ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال هشتم)
شماره ۳۲ / زمستان ۱۴۰۲ / ص ۱۸- ۱

ارائه مدلی مبتنی بر تحلیل احساسات برای حذف هرزنامه‌ها در شبکه‌های اجتماعی

محمد میراحمدی[۱]، محبوبه شمسی[۲]، عبدالرضا رسولی کناری[۳]

[۱] کارشناس ارشد، دانشکده مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه صنعتی قم، قم، ایران.
[۲] دانشیار دانشکده مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه صنعتی قم، قم، ایران.
[۳] استادیار دانشکده مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه صنعتی قم، قم، ایران.

نام نویسنده مسئول:
محمد میراحمدی

تاریخ دریافت: ۱۴۰۲/۱۰/۰۴
تاریخ پذیرش: ۱۴۰۲/۱۲/۱۹

چکیده

هرزنامه یا هرزنامه‌های الکترونیکی در اصطلاحات علم کامپیوتر به ارسال یا دریافت پیام‌های ناخواسته الکترونیکی با استفاده از پست الکترونیکی، پیام‌رسان آنی، وبلاگ‌ها، گروه‌های خبری، شبکه‌های اجتماعی، جستجوی وب، تلفن‌های همراه و غیره اشاره دارد. برای شناسایی هرزنامه در سطح توییت، اغلب ویژگی‌های تعریف شده‌ای وجود دارد و الگوریتم‌های یادگیری ماشین مناسب بکار برده می‌شوند. تنها استفاده از ویژگی‌های حساب کاربری برای تشخیص می‌تواند ضعف‌هایی داشته باشد. لذا در روش پیشنهادی ویژگی‌های احساسی نیز مورد استفاده قرار گرفته و از یک روش تلفیقی به منظور هم افزایی نتایج حاصل از چند روش و دستیابی به نتایج دقیق‌تر استفاده شده است. در این پژوهش از مدل مبتنی بر ویژگی به همراه یک مدل مبتنی بر تحلیل احساسات در ترکیب با الگوریتم‌های یادگیری ماشین استفاده شد. بهبود عملکرد روش پیشنهادی به دلیل استفاده از ویژگی‌های احساسی در تلفیق با ویژگی‌های در سطح کاربر در روش پیشنهادی می‌باشد. انتخاب ویژگی‌های احساسی پیشنهادی نسبت به ویژگی‌های کلاسیک مبتنی بر کاربر و یا متن توییت که در مقاله پایه بررسی شده از کارایی بالاتری برخوردار بود. حداکثر دقت در تحقیق ذکر شده ۹۳ درصد ذکر شده است درحالی‌که در روش پیشنهادی و با استفاده از شبکه عصبی مصنوعی که در این تحقیق بکار گرفته شد، دقت تا ۹۸ درصد برآورد گردید.

واژگان کلیدی: هرزنامه، اسپم، تحلیل احساسات، توییت، یادگیری ماشین، شبکه عصبی مصنوعی.

مجله‌ی علمی پژوهش در علوم رایانه، شماره ۳۲، زمستان ۱۴۰۲، ص ۱۸- ۱

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

شبکه‌های اجتماعی به مجموعه‌ای از افراد گفته می‌شود که به‌صورت گروهی با یکدیگر ارتباط داشته و مواردی مانند اطلاعات، نیازمندی‌ها، فعالیت‌ها و افکار خود را به‌اشتراک می‌گذارند. امروزه شبکه‌های اجتماعی برخط، ابزار بسیار محبوبی برای همکاری و ارتباط به شمار می‌روند که میلیون‌ها نفر از کاربران اینترنت را به‌خود جلب کرده‌اند. شبکه‌های اجتماعی برخط مانند فیسبوک، لینکدین، توییتر و غیره از پرطرفدارترین این برنامه‌ها به‌حساب می‌آیند. این شبکه‌ها، به‌خصوص آن‌هایی که کاربردهای معمولی و غیرتجاری دارند، مکان‌هایی در دنیای مجازی هستند که مردم در آن‌ها خود را به‌طور خلاصه معرفی کرده و امکان برقراری ارتباط بین خود و هم‌فکرانشان را در زمینه‌های مختلف مورد علاقه فراهم می‌کنند. به‌نظر می‌رسد شبکه‌های اجتماعی اینترنتی، در آینده بیش از این هم اهمیت پیدا می‌کنند. این شبکه‌ها روزبه‌روز محبوبتر می‌شوند. با شبکه‌های اجتماعی، دیگر افراد برای پیداکردن هم‌فکران خود در موارد گوناگون تنها نیستند.

همان‌طور که بیان شد، این روزها استفاده از شبکه‌های اجتماعی بسیار افزایش یافته است. افراد مختلف با تحصیلات، سن، جنسیت و شغل‌های مختلف در این شبکه‌ها عضو هستند و کاربران، روزانه میلیون‌ها بار به انتشار مطالب مختلف اقدام می‌کنند. اما زندگی در چنین فضایی آدابی دارد که باید به آن‌ها توجه ویژه‌ای نمود تا از مشکلات جدی جلوگیری شود. میلیون‌ها نفر در سراسر دنیا به‌منظور برقراری ارتباط با دوستان، ملاقات افراد جدید، ایجاد ارتباطات کاری با همکاران و غیره از شبکه‌های اجتماعی استفاده می‌کنند(Wang, 2010).

درحال‌حاضر برمبنای آمار دریافتی بیش از ۶۰۰ میلیون نفر کاربر شبکه اجتماعی توییتر بوده و بالغ بر ۴۰۰ میلیون توییت در روز ارسال می‌کنند و بیش از ۱۶ میلیون جستجو در آن انجام می‌پذیرد. یکی از ویژگی‌های توییتر، محدود بودن به تایپ تنها ۱۴۰ کاراکتر است. در توییتر علاوه بر افزودن متن می‌توان فیلم، عکس و صدا را نیز ارسال کرد. توییتر تا پایان اوت سال ۲۰۱۱ به ۱۱ زبان زنده دنیا ترجمه شده بود لیکن به‌منظور افزایش زبان‌های این شبکه اجتماعی درحال‌حاضر تیمی متشکل از ۲۰۰ هزار نفر در حال ترجمه آن به سایر زبان‌های پرطرفدار می‌باشند. همان‌طور که بیان شد، امروزه هرزنامه باعث ایجاد مشکلات روزافزون برای شبکه‌های اجتماعی شده‌است. برای مثال، تحقیقات نشان داده است که حدود ۴۰٪ از حساب‌های کاربری فیس‌بوک و همچنین حدود ۸٪ از پست‌های ارسال‌شده در این شبکه، هرزنامه هستند. باتوجه به ورود فزاینده هرزنامه در شبکه‌های اجتماعی، موفقیت ابزارهای جستجو و کاوش در زمان واقعی، به توانایی تشخیص پست‌های ارزشمند از هرزنامه بستگی دارد. فیس‌بوک و توییتر برای کاربران خود چند راه به‌منظور گزارش هرزنامه فراهم آورده‌اند (Wang, 2010; Wang et al, 2011).

سیستم ایمنی مورد استفاده در این شبکه‌ها دارای دو مزیت در رویارویی با مهاجمین است: بازخورد کاربر و دانش جهانی. بازخورد کاربر، خود شامل دو روش صریح و ضمنی است. بازخورد صریح، شامل علامت‌گذاری مورد یافت‌شده به‌عنوان هرزنامه و یا گزارش کاربر موردنظر بوده و بازخورد ضمنی، شامل پاک کردن یک پست و یا رد درخواست یک کاربر می‌باشد. هردوی این بازخوردها در مسئله دفاع، ارزشمند و مهم هستند. علاوه‌بر بازخورد کاربر، سیستم دارای دانشی در رابطه با الگوهای کلی مبنی بر این‌که چه چیزی عادی و چه چیزی غیرعادی است می‌باشد که مبتنی بر شناخت، خوشه‌بندی و جمع‌آوری ویژگی‌های ناهنجاری است. به‌طورکلی، سیستم از این روش‌ها به‌منظور تشخیص پست‌های هرزنامه و واکنش به آن استفاده می‌کند.

همان‌طور که بیان شد، یکی از مهم‌ترین فعالیت‌های کاربران در شبکه‌های اجتماعی، ارسال و همچنین مشاهده پست‌ها می‌باشد. این پست‌ها به‌طورکلی می‌تواند شامل مواردی همچون متن، لینک‌های آدرس، عکس و یا ویدیو باشد. اما متاسفانه هرزنامه سازها نیز از طریق ارسال پست می‌تواند به اهداف خود نظیر دسترسی به اطلاعات شخصی کاربران، انتشار ویروس وغیره دست یابند. پس می‌توان پست‌ها را به دو دسته هرزنامه و غیرهرزنامه دسته‌بندی کرد.

با توجه به گسترش روزافزون محبوبیت شبکه‌های اجتماعی، هرزنامه‌ها نیز این بستر را برای گسترش محتوای خود، هدف قرار می‌دهند. توییتر یکی از محبوبترین شبکه‌های اجتماعی است که در آن کاربران با موضوعات مختلف مباحث را مطرح کرده و با هم ارتباط برقرار می‌کنند. اکثر روش‌های فیلتر کردن هرزنامه در توییتر بر شناسایی هرزنامه‌گرها (افرادی که هرزنامه منتشر می‌کنند) و مسدود کردن آن‌ها تمرکز دارند. با این حال، هرزنامه‌گرها می توانند یک حساب کاربری جدید ایجاد کرده و دوباره

ISSN: 2745-3673
http://www.Rcsj.ir

هرزنامه جدید ارسال کنند. در سال ۲۰۱۳، توییتر یکی از ده وبسایت برتر در فهرست محبوب‌ترین وبگاه‌ها اعلام شد و همچنین عنوان پیامک اینترنتی به آن داده شده است (Top Sites, Alexa Internet, 2019).

