتار نفرت‌افکن آنلاین افزایش داده‌اند، جالب‌ترین کارهای پژوهشی نسبتاً متشتت هستند و غالباً یک سنخ واحد از گفتار نفرت‌افکن را روی یک از سکوها در یک زمان مشخص کشف می‌کنند. به علاوه، به خاطر راحتی جمع‌آوری داده‌ها، اکثریت غالب مطالعات با استفاده از داده‌های انگلیسی توییتر انجام شده است و بنابراین ضرورتاً به ما اطلاعات زیادی در مورد دیگر سکوها یا زمینه‌های فرهنگی نمی‌دهد.

1. Context-Specific

اگر بخواهیم پیچیدگی‌های بیشتری اضافه کنیم، تعاریف گفتار نفرت‌افکن و رویکردهای کشف آن به شدت، بالاخص در زمینه‌های اقتداگرایانه و شامل تضاد، سیاسی‌شده[1] هستند. هرچند برخی پژوهش‌ها انواع مختلف گفتار نفرت‌افکن، با استفاده از مجموعه داده‌های متعدد و روی سکوهای مختلف را مورد کاوش قرار داده‌اند یا روندهای گفتار نفرت‌افکن را در طول زمان بررسی کرده‌اند، این مطالعات بیشتر از این که یک قاعده باشند، استثنا هستند.[2] با استفاده از ادبیات غنی تکنیک‌های کشف گفتار نفرت‌افکن در علوم کامپیوتر و علوم اجتماعی، کارهای آینده باید تلاش کنند تا تحلیل‌های تطبیقی نظام‌مندتری را برای بهبود توانایی کشف گفتار نفرت‌افکن آنلاین در انواع گوناگونش به دست دهند.

هرچند با توسعه کمتری نسبت به ادبیات موجود روی تعریف و اندازه‌گیری گفتار نفرت‌افکن آنلاین، کارهای اخیر هم به تولیدکنندگان گفتار نفرت‌افکن آنلاین هم به جامعه هدف آن پرداخته‌اند. حجم بزرگی از ادبیات با استفاده از تحلیل کیفی داده‌ها از وب‌سایت‌های رسمی گروه‌های نفرت‌افکن، به ارزیابی این امر پرداخته است که چگونه گروه‌های نفرت‌افکن به طور راهبردی از اینترنت برای جلب تازه‌واردان و اشاعه حس اجتماع میان اعضاء متفرق استفاده می‌کنند.[3] آثار دیگر مطالعات مشاهدتی، مقیاس بزرگی را از کاربرانی که در سکوهای رسانه‌های اجتماعی اصلی مانند توییتر و ردیت گفتار نفرت‌افکن دارند، انجام داده‌اند که شامل ویژگی‌های جمعیت‌شناختی[4] و ساختارهای شبکه ایشان است. این کاربران عموماً جوان، مذکر، بسیار فعال در رسانه‌های اجتماعی و اعضاء اجتماعات بسیار مرتبطی هستند که در آن تولیدکنندگان گفتار نفرت‌افکن مرتباً پست‌های یکدیگر را لایک و بازتوییت می‌کنند.[5]

1. Politicized
2. Fortuna 2017
3. Selepak 2010
4. Demographic
5. Costello and Hawdon 2018; Ribeiro et al. 2018

در رابطـه بـا جامعه هدف گفتار نفرت‌افکن، پژوهشـگران هم از تحلیل‌های کلان‌داده‌هـای تجربـی و هـم نظرسـنجی‌ها از کاربران آنلایـن هدف‌قرارگرفته اسـتفاده کرده‌انـد تـا نشـان دهنـد کـه هدف‌قرارگرفتـگان گفتـار نفرت‌افکـن غالبـاً کاربـران مطـرح رسـانه‌های اجتماعی بـا دنبال‌کننـدگان زیاد هسـتند.[1] بـه عـلاوه، کارهای کمـی و کیفی نشـان می‌دهند که یک راهبـرد هدف‌گیری بـرای گروه‌هـای کاربـران بـه خوبـی سـامان یافته، انجـام حمـلات نفرت‌افکن هماهنگ‌شـده یا «یورش‌ها» بر روی بلاگرها، سـلبریتی‌ها، خبرنگاران یا دیگر بازیگـران مطـرح اسـت.[2] ایـن می‌توانـد دلیلی باشـد بـرای این که چـرا گفتار نفرت‌افکـن آنلایـن تـا ایـن انـدازه توجهـات را در رسـانه‌های جمعـی به خود جلب کرده اسـت، علی‌رغم این که شـواهد تجربی گواه این هسـتند که گفتار نفرت‌افکـن در واقـع در سـکوهای رسـانه‌های اجتماعی در مجموع نادر اسـت.

