سال نهم/ پاییز ۱۳۹۹

ژورنال فن‌آوری اطلاعات و سیاست

بررسی کلام تنفرآمیز ضعیف و قوی اسلام هراسانه در رسانه‌های اجتماعی

به قلم برتی ویدگن و تاها یاسری

- **نیما اسلامی الکانی**[۱]
- **مریم مقیمی**[۲]

DOR: 20.1001.1.38552322.1399.9.36.7.5

چکیده

سخنان تنفرآمیز اسلام هراسانه در رسانه‌های اجتماعی دغدغه‌ای فزاینده در سیاست و جامعه غربی معاصر است. این امر می‌تواند آسیب‌های قابل‌توجهی بر تک‌تک قربانیان داشته باشد و احساس ترس و محرومیت را در بین جوامع آنان ایجاد و گفتمان عمومی را زهرآلود کند و دیگر شکل‌های افراط‌گرایی و رفتارهای تنفرآمیز را موجب شود. بر این اساس، نیاز زیادی برای ایجاد ابزارهای قوی خودکار و در مقیاس بالا برای شناسایی و طبقه‌بندی سخنان تنفرآمیز اسلام‌هراسی وجود دارد تا به وسیله آن تجزیه‌وتحلیل‌های کمی پایگاه‌های داده متنی بزرگ ازجمله موارد گردآوری‌شده از رسانه‌های اجتماعی امکان‌پذیر باشد. تحقیق قبلی به میزان زیادی به شناسایی خودکار سخنان تنفرآمیز به‌عنوان یک کار باینری نزدیک شد، اما ماهیت گوناگون اسلام‌هراسی به این معنی است که این موضوع غالباً به لحاظ علوم اجتماعی نظری و نیز پایش مؤثر پلتفرم‌های رسانه‌های اجتماعی امری بی‌مورد و نابجاست. ما با ترسیم کار عمیق مفهومی، یک ابزار نرم‌افزاری خودکار ساخته‌ایم که بین محتوای غیر اسلام‌هراسی، اسلام‌هراسی ضعیف و اسلام‌هراسی قوی تمایز قائل می‌شود. میزان دقت این ابزار ۷۷/۶٪ و دقت متوسط آن ۸۳٪ است. این ابزار، امکان تحقیق کمی در زمینه محرک‌ها، میزان گستردگی، رواج موضوع و آثار سخنان تنفرآمیز اسلام‌هراسی در رسانه‌های اجتماعی را در آینده برای ما فراهم می‌کند.

۱مدیریت توسعه زیرساخت، دانشگاه آزاد اسلامی واحد تهران جنوب nima_eslamialkami@yahoo.com

۲دکتری ارتباطات moghimi333@gmail.com

مقدمه

اسلام‌هراسی نگرانی رو به رشدی در سیاست و جامعه غربی معاصر است و تحقیقات نشان میدهد این پدیده هم در سیاست‌های راست افراطی گسترش یافته و هم به جریان اصلی گفتمان سیاسی وارد شده است. به‌خصوص، گسترش سخنان نفرت‌آمیز اسلام هراسانه در رسانه‌های اجتماعی از جانب دولت (یعنی کلیه گروه‌های پارلمانی حزبی در بین مسلمانان انگلستان)، گروه‌های جامعه مسلمانان، دانشگاهیان و خود رسانه‌ها به‌شدت مورد توجه قرار گرفته است. اسلام‌هراسی در پلتفرم‌هایی نظیر توییتر به دلیل آنکه از سوی نمایندگان منتخب برای مراوده و مفاهمه با عامه مردم و برای گفتگوی سیاسی به معنای وسیع‌تر کلمه استفاده می‌شود به‌شدت باعث نگرانی است. بیشتر تحقیقات موجود در زمینه اسلام‌هراسی ماهیتاً به شکل کمی بوده‌اند و نیاز مبرمی برای انجام مطالعات نظری با مقیاس وسیع‌تر کمی وجود دارد تا درک ما از محرک‌های و دینامیک‌های آن را تعمیق بخشد. این امر مستلزم ابزارهای محاسباتی قوی است که بتواند به‌سرعت و به شکل نظام‌مند به کار گرفته شود. مقاله حاضر گزارشی است از خلق و عملکرد یک ابزار نرم‌افزاری خودکار (به‌خصوص یک طبقه کننده یادگیری ماشین) به‌منظور پرداختن به یک هدف واحد پژوهشی یعنی خلق یک طبقه‌بندی کننده برای کشف و ردیابی محتوای اسلام‌هراسی در رسانه‌های اجتماعی که بین نقاط قوت گوناگون پدیده اسلام‌هراسی تمایز قائل شود.

رمز، دستورالعمل‌های تفسیر و حاشیه‌نویسی و مدل مبتنی بر واژه که برای تولید این ابزار استفاده‌شده است در ضمایم آنلاین در دسترس هستند و به‌این‌ترتیب دیگر محققان می‌توانند از این پژوهش استفاده و آن را بسط دهند.

بررسی مفهوم اسلام‌هراسی

سخنان نفرت پراکنی اسلام‌هراسی در فضای آنلاین مشکلات زیادی برای جامعه ایجاد می‌کند. مثلا می‌تواند به تک‌تک قربانیان خود آسیب وارد کند، حس ترس و انزوا در جوامع آنان ایجاد کند، گفتمان عمومی را مسموم کرده و دیگر شکل‌های افراط‌گرایی و رفتارهای نفرت‌انگیز را در قالب یک چرخه «افراط‌گرایی انباشته» برانگیزد. در مورد معنای اصطلاح اسلام‌هراسی اختلاف‌نظر زیادی وجود دارد. این اصطلاحی است که گالیه آن را یک مفهوم الزاماً مورد تنازع می‌نامد. یعنی تعاریف و شرح و بسط بی‌شماری از اسلام‌هراسی وجود دارد، اما اجماع اندکی در

مورد ویژگی‌های اصلی آن در دست است. میر و مودود در سال ۲۰۰۹ آن را شکلی از نژادپرستی تعریف کرده‌اند. موسوی در سال ۲۰۱۵ آن را کلیشه‌سازی خواند. ایمهوف و رکر در سال ۲۰۱۲ آن را تبعیض، کنست، سام و اولبرگ در سال ۲۰۱۳ آن را ترس، و بیراکلی و حافظ در سال ۲۰۱۸ محرومیت نامیدند. در اثر حاضر ما از تعاریف ذکر شده توسط بلییچ در مورد اسلام‌هراسی استفاده می‌کنیم، زیرا استدلال‌های مفهومی دیگر نظریه‌پردازان مشهور در این حوزه را در بر می‌گیرد و مشابه تعاریف ارائه شده توسط تراست رانیمد و گروه پارلمانی متشکل از تمامی احزاب برای مسلمانان بریتانیایی است. بلییچ اسلام‌هراسی را اینگونه تعریف می‌کند: «نگرش‌ها یا احساسات منفی بی‌رویه و معطوف به اسلام یا مسلمانان». ما تعریف بلییچ را برای محتوای رسانه‌های اجتماعی استفاده می‌کنیم:

مفهومی که تولید یا تبادل می‌شود و منفی گرایی بی‌رویه علیه اسلام یا مسلمانان را بیان می‌کند.