از سال ۲۰۱۸، توییتر ماهانه بیش از ۳۲۱ میلیون کاربر فعال دارد (USA Today, 2013). این حجم عظیم کاربر محل جذابی برای تولیدکنندگان هرزنامه می‌باشد تا به شکار قربانیان خود بپردازند. اگرچه تولید و انتشار هرزنامه برای تولیدکنندگان آن‌ها بسیار پرهزینه می باشد، اما روش‌های جلوگیری از انتشار آن‌ها برای شرکت‌های میزبان بسیار پرهزینه‌تر است. بر اساس برآورد قوه مقننه آمریکا هزینه هرزنامه در ایالات متحده بالغ بر ۱۳ میلیارد دلار در سال ۲۰۰۷ بوده که شامل پایین آمدن کارایی، اتلاف تجهیزات و نیروی کار لازم بوده است (Spamlaws, 2013). تاثیرات مالی مستقیم هرزنامه نیز شامل اضافه بار بر سیستم‌های کامپیوتری و منابع شبکه، اتلاف زمان و منابع انسانی است. به علاوه هرزنامه از چندین بعد دارای هزینه است. این هزینه در مورد شرکتی همچون توییتر با میلیون‌ها کاربر از اهمیت بیشتری برخوردار است. بنابراین برای شناسایی هرزنامه‌ها در سطح توییت، نیاز به تکنیک‌های تشخیص هرزنامه قوی وجود دارد. این نوع تکنیک‌ها می‌توانند به صورت بلافاصله از هرزنامه جلوگیری کنند. برای شناسایی هرزنامه در سطح توییت، اغلب ویژگی‌هایی تعریف شده و الگوریتم‌های یادگیری ماشین مناسب بر روی آن‌ها اعمال می‌شود اما به تازگی، روش‌های یادگیری عمیق نتایج موثری در کاربرد پردازش زبان طبیعی نشان داده‌اند.

روش‌های تشخیص هرزنامه مبتنی بر حساب براساس ویژگی‌های (یا ترکیبی از آن‌ها) حساب است. این روش در دیگر شبکه‌های اجتماعی نیز متداول است و به طور موثری حساب‌های کاربری هرزنامه از غیرهرزنامه را تشخیص می‌دهد. تمرکز این روش بر اطلاعات حساب کاربری متمرکز است. به عنوان مثال تعداد دنبال‌کنندگان و دنبال‌شوندگان در حساب‌های عادی بسیار بیشتر از حساب‌های هرزنامه می‌باشد. به عنوان مثالی دیگر، طول عمر یک حساب هرزنامه به مراتب کمتر از یک حساب عادی می‌باشد. ویژگی مهم دیگر Reputation است که در حساب‌های هرزنامه و غیرهرزنامه متفاوت است. ویژگی Reputation در یک حساب تولیدکننده هرزنامه ۱۰۰٪ یا بسیار کم است، درحالی‌که این مقدار در یک حساب عادی چیزی در حدود ۳۰٪ تا ۰۹٪ می‌باشد. این فاکتور در تشخیص حساب هرزنامه از غیر هرزنامه بسیار کارآمد است. اگرچه این روش دارای قدرت تشخیص بالایی می‌باشد اما حساب‌های تولیدکننده هرزنامه‌ای نیز وجود دارند که در موارد استثناء دارای تعداد دنبال‌شوندگان زیادی هستند و به این ترتیب الگوریتم در این موارد دچار اشتباه می‌شود. معمولا این روش‌ها همراه با دیگر روش‌ها مورد استفاده قرار می‌گیرند (حقیقی و کرمانی، ۱۴۰۱).

چن و همکاران (۲۰۱۵)، با بررسی ۶ الگوریتم یادگیری ماشین بهترینF-measure را باRandom Forest بدست آورده‌اند. آن‌ها از ویژگی‌هایی نظیر تعداد دنبال‌کننده و تعداد دنبال‌شونده و ویژگی طول عمر حساب استفاده کرده‌اند. یکی از نقاط ضعف این روش این است که با بسته شدن حساب کاربری تولیدکننده هرزنامه، او مجددا حساب جدیدی ایجاد می‌کند. همچنین تولیدکنندگان هرزنامه به مرور با دور زدن این ویژگی‌ها می‌توانند روش‌های تشخیص را فریب دهند (Chen et al, 2015).

لی و همکاران (۲۰۱۰) پیشنهاد رویکرد مبتنی بر honeypot برای تشخیص هرزنامه‌ها در شبکه‌های اجتماعی. ویژگی‌هایی که آن‌ها برای تشخیص هرزنامه‌ها در نظر می‌گیرند عبارتند از: طول عمر حساب در توییتر، میانگین توییت‌ها در روز، نسبت تعداد دنبال‌کنندگان و تعداد دنبال‌کنندگان، درصد دوستان دوطرفه، نسبت تعداد آدرس‌های اینترنتی در ۲۰ توییت اخیراً ارسال شده، نسبت تعداد آدرس‌های اینترنتی منحصر به فرد در ۲۰ توییت اخیراً ارسال شده، نسبت تعداد نام کاربری در ۲۰ توییت اخیراً ارسال شده و نسبت تعداد نام کاربری منحصر به فرد در ۲۰ مورد اخیر توییت ارسال کرد (Lee et al, 2010).

روش‌های تشخیص هرزنامه بر اساس توییت بر اساس ویژگی‌ها (یا ترکیبی از آن‌ها) از یک توییت است. تمامی روش‌های مبتنی بر حساب و گراف یک مشکل عمده دارند. پس از مسدود شدن حساب کاربری توسط الگوریتم، تولیدکننده هرزنامه حساب جدیدی ایجاد کرده و به فعالیت خود ادامه می‌دهد. به همین منظور تحقیقات اخیر تمرکز خود را بر روی محتوای خود متن توییت معطوف کرده‌اند. در این روش بدون در نظر گرفتن فرستنده هرزنامه، پس از شناسایی توییت هرزنامه، از انتشار آن جلوگیری می‌شود. با توجه به اینکه هرزنامه‌ها از کلمات و موضوعات مخرب مشابهی استفاده می‌کنند، توییت‌های شامل این کلمات و موضوعات می‌توانند هرزنامه باشند. تکنیک‌های تشخیص در این روش مبتنی بر پردازش زبان‌های طبیعی است.

ISSN: 2745-3673
http://www.Rcsj.ir

ویژگی‌های مبتنی برN-gram نیز به سه دسته Uni-gram وBi-gram وTri-gram تقسیم‌بندی شده است. ۵ طبقه‌بندی بر روی این ویژگی‌ها اعمال می‌شود که عبارتند از الگوریتم‌های Decision Tree, SVM, KNN, Naïve Bayes, و Random Forest. بر طبق این تحقیق نتایج بر روی هر دو مجموعه داده با الگوریتم‌های Random Forest و SVM بهترین خروجی را می‌دهد.

ارزیابی عملکرد روش‌های مبتنی بر یادگیری ماشین برای شناسایی هرزنامه در سطح توییت در پژوهش ژانگ و همکاران (۲۰۱۴) شرح داده شده است(Zhang et al, 2014). بررسی روش‌های تشخیص هرزنامه‌های توییتر با تحلیل مقایسه‌ای در پژوهش وو و همکاران (۲۰۱۸) شرح داده شده است (Wu et al, 2018). داده‌های هرزنامه هشتگ محور توییتر توسط سدهای و سان (۲۰۱۵) ایجاد شده است (Sedhai & Sun, 2015). نویسندگان ۱۴ میلیون توییت جمع‌آوری کرده‌اند و داده‌ها را به عنوان HSpam۱۴ نامگذاری کرده‌اند. سدهای و سان (۲۰۱۷) در تحقیق خود یک چارچوب تشخیص هرزنامه را به دست آوردند. آن‌ها از چهار شناسه سبک وزن برای شناسایی هرزنامه در سطح توییت استفاده کرده اند (Sedhai & Sun, 2017).

یک روش مبتنی بر یادگیری عمیق برای شناسایی هرزنامه در پژوهش آلوم و همکاران (۲۰۲۰) ارائه شده است. در این تحقیق از دو روش مبتنی بر شبکه‌های عصبی پیچشی به‌طور همزمان استفاده شده است. یک شبکه عصبی پیچشی وظیفه طبقه‌بندی متن توییت را بر عهده دارد و یک طبقه‌بندی از ویژگی‌های فرا داده استفاده می‌کند (Alom et al, 2020).

در تحقیق لی و میکولو (۲۰۱۴) بردار توییت را با ترکیب بردار سند توییت (که با مدلسازی بردار پاراگراف بدست می‌آید) ساخته‌اند. این بردارهای ترکیبی به عنوان ویژگی‌های ورودی برای الگوریتم‌های یادگیری ماشین عمل می‌کنند (جنگل‌های تصادفی و شبکه‌های عصبی) (Le & Mikolov, 2014).