در واقـع، کارهـای کیفـی‌ای کـه میزان شـیوع گفتار نفرت‌افکـن آنلاین را ارزیابی می‌کنند قائل هسـتند که این پدیده تنها کسـری از درصد کل پست‌ها را در سـایت‌هایی چـون فیسـبوک و توییـرت را بـه خـود اختصـاص می‌دهد.[3]

بـه عـلاوه، مطالعاتـی کـه بـه پویایـی گفتـار نفرت‌افکـن آنلایـن در طول زمـان در توییتـر می‌پردازنـد گویـای ایـن هسـتند کـه ایـن پدیـده نسـبتاً ناگهانـی[4] اسـت و در واکنـش بـه رخدادهـای خشـونت‌بار یـا عاطفـی افزایش پیـدا می‌کنـد و سـپس عمومـاً بـا سـرعت بـه تعـادل مجـدد می‌رسـد.[5]

هرچند گفتار نفرت‌افکن ممکن اسـت نادر باشـد، ولی هنوز عواقب آفلاین جـدی دارد. داده‌های نظرسـنجی نشـان می‌دهند که گفتـار نفرت‌افکن آنلاین اثـر منفی روی بهزیسـتی افرادی می‌گـذارد که در معرض آن قـرار گرفته‌اند و می‌توانـد عواقـب آسـیب‌زایی بـرای روابط گروهی در سـطح اجتماعی داشـته باشـد.[6] حجـم روبه‌رشـدی از شـواهد تجربی نیز گویـای آن هسـتند که گفتار

1. ElSherief, Nilizadeh et al. 2018
2. Mariconti et al. 2018
3. Gagliardone et al. 2016; Siegel et al. 2020
4. Bursty
5. Awan and Zempi 2015; Olteanu et al. 2018; Siegel et al. 2020
6. Tynes et al. 2008

نفرت‌افکن آنلاین می‌توان افراد را به خشونت تحریک کند و این می‌تواند به طور خاص نقش مخربی در دامن‌زدن به حملات علیه مهاجران مسلمان و پناهندگان مسلمان داشته باشد. کارهای اخیری که به پژوهش روی اثر علی گفتار نفرت‌افکن آنلاین روی رویکردها و رفتار آفلاین پرداخته‌اند[1] بایستی تکرار شوند، گسترش یابند و برای این که ما را قادر به فهم بهتر این پویایی‌ها در زمینه‌های دیگر و بازه‌های زمانی طولانی‌تر سازند، تطبیق یابند.

مطالعات علمی همچنین به ارزیابی راهبردهایی پرداخته‌اند که برای مبارزه با گفتار نفرت‌افکن آنلاین بیشترین کارآمدی را دارند. شواهد تجربی نشان می‌دهند که برای مثال ممنوع کردن اجتماعات نفرت‌آلود در رِدیت حجم گفتار نفرت‌افکن را در کل این سکو کاهش داده است.[2] ولی آثار دیگر نشان می‌دهند که کاربرانی که از بحث روی موضوعات مشخص در سکوهای اصلی منع می‌شوند به راحتی به جای دیگری نقل‌مکان می‌کنند تا گفتمان نفرت‌آلودشان را ادامه دهند.[3] به علاوه، ممنوعیت‌های محتوا و حساب کاربری می‌تواند آثار تحریک‌کننده‌ای روی برخی بازیگران افراطی دارد که تحریم را نشان افتخار می‌دانند.[4] خوش‌بینانه‌تر این است که پژوهش‌های تجربی که از ضدگفتار برای مبارزه با گفتار نفرت‌افکن آنلاین استفاده می‌کنند گویای این هستند که دریافت پیام‌های تحریم‌کننده از دیگر کاربران توییتر - علی‌الخصوص اعضاء هم‌گروه شخص، افراد رده بالا یا بازیگران نخبه مورد اعتماد – کاربران را از توییت محتوای نفرت‌آلود دلسرد می‌کند.[5] علاوه بر این، مطالعات تجربی مقیاس بزرگ گویای این هستند که ضدگفتار در فضای آنلاین نسبتاً رایج است و همان رخدادهایی که موجب افزایش گفتار نفرت‌افکن آنلاین می‌شوند غالباً جهش‌های

1. Chan et al. 2015; Muller and Schwarz 2017; Muller and Schwarz 2019
2. Chandrasekharan, Pavalanathan et al. 2017
3. Newell et al. 2016
4. Vidino and Hughes 2015
5. Munger 2017; Siegel and Badaan 2020

بزرگ‌تری در ضدگفتار را نیز رقم می‌زنند.[1] کارهای آینده بایستی به بررسی انواعی از ضدگفتار که می‌توانند بیشترین کارآمدی را در زمینه‌های فرهنگی مختلف و روی سکوهای مختلف داشته باشد و همچنین روی این‌که چگونه ضدگفتار می‌تواند در میان کاربران روزمره شبکه‌های اجتماعی تشویق شود، ادامه دهند. نظر به عواقب آفلاین خطرناک گفتار نفرت‌افکن آنلاین در زمینه‌های جهانی مختلف، دانشگاهیان و سیاست‌گذاران بایستی به جلو بردن ادبیات موجود به منظور بهبود کشف گفتار نفرت‌افکن، به چنگ آوردن فهمی جامع‌تر از چگونگی بروز و انتشار گفتار نفرت‌افکن، توسعه بیشتر فهم ما از عواقب آفلاین گفتار نفرت‌افکن و ساخت ابزارهای بهتر برای مبارزه کارآمد با آن ادامه دهند.