تعریف بلییچ طیف گسترده‌ای از رفتارهای نفرت پراکنانه را به تصویر می‌کشد که به‌صورت آنلاین رخ می‌دهد، ازجمله تهدید، توهین، زبان فاقد خصلت‌های انسانی، سخنان تحقیرآمیز و استفاده از فحش و ناسزا. قدرت اصلی این است که جهت‌گیری مفهومی را برای درک اساس اسلام‌هراسی فراهم می‌کند، به‌جای آنکه فقط فهرستی تجویزی از ویژگی‌های اصلی آن ارائه کند (که می‌تواند به‌سرعت منسوخ شود، زیرا شیوه‌های زبانی تغییر می‌کند و ممکن است در زمینه‌های خاصی نامناسب باشد). به‌طور خاص، تعریف بلییچ به‌جای منفی‌نگری در سطح فردی بر سطح گروه تأکید دارد. به‌عنوان‌مثال، تهدیدی مانند «من می‌خواهم همه مسلمانان را از انگلیس بیرون بیندازم» نفرتی علیه یک شخص خاص ابراز نمی‌کند، بلکه بیزاری و مخالفت با جمع مسلمانان را ابراز می‌کند.

بخش کلیدی بحث‌های مفهومی در مورد اسلام‌هراسی این است که آیا به ضد اسلام‌گرایی مربوط می‌شود (که می‌توان آن را مخالفت علیه اسلام دانست (quareligion / instituion) یا ضد مسلمان‌گرایی (که می‌توان آن را مخالفت علیه گروه اجتماعی مسلمانان دانست) یا هر دو مورد به‌صورت توأمان. تحقیقات روان‌شناسی اجتماعی اغلب تأکید می‌کنند که تعصب صرفاً به درمان افراد/گروه‌ها و نه نهادها یا ایدئولوژی‌ها اشاره دارد. برخی محققان جامعه‌شناسی استدلال مشابهی مطرح می‌کنند. در بحثی مربوط به سخنان نفرت‌آمیز، روزنفلد به‌صراحت می‌گوید «تحقیر دین نمی‌تواند [...] مساوی با تحقیر مذهبی‌ها باشد» (روزنفلد، ۲۰۱۲، ص ۲۷۷).

بااین‌حال، سایر افرادی که اسلام‌هراسی را مطالعه می‌کنند، و به‌ویژه کسانی که از نزدیک با قربانیان کار می‌کنند، اینگونه استدلال می‌کنند که این موضوع باید شامل اسلام‌ستیزی و مسلمان‌ستیزی باشد (آوان و زمپی، ۲۰۱۶؛ چاکرابورتی و گارلند ۲۰۱۲). بلیچ می‌گوید که «اسلام و مسلمانان اغلب به‌طور جدایی‌ناپذیری هستند در درک فردی و عمومی عجین شده‌اند»(بلیچ، ۲۰۱۱) و آلن نیز می‌گوید که اسلام‌هراسان اغلب از اسلام به‌عنوان نماینده و به‌جای انتقاد از مسلمانان انتقاد می‌کنند (سی. آلن، ۲۰۱۷). ازاین‌رو، ما هر دو موضوع ضد اسلام‌گرایی و ضد مسلمان‌گرایی را در تعریف خود از اسلام‌هراسی لحاظ می‌کنیم، گرچه ما به این موضوع قائلیم که این ممکن است برای دیگر تحقیقات مناسب نباشد.

تحقیق در مورد سخنان نفرت‌آمیز اسلام‌هراسی در رسانه‌های اجتماعی

بررسی داده‌های رسانه‌های اجتماعی چالش‌های منحصربه‌فردی را در مقایسه با مطالعه علوم سیاسی سنتی مطرح می‌کند نظیر داده‌های نظرسنجی و آمارگیری و مانیفست‌های سیاسی (مارگتس، ۲۰۱۷). مجموعه داده‌های رسانه اجتماعی اغلب حاوی میلیون‌ها نکته داده‌ای است و در بیشتر موارد رویکردهای کیفی نمی‌تواند از عهده این حجم انبوه برآید. حتی اگر این کار را تیم‌های تحقیقاتی انجام دهند، یا از به اشتراک‌گذاری منابع استفاده شود. این مشکل خصوصاً زمانی شدت می‌گیرد که نفرت‌پراکنی اینترنتی بررسی می‌شود. زیرا رواج آن در «دنیای وحشی غیرمتعارف» بسیار کم است. در مطالعاتی که به‌طور تصادفی از ۱٪ جریان توییت‌ها نمونه‌گیری انجام شد و آن‌ها را تفسیر کرد نوعاً گزارش شد کمتر از یک درصد توییت‌ها حاوی همه انواع نفرت است، و گاهی اوقات به ۱ از هزار مورد می‌رسد (اشمیت و ویگند، ۲۰۱۷). بنابراین ابزارهای محاسباتی در این حوزه بسیار حیاتی هستند تا محققان علوم اجتماعی بتوانند مسائل پیچیده را در چشم‌انداز گسترده و متنوع رسانه‌های اجتماعی بررسی کنند. این موضوع برای پیشبرد حوزه تحقیقات اسلام‌هراسی آنلاین که در مرحله نوزادی است، بسیار مهم است. زیرا بسیاری از پرسش‌های اصلی صرفاً تا حدودی بررسی شده‌اند (بلیوک، فاکنر، جکبوویکز، و مک گارتی، ۲۰۱۸؛ ویدگن و دیگران،۲۰۱۹).

طی دهه گذشته، علوم اجتماعی دستخوش به اصطلاح «چرخش محاسباتی» شده است (بلی و اسمیت، ۲۰۱۷، کوتنه و دیگران ۲۰۱۲). این امر با خوش‌بینی قابل‌توجهی مواجه شد: لازر و همکاران استدلال می‌کنند که علوم اجتماعی محاسباتی منجر به سطح جدیدی از طبقه‌بندی و وسعت دست یافته است، و البته بده بستان اندکی بین آن‌ها وجود دارد (لازر و

همکاران، ۲۰۰۹)، واتس استدلال می‌کند که داده‌های بزرگ می‌توانند با ایجاد «فرآیندهای اجتماعی قابل‌مشاهده» که بررسی آن‌ها در شرایط سازمانی متعارف بسیار دشوارتر است، بینشی در مورد پرسش‌های قدیمی ایجاد کنند (واتس، ۲۰۰۷) و در علوم سیاسی گریمر و استوارت استدلال می‌کنند که «روش‌های محتوای خودکار می‌تواند موارد قبلی را غیرممکن کند» (گریمر و استوارت، ۲۰۱۳). درعین‌حال، بسیاری از محققان هشدار می‌دهند که داده‌های بزرگ به‌علاوه محاسبه همیشه برابر با تحقیقات علمی خوب نیست. در رابطه با قوم‌نگاری، مانوویچ استدلال می‌کند «الگوریتم‌های استفاده شده از سوی دانشمندان علوم رایانه [...] هرگز به همان بینش و درک یکسان از مردم و پویایی در جامعه نمی‌رسند»(مانوویچ، ۲۰۱۱، پ. ۸) بوید و کرافورد به همین شکل اشاره می‌کنند که برخی از داستان‌ها و فرآیندها را نمی‌توان صرفاً «با ایجاد میلیون‌ها حساب فیسبوکی یا توییتری» کشف کرد، اما در عوض، به کار کیفی عمیق و قوم‌نگاری نیاز دارند (بوید و کرافورد، ۲۰۱۲). دیگران توجه خود را به این موضوع جلب کرده‌اند که چگونه علوم اجتماعی محاسباتی اغلب بیشتر پیش‌زمینه جنبه محاسباتی است تا اجتماعی (کولز و شرودر، ۲۰۱۵)؛ همان‌طور که سیهون و یاسری می‌گویند که، «علی‌رغم نام آن، [علوم اجتماعی محاسباتی] از دانشمندان کامپیوتر، ریاضیدانان و فیزیکدانان گرفته شده است تا دانشمندان علوم اجتماعی»(سیهون و یاسری، ۲۰۱۶).

