ارائه روشی جهت بهبود تشخیص هرزنامه‌نویسان در شبکه‌های اجتماعی‌آنلاین*

سیما سالاری سروری[۱]، هادی خسروی فارسانی[۲]، محمدرضا خیام‌باشی[۳]

[۱]دانشجوی مقطع کارشناسی ارشد، دانشکده مهندسی کامپیوتر، مؤسسه آموزش عالی صفاهان، اصفهان
s.salari@safahan.ac.ir

[۲]استادیار، دانشکده مهندسی کامپیوتر، دانشگاه شهرکرد ، شهرکرد
khosravi@eng.sku.ac.ir

[۳]عضو هیت علمی، دانشکده مهندسی کامپیوتر ، دانشگاه اصفهان، اصفهان
m.r.khayyambashi@eng.ui.ac.ir

چکیده

امروزه شبکه‌ها و اینترنت به بخشی جدایی‌ناپذیر از زندگی بشر تبدیل شده است. این موضوع در کنار مزایای بسیار، می‌تواند انواع تهدیدات امنیتی و ناهنجاری‌های اجتماعی را نیز به دنبال داشته باشد. رشد روز افزون شبکه‌های اجتماعی آنلاین[۱] به دلیل محبوبیت وکاربری آسان، آنها را به عنوان اصلی‌ترین هدف برای هرزنامه‌نویسان[۲] تبدیل کرده است. در این میان بسیاری از مطالعات موجود از روش‌های یادگیری ماشین[۳] برای شناسایی هرزنامه‌نویس‌ها استفاده نموده‌اند. در حالی که هرزنامه‌نویسان همواره برای فرار از ویژگی های تشخیص موجود، در حال ابداع روش‌های جدید هستند. در این مقاله، با درک عمیق از اثربخشی و مشکلات استفاده از ویژگی‌های یادگیری ماشین برای شناسایی هرزنامه‌نویسان، ویژگی‌های موثرتری طراحی شده و نتایج بدست آمده از آن با کارهای پیشین، مورد ارزیابی قرار می‌گیرد. با توجه به آزمایش‌های صورت گرفته، نشان داده می‌شود که ویژگی‌های طراحی شده جدید، قادر هستند بسیار موثرتر برای تشخیص هرزنامه‌نویسان توییتر مورد استفاده قرار گیرند. در ادامه به طور خاص، از الگوریتم ساختاری (SRank[۴]) به عنوان یک ویژگی موثر به منظور محاسبه شباهت بین کاربران و همسایگانشان بهره برده شده است. با توجه به ارزیابی صورت گرفته، با پایین ماندن حداقل نرخ مثبت کاذب[۵]، سرعت کشف و شناسایی هرزنامه‌نویسان همچنان با استفاده از ویژگی‌های جدید نیز به طور قابل توجهی بالاتر از کارهای موجود است.

کلمات کلیدی

شبکه‌های اجتماعی، گراف اجتماعی، هرزنامه‌نویس، طبقه‌بندی

۱- مقدمه

وب‌سایت‌های شبکه‌های اجتماعی آنلاین (OSNs)، مانند توییتر و فیس‌بوک، امروزه به بخش مهمی از زندگی روزمره بسیاری از مردم تبدیل شده‌اند: از آنجایی که مردم بیشتر از همیشه درگیر این وب سایت‌ها شده‌اند، این وب سایت‌ها به عنوان یک پلت‌فرم برای هرزنامه‌نویسان به منظور توزیع پیام‌های هرزنامه[۶] از جمله لینک‌های مخرب، تبلیغ‌ها، محتوای مستهجن و غیره تبدیل می‌شوند[1]. همه این رفتارهای مخرب ممکن است ضررهای اقتصادی قابل توجهی را در جامعه منجر شود و حتی ممکن است امنیت ملی را تهدید کند. توییتر، یک سرویس میکروبلاگینگ است که در سال ۲۰۰۶ تاسیس شد، و یکی از محبوب‌ترین شبکه‌های اجتماعی آنلاین است که با بیش از ۲۰۰ میلیون حساب کاربری به همراه ۶۵ میلیون بار توییت در روز، سریع‌ترین رشد را در بین آنها داشته است[1]. پیام‌های هرزنامه، فضای ذخیره‌سازی

*این مقاله برگرفته از پایان‌نامه تحصیلات تکمیلی در موسسه آموزش عالی صفاهان است.