1. Magdy et al. 2016; Olteanu et al. 2018

منابع

• Albadi, N., Kurdi, M., & Mishra, S. (2019). Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. Social Network Analysis and Mining, 19–1 ,(41)9.

• Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing, 522–501 ,102. https://doi.org/10.1007/s0-00745-019-00607

• Adams, J., & Roscigno, V. J. (2005). White supremacists, oppositional culture and the World Wide Web. Social Forces, 778–759 ,(2)84.

• Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., & Zhuravskaya, E. (2015). Radio and the rise of the Nazis in prewar Germany. The Quarterly Journal of Economics, 1939–1885 ,(4)130.

• Agarwal, S., & Sureka, A. (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr microblogging website. arXiv.org. https://arxiv.org/abs/1701.04931

• Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). Cyber hate classification: "Othering" language and paragraph embedding. arXiv.org. https://arxiv.org/pdf/1801.07495.pdf

• Alvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. European Sociological Review, 237–223 ,(3)34.

• Aswad, E. (2016). The role of US technology companies as enforcers of

Europe's new Internet hate speech ban. HRLR Online, 14–1 ,(1)1.

• Aulia, N., & Budi, I. (2019). Hate speech detection on Indonesian long text documents using machine learning approach. In Proceedings of the 5 2019th International Conference on Computing and Artificial Intelligence (pp. 169–164). New York: ACM.

• Awan, I., & Zempi, I. (2015). We fear for our lives: Offline and online experiences of anti-Muslim hostility. Tell MAMA, October. www. tellmamauk.org/wp-content /uploads/resources/We20%Fear20% For20%Our20%Lives.pdf

• Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 760–759). Geneva: International World Wide Web Conferences Steering Committee.

• Bailard, C. S. (2015). Ethnic conflict goes mobile: Mobile technology's effect on the opportunities and motivations for violent collective action. Journal of Peace Research, 337–323 ,(3)52.

• Barnidge, M., Kim, B., Sherrill, L. A., Luknar, Ž., & Zhang, J. (2019). Perceived exposure to and avoidance of hate speech in various communication settings. Telematics and Informatics, 101263 ,44.

• Bartlett, J., Krasodomski-Jones, A., Daniel, N., Fisher, A., & Jesperson, S. (2015). Social Media for Election Communication and Monitoring in Nigeria. Demos report, London.

• Basile, V., Bosco, C., Fersini, E. et al. (2019). Semeval2019- task 5: Multilingual detection of hate speech against immigrants and women in

twitter.InJ.May,E.Shutova,A.Herbelot,X.Zhu,M.Apidianaki,&S.M.Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation (pp.63–54). Minneapolis: Association for Computer Linguistics.

• Beauchamp, N., Panaitiu, I., & Piston, S. (2018). Trajectories of hate: Mapping individual racism and misogyny on Twitter. Unpublished working paper.

• Benesch, S. (2013). Dangerous speech: A proposal to prevent group violence. Voices That Poison: Dangerous Speech Project proposal paper. February 23. https://dangerousspeech.org/wp-content/uploads/01/2018/Dangerous-Speech-Guidelines2013-.pdf

• (2014a). Countering Dangerous Speech to Prevent Mass Violence During Kenya's 2013 Elections. United States Institute of Peace final report, February 28. https://ihub.co.ke/ihubresearch/jb_BeneschCFPReportPeacebuildingInKenyapdf41-08-07-25-3-2014.pdf

• (2014b). Defining and diminishing hate speech. In P. Grant (Ed.), Freedom from Hate: State of the World's Minorities and Indigenous Peoples 2014 (pp. 25–18). London: Minority Rights Group International. https://minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities2014-.pdf

• Berger, J. M., & Perez, H. (2016). The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters. Occasional paper. Program on Extremism at George Washington University.

• Black, E. W., Mezzina, K., & Thompson, L. A. (2016). Anonymous social

media: Understanding the content and context of Yik Yak. Computers in Human Behavior, 57(C), 22–17.

• Bowman-Grieve, L. (2009). Exploring Stormfront: A virtual community of the radical right. Studies in Conflict and Terrorism, 1007–989 ,(11)32.

• Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science, 15–1 ,(1)5.

• Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. Journal of Communication, 19–4 ,(1)47.

• Castle, T. (2012). Morrigan rising: Exploring female-targeted propaganda on hate group websites. European Journal of Cultural Studies, 694–679 ,(6)15.