در تحقیق انجام شده توسط مادیستی و دسارکار (۲۰۱۸) همچنین از دو ویژگی n-gram نیز استفاده شده است (Madisetty & Desarkar, 2018). با ویژگی Uni-grams و Bi-.grams آن‌ها نتایج مدل پیشنهادی تحقیق خود را با تحقیق وانگ و همکاران (۲۰۱۵) مقایسه کرده‌اند. با وجود زمان اجرای بالاتر، روش پیشنهادی به طور قابل ملاحظه‌ای نتایج تشخیص هرزنامه را بهبود داده است (Wang et al, 2015).

روش‌های مبتنی بر لیست سیاه زیرمجموعه روش‌های مبتنی بر توییت است که برای محافظت از کاربران بسیار کند است زیرا قبل از وارد شدن آدرس‌های اینترنتی مخرب در پایگاه داده تأخیر وجود دارد. شبیه به ویژگی‌های مبتنی بر حساب، ویژگی‌های مبتنی بر توییت به اندازه کافی سبک هستند که می‌توانند برای تشخیص هرزنامه‌های زمان واقعی استفاده شوند که نیاز به تجزیه و تحلیل فوری دارد.

بنابر تحقیقات انجام شده در پژوهش گریر و همکاران (۲۰۱۰) حدود ۰۹٪ کلیک‌ها بر روی آدرس‌های URL هرزنامه در همان دو روز اول انجام می‌گیرد در حالیکه به طور متوسط حدود ۴ روز طول می‌کشد تا URL جدید در لیست سیاه قرار گیرد که تاخیر زیادی در بروز رسانی لیست سیاه می‌باشد و در این زمان هرزنامه به سرعت گسترش می‌یابد و این از نقاط ضعف بزرگ این روش می‌باشد (Grier et al, 2010).

تحقیقات زیادی در این حوزه انجام شده است. به عنوان مثال در پژوهش پاتیل (۲۰۱۸) از درخت تصمیم و ویژگی‌های آماری جهت تشخیص URLهای مخرب استفاده کرده‌اند. برخی ویژگی‌های آن‌ها شامل طول آدرس URL، وجود IP آدرس در Hostname می‌باشد (Patil & Patil, 2018).

روش‌های تشخیص هرزنامه مبتنی بر نمودار از ساختار داده‌های گراف برای مدلسازی ویژگی‌های توییتر به عنوان گره و لبه استفاده می‌کند. مدل‌های داده‌ای گراف راه حل مناسبی برای نمایش داده‌ها هستند، جایی که اطلاعات مربوط به اتصال داده‌ها یا توپولوژی حداقل به اندازه خود داده‌ها اهمیت دارد (Angles & Gutierrez, 2008). بنابراین، نمودارها معمولاً توسط شبکه‌های اجتماعی مانند فیس بوک، توییتر که بیشتر بر اساس کاربران، موضوعات و تعاملات دو طرفه ساخته می‌شوند، استفاده می‌شود (Talha Kabakus & Izzet Baysal, 2017).

این روش، ویژگی‌ها را بر اساس گراف‌های اجتماعی کاربران توییتر بر مبنای روابط بین دنبال‌کنندگان و دنبال‌شوندگان استخراج می‌کند. در این حوزه تحقیقات زیادی در شبکه‌های اجتماعی انجام گرفته است. در روش‌های مبتنی بر گراف که تا

ISSN: 2745-3673
http://www.Rcsj.ir

حدودی به روش‌های مبتنی بر حساب کاربری شباهت دارند، هر حساب کاربری به عنوان یک گره در نظر گرفته می‌شود و درجه ورودی گره نشانگر تعداد دنبال‌کنندگان و درجه خروجی نمایانگر تعداد دنبال‌شوندگان می‌باشد. همچنین ویژگی‌های مبتنی بر همسایگی نیز در این حوزه قرار می‌گیرند. این ویژگی‌ها در طبقه بندهای یادگیری ماشین استفاده می‌شوند.

سونگ و همکاران (۲۰۱۱) فاصله و ارتباط بین فرستنده و یادداشت‌های توییت را استخراج می‌کنند. درحالی‌که فاصله طول کوتاهترین مسیر بین فرستنده توییت و موارد ذکر شده را تعریف می‌کند، اتصال قدرت ارتباط بین کاربران را مشخص می‌کند (Song et al, 2011).

برخلاف ویژگی‌های مبتنی بر حساب و توییت، دستکاری در ویژگی‌های مبتنی بر نمودار سخت است (Gowri & Mohanraj, 2014). با این حال، استخراج این ویژگی‌ها نیاز به تجزیه و تحلیل عمیق در نمودار عظیم و پیچیده توییتر دارد که زمان و منابع زیادی را می‌طلبد. بنابراین، برخلاف ویژگی‌های مبتنی بر حساب و توییت، ویژگی‌های مبتنی بر نمودار برای تشخیص هرزنامه در زمان واقعی به اندازه کافی سبک نیستند.

روش‌های تشخیص هرزنامه ترکیبی از ترکیبی از روش‌های تشخیص هرزنامه که در بخش‌های قبلی توضیح داده شده است استفاده می‌کنند تا بتوانند تشخیص هرزنامه قوی‌تری را ارائه دهند که امکان اسپم را به صورت جامع‌تری بررسی می‌کنند.

وانگ و همکاران یک روش تشخیص هرزنامه را بر اساس حساب، توییت، پردازش زبان طبیعی (NLP)[1] و ویژگی‌های احساس پیشنهاد می‌کند. برخی از ویژگی‌های منحصر به فردی که هنگام تشخیص هرزنامه‌ها استفاده می‌کنند عبارتند از: طول نام نمایه، واژه نامه‌های احساسی به صورت خودکار یا دستی، تعداد علامت تعجب، تعداد علامت سوال، حداکثر طول کلمه، طول کلمه متوسط، تعداد کلمات بزرگ، تعداد فضاهای سفید و بخشی از برچسب‌های گفتار (POS)[2] در هر توییت (Wang et al, 2015).

لی و همکاران یک honeypot اجتماعی ارائه می‌دهند که می‌تواند پروفایل‌های اسپم را از جوامع شبکه‌های اجتماعی جمع آوری کند. هر بار که مهاجم سعی می‌کند با honeypot ارتباط برقرار کند، از یک ربات خودکار برای بازیابی برخی ویژگی‌های قابل مشاهده، مانند تعداد دوستان، از کاربران مخرب استفاده می‌شود. سپس، این مجموعه تجزیه و تحلیل می‌شود تا نمایه ای از هرزنامه‌ها ایجاد کرده و طبقه‌بندی کننده‌های مربوطه را آموزش دهد (Lee et al, 2010).

جدول (۱) خلاصه ای از تحقیقات صورت گرفته روی این موضوع را نشان می‌دهد.

جدول (۱): خلاصه پیشینه تحقیق

روش کلی	ویژگی ها	مزایا و معایب	الگوریتم مورد استفاده	محققین
تشخیص هرزنامه مبتنی بر حساب	تعداد دنبال‌کننده، تعداد دنبال‌شونده، طول عمر حساب	✓سبک بودن برنامه ✓سرعت بالا و زمان پردازش کم ✓کاهش تاخیر اجرا × امکان ایجاد حساب جدید توسط هرزنامه نویسان پس از بسته شدن حساب × امکان فریب روش‌های تشخیص با دور زدن ویژگی‌ها	استفاده از شش الگوریتم مختلف یادگیری ماشین	Chen et al, 2015
	طول عمر حساب، میانگین توییت در روز، نسبت تعداد دنبال‌کننده به دنبال‌شونده، درصد دوستان دو طرفه و ...	✓مشابه پژوهش قبل × مشابه پژوهش قبل	رویکرد مبتنی بر honeypot	Lee et al, 2010

[1] Natural Language Processing
[2] Part Of Speech

ISSN: 2745-3673
http://www.Rcsj.ir

	نسبت دنبال‌کننده به دنبال‌کننده، تعداد توییت ها به نسبت طول عمر حساب، میانگین زمان بین پست ها، تغییرات زمان ارسال، حداکثر ساعات بیکار	✓مشابه پژوهش قبل × مشابه پژوهش قبل × امکان ایجاد خطا به دلیل روش دستی گزارش هرزنامه	شبکه عصبی مصنوعی	Gee & Hakson, 2010
تشخیص هرزنامه مبتنی بر توییت	تعداد دنبال‌شوندگان و دنبال‌کنندگان، طول نام پروفایل کاربر، طول توضیحات پروفایل، عمر اکانت کاربر بر حسب ساعت، تعداد کلمات، تعداد کاراکترها و...	✓سبک بودن برنامه ✓سرعت بالا و زمان پردازش کم ✓کاهش تاخیر اجرا × امکان فریب روش‌های تشخیص با دور زدن ویژگی‌ها	روش‌های ترکیبی مبتنی بر کاربر، مبتنی بر محتوای توییت، و N-gram	Wang et al, 2015
	چهار شناسه سبک وزن برای شناسایی هرزنامه در سطح توییت	✓مشابه پژوهش قبل × مشابه پژوهش قبل	الگوریتم ژنتیک	Zhang et al, Wu et 2014; al, 2018
	تعداد دنبال‌شوندگان و دنبال‌کنندگان، طول نام پروفایل کاربر	✓مشابه پژوهش قبل × مشابه پژوهش قبل	شبکه عصبی پیچشی	Alom et al, 2020
	ترکیب بردار توییت با بردار سند توییت (که با مدلسازی بردار پاراگراف بدست می‌آید)	✓مشابه پژوهش قبل × مشابه پژوهش قبل	شبکه عصبی مصنوعی	Le & Mikolov, 2014
	استفاده از ویژگی برداری کلمات هر توییت	✓بهبود قابل ملاحظه نتایج تشخیص هرزنامه × زمان اجرای بالاتر	شبکه عصبی پیچشی	Madisetty & Desarkar, 2018
	طول آدرس URL، وجود IP آدرس در Hostname و ...	× تاخیر زیاد در بروز رسانی لیست سیاه و گسترش سریع هرزنامه در این زمان	درخت تصمیم و ویژگی‌های آماری جهت تشخیص URLهای مخرب	Patil & Patil, 2018
تشخیص هرزنامه بر اساس نمودار	چگالی گراف و میانگین کوتاهترین مسیر	✓راه حل مناسبی برای نمایش داده‌ها ✓سخت بودن دستکاری در ویژگی‌های مبتنی بر نمودار × نیاز به زمان و منابع زیاد به دلیل لزوم تجزیه و تحلیل عمیق در نمودار عظیم و پیچیده توییتر برای استخراج ویژگی‌ها × تشخیص اشتباه حساب های کاربری افراد مشهور به عنوان هرزنامه × گمراهی سیستم تشخیص به‌جهت تطبیق هرزنامه‌نویسان با ویژگی‌های جدید مبتنی بر گراف	روشی مبتنی بر مبتنی بر گراف	Yang et al, 2013