بنابراین، در حالی که روش‌های محاسباتی می‌توانند روش‌های جدید و ابتکاری رفتار مطالعه در مورد رسانه‌های اجتماعی را مطرح کنند، خطر کاهش بیش‌ازحد موضوع و در نظر نگرفتن بینش‌های مهم اجتماعی را دارند. این مسئله زمانی اهمیت ویژه‌ای می‌یابد که سخنان نفرت‌پراکنی آنلاین بررسی شود، زیرا آنچه منفور تلقی می‌شود به‌شدت مورد مناقشه است و برحسب گروه هدف می‌تواند متفاوت باشد (بسیاری افراد بیشتر با نژادپرستی دمخور هستند تا اسلام‌هراسی (رانی مد تراست، ۲۰۱۷)، چشم‌انداز ذهنی فرد (که می‌تواند تحت تأثیر پیشینه آن‌ها، تجربیات زندگی و ارزش‌ها باشد (سلمینن، ورونسی، آلمرخی، یونگ و یانسن ۲۰۱۸) و متن یا زمینه (چه کسی صحبت می‌کند، با چه اختیاراتی و با چه کسی همه می‌تواند بر معنای هر بیت محتوا تأثیرگذار باشد (لیدر مینارد و بنش ۲۰۱۶).

مداخلات در روان‌شناسی اجتماعی نیز به ماهیت چندوجهی تعصب اشاره دارد که شامل رفتارهای صریح، آشکار و مستقیم و همچنین موارد ضمنی، پنهانی و غیرمستقیم است (پتیگرو و میرتنز، ۱۹۹۵). بسیاری از تحقیقات اخیر بر «تعصبات روزمره» و اقدامات نفرت‌پراکنی

(موسوی، ۲۰۱۵) و نیز «پرخاشگری‌های بسیار کوچک» (هاکو، توبس، کاهوموکو فسلر و بروان ۲۰۱۸) تمرکز دارند. بااین‌حال، این تمایزها بین انواع مختلف گفتارها در زمینه سخنان نفرت پراکنانه آنلاین کامل بررسی نشده است. با توجه به اینکه این تمایزات مزایای نظری و تجربی قابل‌توجهی در زمینه استفاده از یک مقوله منفرد اسلام‌هراسی ارائه می‌کنند، این موضوع عجیب است. ایجاد تمایزهای کوچک‌تر محققان را قادر می‌کند درک بهتری از دینامیک زمانی، جغرافیایی و شبکه‌ای اسلام‌هراسی داشته باشند (که ممکن است ازنظر نقاط قوت متفاوت باشند)، فرایندهای رادیکال‌سازی را بررسی کنند (که به وسیله آن افراد از اسلام‌هراس ضعیف تا قوی پیشرفت می‌کنند) و بدانند چگونه اسلام‌هراسی وارد گفتمان‌های اصلی می‌شود و با استقبال گسترده‌ای روبرو می‌شود. همچنین می‌تواند پلتفرم‌ها و دولت‌ها را قادر کند بهتر شبکه‌های اجتماعی را نظارت و پایش کنند و راهبردهای مداخله و تغییر محتوای ظریف‌تری توسعه دهند (ژیلسپی، ۲۰۱۸).

برای تحقیق بهتر نفرت‌پراکنی و اسلام‌هراسی در رسانه اجتماعی به ابزارها و روش‌هایی نیاز داریم که بتوانند به دو مسئله‌ای که در اینجا بحث شده است بپردازند. اول، ابزارها باید مقیاس‌پذیر باشد و توانایی بررسی حجم زیاد محتوا را داشته باشد. دوم، آن‌ها باید براساس بینش‌های تئوریک ترسیم شوند. مانند قدرت متغیر نفرت تا بتوان از آن‌ها برای تجزیه‌وتحلیل علمی اجتماعی استفاده کرد. این موضوع در هدف این مقاله منعکس‌شده است، که در مقدمه مطرح شد:

برای ایجاد یک طبقه‌بندی کننده تشخیص محتوای اسلام‌هراسی در شبکه‌های اجتماعی که بین نقاط قوت گوناگون اسلام‌هراسی تمایز قائل شود.

تعریف اسلام‌هراسی ضعیف و قوی

براساس تعریف بلیچ از اسلام‌هراسی و نیز کار انجام‌شده با قربانیان اسلام‌هراسی توسط سازمان‌های خیریه مربوطه (اینگهام بارو ۲۰۱۸)، اسلام‌هراسی قوی در رسانه‌های اجتماعی به‌صورت زیر تعریف می‌شود:

محتوایی که به‌طور آشکار منفی‌گرایی علیه مسلمانان را بیان می‌کند.

اما اسلام‌هراسی ضعیف در رسانه‌های اجتماعی به شکل زیر تعریف می‌شود:

محتوایی که به‌طور تلویحی منفی‌گرایی علیه مسلمانان را بیان می‌کند.

برای به دست آوردن روش‌های چندوجهی آشکار کردن اسلام‌هراسی به شکل اخراج سرد و بی‌روح و تعصب بدیهی تا سخنان گزنده و خشمناک ما به‌جای «نفرت» از اصطلاح «منفی-گرایی» استفاده می‌کنیم (آوان و زمپی ۲۰۱۶). تمایز بین منفی‌گرایی صریح و ضمنی براساس تحقیقات در خصوص سخنان نفرت پراکنانه قبلی مانند واسم و همکاران (واسم، دیویدسون، وارمسلی و وبر، ۲۰۱۷) و کومار و دیگران (Kumar، Ojha، Malmasi، Zampieri & 2018) استوار است. اصطلاح «ضمنی» و «صریح» به‌سادگی نشان می‌دهد که منفی‌گرایی تا چه حد آشکار است. آن‌ها شامل هیچ‌گونه ارزیابی از قصد بازیگران یا تأثیر سخنان بر قربانیان نمی‌شود. علی‌رغم استحکام تحلیلی این مفهوم‌سازی دوبخشی، صراحت و ضمنی بودن همچنان مفاهیمی بسیار ذهنی است؛ هر توییتی را می‌توان به طور گوناگون تفسیر کرد. در زیر نمونه‌هایی از هر دو اسلام‌هراسی ضعیف و قوی ارائه می‌شود تا این تمایز روشن شود. این نمونه‌ها به دلیل ماهیت بسیار حساس موضوع و چالش‌های اخلاقی ناشی از گزارش کلمه به کلمه داده‌های رسانه‌های اجتماعی براساس داده‌های واقعی، اما مصنوعی هستند (ویلیامز، برناپ و اسلوآن ۲۰۱۷).

نمونه‌هایی از اسلام‌هراسی قوی

اسلام‌هراسی قوی شامل ابراز دیدگاه‌های صریحاً منفی‌گرایانه است نظیر وصف مسلمانان به‌عنوان بربرها و خواستار اقدامات متعصبانه نظیر درخواست منع ورود مسلمانان به انگلستان و ابراز احساسات منفی در مورد مسلمانان نظیر خشم و بی‌اعتمادی که اغلب با ناسزاگویی و بی‌حرمتی ابراز می‌شود. مثال‌هایی از توییت‌های اسلام‌هراسی قوی عبارت‌اند از:

۱. مردان مسلمان با کودکان ازدواج و به آن‌ها تجاوز می‌کنند.
۲. مادران مسلمان از ختنه آلت تناسلی زنان حمایت می‌کنند!
۳. یک مسلمان وحشی دیگر خودش را منفجر کرد.
۴. لعنت به هر چه مسلمان است.
۵. تهاجم مسلمانان آن‌ها می‌خواهند انگلستان را اشغال کنند.
۶. وکیل طراز اول اروپایی می‌گوید که مسلمانان از حاکمیت قانون تبعیت نمی‌کنند و نباید اجازه ماندن در اروپا به آن‌ها داده شود. چون تهدید محسوب می‌شوند.
۷. پلیس مسلمانان را هدف قرار می‌دهد، زیرا آن‌ها باعث دردسر هستند.
۸. تجمع بزرگ علیه مسجد لافبرو – بیایید کشور خود را پس بگیریم.