• Cederman, L.-E., Wimmer, A., & Min, B. (2010). Why do ethnic groups rebel? New data and analysis. World Politics, 119–87 ,(1)62.

• Chan, J., Ghose, A., & Seamans, R. (2015). The Internet and racial hate crime: Offline spillovers from online access. MIS Quarterly, 403–381 ,(2)14.

• Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (pp. 1213–1201). New York: ACM.

• Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017a). The bag of communities: Identifying abusive behavior online with preexisting internet data. In Proceedings of the 2017 CHI Conference on

Human Factors in Computing Systems (pp. 3187–3175). New York: ACM.

• Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017b). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In Proceedings of the ACM on Human-Computer Interaction, Vol. 1 (CSCW) (pp. 22–1). New York: Association for Computing Machinery.

• Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. In Proceedings of the 2017 ACM on Web Science Conference (pp. 22–13).

• Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. International Journal of Human-Computer Studies, 70–57 ,(1)65.

• Chess, S., & Shaw, A. (2015). A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity. Journal of Broadcasting and Electronic Media, 220–208 ,(1)59.

• Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. Aggression and Violent Behavior, 118–108 ,40.

• Chowdhury, A. G., Didolkar, A., Sawhney, R., & Shah, R. (2019). ARHNet: Leveraging community interaction for detection of religious hate speech in Arabic. In F. AlvaManchego, E. Choi, & D. Khashabi (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics: Student Research Workshop (pp. 280–273). Florence: Association for Computational Linguistics.

• Chyzh, O., Nieman, M. D., & Webb, C. (2019). The effects of dog-whistle

politics on political violence. Political Science Publications, 10–1. https://lib.dr.iastate.edu/pols_pubs/59/

• Citron, D. K. (2011). Misogynistic cyber hate speech. Written Testimony and Statement of Danielle Keates Citron, Professor of Law, Boston University School of Law hearing on "Fostering a Healthier Internet to Protect Consumers" before the House Committee on Energy and Commerce, October 2019 ,16, Washington, DC.

• (2014). Hate Crimes in Cyberspace. Cambridge, MA: Harvard University Press.

• Cohen-Almagor, R. (2011). Fighting hate and bigotry on the Internet. Policy and Internet, 26–1 ,(3)3.

• (2017). Why confronting the Internet's dark side? Philosophia, ,(3)45 929–919.

• (2018). When a ritual murder occurred at Purim: The harm in hate speech. El Profesional de la Informacion, 681–671 ,(3)27.

• Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. Violence and Gender, 60–55 ,(1)5.

• Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. Sociological Inquiry, –427 ,(3)89 452.

• Costello, M., Rukus, J., & Hawdon, J. (2018). We don't like your type around here: Regional and residential differences in exposure to online hate material targeting sexuality. Deviant Behavior, 17–1 ,(3)40.

• Czapla, P., Gugger, S., Howard, J., & Kardas, M. (2019). Universal language model finetuning for Polish hate speech detection. Paper presented at the Proceedings of the PolEval 2019 Workshop: 149. May 31, Warsaw, Poland.

• Dadvar, M., de Jong, F. M. G., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In Proceedings of the Twelfth DutchBelgian Information Retrieval Workshop (DIR 2012) (pp. 25–23). Ghent: University of Ghent.

• Daniels, J. (2017). Twitter and white supremacy: A love story. Dame Magazine, October 19. www.damemagazine.com/19/10/2017/twitter-and-white-supremacylove-story/

• Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. arXiv.org. https://arxiv.org/pdf/1703.04009.pdf

• De Smedt, T., De Pauw, G., & Van Ostaeyen, P. (2018). Automatic Detection of Jihadist Online Hate Speech. CLiPS Technical Report No. 7 Computational Linguistics & Psycholinguistics Technical Report Series, Ctrs007-, February. www.uantwerpen.be/clips

• Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi. M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In A. Armando, R. Baldoni, & R. Focardi (Eds.), Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), (pp. 95–86). Venice: CEUR.

• Delgado, R. (1982). Words that wound: A tort action for racial insults, epithets, and name calling. Harvard Civil Rights-Civil Liberties Review, 181–133 ,17.

• Delgado, R., & Stefancic, J. (2014). Hate speech in cyberspace. Wake Forest Law Review, 319 ,49. https://ssrn.com/abstract=2517406

• DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., & Zhuravskaya, E. (2014). Cross-border media and nationalism: Evidence from Serbian radio in Croatia. American Economic Journal: Applied Economics, 132–103, (3)6.

• Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying. The Social Mobile Web, 17–11, (02)11.

• Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), Proceedings of the 24th International Conference on World Wide Web (pp. 30–29). New York: ACM.

• Douglas, K. M. (2007). Psychology, discrimination and hate groups online. In The Oxford Handbook of Internet Psychology (pp. 163–155). Oxford: Oxford University Press.