وب‌سایت‌ها را به هدر می‌دهند و همچنین کاربران عادی را آزار می‌دهند. از طرفی، سیستم‌های کاربران نیز ممکن است با بازکردن لینک‌های مخرب تعبیه شده در توییت‌ها، آلوده شوند[2,3]. هرزنامه در وب سایت‌های شبکه‌های اجتماعی، مانند هرزنامه ایمیل، یک مشکل محسوب می‌شود. بنابراین، توسعه یک سیستم موثر برای طبقه‌بندی حساب‌های هرزنامه در وب سایت‌های شبکه‌های اجتماعی، توجه زیادی را در سال‌های اخیر به خود جلب کرده است.

اکثر مطالعات موجود، هرزنامه‌نویسان را در وب سایت‌های شبکه‌های اجتماعی با توجه به ویژگی‌های مربوط به پروفایل‌های کاربری و یا محتوایی، با روش‌های یادگیری ماشین شناسایی می‌کنند. ویژگی‌هایی از جمله تعداد فالوورها[۷] و فالوینگ‌ها[۸]، تعداد توییتها[۹] در روز، و غیره[4-7]. برخی از مطالعات، ارتباط بین کاربران را در گراف توصیف می‌کند، و از یک الگوریتم مبتنی بر گراف برای استخراج ویژگی‌ها به منظور تشخیص هرزنامه‌نویسان، استفاده می‌کنند[8,9]. مطالعات انجام شده در [10,11] حساب‌های honey-profile را که توسط هرزنامه‌نویسان فالو می‌شوند، ایجاد کرده و سپس رفتار هرزنامه‌نویسان را تجزیه و تحلیل می‌کنند.

با این حال، هرزنامه‌نویسان توییتر می‌توانند از ویژگی‌های تشخیص موجود، به عنوان مثال با خرید فالورها [8] فرار کنند و یا با استفاده از ابزارهایی شروع به ارسال خودکار توییت‌ها با یک معنای یکسان اما کلمات مختلف نمایند[8]. مانند بسیاری از مطالعات قبلی، این کار نیز به منظور تشخیص هرزنامه‌نویسان در توییتر می‌باشد. توییتر، گزارش داده است که بیش از ۲۰۰ میلیون کاربر دارد[1]، و روش مختص خود را در قالب قوانین توییتر برای شناسایی و تعلیق حساب‌های هرزنامه دارد[12]. هر حسابی که شامل فعالیت‌های غیرطبیعی باشد، به صورت موقت و یا حتی دائمی توسط توییتر به تعلیق می‌افتد. با تمامی این دلایل، همچنان هنوز هم تعدادی از حساب‌های هرزنامه برای یک مدت طولانی زنده مانده‌اند و می‌توانند به فعالیت‌های مخرب خود ادامه دهند. چنین پدیده‌ای به محققان این انگیزه را می‌دهد که روش جدید و موثرتری برای تشخیص هرزنامه‌نویسان توییتر طراحی کنند.

۲- روش پیشنهادی

آن‌چه که در این تحلیل، پایه و اساس محسوب می‌شود، نظریه‌ی گراف‌ها است، که در کنار کاربردهای بیشمار آن، در تحلیل شبکه‌های اجتماعی نیز نقش مهمی ایفا می‌کند.

برای رسیدن به این هدف، علاوه بر طراحی چند ویژگی تشخیص جدید، همچنین از معیار شباهت برای تشخیص هرزنامه‌نویس‌ها در این مدل استفاده می‌شود. یک دیدگاه گراف گونه برای تشخیص هرزنامه‌نویس‌ها از تجزیه و تحلیل کل شبکه تا همسایه کاربران وجود دارد. به این منظور، از یک مجموعه داده که شامل ۱۱۰۰۰حساب توییتر و بیش از ۱ میلیون توییت جمع‌آوری شده از کار [8] استفاده کرده و بر این اساس رفتارهای کاربران، تجزیه و تحلیل می‌شوند. مقایسه نتایج در روش پیشنهادی در مقابل کارهای پیشین، نشان‌دهنده برتری مدل پیشنهادی است.

به طور خلاصه، مراحل روند این مقاله به شرح زیر است:

به منظور درک بهتر، ابتدا اثربخشی وکاستی‌های ویژگی‌های موجود برای تشخیص هرزنامه‌نویسان در شبکه‌های اجتماعی توییتر، به صورت خلاصه بررسی می‌شوند.