ISSN: 2745-3673
http://www.Rcsj.ir

		× امکان ایجاد حساب جدید توسط هرزنامه‌نویسان پس از بسته شدن حساب		
	استخراج فاصله و ارتباط بین فرستنده و یادداشت‌های توییت	✓مشابه پژوهش قبل × مشابه پژوهش قبل	نمودار ساختار داده‌های گراف	Song et al, 2011
تشخیص هرزنامه ترکیبی	نسبت تعداد درخواست‌های دوستی، تعداد کل توییت‌های کاربر، شباهت توییت‌های ارسال شده توسط کاربر، تعداد توییت‌های ارسال شده توسط کاربر، تعداد دوستان کاربر	✓امکان بررسی جامع تر اسپم‌ها و تشخیص هرزنامه قوی‌تر × سرعت کمتر و زمان بیشتر نسبت به روش‌های مبتنی بر حساب	رویکردی بر اساس ویژگی‌های مبتنی بر حساب و توییت	Stringhini et al, 2010
	تعداد پیوندهای دو طرفه، نسبت پیوندهای دو جهته، مرکزیت، ضریب خوشه بندی در کنار ویژگی‌های توییتی و حساب محور مانند تعداد دنبال‌کنندگان	✓مشابه پژوهش قبل × مشابه پژوهش قبل	یک روش تشخیص هرزنامه توییتر بر اساس ترکیبی از ویژگی‌های مبتنی بر نمودار، توییت و حساب	Yang et al, 2011
	طول نام نمایه، واژه نامه های احساسی به صورت خودکار یا دستی، تعداد علامت تعجب، تعداد علامت سوال، حداکثر طول کلمه، طول کلمه متوسط، تعداد کلمات بزرگ، تعداد فضاهای سفید و بخشی از برچسب های گفتار	✓مشابه پژوهش قبل × مشابه پژوهش قبل	یک روش تشخیص هرزنامه را بر اساس حساب، توییت، پردازش زبان طبیعی	Wang et al, 2015; Tolosana et al, 2020; Majeed et al, 2020
	تعداد دنبال‌شوندگان و دنبال‌کنندگان، طول نام پروفایل کاربر	✓مشابه پژوهش قبل × مشابه پژوهش قبل	یک روش تشخیص هرزنامه را بر اساس حساب، توییت، پردازش زبان طبیعی	Lee et al, 2010

تحلیل احساسات

یکی از زمینه‌های جدید در تحقیقات مبتنی بر متن شبکه‌های اجتماعی استفاده از تحلیل احساسات می‌باشد تجزیه و تحلیل احساسات، که همچنین به‌عنوان افکار اندیشی یا عقیده کاوی نیز خوانده می‌شود، یکی از مهم‌ترین زیرمجموعه‌های پردازش زبان طبیعی می‌باشد که به طور گسترده‌ای در زمینه داده‌کاوی، وب‌کاوی و متن‌کاوی مورد استفاده قرار می‌گیرد. سیستم‌های تحلیل احساسی تقریباً در هر کسب وکار و حوزه اجتماعی به کار گرفته می‌شوند، زیرا عقاید در تمام فعالیت‌های انسانی نقش اساسی دارد و از تأثیرگذارترین رفتارهای ما می‌باشد. اعتقادات و برداشت ما از واقعیت و انتخاب‌هایی که انجام می‌دهیم، تا حد زیادی مشروط به این است که دیگران چگونه دنیا را می‌بینند و ارزیابی می‌کنند. به همین دلیل، زمانی که ما نیاز به تصمیم‌گیری داریم، اغلب به دنبال عقاید دیگران هستیم. این نه تنها برای افراد بلکه برای سازمان‌ها نیز صادق است.

در تحقیقات راسل (۱۹۹۹) یک مدل مدور اثر توصیفی با دو بعد را توسعه داد، خوشایندی/ ناخوشی و برانگیختگی (میزان واکنش پذیری به محرک). ابعاد قطب‌ها به درجه مثبت یامنفی احساس ارجاع دارد بطوری که بعد برانگیختگی به درجه آرامش

ISSN: 2745-3673
http://www.Rcsj.ir

یا هیجان مرتبط است. محدوده هر دو بعد از ۱ (کاملا منفی یا آرام) تا ۹ (کاملا مثبت یا هیجانی) قرار می‌گیرد. در نتیجه اکثر تحقیقات در زمینه تحلیل احساسات به عامل خوشایندی/ناخوشی اختصاص یافتند (Russell, 1999).

تحقیقات بسیاری به رابطه میان احساسات مختلف پرداخته‌اند که برای تحقیقات در مورد احساسات به صورت خاص، مطالعه آن‌ها توصیه می‌گردد (Kuppens & Tuerlincks, 2017). امروزه تحقیقات زیادی بر روی تحلیل احساسات متن انجام گرفته است. از این رو کتابخانه‌های آماده فراوانی جهت تحلیل احساسات ایجاد شده است، با این وجود در زمینه تحلیل احساسات در حوزه تشخصیص هرزنامه در توییتر تحقیقات کمتری صورت گرفته است.

عادل مجید و همکاران (۲۰۲۰) بر روی تشخیص احساس از متن به زبان اردو رومی متمرکز است. یک مجموعه جامع جمله ای توسعه داده شده که از حوزه‌های مختلف جمع شده و آن را با شش کلاس (شاد، غم، خشم، ترس، عشق و خنثی) مختلف حاشیه نویسی می‌کنند. برای استخراج ویژگی از Word2Vec استفاده شده است. برای طبقه‌بندی از الگوریتم‌های مختلف پایه مانند کی-نزدیکترین همسایه، درخت تصمیم، ماشین بردار پشتیبان و جنگل تصادفی را بر روی مجموعه اعمال شد. پس از آزمایش و ارزیابی، بر طبق ادعای این پژوهش مدل ماشین بردار پشتیبان نسبت به بقیه‌ی الگوریتم‌های طبقه‌بندی با دقت ۵۴.۶۹٪ به نتایج بهتری دست یافته است (Majeed et al, 2020).

مقاله گیوالا و پاتل (۲۰۱۸) چهارچوبی کلی برای تشخیص هرزنامه در توییت را ارائه داده است. این چهارچوب که در تمامی تحقیقات تقریبا یکسان است با تغییراتی اندک و با بهره گیری از رویکرد ترکیبی در شکل (۱) نشان داده شده است.

شکل (۱): چهارچوب کلی تشخیص هرزنامه با رویکرد ترکیبی

یادگیری ماشینی[۳] مطالعه الگوریتم‌ها و مدل‌های آماری مورد استفاده سیستم‌های کامپیوتری است که به‌جای استفاده از دستورالعمل‌های واضح، از الگوها و استنباط برای انجام وظایف استفاده می‌کنند (Hastie & Tibshirani, 2009).

[3] Machine learning

ISSN: 2745-3673
http://www.Rcsj.ir

مواد و روش ها

مجموعه داده

در این تحقیق از مجموعه‌داده creaci-2017 برای آزمایشات استفاده شده است. مجموعه داده‌های اصلی شامل ۷ میلیون توییت است و روند جمع‌آوری مجموعه داده‌ها به مدت دو ماه انجام گرفته است.

در این مجموعه داده که مربوط به شبکه اجتماعی توییتر می‌باشد، دو دسته اصلی وجود دارد که شامل کاربران سالم و کاربران ربات می‌باشد. همچنین در دو بخش دیگر از این داده‌ها توییت‌های سالم و اسپم قابل تفکیک می‌باشند.