• Duarte, N., Llanso, E., & Loup, A. (2018). Mixed messages? The limits of automated social media content analysis. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR Vol. 81 (pp. 106). New York: Association for Computing Machinery.

• ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. arXiv.org. https://arxiv.org/abs/1804.04257

• ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. arXiv.org.

https://arxiv.org/abs/1804.04649

• Facebook. (2018). Facebook Community Standards. www.facebook.com /communitystandards/introduction

• Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). Understanding harmful speech online. Berkman Klein Center Research Publication No. 21-2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882824

• Farkas, J., & Neumayer, C. (2017). Stop fake hate profiles on Facebook: Challenges for crowdsourced activism on social media. First Monday, 9)22). http://firstmonday.org/ojs/index.php/fm/article/view/8042

• Fleishman, C., & Smith, A. (2016). Exposed: The secret symbol Neo-Nazis use to target Jews online. Mic, June 1. https://mic.com/articles/144228/echoes-exposed-thesecret-symbol-neo-nazis-use-to-target-jews-online

• Flores-Yeffal, N. Y., Vidales, G., & Plemons, A. (2011). The Latino cyber-moral panic process in the United States. Information, Communication and Society, 589–568 ,(4)14.

• Fortuna, P. C. T. (2017). Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. U.Porto. https://repositorioaberto.up.pt/handle/106028/10216

• Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 85 ,(4)51.

• Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering Online Hate Speech. Paris: UNESCO Publishing.

• Gagliardone, I., Pohjonen, M., Beyene, Z. et al. (2016). Mechachal: Online debates and elections in Ethiopia: From hate speech to engagement in

social media. Working paper. https://eprints.soas.ac.uk/30572/

• Gagliardone, I., Patel, A., & Pohjonen, M. (2014). Mapping and analysing hate speech online: Opportunities and challenges for Ethiopia. University of Oxford Comparative Media, Law & Policy website. https://pcmlp.socleg.ox.ac.uk/mapping-and-analysing-hate-speech-online-opportunities-and-challenges-forethiopia/

• Gerstenfeld, P. B. (2017). Hate Crimes: Causes, Controls, and Controversies. London:Sage.

• Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 230–215 ,(4)10.

• Greenawalt, K. (1996). Fighting Words: Individuals, Communities, and Liberties of Speech. Princeton: Princeton University Press.

• Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 469–468). New York: ACM.

• Gross, T. (2017). Attacked by alt-right trolls: A Jewish journalist links Trump to the rise of hate. NPR: Fresh Air, March 19. www.npr.org/594894657/19/03/2018/attacked-by-alt-right-trolls-ajewish-journalist-links-trump-to-the-rise-of-hate

• Haraszti, M. (2012). Foreword: Hate speech and coming death of the international standard before it was born (complaints of a watchdog). In M. Herz & P. Molnar (Eds.), The Content and Context of Hate Speech: Rethinking

Regulation and Responses (pp. xiii–xviii). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.001

• Hawdon, J., Oksanen, A., & Rasänen, P. (2014). Victims of online hate groups: American youths exposure to online hate speech. In J. Hawdon, J. Ryan, & M. Lucht (Eds.), The Causes and Consequences of Group Violence: From Bullies to Terrorists (pp. 182–165). Lanham, MD: Lexington Books.

• Hawdon, J., Oksanen, A., & Rasänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. Deviant Behavior, 266–254, (3)38.

• Henson, B., Reyns, B. W., & Fisher, B. S. (2013). Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. Journal of Contemporary Criminal Justice, 497–475 ,(4)29.

• Herek, G. M., Berrill, K. T., & Berrill, K. (1992). Hate Crimes: Confronting Violence Against Lesbians and Gay Men. London: Sage.

• Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. Journal of School Violence, 112–89 ,(3)6.

• Hine, G. E., Onaolapo, J., De Cristofaro, E. et al. (2016). Kek, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. arXiv.org. https://arxiv.org/abs/1610.03452

• Holtz, P., & Wagner, W. (2009). Essentialism and attribution of monstrosity in racist discourse: Right-wing Internet postings about Africans and Jews. Journal of Community and Applied Social Psychology, 425–411 ,(6)19.

• Howard, J. W. (2019). Free speech and hate speech. Annual Review of Political Science, 109–93 ,22.

• Isaac, M. (2016). Twitter bars Milo Yiannopoulos in wake of Leslie Jones's reports of abuse. New York Times, July 20. www.nytimes.com/20/07/2016/technology/twitter-bars-milo-yiannopoulos-in-crackdown-on-abusive-comments.html

• Isbister, T., Sahlgren, M., Kaati, L., Obaidi, M., & Akrami, N. (2018). Monitoring targeted hate in online environments. arXiv.org. https://arxiv.org/abs/1803.04757

• Izsak, R. (2015). Hate speech and incitement to hatred against minorities in the media. UN Humans Rights Council. arXiv.org. www.ohchr.org/EN/Issues/Minorities/SRMinorities/Pages/Annual.aspx

• Jackson, S. (2019). The double-edged sword of banning extremists from socialmedia.https://osf.io/preprints/socarxiv/2g7yd/

• Kaakinen, M., Räsänen, P., Näsi, M., Minkkinen, J., Keipi, T., & Oksanen, A. (2018). Social capital and online hate production: A four country survey. Crime, Law and Social Change, 39–25 ,(1)69.