در مثال شماره ۶، گوینده درحال گزارش ادعاهای فرد دیگری است (به یک وکیل اروپایی طراز اول اشاره می‌کند) اما با این حال خود گوینده درگیر اسلام‌هراسی است، زیرا او این محتوا را به اشتراک گذاشته‌است. همچنین برای تعیین این که آیا توییت اسلام‌هراسی است یا نه (قوی است یا ضعیف) حقیقت محتوای آن ارزیابی نمی‌شود. حتی اگر ادعایی به پشتوانه شواهد مفهومی حمایت شود، اسلام‌هراسی را همچنان می‌توان ابراز کرد. در هر زمینه و شرایطی، حقیقت همواره مورد اختلاف است و هیچ موضع عینی خنثایی وجود ندارد که از طریق آن بتوان اعتبار و روایی معرفت‌شناسی هر ادعایی را مورد قضاوت قرارداد (آلن ۱۹۹۶). به‌این‌ترتیب درحالی‌که به لحاظ حسی این موضوع به بسیاری توییت‌های اسلام‌هراسی حاوی دروغ شباهت دارد، اما این مبنای مفهومی نیست که براساس آن تعیین کنیم که آن‌ها اسلام‌هراسی هستند یا خیر.

نمونه‌هایی از اسلام‌هراسی ضعیف

اسلام‌هراسی ضعیف را می‌توان براساس اینکه آیا منفی‌گرایی ضمنی و تلویحی است از نوع قوی آن متمایز کرد. دو نوع اصلی منفی‌گرایی ضعیف وجود دارد. نخست تأکید بر تفاوت‌های بین مسلمانان و دیگر افراد جامعه نظیر نسبت دادن روش‌های عجیب و غیرمعمول به مسلمانان. اینگونه محتوا مسلمانان را به شیوه‌ای موذیانه محروم و به حاشیه می‌برد؛ مسلمانان به‌طور صریح و آشکار هدف حمله نیستند، اما ناسازگاری برجسته می‌شود. این امر می‌تواند به‌طور ضمنی منفی تلقی شود، زیرا اختلافات درک شده ستوده نمی‌شوند، بلکه مسئله‌ساز تلقی می‌شود (بالمر و سالوموس ۲۰۱۵). نمونه‌هایی از این دست عبارت‌اند از:

۱) مسلمانان فقط متفاوت هستند!

۲) غذای مسلمانان بوی خیلی عجیبی می‌دهد.

۳) پوشیدن روبنده خیلی انگلیسی نیست.

شکل دوم منفی گرایی ضعیف گرفتن استعارات مربوط به منفی‌گرایی قوی است (نظیر ادعای اینکه مسلمانان تروریست، بربر یا بی‌سواد هستند) و به‌ظاهر آن‌ها را فقط به یک گروه فرعی کوچک مسلمانان پیوند دادن (مثلاً به یک تروریست یا مسلمانان که فقط در یک منطقه کوچک جغرافیایی زندگی می‌کنند نظیر روترهام) و با این کار به‌طور تلویحی ارتباطی بین استعاره منفی و کل مسلمانان ایجاد کردن با استفاده از اصطلاح مسلمانان یا اسلام، حتی با هشدارها و اخطارها برای برجسته کردن ویژگی خاص (نظیر این تروریست مسلمان یا مردان

مسلمان در روترهام)، یعنی یک ارتباط ضمنی با کل مسلمانان ایجاد کردن. نکته کلیدی در اینجا این است که گفتمان‌های مربوط به بچه‌بازها، تروریست‌ها یا کسانی که آلت تناسلی زنان را ختنه می‌کنند را غالبا می‌توان بدون نیاز به اشاره به هویت مسلمان ابراز کرد. مثال‌هایی از این نوع اسلام‌هراسی ضعیف ذیلاً آمده است. در تمامی موارد، به نظر می‌رسد گوینده مورد خاص را تفسیر و بیان می‌کند، اما همچنان به‌طور تلویحی ارتباطی با استعاره منفی و کل مسلمانان برقرار می‌کنند.

۴) تروریست‌های مسلمان به لوندون بریج حمله کردند.

۵) تندروهای مسلمان در بیابان پناهنده مسیحی را کشت.

۶) بچه‌باز مسلمان یک بیمار روانی است.

آثار قبلی

اثر قبلی در زمینه بررسی سخنان نفرت‌پراکنانه حاکی از چالش‌ها و درعین‌حال پتانسیلی برای ایجاد یک سیستم طبقه‌بندی است که بین سخنان اسلام‌هراسی ضعیف و قوی تمایز قائل می‌شود. تا آنجا که می‌دانیم هیچ تحقیق قبلی به‌طور ویژه بر این موضوع متمرکز نشده است. در عوض طبقه‌بندی کننده‌های باینری برای اسلام‌هراسی ارائه می‌شوند (ویدگن و همکاران ۲۰۱۹). محققان در دموس محتوای اسلام‌هراسی در توییتر با ایجاد چندین طبقه‌بندی کننده ردیابی کردند که هر یک از آن‌ها یک روی سکه اسلام‌هراسی را آشکار می‌کند. این‌ها بر مبنای ارتباطات بین یونیگرام‌ها و بیگرام‌ها هستند و وقتی با هم ترکیب می‌شوند یک مقیاس واحد از اسلام‌هراسی در توییت‌ها به دست می‌دهند (دموس ۲۰۱۷). برناپ و ویلیامز تنفر علیه گروه‌های سیاه‌پوست و اقلیت‌های قومی و گروه‌های دینی را در جریان حمله تروریستی وولویچ در سال ۲۰۱۳ بررسی کردند. آنان از یک مجموعه طبقه‌بندی کننده استفاده کردند که نتایج چندین طبقه‌بندی کننده را در یکجا مجتمع می‌کند و هر یک از تجزیه وابستگی و یک فرهنگ اصطلاحات نفرت پراکنانه به‌عنوان درون‌داد اصلی استفاده می‌کند. دقت این روش ۰/۹۵ و میزان فراخوانی آن ۰/۹۵ است. عملکرد بالای آن‌ها تا حدی به دلیل مجموعه داده‌های نامتوازن آن است. یعنی از مجموع هزار و ۹۰۱ توییت، تعداد هزار و ۶۷۹ مورد (یعنی ۸۸ درصد) ملایم و فقط ۲۲ مورد (یعنی ۱۲ درصد) نفرت‌پراکنانه است و طبقه‌بندی کننده آن‌ها در مورد توییت‌های ملایم بهتر عمل می‌کند.