در این تحقیق این دو دسته مختلف با یکدیگر ترکیب شده‌اند و یک بخش از این داده‌ها برای آموزش دخالت داده شده‌اند. برای این منظور در پایتون از دستورات ساده پایتون مانند pandas، numpy، tqdm، glob، matplotlib برای خواندن دیتاست استفاده شده است. دیتاست مورد استفاده در این پژوهش از لحاظ تعداد و حجم داده نسبتا بزرگ است، لذا نتایج آن می‌تواند قابل استناد باشد. حجم داده دریافت شده در قالب CSV بالغ بر ۴۴۰ مگابایت می‌باشد. بطور مثال این مجموعه داده شامل ۳۴۷۴ کاربر سالم و حدود ۳ میلیون توییت از این کاربران و ۴۹۱۲ کاربر هرزنامه‌نویس (احتمالا ربات) با حدود ۳ میلیون و پانصد هزار توییت اسپم در این مجموعه داده در دسترس می‌باشد. برای هر یک از کاربران (شامل انسان و ربات) ۴۳ ویژگی (فیچر[4]) بصورت خام تعریف شده است

اطلاعات این کاربران و توییت‌ها که همگی به زبان انگلیسی نگارش شده‌اند شامل ID، Name، screenname، followers، friends count، language و... به‌صورت فشرده در این مجموعه داده در دسترس می‌باشد.

بعد از تلفیق تمام داده‌ها، بصورت کلی چهار دسته داده وجود خواهد داشت، شامل کاربرهای سالم، توییت‌های کاربران سالم، کاربران اسپم و توییت‌های کاربران اسپم. ادغام همه این‌ها تعداد کاربرهای سالم و اسپم می‌شود. در مجموع ۸۳۸۶ کاربر و۴۳ فیچر وجود خواهد داشت که مبنای تحلیل‌ها قرار خواهد گرفت.

با توجه به اینکه تمرکز تحقیق بر روی استفاده از NLP و پردازش زبان طبیعی می‌باشد، بایستی هر جا داده متنی هم در دسترس است از آن استفاده نمود. بطور مثال برای تحلیل کاربران، با توجه به اینکه هر کاربر در توییتر یک توصیف (description) دارد، علاوه بر اینکه روی زمان ساخت اکانت و تعداد دنبال‌کننده‌ها باید بررسی صورت پذیرد، می‌توان روی توصیف هر کاربر نیز تحلیل لازم را صورت داد. مثلاً توصیف ربات‌ها چگونه نوشته شده است و کاربران انسانی و حقیقی چگونه توصیف اکانت خود را می‌نویسند. حتی بر روی متن توییت‌ها باید تحلیل صورت پذیرد. احتمالا تعداد محبوبیت هر پست (لایک)، تعداد پاسخ‌ها وگفتگوهای مرتبط با هر نوشته (منشن)، تعداد هشتگ‌های بکار رفته در هر پست موارد موثری در شناسایی هرزنامه‌نویس‌ها خواهد بود؛ همچنین بررسی متنی هر پست نیز می‌تواند موثر واقع شود. ترکیب ویژگی‌های عددی و ویژگی‌های متنی در نهایت می‌تواند موجب بهبود شناسایی هرزنامه‌ها (پست‌های اسپم) گردد.

در یک بخش از این پژوهش با بررسی طول شناسه هر کاربر (screen name) که بعبارتی نقش URL یا ID را دارد، بعنوان یک ویژگی جدید ایجاد شده در مجموعه داده نسبت به شناسایی کاربران غیرانسانی (ربات) اقدام شد. نام این ویژگی در مجموعه داده user_name_length تعریف گردید.

یک ویژگی مهم دیگر مدت زمان حضور یک کاربر در یک رسانه اجتماعی است. احتمالا کسانی که تازه عضو توییتر شده‌اند، شانس بیشتری برای شناسایی بعنوان ربات خواهند داشت. یا احتمالا کسانی که جدیدتر حساب کاربری در توییتر ایجاد نموده‌اند، خیلی کم‌تر توییت می‌کنند و بیشتر اقدام‌شان پسندیدن نوشته‌های دیگران است (لایک کردن پست‌های سایر کاربران)، یا بیشتر گفتگو (منشن) و باز نشر یک نوشته (ری‌توییت) می‌کنند. با شناسایی و تعریف برخی رفتارهای ربات‌گونه می‌توان مسیر شناسایی ربات‌ها و پست‌های اسپم را در این رسانه بهتر تعریف نمود. بطور مثال داشتن تصویر برای یک حساب کاربری شانس ربات بودن را کاهش خواهد داد. بنابراین علاوه بر ۴۳ ستون (فیچر) خام در مجموعه داده، یک‌سری ستون به عنوان ویژگی‌های جدید به مجموعه داده اضافه می‌گردد. برخی ستون‌ها هم که بنظر حاوی ویژگی‌های مهمی می‌باشند را با محاسبات ساده ریاضی بعنوان یک ویژگی جدید به مجموعه داده اضافه خواهیم نمود بطور مثال سابقه ایجاد حساب کاربری برحسب روز (که از

[4] Feature

ISSN: 2745-3673
http://www.Rcsj.ir

تفریق زمان حال از تاریخ ایجاد حساب بدست می‌آید). برخی ویژگی‌ها هم همانطور که اشاره شد با استفاده از بررسی متون قبلی و همچنین بصورت ابتکاری در طی تحقیق به مجموعه داده افزوده شد (بطور مثال طول رشته‌ی screen_name یا تعداد اعداد بکاررفته در screen_name).

تعداد سال‌های فعالیت در رسانه با استفاده از رابطه (۱) بعنوان یک ویژگی جدید به داده‌ها افزوده شد.

[account_age]=all_data[account_age(days)]/365 رابطه (۱)

ویژگی دیگر حاصل تقسیم follower_ccount بر account_age است که بعنوان followers _ growth _rate نامگذاری شد. این ویژگی بسیار ساده و جالبی است که بر اساس آن تلاش شد کابران ربات شناسایی شوند. اینکه یک کاربر چند سال در توییتر بوده و در این چند سال چند فالوور دارد می‌تواند از این جهت حائز اهمیت باشد که معمولاً کسانی که سابقه حساب‌شان کم است اما تعداد دنبال‌کننده‌های (فالوورهای) خیلی بالایی دارند احتمالا یا ربات هستند یا افراد مشهور (سلبریتی) که ناگهان وارد توییتر شده‌اند و رشد حساب‌شان بالا بوده است.

سرعت رشد دنبال نمودن (following) در یک اکانت هم ویژگی بسیار مهم دیگری است که بر اساس رابطه account_age/ friends_count محاسبه گردید.

ویژگی دیگر ابداعی محاسبه محبوبیت یک حساب کاربری است که با نام popularity به مجموعه داده افزوده شد و با رابطه زیر محاسبه گردید (حاصل تقسیم تعداد دنبال‌کننده‌ها بر تعداد دنبال‌کننده‌ها بعلاوه دنبال‌شونده‌ها):

all_data[friends_count] +[followers _count]/all_data[followers _count] [popularity]=all_data

این فاکتور میزان محبوبیت را برای ما مشخص می‌کند. به عنوان مثال یک کاربر ده هزار فالوور دارد ولی فقط ۵ نفر را فالو کرده است، پس احتمالا محبوبیت بالایی دارد. است. هر چقدر فاکتور تعریفی (popularity) به یک نزدیک‌تر باشد، محبوبیت بیشتر است.

ویژگی دیگر موقعیت مکانی (location) است که بر این اساس بررسی می‌شود یک کاربر لوکیشن خود را در حساب تعریف کرده است یا خیر. همانطور که پیش از این در مورد داشتن تصویر برای یک حساب کاربری و محسوب شدن آن بعنوان شاخص انسانی عنوان شد، درج موقعیت مکانی صاحب حساب نیز می‌تواند موید این مطلب باشد.

این ویژگی‌ها در سطح کاربر تعریف گردید. اما بخش دیگر و مهم در شناسایی پست‌های اسپم بررسی متن توییت‌ها است. این موارد براساس ویژگی‌های اکانت است، در ابتدا از این حیث یک کاربر بررسی می‌شود، مثلاً میزان محبوبیت آن تخمین زده می‌شود و سپس وارد فاز بررسی توییت‌ها خواهیم شد. وجه تمایز این تحقیق ترکیب یوزر و پروفایل و توییت‌ها است، یعنی هم پروفایل و هم توییت‌های یک کاربر بررسی می‌گردد.

تعداد unique_mention، unique_URL و... در تحلیل‌ها مورد استفاده قرار گرفت. به عنوان مثال اگر فردی پشت سر هم توییت بگذارد، احتمالاً ربات است اما فاصله زمانی که انسان برای گذاشتن توییت می‌گذارد کمی بیشتر است. همچنین تعداد unique_URL که در یک توییت بکار می‌رود مهم است. (انسان وقتی توییت می‌گذارد، احتمالا تعداد نشانی‌های اینترنتی کمتری استفاده می‌کند اما ربات‌ها در توییت‌ها از تعداد نشانی اینترنتی بیشتری استفاده می‌کنند). یا مثلاً کاربران انسانی تعداد هشتگ‌های کمتری در پست‌هایشان استفاده می‌کنند. یا مثلاً کاربران انسانی هنگامی که توییت می‌گذارد، معمولا لایک می‌گیرند، ریپلای می‌شود، بعضی‌ها پاسخ می‌دهند، ریتوییت می‌کنند، منشن می‌کنند، ولی پست‌هایی که ربات می‌گذارد، معمولا نه لایک می‌شود، نه کسی ریپلای می‌کند و نه منشن می‌شود.

برخی ویژگی‌های استفاده شده در این تحقیق بر اساس تحقیق صورت گرفته توسط Rodríguez-Ruiz و همکاران (۲۰۲۰) است که در جدول (۲) و جدول (۳) نشان داده شده است.