در مرحله بعد گراف اجتماعی کاربران، تجزیه و تحلیل شده و میزان شباهت کاربران، از طریق الگوریتم ساختاری SRank[13]، در قالب یک ویژگی مبتنی بر گراف، و همچنین ویژگی نرخ همسایگی مشترک برای کاربران موجود در گراف اجتماعی محاسبه می‌شوند. علاوه بر این، طراحی چند ویژگی ساده ولی در عین حال موثر(نرخ پاسخ–صحیح و نرخ تنوع API) به منظور شناسایی و طبقه‌بندی حساب‌های هرزنامه توییتر با استفاده از مدل‌های داده‌کاوی صورت می‌گیرد. با بررسی نتایج، مشخص می‌شود که نرخ تشخیص[۱۰] با استفاده از ویژگی‌های جدید، به میزان قابل توجهی افزایش یافته است.

۳) در آخر، میزان اثربخشی از کارهای موجود و این مقاله را بررسی نموده، و نتایج حاصل با نتایج روش‌های پیشین، مقایسه می‌شوند.

۳- کارهای مرتبط

محبوبیت شبکه‌های اجتماعی الهام‌بخش بسیاری از مطالعات علمی انجام شده در هر دو بحث شبکه و امنیت می‌باشد. عملکرد سریع سیستم‌ها برای شناسایی فعالیت‌های مخرب در شبکه‌های اجتماعی با تمرکز در شناسایی حساب‌های هرزنامه و پیام‌های هرزنامه می‌باشد که محققان از طریق بهره‌گیری از روش‌های مختلف به نتایج مختلفی دست یافته‌اند: در این راستا، Wang، سیستمی جهت شناسایی هرزنامه بر اساس ویژگی‌های مبتنی بر محتوای پیام و مبتنی بر گراف، کمک گرفته است که ارتباط دنبال‌کنندگان و دوستان در این شبکه با استفاده از مدل گراف اجتماعی مورد بررسی قرار گرفته است[14]. Benevento و همکاران، روش‌های یادگیری ماشین را برای شناسایی هرزنامه‌نویسان در شبکه‌های اجتماعی ویدئویی مانند یوتیوب به کار گرفته‌اند[15]. همچنین آنها در سیستم دیگری، هرزنامه‌نویسان را بر اساس محتوای توییت و ویژگی‌های مبتنی بر کاربر شناسایی نمودند[16]. Gao و همکاران نیز، یک مطالعه بر روی تشخیص و توصیف کمپین‌های هرزنامه اجتماعی ارائه داده‌اند[17]. در سیستم دیگری Yang و همکاران، ۱۰ ویژگی جدید تشخیص از جمله سه ویژگی مبتنی بر گراف، سه ویژگی مبتنی بر همسایه[۱۱]، سه ویژگی مبتنی بر کنترل خودکار[۱۲] و یک ویژگی مبتنی بر زمان‌بندی را برای تشخیص هرزنامه‌نویسان در توییتر طراحی نمودند[8]. همچنین از کارهای پیشین، می‌توان یک رویکرد جدید یکپارچه به نام "Combines" را نام برد که در آن سه الگوریتم یادگیری ماشین با هدف بهبود دقت تشخیص هرزنامه‌نویسان ترکیب شده است[18]. به صورت مشابه، Lee و همکاران، هانی‌پات‌های اجتماعی را به منظور شناسایی هرزنامه‌نویسان در شبکه‌های اجتماعی مستقر نمودند و روش‌های یادگیری ماشین را برای طبقه‌بندی هرزنامه‌نویسان، با طراحی ویژگی‌های خود، مانند ویژگی‌های مربوط به محتوا به کار گرفته‌اند[11].

درمقایسه با کارهای مرتبط موجود، انگیزه تجزیه و تحلیل این است که هرزنامه‌نویسان همواره در حال تکامل هستند. اگرچه بتوانند از بسیاری از ویژگی‌های موجود فرار کنند و برای مدت زمان زیادی زنده بمانند، با این حال تغییر دادن موقعیت خود در گراف اجتماعی، همچنان برای آنها دشوار خواهد بود. این روش، در طراحی ویژگی‌هایی تمرکز دارد که بتواند با توجه به ویژگی‌های مبتنی بر محتوا و مبتنی بر گراف، هرزنامه‌نویسان را به صورت موثرتر شناسایی کند.