عملکرد طبقه‌بندی کننده در روش‌های بررسی سخنان نفرت‌آلود چندطبقه‌ای که به دیگر اهداف سوءاستفاده نظر دارد اغلب به‌مراتب پایین‌تر است. همان‌گونه که سالمینن و همکاران اشاره می‌کنند، آثار موجود با استفاده از طبقه‌بندی چند عنوانی برای سخنان نفرت‌آلود آنلاین فوق‌العاده نادر هستند و ما نمی‌توانیم کار قبلی را که نتایج خوبی به دست آورده ملاک قرار دهیم (سالمینن و همکاران ۲۰۱۸). علاوه براین بیشتر تحقیقات موجود درخصوص طبقه‌بندی چند گروهی بر تمایز بین اهداف گوناگون نفرت به‌جای نقاط قوت گوناگون متمرکز بوده است. طبقه‌بندی محتوا براساس نقطه قوت به‌جای هدف چالش‌های دیگری را ایجاد می‌کند زیرا تنوع و تفاوت اندکی بین طبقات و گروه‌ها وجود دارد.

برناپ و ویلیامز یک طبقه‌بندی کننده را برای تمایز بین سطوح مختلف نفرت سایبری آموزش دادند (که به گروه‌های میانه‌رو و افراطی تقسیم شدند) و هدف آن علیه اقلیت سیاه‌پوست و گروه‌های مذهبی در توییتر بود (برناپ و ویلیامز ۲۰۱۶)، که این بررسی با دقت ۰/۷۷ به پایان رسید. مالماسی و زامپیری بین سخنان نفرت‌پراکنانه، سخنان توهین‌آمیز و سخنان قابل‌قبول تمایز قائل شدند. آنان دقت ۷۸ درصدی به دست آوردند، اما درمجموع داده‌هایی که یکدست نبودند یعنی بیش از نیمی از موارد قابل‌قبول بود. مدل آنان نمی‌تواند بین محتوان غیرقابل‌قبول تمایز قائل شود؛ از دوهزار و ۳۹۹ مثال نفرت‌پراکنانه در مجموعه داده‌های آنان هزار و ۵۰ مورد به‌درستی طبقه‌بندی شدند و هزار و ۱۱۳ مورد به‌طور نادرست به‌عنوان توهین‌آمیز طبقه‌بندی شدند و ۲۳۶ مورد به‌عنوان قابل‌قبول. آنان همچنین مدل خود را در داده‌های مشاهده نشده آزمایش نکردند بلکه فقط نتایج اعتبارسنجی را گزارش کردند که احتمال خطا دارد.

دیویدسون و همکاران مدلی را برای تمایز بین سخنان نفرت‌پراکنانه و سخنان توهین‌آمیز و سخنان غیر توهین‌آمیز در توییت‌ها آزمودند. آن‌ها نتایج جالبی را گزارش کردند که دقت آن ۰/۹۱ و بازخوانی ۰/۹۰ و نمره F1 ۰/۹۰ را نشان می‌داد. اثر آنان پتانسیل موجود برای طبقه‌بندی چند گروهی را اثبات کرد و یک استدلال نظری مهم در موردنیاز به تفکیک انواع مختلف محتوا را نشان می‌دهد، اما همان‌طور که آنان اشاره‌کرده‌اند مدل آن‌ها در مورد سخنان نفرت‌آلود ضعیف عمل می‌کند و تقریباً ۴۰ درصد به غلط طبقه‌بندی می‌شود. امتیاز بالای آن به دلیل این است که طبقات و گروه‌های آنان بسیار ناهموار است (۷۶ درصد داده‌ها در مقوله

سخنان توهین‌آمیز). آنان همچنین طبقه‌بندی کننده خود را براساس یک مجموعه واحد داده‌ها آموزش و آزمون کرده‌اند.

ایجاد طبقه‌بندی کننده

ما از قدرت تجزیه‌وتحلیل محاسباتی به‌ویژه یادگیری ماشین برای ایجاد یک ابزاری نرم‌افزاری خودکار استفاده می‌کنیم که می‌تواند بین اسلام‌هراسی ضعیف، اسلام‌هراسی قوی و محتوای غیر اسلام‌هراسی را تمیز بدهد. این موضوع شامل پنج مرحله است که در ادامه این بخش ارائه می‌شود. ابتدا داده‌های مربوطه را گردآوری کنید. دوم، داده‌ها را برای ایجاد یک مجموعه آموزشی داده‌ها تفسیر کنید. سوم، ویژگی‌های اطلاعات ورودی را استخراج‌کنید چهارم، الگوریتم بهینه را شناسایی کنید. پنجم، نتایج را آزمایش و اعتبار سنجی کنید.

گردآوری داده‌ها

با توجه به اینکه گسترش اسلام‌هراسی در احزاب سیاسی انگلیس یک نگرانی جدی در جریان گفتمان مدنی و سیاسی محسوب می‌شود، ما طبقه‌بندی کننده خود را روی مجموعه داده‌هایی که در این دامنه قرار دارد آموزش می‌دهیم. برای ایجاد مجموعه داده‌ای چهار مرحله را دنبال می‌کنیم. اول فهرستی از ۵۰ هزار کاربر توییتر تهیه کنید که هرکدام از آن‌ها حداقل هوادار یکی از شش حزب سیاسی برجسته انگلستان هستند. این احزاب مشتمل بر انواع و اقسام ترکیب ایدئولوژیک هستند، ازجمله احزاب جریان اصلی با پایگاه‌های سراسری حامی و سیاست‌های کاملاً لیبرال (محافظه‌کاران، لیبرال دموکرات‌ها، و حزب کارگر)، یک حزب راستگرای پوپولیست که به ابراز نظر سرسختانه ضد اتحادیه اروپا (UKIP) معروف است، و دو حزب راست افراطی (BNP و اول انگلیس) (فورد و گودوین، ۲۰۱۴؛ گلدر، ۲۰۱۶؛ جان و مارگتس، ۲۰۰۹). با توجه به ارتباط نزدیک اسلام‌هراسی با سیاست راست افراطی، ما بیشتر توییت‌ها را از بازیگران راست افراطی با ۴۵ حساب طراز اول راست افراطی در فهرست ۵۰ هزار کاربر تضمین می‌کنیم. این ۴۵ حساب راست افراطی شناسایی شده است

از گزارش‌های امید به‌جای تنفر شناسایی شده‌اند. دوم، ما همه توییت هایی که این ۵۰ هزار حساب کاربری بین ژانویه ۲۰۱۷ و ژوئن ۲۰۱۸ تولید شده است را با استفاده از API «جستجو» توییتر گردآوری می‌کنیم. به‌این‌ترتیب یک مجموعه داده با ۱۴۰ میلیون توییت ایجاد می‌شود. سوم، ما از ۱۴۰ میلیون توییت نمونه‌گیری کردیم تا مجموعه داده‌های آموزشی

۴هزار توییت را ایجاد کنیم و نمونه را برای لحاظ کردن توییت‌ها از هواداران هر یک از احزاب سیاسی طبقه‌بندی کردیم. ما تأثیر ربات‌ها و کاربران آنلاین ناشناس را (که ممکن است الگوهای غیرمعمول توییت کردن داشته باشند) با محدود کردن حداکثر تعداد توییت‌هایی که هر کاربر می‌تواند به مجموعه داده‌ها بدهد به فقط سه مورد کاهش دادیم.

ایجاد یک مجموعه داده آموزشی با نمونه‌های کافی از محتوای نفرت پراکنانه تلاشی زمان بر است و مهم‌ترین دلیل آن این نیست که در بیشتر متون آنلاین رواج نفرت‌پراکنی در کل نسبتاً کم است (اشمیت و ویگند، ۲۰۱۷). برای بهبود این معضل، واسم و هووی توصیه می‌کنند که افزایش و رواج گفتار نفرت‌پراکنانه با استفاده از داده‌های نمونه‌گیری به‌احتمال‌زیاد مرتبط است. این رویکرد تا حدی در اینجا اتخاذ شده است؛ از ۴ هزار توییت موجود در این نمونه، هزار مورد با استفاده از جستجوی کلمات کلیدی برای «مسلمانان» و «اسلام» شناسایی می‌شوند.