ISSN: 2745-3673
http://www.Rcsj.ir

جدول (۲): ویژگی‌های استفاده شده در تحقیق‌های پیشین (Rodriguez-Ruiz et al, 2020)

ویژگی تعریف شده	شرح
retweets	نسبت بین تعداد ریتوییت و تعداد توییت
replies	نسبت بین تعداد پاسخ‌ها
favoriteC	نسبت بین توییت مورد علاقه و تعداد توییت
hashtah	نسبت بین تعداد هشتگ‌ها و تعداد توییت‌ها
url	نسبت بین تعداد url و تعداد توییت
mentions	نسبت بین تعداد اشاره‌ها و تعداد توییت‌ها
intertime	میانگین ثانیه بین پست‌ها
ffration	نسبت دوستان به دنبال‌کنندگان
favorites	تعداد توییت‌های مورد علاقه در این حساب
listed	تعداد توییت‌های فهرست‌شده در حساب
uniqueHashtags	نسبت بین تعداد هشتگ‌های منحصربه‌فرد و تعداد توییت
uniquementions	نسبت بین تعداد اشاره منحصر به فرد و تعداد توییت
uniqueURL	نسبت بین تعداد urlهای منحصر به فرد و تعداد توییت

جدول (۳): ویژگی‌های استفاده شده در تحقیق حاضر (در سطح توییت و سطح حساب کاربری)

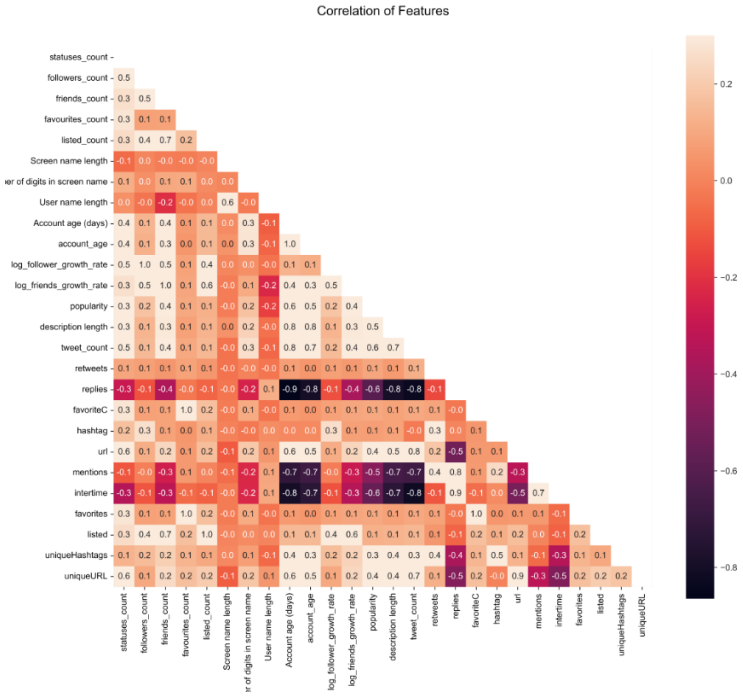
ویژگی تعریف شده	شرح
retweets	نسبت بین تعداد ریتوییت و تعداد توییت
replies	نسبت بین تعداد پاسخ‌ها
favoriteC	نسبت بین توییت مورد علاقه و تعداد توییت
hashtag	نسبت بین تعداد هشتگ‌ها و تعداد توییت‌ها
url	نسبت بین تعداد url و تعداد توییت
mentions	نسبت بین تعداد اشاره‌ها و تعداد توییت‌ها
intertime	میانگین ثانیه بین پست‌ها
Tweet_count	تعداد توییت ها
Description length	طول توضیحات حساب کاربری
Popularity	محبوبیت
Log_friends_growth_rate	نسبت دنبال‌کنندگان به عمر حساب
Log_followers_growth_rate	نسبت دنبال‌شوندگان به عمر حساب
Account_age	عمر حساب (روز)
User name length	طول رشته نام کاربری
Number of digits in screen name	تعداد ارقام بکار رفته در نام نمایشی
Screen name length	تعداد کاراکترهای نام نمایشی
Listed_count	تعداد توییت های ذخیره شده
Favorites count	تعداد پست های ذخیره شده
Friends count	تعدا دوستان
Followers count	تعداد دنبال‌کنندگان
Statuses count	تعداد دنبال‌شوندگان
favorites	تعداد توییت‌های مورد علاقه در این حساب
listed	تعداد توییت‌های فهرست‌شده در حساب

ISSN: 2745-3673
http://www.Rcsj.ir

نسبت بین تعداد هشتگ‌های منحصربه‌فرد و تعداد توییت	uniqueHashtags
نسبت بین تعداد urlهای منحصر به فرد و تعداد توییت	uniqueURL

شاخص بسیار مهم دیگر که در این تحقیق مورد استفاده قرار گرفت تحلیل احساسات (sentiment analysis) می‌باشد. برای هر توییت منتشر شده اقدام به تحلیل شاخص‌های مرتبط با محتوای توییت‌ها گردید. برخی از این شاخص‌ها شامل تعداد ریتوییت، لایک، منشن، فاصله زمانی میان توییت‌ها و. می‌باشد. برای تحلیل احساسات از کتابخانه NLTK در پایتون استفاده شد. کتابخانه NLTK[5] یکی از جامع‌ترین و قدیمی‌ترین کتابخانه‌های پردازش زبان طبیعی در پایتون است. این کتابخانه پایه و استانداردی برای کتابخانه‌های پردازش متن محسوب شده و برای کاربردهای پژوهشی فوق‌العاده است. یکی از ویژگی‌های خوب این کتابخانه امکان اتصال به پیکره‌های مختلف متنی است. که در شناسایی پست‌های اسپم می‌تواند بسیار مفید باشد. خروجی این ابزار شاخصی به نام SA[6] خواهد بود که قطبیت یک توییت را نشان می‌دهد که یکی از اعداد مثبت یک، صفر و یا منفی یک خواهد بود. برای هر کاربر اقدام به محاسبه میانگین این شاخص در میان تمام توییت‌های آن کاربر می‌باشد. عدد مثبت یک نشان دهنده شناسایی آن توییت بعنوان یک توییت غیراسپم، منفی یک به معنای شناسایی آن توییت بعنوان اسپم و صفر نشان‌دهنده بی‌تصمیمی است. ماژول Vader[7] در ابزار NLTK وجود دارد و با توجه به نتایج تحقیق Rodríguez-Ruiz و همکاران (۲۰۲۰) در این تحقیق از این ماژول در پایتون استفاده شد.

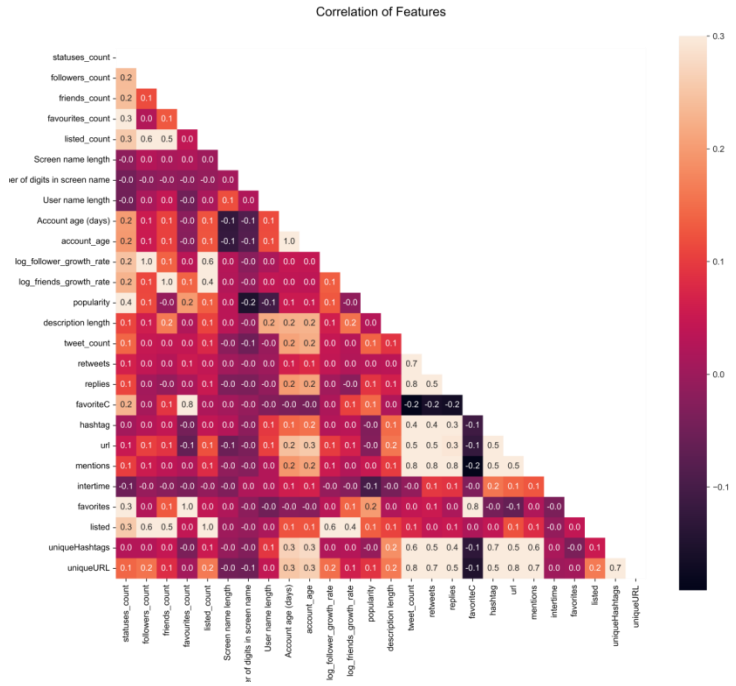
شکل (۲)، (۳) و (۴) میزان همبستگی میان هر جفت از ویژگی‌ها را در میان کاربران ربات، کاربران واقعی و برای کلیه کاربران نشان می‌دهد. میزان ضریب همبستگی بر اساس رابطه پیرسون که نسبت کواریانس به ضرب انحراف معیار هر جفت ویژگی می باشد محاسبه گردید.