• Kang, S., Kim, J., Park, K., & Cha, M. (2018). Classification of hateful comments in a Korean news portal. www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07503234

•Kennedy,B.,Kogon,D.,Coombs,K.etal.(2018).Atypologyandcodingmanual for the study of hate-based rhetoric. arXiv.org. https://psyarxiv.com/hqjxn/

• Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In R. E. Kraut & P. Resnick (Eds.), Building

Successful Online Communities: Evidence-Based Social Design (pp. –125 177). Cambridge, MA: MIT Press.

• Klubicka, F., & Fernandez, R. (2018). Examining a hate speech corpus for hate speech detection and popularity prediction. arXiv.org. https://arxiv.org/abs/1805.04661

• Kogen, L. (2013). Testing a media intervention in Kenya: Vioja Mahakamani, dangerous speech, and the Benesch guidelines. University of Pennsylvania Scholarly Commons. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1000&context=africaictresearch

• Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In Proceedings of the 2018 World Wide Web Conference on World Wide Web (pp. 943–933). Geneva: International World Wide Web Conferences Steering Committee.

• Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting Tweets against Blacks. In AAAI'13 Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 1622–1621). Bellevue, WA: AAAI Press.

• Lapin, T. (2018). Facebook flagged Declaration of Independence as hate speech. New York Post, 5 July. https://nypost.com/05/07/2018/facebook-flaggeddeclaration-of-independenceas-hate-speech/

• Laub, Z. (2019). Hate speech on social media: Global comparisons. Council on Foreign Relations Backgrounder, June 7. www.cfr.org/backgrounder/hate-speech-socialmedia-global-comparisons

• Leetaru, K. (2017). Fighting social media hate speech with AI-powered bots. Forbes, February 4. www.forbes.com/sites/

kalevleetaru/04/02/2017/fighting-socialmediahate-speech-with-ai-powered-bots/5a22dfa327b1

• Lima, L., Reis, J. C., & Melo, P. (2018). Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 522–515). IEEE.

• Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2019). Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. Behaviour and Information Technology, 11–1.

• Liu, S., & Forss, T. (2014). Combining n-gram based similarity analysis with sentiment analysis in web content classification. In Proceedings of the InternationalJointConferenceonKnowledgeDiscovery,KnowledgeEngineering and Knowledge Management, Vol. 1. (pp. 537–530). Setúbal: SciTePress.

• (2015). New classification models for detecting Hate and Violence web content. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Vol. 1. (pp. 495–487). Lisbon: IEEE.

• Lizza, R. (2016). Twitter's anti-Semitism problem. The New Yorker, October 19. www .newyorker.com/news/news-desk/twitters-anti-semitism-problem

• Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A., & Baldwin, T. (2016). #isisisnotislamor#deportallmuslims?Predictingunspokenviews.InProceedings of the 8th ACM Conference on Web Science (pp. 106–95). New York: ACM.

• Magu, R., Joshi, K., & Luo, J. (2017). Detecting the hate code on social media. arXiv.org. https://arxiv.org/abs/1703.05443

• Mariconti, E., Suarez-Tangil, G. Blackburn, J. et al. (2018). "You know what to do": Proactive detection of YouTube videos

• Marwick, A. (2017). Are there limits to online free speech? Data & Society Research Institute (blog), January 5. https://points.datasociety.net/are-_there-_limits-_to-_online-_free-_speech-_14dbb7069aec

• Marwick, A., & Lewis, R. (2017). Media Manipulation and Disinformation Online. New York: Data & Society Research Institute.

• Mathew, B., Kumar, N., Goyal, P., & Mukherjee, A. (2018). Analyzing the hate and counter speech accounts on Twitter. arXiv.org. https://arxiv.org/abs/1812.02712

• Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In Proceedings of the 10th ACM Conference on Web Science (pp. 182–173). New York: ACM.

• Mathew, B., Saha, P., & Tharad, H. et al. (2019). Thou shalt not hate: Countering online hate speech. arXiv.org. https://arxiv.org/abs/1808.04409

• Matsuda, M. J. (2018). Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment. London: Routledge.

• McMillin, S. E. (2014). Ironic outing: The power of hate group designations to reframe political challenges to LGBT rights and focus online advocacy efforts. Journal of Policy Practice, 100–85 ,(2)13.

• McNamee, L. G., Peterson, B. L., & Peña, J. (2010). A call to educate,

participate, invoke and indict: Understanding the communication of online hate groups. Communication Monographs, 280–257 ,(2)77.

• Meza, R. M. (2016). Hate-speech in the Romanian online media. Journal of Media Research, 55 ,(3)9.