نمونه ما بسیار ناهمگن است و شامل توییت‌های تعداد زیادی از کاربران می‌شود که طیف متنوعی از گزارش‌های سیاسی را دنبال می‌کنند و این گزارش‌ها برای اطمینان از پوشش موضوع برای یک سال کامل طبقه‌بندی‌شده‌اند. ما درصددیم که ناهمگنی را به حداکثر برسانیم تا کاربرد طبقه‌بندی ما را در شرایط مختلف دنیای واقعی به‌منظور کاربرد طبقه‌بندی در موارد مختلف افزایش دهد. این امر همچنین خطر پوشش بیش‌ازحد را با اطمینان از طیف متنوعی از سبک‌های زبانی که در مجموعه داده گنجانده‌شده کاهش می‌دهد (Waseem & Hovy، 2016). با این حال، ما جانب احتیاط را در نظر می‌گیریم که با تمرکز بر مجموعه‌هایی که تحقیق قبلی نشان می‌دهد مجموعه داده ما نمونه «نماینده» بازیگران سیاسی آنلاین نیست. همان‌طور که ما عملکرد طبقه‌بندی کننده را با تمرکز روی شرایطی بهینه‌سازی کرده‌ایم که تحقیقات قبلی نشان می‌دهد احتمالاً حاوی نفرت اسلام‌هراسی (مانند راست افراطی) است. به‌این‌ترتیب، طبقه‌بندی کننده ما ممکن است هنوز در همه زمینه‌ها قابل‌استفاده نباشد.

تفسیر داده‌ها

کل ۴هزار توییت در مجموعه داده‌های آموزشی توسط سه مفسر انسانی بدون بصیرت تفسیر شده‌اند که کارشناسان سیاست انگلستان و مطالعه در زمینه تبعیض و تعصب بودند. براساس تعاریفی که پیش‌تر از اسلام‌هراسی ارائه شد ما دستورالعمل‌هایی برای مفسران ارائه کردیم، و آن‌ها را از طریق دو مطالعه مقدماتی بسط دادیم، که هرکدام شامل ۲۰۰ توییت است. در

تمامی ۴هزار توییت، توافق پایایی بالا بود. توافق درصدی ۸۹,۹٪، کاپای فلیس ۰,۸۳۷ و آلفای کریپندورف ۰/۸۹۵ است. ما همچنین امتیازات برمبنای مقوله‌ها را برای کاپای فلیس محاسبه می‌کنیم که از ۰,۷۳۷ برای اسلام‌هراسی ضعیف تا ۰,۹۰۷ برای اسلام‌هراسی قوی متغیر است. سازگاری این نتایج اعتبار داخلی دستورالعمل‌های تفسیر را نشان می‌دهد.

در مواردی که مفسران اختلاف‌نظر دارند، توییت ها براساس تصمیم اکثریت به طبقات گوناگون اختصاص داده می‌شود. در مجموعه داده نهایی، ۳ هزار و ۱۰۶ توییت به‌عنوان غیر اسلام‌هراسی، ۴۸۴ توییت به‌عنوان ضعیف و ۴۱۰ توییت به‌عنوان قوی طبقه‌بندی‌شده‌اند. برای ایجاد یک مجموعه داده متوازن و یکنواخت، تعداد توییت‌های غیر اسلام‌هراسی به‌طور تصادفی به ۴۷۰ مورد کاهش یافت. این یک مجموعه داده آموزشی نهایی از ۱،۳۶۴ توییت ایجاد می‌کند.

انتخاب ویژگی

انتخاب ویژگی به انتخاب متغیرهای ورودی مورد استفاده برای آموزش طبقه‌بندی کننده اشاره دارد. برای مقایسه، ابتدا یک مدل پایه ایجاد می‌کنیم که فقط تعداد هر اصطلاح در هر توییت را به‌عنوان یک ورودی استفاده می‌کند. دوم اینکه مدلی را با استفاده از ۵۰ ویژگی سطحی و ویژگی‌های اضافی مشتق شده ایجاد می‌کنیم. این‌ها شامل احساسات و تضاد، تعداد کلمات به‌کاررفته برای سوگند خوردن (Ipsos، 2016)، بخش‌هایی از گفتار و نهادها و شخصیت‌های نامبرده است. همچنین دو ویژگی ورودی جدید به دست می‌آوریم: (۱) ذکر نام مسلمانان و (۲) ذکر نام مساجد، که هر دو از صفحات مربوط به ویکی پدیا گرفته شده است. سوم، یک مدل ترکیبی خلق می‌کنیم که از هر دو مورد کدگذاری داغ و همه ۵۰ ویژگی غیر متنی استفاده می‌کند. چهارم، با استفاده از کلمه gloVe پیش آموزش دیده یک مدل ایجاد می‌کنیم که برای دو میلیارد توییت آموزش دیده است (استنفورد، ۲۰۱۸). پنجم، ما یک مدل gloVe ایجاد می‌کنیم

با استفاده از کلمه جدیداً آموزش داده شده که در ۱۴۰ میلیون توییت خود ما تعبیه شده است ایجاد می‌کنیم که از روش پنینگتون و همکاران استفاده می‌شود (پنینگتون، سوچر و منینگ، ۲۰۱۴). نهایتاً و در مرتبه ششم ما مدلی ایجاد می‌کنیم که از کلمه تازه آموزش داده شده و نیز تمامی ۵۰ مورد ویژگی‌های غیرمتنی استفاده می‌کند. برای آزمایش ما از اعتبارسنجی متقاطع ۱۰ لایه در مورد مجموعه داده‌های تفسیر شده با استفاده از الگوریتم

Naïve Bayes بهره می‌بریم. در جدول شماره ۱ نتایج نشان داده شده است. ما عملکرد را با دقت که درصدی از توییت هاست و به طبقه درست مورد نظر اختصاص داده می‌شوند اندازه‌گیری می‌کنیم.

مدل پایه که فقط از تعداد هر اصطلاح به‌عنوان اطلاعات درون‌داد استفاده می‌کند فقط دارای ۳۰,۷٪ دقت است، که از تعیین تکلیف تصادفی بدتر است (که ۳۳,۴۹٪ دقت دارد) تعیین تکلیف با قانون صفر (که ۳۶,۰۹٪ دقت دارد). عمل هر پنج مدل از آن بهتر است. این نتیجه دشواری استخراج سیگنال‌های مربوطه برای طبقه‌بندی سخنان نفرت‌انگیز اسلام‌هراسی را نشان می‌دهد و انتخاب ما برای معماری الگوریتم پیچیده‌تر را توجیه می‌کند. بهترین مدل اجرایی فقط کلمه تازه آموزش داده شده است (مدل ۵). جالب‌توجه این است که این مدل به‌طور قابل‌توجهی از دقت مدل کلمه از قبل آموزش داده شده بهتر عمل می‌کند (۵,۹ درصد امتیاز، ۶۹,۱۳ درصد در مقایسه با ۶۳,۲). این نشان می‌دهد که مزیت‌های آموزش کلمه تعبیه شده در توییت ها که به‌طور خاص تنظیم می‌شوند از هزینه داشتن یک مجموعه داده کوچک‌تر بیشتر است. این امر در راستای کار قبلی در خصوص تعبیه کلمات است، مانند لای و دیگران، که طبق گزارش آن‌ها «دامنه مجموعه نوشته از اندازه مجموعه نوشته‌ها مهم‌تر است» (لای، لئو، او و ژائو،۲۰۱۶). ما یک مدل نهایی (مدل ۷) را ایجاد می‌کنیم که کلمه تعبیه و تازه آموزش و استفاده شده در مدل ۵ با درج برخی ویژگی‌های اضافی بهینه شده است. از طریق آزمایش هر ترکیبی از ویژگی‌های اضافی، شش مورد را شناسایی می‌کنیم که دقت را در مدل ۷ به حداکثر می‌رساند:

تعبیه کلمات + شمارش ذکر مساجد + حضور HTML+ حضور RT+ بخشی از سخنان: حرف ربط + شناسایی رسمی نهاد نامبرده: مکان + شناسایی رسمی نهاد نامبرده: سازمان‌دهی

مدل ۷ صرفاً مشتمل بر یکی از دو ویژگی است که ما مهندسی کرده‌ایم (شمارش ذکر مساجد)، و نه استفاده از نام‌های مسلمانان. این نشان‌دهنده این حقیقت است که طی دوره گردآوری داده‌ها، چندین مسجد هدف احساسات ضد مسلمانان قرار گرفته است ازجمله چندین تهاجم توسط فعالان راست افراطی و یک حمله تروریستی. در مجموعه داده‌های ما، توییت هایی که حاول ذکر نام مساجد هستند احتمالاً ۵ مرتبه بیشتر به عنوان اسلام‌هراسی (خواه ضعیف یا قوی) طبقه‌بندی می‌شوند. وجود HTML و Retweets نیز با توئیتی مرتبط است که با احتمال بیشتری به عنوان اسلام‌هراسی طبقه‌بندی می‌شود که می‌تواند بازتاب چند رسانه‌ای

و ماهیت وایرال محتوای منفی رسانه‌های اجتماعی باشد. دو نهاد نامبرده، مکان و تشکیلات در مدل نهایی لحاظ شده‌اند. تحلیل کیفی نشان می‌دهد که در بسیاری موارد، تشکیلات و مکان‌های شناسایی شده یا به دین اسلام مربوط است یا به رهبران و احزاب سیاسی راست افراطی. از اینرو، لحاظ نمودن اینگونه ویژگی‌ها در مدل می‌تواند به لحاظ نظری توجیه شود. این موضوع مهمی است زیرا به این معناست که طبقه‌بندی‌های تولید شده توسط طبقه‌بندی کننده ما را می‌توان به‌صورت محدود شرح داد. درج حرف ربط به لحاظ نظری و تئوریک کمتر توجیه می‌شود و تحلیل بیشتری در کار آینده لازم است تا به این درک برسیم که چرا این ویژگی به نفرت ونفرت پراکنی مربوط است. یک راه بررسی ساختارهای دستور زبانی استفاده شده در توییت های نفرت پراکنانه است با توجه به تحقیق قبلی که نشان می‌دهد که این رامی توان برای شناسایی محتوای توهین‌آمیز استفاده کرد.

گزینه الگوریتم

برای سهولت کار ما الگوریتم‌ها را در مدل شماره ۵ آزمایش می‌کنیم (که مشتمل بر فقط واژه‌های تعبیه شده جدیداً آموزش داده شده است). ما شش الگوریتم گوناگون را بررسی می‌کنیم که براساس تحقیق قبلی انتخاب‌شده‌اند (وینر ۲۰۱۶): Naïve-Bayes، رندوم فارستس (با درختان= ۱۰، ۱۰۰ و ۱۰۰۰)، لاجستیک رگرشن، دسیژن تریز، ساپورت وکتور ماشینز و دیپ لرنینگ (SVM). ما از یکSVM چند طبقه‌ای با یک راهبرد یکی علیه یکی استفاده می‌کنیم. ما ابرپارامترهای طبقه‌بندی کننده SVM را با یک هسته رادیال بهینه‌سازی می‌کنیم. C عدد ۲ است و گاما ۰/۰۱. همچنین مدل دیپ لرنینگ را بهینه می کنمی و برای عملکرد فعال‌سازی، عملکرد بهینه‌سازی، سرعت یادگیری، و تعدادی از دوره‌ها آزمایش می‌کنیم. نتایج ازجمله ابرپارامترهای بهینه‌سازی شده در جدول شماره ۲ آمده‌اند.

تمامی شش الگوریتم به‌خوبی عمل می‌کنند و دقت عملکرد آن‌ها از ۶۱/۲۳ درصد تا ۷۲/۱۷ درصد متغیر است. دو مورد که بالاترین عملکرد را دارند SVM و دیپ لرنینگ هستند (با استفاده از فقط یا معماری چندلایه کم‌عمق گیرنده) که دقت SVM ۷۲/۱۷ درصد است و به میزان ۱/۰۳ درصد بهتر از دیپ لرنینگ عمل می‌کند. به این شکل برخلاف انتظارات اولیه ما، از SVM برای طبقه‌بندی کننده استفاده می‌کنیم. این موضوع در راستای کار انجام‌شده توسط مک آونی و همکاران است که یافتند که SVM در مورد یافتن و رهگیری سخنان نفرت پراکنانه بهتر از دیپ لرنینگ عمل می‌کند. ابرپارامترهای SVM برای به حداکثر رساندن درجه

تعمیم‌پذیری تنظیم می‌شوند (یعنی C و ارزش‌های گامای پایین) که طبقه‌بندی کننده را برای استفاده در شرایط و زمینه‌های تجربی گوناگون بسیار مناسب می‌کند.

ارزیابی عملکرد و محدودیت‌ها

طبقه‌بندی کننده نهایی شامل مدل شماره ۷ اجراشده با یک SVM تنظیم‌شده است. برای ارزیابی عملکرد آن ما طبقه‌بندی کننده را در آموزش مجموعه داده‌ها اعتبارسنجی متقابل می‌کنیم (n= ۱۳۴۱ توییت) با استفاده از یک طبقه‌بندی ده لایه. نتایج در جدول شماره ۳ آمده است.

برای دقت عمل، نمرات بازخوانی و دقت و F1 ما از راهبرد تجمیع کلان شرح داده شده توسط سوکولووا و لاپالمه استفاده می‌کنیم که طبق آن ارزش‌ها برای هر طبقه محاسبه می‌شوند و سپس میانگین توافق پیش از طبقه‌بندی اندازه‌گیری می‌شود، درحالی‌که با هر طبقه به‌طور مساوی برخورد می‌شود. طبقه‌بندی کننده ازنظر بازخوانی و دقت عمل مشابه عمل می‌کند (به ترتیب ۰/۷۴۱ و ۰/۷۳۹) و از اینرو نمره F1 آن قابل‌مقایسه است (۰/۷۴). دقت متوازن ۰/۸۰۷ است. این نتایج نشان می‌دهد که طبقه‌بندی کننده در ایجاد توازن نیاز به شناسایی موارد مربوطه با به حداقل رساندن طبقه‌بندی‌های غلط خوب عمل می‌کند.

یک‌صد توییت به‌طور تصادفی از توییت های مورد استناد به هر یک از سه گروه نمونه‌برداری شد (اسلام‌هراسی غیر ضعیف و قوی) تا یک مجموعه داده‌ها از ۳۰۰ توییت مشاهده نشده و طبقه‌بندی‌شده به‌صورت خودکار ایجاد شود. همان سه تفسیرکننده این ۳۰۰ مورد را تفسیر می‌کند. مثل قبل ما تصمیم اکثریت را برای تصمیم‌گیری در مورد تفسیر در نظر می‌گیریم (در ۹۵٪ موارد کلیه سه تفسیرکننده‌ها موافق هم هستند). عملکرد نتیجه طبقه‌بندی کننده در این مورد در جدول ۴ آمده است.