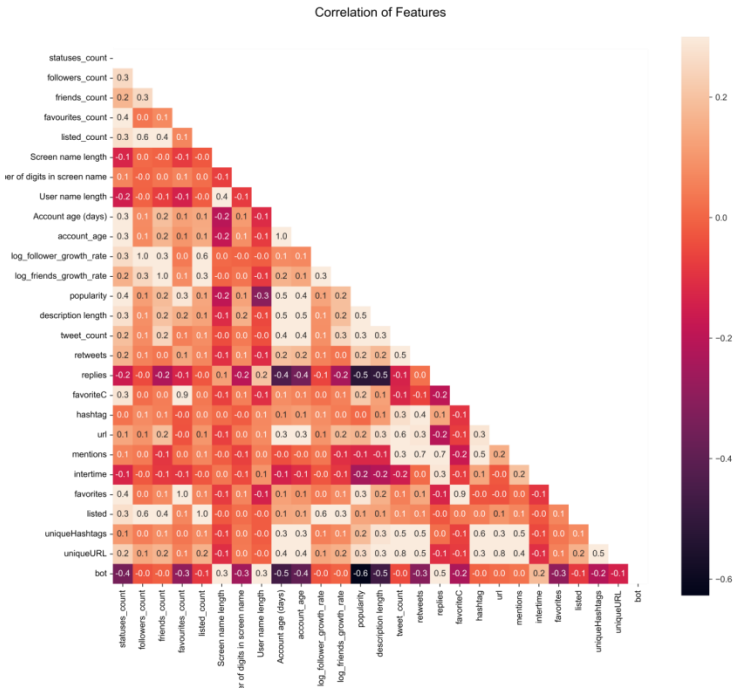
شکل (۲): ضرایب همبستگی میان ویژگی‌های مختلف در میان کاربران هرزنامه‌نویس

[5] Natural Language ToolKit
[6] Sentiment Analysis
[7] Valence Aware Dictionary and Sentiment Reasoner

ISSN: 2745-3673
http://www.Rcsj.ir



شکل (۳): ضرایب همبستگی میان ویژگی‌های مختلف در میان کاربران واقعی



شکل (۴): ضرایب همبستگی میان ویژگی‌های مختلف در میان تمام کاربران

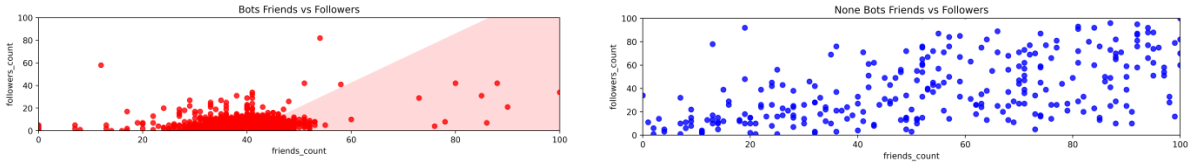
بر اساس نتایج همبستگی‌های میان ویژگی‌های مختلف تعریفی مشاهده می‌شود، بطور مثال همبستگی میان تعداد هشتگ-های بکار رفته با تعداد توییت‌های منشتر شده توسط ربات‌ها صفر و میان کاربران واقعی ۰.۴ می‌باشد. نتایج این تحلیل در دسته‌بندی‌های کاربران و پست‌های منتشره بسیار مفید خواهد بود. بر پایه سعی و خطا حد آستانه[8] ۰.۲ برای جدایش

[8] Treshold

ISSN: 2745-3673
http://www.Rcsj.ir

همبستگی‌های معنی دار از بی‌معنی تعیین گردید. یعنی چنانچه ضریب همبستگی در بازه (+0.2,+1) و یا (-0.2 ,-1) قرار گیرد، دارای همبستگی معنی‌دار خواهد بود.

شکل دیگر نمایش داده‌ها نیز در تحلیل همبستگی مفید می‌باشد. به‌طور مثال در شکل (۵) تفاوت نسبت دنبال‌کننده به دنبال‌شونده را در دو دسته ربات و غیرربات نشان می‌دهد. نقاط قرمز رنگ برای دسته ربات و نقاط آبی رنگ برای دسته غیرربات مشخص شده است.



شکل (۵): تفاوت نسبت دنبال‌کننده به دنبال‌شونده را در دو دسته ربات و غیرربات (نقاط قرمز رنگ برای دسته ربات و نقاط آبی رنگ برای دسته غیرربات می‌باشد)

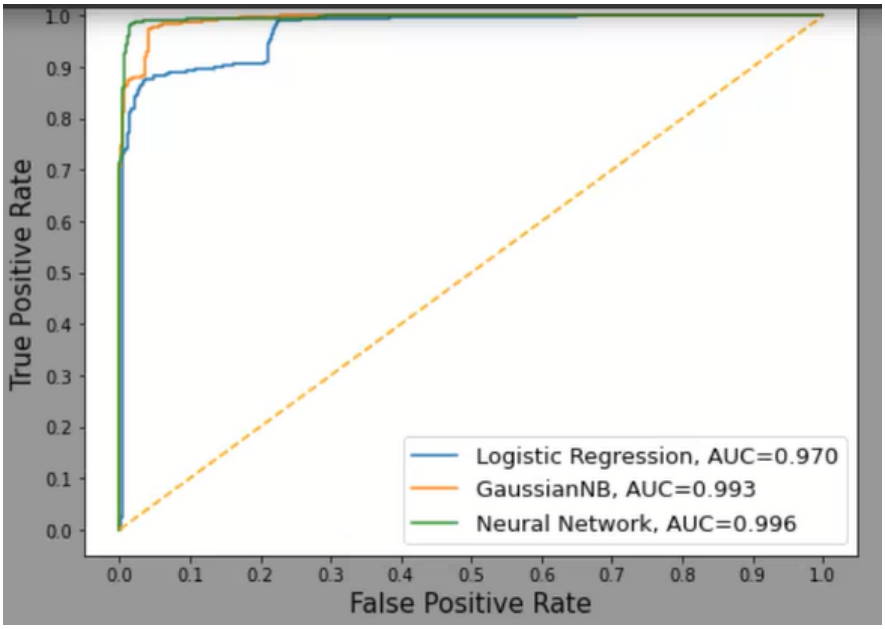
جدول (۴) مقایسه نتایج دسته‌بندی به سه روش مختلف برای داده‌های در دسترس می‌باشد. نشان داده شده است که به ترتیب روش‌های شبکه عصبی مصنوعی، بیزین ساده و رگرسیون لجستیک بهترین عملکرد را در طبقه‌بندی داشته‌اند. Precisionیا صحت به معنای درصدی از پیش‌بینی‌های مدل که مرتبط هستند و Recall یا پوشش اشاره به درصدی از کل پیش‌بینی‌هایی که توسط مدل درست دسته‌بندی‌شده‌اند می‌باشد.

جدول (۴): مقایسه فاکتورهای ارزیابی سه روش پیاده شده در تحقیق

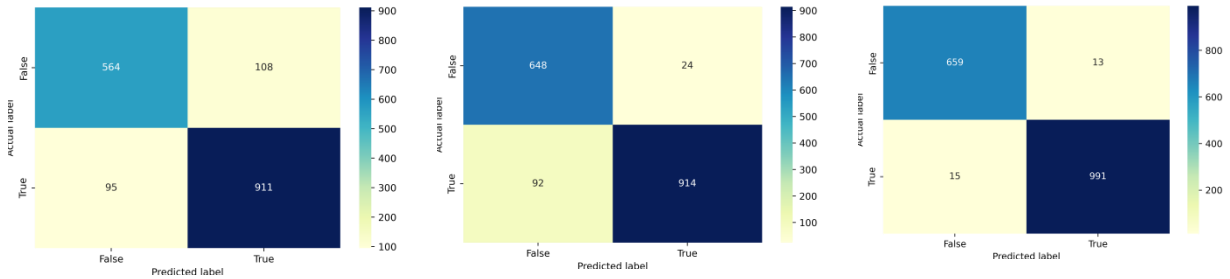
Time (s)	F1	Precision	Recall	Accuracy	روش
0.0007	0.878	0.877	0.878	0.879	رگرسیون لجستیک
0.0013	0.931	0.934	0.931	0.930	بیزین ساده گاوسی
0.1765	0.980	0.983	0.982	0.983	شبکه عصبی مصنوعی

همچنین شکل (۶) نمودار ROC هر سه روش را نشان می‌دهد که کاملا مشخص است عملکرد مدل پیاده‌سازی شده بر پایه شبکه عصبی مصنوعی در دسته‌بندی نسبت به دو روش دیگر مطلوب بوده است. میزان دقت طبقه‌بندی بیش از ۹۸ درصد بوده است که نشان دهنده کارایی روش پیشنهادی بر اساس تلفیق ویژگی‌های درسطح کاربر و ویژگی‌های در سطح توییت و با توجه به تحلیل احساس با استفاده از شبکه عصبی مصنوعی است.