• Mossie, Z., & Wang, J.-H. (2018). Social network hate speech detection for Amharic language. Paper presented at the Fourth International Conference on Natural Language Computing (NATL). April 29–28, Dubai, UAE.

• Muller, K., & Schwarz, C. (2017). Fanning the flames of hate: Social media and hate crime. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972

• Muller, K., & Schwarz, C. (2019). Making America hate again? Twitter and hate crime under Trump. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103

• Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior, 649–629 ,(3)39.

• Nedig, H. (2017). Twitter launches hate speech crackdown. The Hill. December, 18. https://thehill.com/policy/technology/-365424twitter-to-begin-enforcing-new-hate-speech-rules

• Newell, E., Jurgens, D., Saleem, H. M. et al. (2016). User migration in online social networks: A case study on Reddit during a period of community unrest. In Proceedings of the Tenth International AAAI Conference on Web and Social Media (pp. 288–279). Palo Alto, CA: AAAI Press.

• Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. arXiv.org. https://arxiv.org/abs/1804.05704

• Paluck, E. L. (2009). Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. Journal of Personality and Social Psychology, 587–574 ,(3)96.

• Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), The Content and Context of Hate Speech: Rethinking Regulation and Responses (pp. 56–37). Cambridge: Cambridge University Press.

• Parenti, M. (2013). Extreme right organizations and online politics: A comparative analysis of five Western democracies. In P. Nixon, R. Rawal, & D. Mercea (Eds.), Politics and the Internet in Comparative Context (pp. 173–155). London: Routledge.

• Phillips, W. (2015). This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture. Cambridge, MA: MIT Press.

• Pierskalla, J. H., & Hollenbach, F. M. (2013). Technology and collective action: The effect of cell phone coverage on political violence in Africa. American Political Science Review, 224–207 ,(2)107.

• Posner, R. A. (2001). The speech market and the legacy of Schenck. In G. R. Stone & L. C. Bollinger (Eds.), Eternally Vigilant: Free Speech in the Modern Era (pp. 152–121). Chicago: University of Chicago Press.

• Potok, M. (2015). The Year in Hate and Extremism. Southern Poverty Law Center intelligence report. www.splcenter.org/fighting-hate/intelligence-report/2015/yearhate-and-extremism0-

• Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W., Jr.

(2018). Characterizing and detecting hateful users on Twitter. arXiv.org. https://arxiv.org/pdf/1803.08977.pdf

• Richards, R. D., & Calvert, C. (2000). Counterspeech 2000: A new look at the old remedy for bad speech. BYU Law Review, 586–553 ,(2)2000.

• Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. arXiv.org. https://arxiv.org/abs/1701.08118

• Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. In Proceedings of the 11th ACM Conference on Web Science (pp. –255 264). New York: ACM.

• Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. arXiv.org. https://arxiv.org/abs/1709.10159

• Saleem, H. M., & Ruths, D. (2019). The aftermath of disbanding an online hatefulcommunity.arXiv.org.https://arxiv.org/pdf/1804.07354.pdf

• Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S.-g., & Jansen, B. J. (2019). Online hate ratings vary by extremes: A statistical analysis. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (pp. 217–213).

• Santosh, T. Y. S. S., & Aravind, K. V. S. (2019). Hate Speech Detection in Hindi-English Code-Mixed Social Media Text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of

Data (pp. 313–310). New York: ACM.

• Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. Paper presented at the 66th Annual Conference of the International Communication Association: Communicating with Power, June 13–9, Fukuoka, Japan.

• Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Proceedings of the 5th International Workshop on Natural Language Processing for Social Media (pp. 10–1). Stroudsburg, PA: Association for Computational Linguistics.

• Selepak, A. (2010). Skinhead Super Mario Brothers: An examination of racist and violent games on White supremacist web sites. Journal of Criminal Justice and Popular Culture, 47–1 ,(1)17.

• Sellars, A. (2016). Defining hate speech. SSRN. https://papers.ssrn.com/sol3/papers.cfm ?abstract_id=2882244

• Siapera, E., Moreo, E., & Zhou, J. (2018). Hate Track: Tracking and Monitoring Online Racist Speech. Dublin: Irish Human Rights and Equality Commission.

• Siegel, A. (2015). Sectarian Twitter Wars: Sunni-Shia Conflict and Cooperation in the Digital Age, Vol. 20. Washington, DC: Carnegie Endowment for International Peace.

• Siegel, A., Nitikin, E., & Barberá, P. (2020). Trumping Hate on Twitter: Online HateSpeech in the 2016 Presidential Election Campaign and Its Aftermath. Unpublished manuscript.

• Siegel, A., Tucker, J., Nagler, J., & Bonneau, R. (2018). Socially Mediated

Sectarianism. Unpublished manuscript.

• Siegel, A., & Badaan, V. (2020). No2Sectarianism: Experimental Approaches to Reducing Online Hate Speech. Forthcoming in the American Political Science Review.

• Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. arXiv.org. https://arxiv.org/abs/1603.07709v1

• Sonnad, N. (2016). Alt-right trolls are using these code words for racial slurs online. Quartz, October 1. https://qz.com/798305/alt-right-trolls-are-using-googlesyahoos-skittles-andskypes-as code-words-for-racial-slurs-on-twitter/

•Soral,W.,Bilewicz,M.,&Winiewski,M.(2018).Exposuretohatespeechincreases prejudice through desensitization. Aggressive Behavior, 146–136 ,(2)44.

• Staub, E., Pearlman, L. A., & Miller, V. (2003). Healing the roots of genocide in Rwanda. Peace Review, 294–287 ,(3)15.

• Stecklow, S. (2018). Why Facebook Is Losing the War on Hate Speech in Myanmar. Reuters Special Report, August 15. www.reuters.com/investigates/special-report/myanmar-facebook-hate/

• Stephens-Davidowitz, S. (2017). Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are. New York: HarperCollins.

• Tsesis, A. (2002). Destructive Messages: How Hate Speech Paves the Way For Harmful Social Movements, Vol. 778. New York: New York University Press.

• Tuckwood, C. (2014). The state of the field: Technology for atrocity response. Genocide Studies and Prevention: An International Journal, 86–81 ,(3)8.

• Twitter. (2017). Twitter Rules and Policies. https://help.twitter.com/en/rules-andpolicies/violent-groups

• (2018). The Twitter Rules. https://support.twitter.com/articles/18311

• Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online racial discrimination and psychological adjustment among adolescents. Journal of Adolescent Health, 569–565 ,(6)43.

• Tynes, B. M., & Markoe, S. L. (2010). The role of color-blind racial attitudes in reactions to racial discrimination on social network sites. Journal of Diversity in Higher Education, 13–1 ,(1)3.

• Unsvåg, E. F., & Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) (pp. 85–75). Stroudsburg, PA: Association for Computational Linguistics.

• Van Hee, C., Lefever, E., Verhoeven, B. et al. (2015). Detection and fine-grained classification of cyberbullying events. In International Conference Recent Advances in Natural Language Processing (RANLP) (pp. 680–672). Shumen: INCOMA.

• Vidino, L., & Hughes, S. (2015). ISIS in America: From Retweets to Raqqa. Program on Extremism, George Washington University report. https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/ISIS20%in20%America20%-20%Full20%Report.pdf

• Vindu G., Kumar, H., & Frenkel, S. (2018). In Sri Lanka, Facebook contends with shutdown after mob violence. New York Times, March 8. www.nytimes.com/08/03/2018/technology/sri-lanka-facebook-shutdown.html

• Vollhardt, J., Coutin, M., Staub, E., Weiss, G., & Deflander, J. (2007). Deconstructing hate speech in the DRC: A psychological media sensitization campaign. Journal of Hate Studies, 35–15 ,(15)5.

• Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In S. Owsley Sood, M. Nagarajan, & M. Gamon (Eds.), Proceedings of the Second Workshop on Language in Social Media (pp. 26–19). New York: ACL.

• Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In D. Bamman, A. Seza Dog˘ruöz, J. Eisenstein et al. (Eds.), Proceedings of the First Workshop on NLP and Computational Social Science (pp. 142–138). Stroudsburg, PA: Association for Computational Linguistics.

• Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In M. Sahlgren & O. Knutsson (Eds.), Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics (pp. 93–88). Stroudsburg, PA: Association for Computational Linguistics.

• Weaver, S. (2013). A rhetorical discourse analysis of online anti-Muslim and anti-Semitic jokes. Ethnic and Racial Studies, 499–483 ,(3)36.

• Weidmann, N. B. (2009). Geography as motivation and opportunity:

Group concentration and ethnic conflict. Journal of Conflict Resolution, 543–526 ,(4)53.

• (2015). Communication networks and the transnational spread of ethnic conflict. Journal of Peace Research, 296–285 ,(3)52.

• Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. The British Journal of Criminology, 117–93 ,(1)60.

• Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandangenocide.TheQuarterlyJournalofEconomics,1994–1947,(4)129.

• YouTube. (2018). Community Guidelines. www.youtube.com/yt/policyandsafety/communityguidelines.html

• Yuan, S., Xintao W., & Xiang, Y. (2016). A two phase deep learning model for identifying discrimination from tweets. In EDBT: 19th International Conference on Extending Database Technology (pp. 697–696). OpenProceedings.org.

• Zannettou, S., Bradlyn, B., De Cristofaro E. et al. (2018). What is Gab? A bastion of free speech or an alt-right echo chamber? arXiv.org. https://arxiv.org/abs/1802.05287

• Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In A. Gangemi, R. Navigli, M.-E. Vidal et al. (Eds.), The Semantic Web: ESWC 2018 (pp. 760–745). Cham: Springer.

csri.majazi.ir