طبقه‌بندی کننده در تمام معیارهای موجود در ۳۰۰ توییت دیده نشده، با دقت ۷۷٫۳٪عملکرد بهتری دارد.

بهتر کردن عملکرد و ثبات نتایج قدرت رویکرد ما و تعمیم‌پذیری آن را نشان می‌دهد که به‌احتمال‌زیاد به دلیل انتخاب ما از ویژگی‌های ورودی با آگاهی نظری است. مهم‌تر اینکه این نتایج نشان می‌دهد که طبقه‌بندی برای پیاده‌سازی در تحقیقات تجربی مناسب است همان‌طور که دقت بسیار بالاتر از حداقل ۷۰٪ توصیه‌شده توسط ون رایزبرگن (ون رایزبرگن، ۱۹۷۹) است. طبقه‌بندی‌های طبقه‌بندی کننده در جدول ۵ نشان داده شده است.

طبقه‌بندی کننده در تشخیص تفاوت غیر اسلام‌هراسی از اسلام‌هراسی قوی عملکرد خوبی دارد. با این حال، با تمایز توییت‌های ضعیف از قوی و غیر اسلام‌هراسی در کشاکش است. به‌عنوان‌مثال، از ۱۰۰ توییت با برچسب اسلام‌هراسی قوی، در واقع ۲۳ مورد ضعیف است. به همین ترتیب، از ۱۰۰ توییت با برچسب اسلام‌هراسی ضعیف، ۲۲ مورد در واقع غیر اسلام-هراسی است.

بررسی کیفی ۳۰۰ توییت دیده نشده نشان می‌دهد که در بسیاری از موارد طبقه‌بندی نامناسب توییت های غیر اسلام‌هراسی، نفرت و تعصب علیه گروه‌های دیگری غیر از مسلمانان، مانند مهاجران را ابراز می‌کند. این بدان معنی است که آن‌ها بسیاری از سیگنال‌های مشابه، مانند استفاده از اصطلاحات توهین‌آمیز و جملات تحقیرآمیز، مانند «وحشی ...» یا جملاتی، مانند «من متنفرم ...» را استفاده می‌کنند. برخی نیز درباره مسلمانان و اعمال اسلامی بحث می‌کنند اما بدون ابراز هرگونه منفی گرایی. در این موارد، توییت های ضعیف و غیر اسلام-هراسی سیگنال‌های مشابهی را نشان می‌دهد، مانند اصطلاحات مربوط به مساجد، مسلمانان و نماز. در واقع، مقوله میانی اسلام‌هراسی ضعیف باید با بیشترین تنوع درون گروه و بیشترین سیگنال‌های مختلط قانع و خشنود باشد و این کار جداکردن آن از دو مورد دیگر را دشوارتر می‌کند. این کار بازتاب چالش‌های مشابه در کارهای محاسباتی قبلی است مانند عدم توانایی اکثر طبقه‌بندی کننده‌ها در بررسی نمونه‌های طنز و شوخ‌طبعی و یا به‌صراحت طراحی نقش کاربران و زمینه‌های اجتماعی (اشمیت و ویگند، ۲۰۱۷). در آثار آینده، نوآوری‌های اخیر در زبان طبیعی را می‌توان برای پرداختن به این خطاهای طبقه‌بندی از طریق تعبیه کلمه متنی استفاده کرد. کلمات متنی یک ارائه برداری منحصربه‌فرد برای هر شرایط و زمینه منحصربه‌فرد که هر اصطلاح در آن استفاده می‌شود ایجاد می‌کند. برای مثال، استفاده از اصطلاح «مسلمان» در یک شرایط تهاجمی، مانند «لعنت به مسلمانان ...»، ارائه بردار دیگری دارد و به همین ترتیب می‌تواند از استفاده از اصطلاح «مسلمان» در شرایط و فضای مثبت‌تر یا خنثی مانند «مسلمانان به جامعه کمک می‌کنند.... » جدا شود. بردارهای منحصربه‌فرد برای هر اصطلاح می‌تواند یا به عنوان ویژگی‌های ورودی یعنی خوشه‌ای شده برای شناسایی تفاوت‌های معنایی گسترده استفاده می‌شود (که می‌تواند به عنوان ورودی استفاده می‌شود)، یا به عنوان یک لایه ورودی در پیوند با یک تعبیه غیر زمینه‌ای استفاده می‌شود، مانند مدل GloVe که در اینجا استفاده می‌شود (دولین، چانگ، لی و توتانوا، ۲۰۱۸). اجرای اینگونه مدل‌ها اگر با مجموعه

داده‌های کوچک استفاده شود می‌تواند ریسک تناسب بیش‌ازحد را بالا ببرد، و به همین ترتیب نسبت به کار حاضر حجم بیشتری از داده‌ها نیاز به تفسیر دارند.

کاربرد

برای نشان دادن کاربردهای بالقوه طبقه‌بندی کننده خود برای تحقیق علمی اجتماعی ما آن را در مورد یک مجموعه داده دیده نشده از ۷۳ هزار و ۳۱۱ توییت تولید شده توسط ۴۵ حساب توییتری راست افراطی در سال ۲۰۱۷ به کار بردیم. نتایج در شکل شماره ۱ زیر نشان داده شده‌اند.

نتیجه‌گیری

در مقاله حاضر، ما تلاش کردیم تحقیقی با هدف‌های زیر را انجام دهیم و برای شروع: یک طبقه‌بندی کننده جهت رهگیری و یافتن محتوای اسلام‌هراسی در رسانه‌های اجتماعی ایجاد کنیم که بین نقاط قوت گوناگون اسلام‌هراسی تمایز قائل شود. طبقه‌بندی کننده سخنان نفرت پراکنانه اسلام‌هراسی چند گروهی که ما توضیح می‌دهیم گامی به جلو در جهت توسعه ابزارهای کمی و نظری برای تحقیق در مورد اسلام‌هراسی در رسانه‌های اجتماعی است و هدف این تحقیق را برآورده می‌کند. می‌توان از آن برای تحلیل‌های با مقیاس بالای کمی در نفرت پراکنی اینترنتی تا بررسی و تحقیق عناوین جدید و پاسخ به پرسش‌های سرگشاده در نوشته‌های موجود استفاده کرد نظیر شناخت دینامیک‌های زمانی نفرت پراکنی که به‌طور مختصر در بخش ۳,۶ به آن پرداخته‌شده است. خط طبقه‌بندی ارائه شده ما همچنین با (۱) بررسی محتوای اسلام‌هراسی از دیگر پلتفرم‌های رسانه‌های اجتماعی نظیر فیس‌بوک یا گب و (۲) طبقه‌بندی و مطالعه دیگر شکل‌های نفرت پراکنی آنلاین نظیر زن‌ستیزی، نژادپرستی و ضد یهودیت مربوط است. اما همچنین مراقب هستیم که کار بیشتری باید انجام شود به‌خصوص برای تمایز ظریف‌تر بین نقاط قوت گوناگون و اینکه یافتن و رهگیری نفرت پراکنی یک حوزه مستمر در امر تحقیق است که لازم است مرتباً مورد بازبینی قرار گیرد همان‌گونه که ماهیت توهین آنلاین تغییر می‌کند. این اثر همچنین خوراک بحث‌های گسترده‌تر در حوزه علوم اجتماعی در خصوص کاربرد ابزارهای محاسباتی خودکار با ارائه نمونه کاری از یک ابزار است که برای یک کار دقیق و چالش‌برانگیز به‌طور اختصاصی تهیه‌شده است. ما از آثار آینده که ابزار شرح داده شده در اینجا را بسط و توسعه دهد استقبال می‌کنیم.