ISSN: 2745-3673
http://www.Rcsj.ir



شکل (۶): نمودار ROC Curve برای هر سه روش پیاده‌سازی شده



رگرسیون لجستیک **بیزین ساده گاوسی** **شبکه عصبی مصنوعی**

شکل (۷): ماتریس ارزیابی طبقه‌بندی برای سه روش پیاده سازی شده

نتیجه‌گیری

در این تحقیق به روش‌های شناسایی هرزنامه‌ها و انواع آن پرداخته شد. هرزنامه یا هرزنامه‌های الکترونیکی در اصطلاحات علم کامپیوتر به ارسال یا دریافت پیام‌های ناخواسته یا بدون درخواست‌شده الکترونیکی با استفاده از پست الکترونیکی، پیام‌رسان آنی، وبلاگ‌ها، گروه‌های خبری، شبکه‌های اجتماعی، جستجوی وب، تلفن‌های همراه و غیره اشاره دارد. هرزنامه امروزه تقریباً در تمام انواع ارتباطات آنلاین اجتناب ناپذیر است و به عنوان مانع بهره وری از محیطی که در آن به نظر می‌رسد؛ شناخته شده است. اقدامات مختلف برای بهبود استحکام رسانه‌های مختلف الکترونیکی در مقابل یک سری از حملات هرزنامه صرف شده است. این اقدامات بیشتر به عنوان تکنیک‌های ضد هرزنامه یا تکنیک‌های مبارزه هرزنامه شناخته شده است. در این تحقیق به به بیان مروری بر تحقیقات انجام شده در حوزه حذف هرزنامه‌ها در شبکه‌های اجتماعی و توییتر پرداخته و روش‌های تشخیص هرزنامه بررسی شد و همچنین تحقیقات زیادی که بر روی تحلیل احساسات از متن انجام گرفته بود مورد بررسی قرار گرفت. این روش‌ها شامل روش‌های تشخیص هرزنامه مبتنی بر حساب، روش‌های تشخیص هرزنامه مبتنی بر توییت، روش‌های تشخیص هرزنامه مبتنی بر نمودار، روش‌های ترکیبی تشخیص هرزنامه، شناسایی هرزنامه‌های چندرسانه ای (Deepfakes) با قابلیت هوش مصنوعی و روش‌های مبتنی بر یادگیری ماشین می‌باشد.

همانند نتایج موجود در مقاله Madisetty & Desarkar (2018)، با ویژگی‌های مقاله پایه و ویژگی‌های روش پیشنهادی برای دقت و معیار F-measure، نتایج طبقه‌بندی بر اساس تلفیق اطلاعات در سطح کاربر و همچنین در سطح توییت بر پایه

ISSN: 2745-3673
http://www.Rcsj.ir

شبکه عصبی مصنوعی عملکرد بهتری دارد. درحالی‌که از جهت فراخوانی و سرعت پردازش، عملکرد الگوریتم رگرسیون لجستیکی بهتر است. با این وجود عملکرد روش پیشنهادی در این تحقیق نسبت به تمامی روش‌های ارائه شده در مقاله Madisetty & Desarkar (2018) بهتر عمل کرده است. این نشان می‌دهد انتخاب ویژگی‌های احساسی پیشنهادی نسبت به ویژگی‌های کلاسیک مبتنی بر کاربر و یا متن توییت که در مقاله پایه بررسی شده از کارایی بالاتری برخوردار است. حداکثر دقت در تحقیق ذکر شده ۹۳ درصد ذکر شده است درحالی‌که در روش پیشنهادی در این تحقیق دقت تا ۹۸ درصد برآورد شده است.

از طرف دیگر فرآیند آموزش تحت تاثیر این مجموعه‌داده نامتعادل قرار گرفت و چون تعداد بیشتری از نمونه‌های آموزش غیراسپم بوده و تعداد نمونه‌های کمتری اسپم هستند، علیرغم نتایج ضعیف طبقه بندهای مبتنی بر ویژگی، مدل‌های ارائه شده برپایه تلفیق اطلاعات کاربر و توییت کارایی قابل قبولی داشته‌اند. بهبود عملکرد روش پیشنهادی به دلیل استفاده از ویژگی‌های احساسی در تلفیق با ویژگی‌های در سطح کاربر در روش پیشنهادی می‌باشد. این نشان می‌دهد ویژگی‌های مورد استفاده در مقاله پایه تاثیرگذاری کمتری نسبت به ترکیب ویژگی‌های احساسی با ویژگی‌های صرفا مبتنی بر کاربر و صرفا مبتنی بر متن دارد. با این حال، وجود ویژگی‌های احساسی باعث افزایش بیشتر زمان اجرای روش پیشنهادی نسبت به مقاله پایه می‌شود. به طور کلی کمترین زمان اجرا متعلق به ویژگی‌های مبتنی بر حساب کاربر می‌باشد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Alom, Z., Carminati, B., & Ferrari, E. (2020). A deep learning model for Twitter spam detection. Online Social Networks and Media, 18, 100079.

[2] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive bayesian anti -spam filtering. arXiv preprint cs/0006013. differentiation for combating link-based Web spam," ACM Trans. Web, vol. 8, no. 3, 2014, Art. no. 15 pp. 37-40.

[3] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." In Lrec, vol. 10, no. 2010, pp. 2200-2204. 2010.

[4] Chen, C., Zhang, J., Xie, Y., Xiang, Y., Zhou, W., Hassan, M. M., & Alrubaian, M. (2015). A performance evaluation of machine learning-based streaming spam tweets detection. IEEE Transactions on Computational social systems, 2(3), 65-76.

[5] Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015, June). 6 million spam tweets: A large ground truth for timely Twitter spam detection. In 2015 IEEE international conference on communications (ICC) (pp. 7065-7070). IEEE.

[6] Chu Z, Gianvecchio S, Wang H, Jajodia S. Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 2012;9(6):811-24.

[7] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv: 1703.04009.

[8] Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.

[9] Dustin Hillard, N-gram Language Modeling Tutorial, Lecture notes courtesy of Prof. Mari Ostendorf. http://ssli.ee.washington.edu/WS07/notes/ngrams.pdf

[10] Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC (Vol. 6, pp. 417-422).

[11] Gheewala, S., & Patel, R. (2018, February). Machine learning based Twitter Spam account detection: a review. In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC) (pp. 79-84). IEEE.

[12] Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010, October). spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security (pp. 27-37).

[13] Gupta, S., Khattar, A., Gogia, A., Kumaraguru, P., & Chakraborty, T. (2018, April). Collective classification of spam campaigners on Twitter: A hierarchical meta-path based approach. In Proceedings of the 2018 World Wide Web Conference (pp. 529 -538).

[14] Jorge Rodríguez-Ruiz, Javier Israel Mata-Sánchez, Raúl Monroy, Octavio Loyola-González, Armando López-Cuevas (2020) A one-class classification approach for bot detection on Twitter, Computers & Security, Volume 91, 101715, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2020.101715.

[15] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint:1408.5882.

[16] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[17] Lakshmana Phaneendra Maguluri; R. Ragupathy; Sita Rama Krishna Buddi; Vamshi Ponugoti Tharun Sai KalimilAdaptive Prediction of Spam Emails: Using Bayesian Inference, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).

[18] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).

[19] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[20] Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in Proc. ICWSM, 2011, pp. 185-192.

[21] Maâli Mnasri, How to train word embeddings using small datasets?. 2019. https://medium.com/opla/how-to-train-word-embeddings-using-small-datasets-9ced58b58fde

ISSN: 2745-3673
http://www.Rcsj.ir

[22] Madisetty, S., & Desarkar, M. S. (2018). A neural network-based ensemble approach for spam detection in Twitter. IEEE Transactions on Computational Social Systems, 5(4), 973-984.
[23] Martinez-Romo, J., & Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40(8), 2992-3000.
[24] Niculescu-Mizil, A., Perlich, C., Swirszcz, G., Sindhwani, V., Liu, Y., Melville, P & Shang, W. X. (2009, December). Winning the KDD cup orange challenge with ensemble selection. In KDD-Cup 2009 Competition (pp. 23-34).
[25] Ohana, B., & Tierney, B. (2009, October). Sentiment classification of reviews using SentiWordNet. In 9th. it & t conference (Vol. 13, pp. 18-30).
[26] oriented spam research. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 223-232).
[27] Patil, D. R., & Patil, J. B. (2018). Malicious URLs detection using decision tree classifiers and majority voting technique. Cybernetics and Information Technologies, 18(1), 11-29.
[28] Perveen, N., Missen, M. M. S., Rasool, Q., & Akhtar, N. (2016). Sentiment based twitter spam detection. International Journal of Advanced Computer Science and Applications (IJACSA), 7(7), 568- 573.
[29] Santos, I., Miñambres-Marcos, I., Laorden, C., Galán-García, P., Santamaría-Ibirika, A., & Bringas, --CISIS' P. G. (2014). Twitter content-based spam filtering. In International Joint Conference SOCO' (pp. 449-458). Springer, Cham. ICEUTE.
[30] Sedhai, S., & Sun, A. (2017). Semi-supervised spam detection in Twitter stream. IEEE Transactions on Computational Social Systems, 5(1), 169-175.
[31] Shen, H., Ma, F., Zhang, X., Zong, L., Liu, X., & Liang, W. (2017). Discovering social spammers from multiple views. Neurocomputing, 225, 49-57.
[32] Töscher, A., Jahrer, M., & Bell, R. M. (2009). The bigchaos solution to the netflix grand prize. Netflix prize documentation, 1-52.
[33] Valerie Niechai, How to Use TF -IDF for SEO, 2019, How to use TF-IDF tools for semantic SEO (link-assistant.com).
[34] Van de Vegte, J. (1990). Feedback Control Systems (2nd ed.). Prentice Hall.
[35] Wang, A. H. (2010, July). Don't follow me: Spam detection in twitter. In 2010 international conference on security and cryptography (SECRYPT) (pp. 1 -10). IEEE.
[36] Wang, B., Zubiaga, A., Liakata, M., & Procter, R. (2015). Making the most of tweet-inherent features for social spam detection on Twitter. arXiv preprint arXiv:1503.07405.
[37] X. Zhang, Y. Feng, H. Shen, and W. Liang, "Differential trust propagation with community discovery for link-based Web spam demotion," in Proc. Int. Conf. Web- Age Inf. Manage. Cham, Switzerland: Springer, 2015, pp. 452-456.
[38] Yang C, Harkreader R, Gu G. Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans InfForensics Secur 2013;8(8):1280-93.