٤- تشخیص هرزنامه در توییتر

٤-١- تجزیه و تحلیل ویژگی‌های موجود

در این بخش، ویژگی‌های مشترک در مطالعات قبلی، برای هر دو کاربر عادی و هرزنامه‌نویس انتخاب و مقایسه می‌شود و سپس چگونگی موثر بودن ویژگی‌ها در تشخیص حساب‌های هرزنامه بررسی می‌شود. در مقایسه‌ای که صورت گرفت، ویژگی‌های مبتنی بر پروفایل، مبتنی بر محتوا و مبتنی بر همسایه به دلیل اینکه هرزنامه‌نویسان همواره برای فرار و عدم شناسایی توسط سیستم‌های فعلی در حال تکامل هستند، از این رو زیاد موثر نمی‌باشند. از طرفی دیگر، ویژگی‌های مبتنی بر زمان‌بندی به دلیل دشوار بودن محاسبه زمان از میزان فعالیت‌هایی که یک حساب کاربری دارد، برای پیاده‌سازی موثر نمی‌باشد. با این حال از ویژگی‌های مبتنی بر گراف به دلیل ساختار سلسله مراتبی، و همچنین از تجزیه و تحلیل تعاملات کاربران، می‌توان مدل بهتر و دقیق‌تری را برای تشخیص هرزنامه نویسان ارائه نمود. بنابراین در این مقاله، علاوه بر اضافه نمودن دو ویژگی مبتنی بر محتوا، دو ویژگی جدید دیگر در بحث گراف اجتماعی، به منظور بهره‌وری بیشتر استفاده می‌شود. برای اثبات نتیجه‌گیری، این روش بر روی مجموعه داده‌ای شامل ۱۰۰۰۰ حساب کاربری عادی و ۱۰۰۰ حساب هرزنامه مورد ارزیابی قرار می‌گیرد.

٤-٢- طراحی ویژگی‌های جدید

در این بخش چند ویژگی جدید تشخیص، شامل ۲ ویژگی میتنی بر محتوا و ۲ ویژگی مبتنی بر گراف ارائه داده می‌شود.

٤-٢-١- ویژگی‌های مبتنی بر محتوا

- **نرخ پاسخ-صحیح**

در این کار، از نرخ پاسخ صحیح به عنوان یک ویژگی موثر استفاده می‌شود، چرا که اکثر حساب‌های کاربری عادی معمولا با دوستان خود تعامل دارند. در مقابل، هرزنامه‌نویس‌ها نسبت به کاربران عادی، دارای تعاملات متقابل کمتری هستند و ارتباط آنها با دیگر اعضای شبکه معمولا به صورت حجم زیادی از ارتباطات یک‌طرفه است و معمولا فقط لینک‌های URL را ارسال می‌کنند.

فیلد in_reply_to_user_id در مقدار بازگشتی از تابع user TimeLine ، برای بررسی این است که آیا یک حساب کاربری، پاسخ‌هایی که به کاربران دیگر داده است، به دوستان خود بوده است یا خیر؟. اگر کاربر در لیست دوستان او باشد، سپس شامل یک پاسخ صحیح می‌شود. بنابراین، نرخ پاسخ صحیح ($R_{Reply-Correct}$) با رابطه (۱) نشان داده می‌شود:

$$R_{Reply-Correct} = \frac{N_{reply-correct}(t)}{T(t)} \quad (۱)$$

که در آن $N_{reply-correct}(t)$، برابر با تعداد پاسخ‌های صحیح یک کاربر و $T(t)$، برابر با تعداد کل توییت‌های ارسالی از یک کاربر می‌باشد. در مقایسه با کاربران عادی، این نرخ برای حساب‌های هرزنامه بسیار کمتر است.

- **نرخ تنوع API**

هرزنامه‌نویسان معمولا برای ارسال پیام‌های بیشتر و مدیریت حساب‌های هرزنامه خود، از یک برنامه سفارشی مربوط به API Twitter استفاده می‌کنند. اگرچه کاربران عادی نیز بسته به نیاز خود از انواع این برنامه‌ها بهره می‌برند، اما هرزنامه‌نویسان به منظور سادگی، تنها از چند نوع خاص از APIها استفاده می‌کنند، بنابراین در مقایسه با کاربران عادی این نرخ برای آنها کمتر است. $R_{API-Variety}$ برای محاسبه تنوع API های استفاده شده توسط کاربران است و با رابطه (۲) نشان داده می‌شود:

$$R_{API-Variety} = \frac{N_{API-Variety}}{T(N_{API})} \quad (۲)$$

که در آن $N_{API-Variety}$، تعداد APIهای منحصربه‌فرد به کار برده شده از طرف کاربر است و $T(N_{API})$ تعداد دفعاتی است که از API در توییت‌های ارسالی استفاده می‌شود.

٤-٢-٢- ویژگی‌های مبتنی بر گراف

به طور ویژه، ویژگی‌های مبتنی بر گراف، عمدتا برای شناسایی هرزنامه‌نویسانی است که تلاش برای فرار از ویژگی‌های مبتنی بر پروفایل (به عنوان مثال، افزایش تعداد فالورها و یا توییتها) دارند. یکی از ویژگی‌های غیر قابل تقلید که تا به حال مورد استفاده قرار نگرفته است، محاسبه شباهت کاربران با همسایگان خود از طریق الگوریتم SRank می‌باشد. از آنجایی که احتمال ارتباط هریک از کاربران با دیگر کاربران در تجزیه و تحلیل شبکه اجتماعی به صورت سلسله مراتبی است، این ویژگی نقش به سزایی در شناسایی هرزنامه‌نویس‌ها دارد.

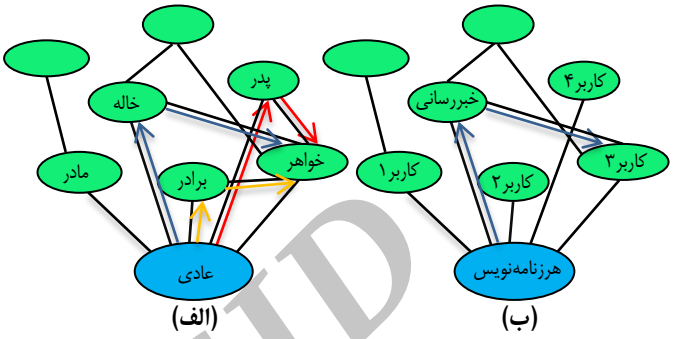
اگر هر حساب توییتر i ، به عنوان یک گره یا نود و هر رابطه دوستی، به عنوان یک لبه مستقیم e درنظر گرفته شود، به این صورت می‌توان روابط را به عنوان یک گراف G=(V,E) مشاهده کرد. حتی اگر هرزنامه‌نویسان بتوانند رفتار توییت زدن یا فالوینگ‌شان را تغییر دهند، باز برای آنها دشوار خواهد بود تا بتوانند موقعیت خود را در گراف تغییر دهند. با توجه به این بینش، دو ویژگی مبتنی بر گراف طراحی می‌شوند:

- **شباهت از طریق الگوریتم SRank**[13]

ایده اصلی در روش SRank، به این صورت بیان می‌شود که «دو گره در یک گراف جهت دار در صورتی به یکدیگر شباهت دارند، که توسط چندین مسیر با طول کوتاه به یکدیگر متصل شوند»[13]. هر چه تعداد این مسیرهای کوتاه بیشتر باشد، بنابراین دو گره شباهت بیشتری به یکدیگر خواهند داشت. برای مثال، $SRank_2$ بین دو گره a و b، شامل تمامی مسیرهایی است که از گره a شروع و به گره b ختم شود و تنها از مسیرهای به طول دو استفاده کند.

SRank، مسیرهای متفاوت بین دو گره را در نظر گرفته و یک عدد به عنوان میزان شباهت دو گره را بر می‌گرداند که بر اساس تعداد این مسیرها و همچنین طول هر کدام از مسیرها است.

بنابراین، می‌توان میزان شباهت یک کاربر با همسایگان خود را از طریق محاسبه این مقدار بدست آورد. بدیهی‌ست از آنجا که کاربران عادی معمولا حساب‌هایی را فالو می‌کنند که دوستان، همکاران و یا اعضای خانواده آنها



شکل (۱): تفاوت تعداد مسیرهایی با طول ۲ بین یک کاربر عادی و هرزنامه‌نویس. (الف) کاربری عادی. (ب) کاربر هرزنامه‌نویس.

هستند، و این حساب‌ها به احتمال زیاد دارای یک ارتباط با یکدیگر هستند. اما، هرزنامه‌نویس‌ها معمولا کورکورانه دیگر کاربران را فالو می‌کنند، که این حساب‌ها معمولا یکدیگر را نمی‌شناسند. بنابراین، در مقایسه با حساب‌های عادی، هرزنامه‌نویسان میزان شباهت کمتری به همسایگان خود دارند. این بینش در شکل (۱)- (الف) و (ب) نشان داده شده است.

به طور کلی، میزان شباهت یک کاربر با همسایگان خود در یک گراف اجتماعی، توسط دو شرط زیر تاثیرگذار خواهد شد:

- تعداد مسیرهای کوتاه از a تا b.
- طول هر کدام از این مسیرهای کوتاه.

برای محاسبه شباهت ابتدا مقدار دسترسی تعریف می‌شود. فرض می‌شود P_p ماتریس احتمال انتقال N*N برای گراف G باشد. طول این ماتریس p است. مقدار دسترسی از a به b به صورت رابطه (۳) و (۴) محاسبه می‌شود:

$$H(a,b) = w_1 * P^1_{a,b} + \ldots + w_p * P^p_{a,b} + \ldots + w_{n-2} * P^{n-2}_{a,b} \qquad (۳)$$

که در آن w_i وزن همه مسیرهای به طول i است. $P^p_{a,b}$ برابراحتمال رفتن از a به b توسط مسیرهایی به طول p است.

$$P^p_{a,b} = \frac{k_p(a,b)}{\sum_{\forall x \in G-\{a\}} k_p(a,x)} \qquad (۴)$$

که در آن $P^p_{a,b}$، برابر است با تعداد مسیرهای به طول p از گره a به b ($k_p(a,b)$)، تقسیم بر تعداد کل مسیرهای به طول p، شروع شده از گره a ($k_p(a,x)$) به تمام گره‌ها است. به دست آوردن مقدار دسترسی با در نظر گرفتن تمام مسیرها با طول‌های متفاوت درگراف، به زمان زیادی نیاز دارد. بنابراین H(a, b) با H_s(a, b) جایگزین خواهد شد و به صورت رابطه (۵) تعریف می‌شود.

$$H_s(a,b) = w_1 * P^1_{a,b} + \ldots + w_s * P^s_{a,b} \qquad 1 \le s \le n-2 \qquad (۵)$$

رای رسیدن به نتایج صحیح می‌بایست وزنی که به مسیرهای کوتاه داده می‌شود بیشتر از وزنی باشدکه به مسیرهای بلند اختصاص داده می‌شود. این از طریق رابطه (۶) بدست می‌آید.

$$w_p = 2^{s-p} \qquad (۶)$$

به منظور سادگی در بدست آوردن میزان شباهت دو گره a و b در گراف داده شده، H_s(a, b) با توجه به حداقل میزان شباهت(H_{Min}) و حداکثر آن (H_{Max})، نرمال می‌شوند. بنابراین از رابطه (۷) برای تخمین شباهت بین گره‌ها استفاده می‌شود. اندیس s در این رابطه به معنای مسیر کوتاه است.

$$SRank_s(a,b) = \frac{H_s(a,b) - H_{Min}}{H_{Max} - H_{Min}} \qquad (۷)$$

در پیاده‌سازی انجام شده، برای محاسبه SRank، در ابتدا، دوستان یک کاربر به عنوان همسایه‌های آن جمع‌آوری شده‌اند. سپس، یک گراف کوچک اجتماعی بر اساس این حساب و همسایگان آن ایجاد می‌شود. در نهایت، میزان شباهت هر جفت حساب، بر اساس تعداد مسیرهایی با طول یک و دو، به ازای هر همسایه محاسبه می‌شود. بررسی میانگین شباهت‌های بدست آمده از کلیه همسایگان یک کاربر، نشانگر این است که این مقدار برای هرزنامه‌نویسان بسیار کمتر از کاربران عادی است.

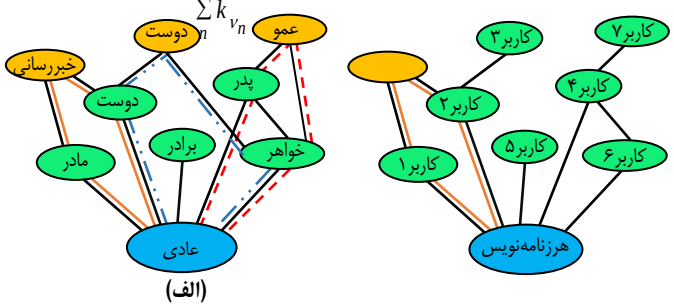
- **نرخ همسایگی مشترک (R_{cn})**

یکی دیگر از ویژگی‌های مبتنی بر گراف طراحی شده دراین مقاله، نرخ همسایگی مشترک است که نشان‌دهنده تعداد دوستان مشترک مربوط به همسایگان یک کاربر است. روابط بین همسایگان کاربران عادی، به صورت پیوسته می‌باشد. در مقابل، از آنجایی که همسایگان هرزنامه‌نویسان شناختی نسبت به یکدیگر ندارند، بنابراین گراف اجتماعی نودهای همسایه آنها به صورت خوشه‌های جدا از یکدیگر می‌باشد.

با محاسبه نرخ همسایگی مشترک، واضح است که نرخ R_{cn} برای هرزنامه‌نویسان بسیار کمتر از کاربران عادی است و این ویژگی موثری برای شناسایی آنها محسوب می‌شود که با رابطه (۸) محاسبه می‌شود:

$$R_{common\ neighborhood} = \frac{\sum_{cn} v_n}{k_v * \sum_n k_{v_n}} \qquad (۸)$$

که در آن $\sum_{cn} v_n$ برابر با تعداد همسایه‌های مشترک از همسایگان رأس $\sum_n k_{v_n}$



شکل (۲): تفاوت تعداد همسایه‌های مشترک بین همسایه‌های یک کاربر عادی و هرزنامه‌نویس. (الف) کاربر عادی. (ب) کاربر هرزنامه‌نویس.

ν است. k_ν، مجموع همسایه‌های رأس ν و ، برابر با مجموع همسایه‌های مربوط به هر همسایه راس ν است. همانطور که در شکل (۲) مشاهده می‌شود تعداد دوستان مشترک بین همسایگان، برای یک کاربر عادی بیشتر است. بنابراین، در مقایسه با حساب‌های هرزنامه، روابط اجتماعی قویتری بین حساب‌های عادی که تمایل به شکل چهار ضلعی دارند، وجود دارد.

۵- طبقه‌بندی و ارزیابی

در این بخش، مانند مطالعات قبل از یادگیری نظارت شده برای طبقه‌بندی حساب‌های توییتر، استفاده می‌شود. ۱۱۰۰۰ حساب برچسب خورده شده، به عنوان مجموعه آموزش و تست انتخاب می‌شوند، که در آن ۱۰۰۰۰ کاربر عادی و ۱۰۰۰ کاربر هرزنامه‌نویس وجود دارد. سپس ویژگی‌های طراحی شده برای هر حساب محاسبه می‌شود. در نهایت، حساب‌های موجود در مجموعه آموزشی به صورت دستی مورد بررسی قرار می‌گیرند تا اطمینان حاصل شود که هیچ حسابی به اشتباه طبقه‌بندی نشده باشد. الگوریتم درخت تصمیم[۱۳]، جنگل تصادفی[۱۴]، شبکه بیزین[۱۵] و تزئین[۱۶] پیاده‌سازی شده در WEKA[۱۷]، ابزارهای یادگیری ماشینی هستند[19]، که برای انجام فرایند آموزش انتخاب شده‌اند. به منظور آزمایش و بررسی قابلیت اطمینان از مدل ایجاد شده برای هر طبقه‌بندی، از (10fold cross-validation) استفاده می‌شود، به طوری که نتایج طبقه‌بندی شده به صورت یک ماتریس confusion، مطابق جدول (۱) نشان داده می‌شود.

جدول (۱) : ماتریس confusion

	پیش‌بینی شده	
	هرزنامه‌نویس	عادی
هرزنامه‌نویس	۹۸۵	۱۵
عادی	۲۹	۹۹۷۱

به منظور بررسی میزان اثربخشی از طبقه‌بندی صورت گرفته بر اساس ویژگی‌های پیشنهادی جدید، از سه معیار عملکرد، استفاده می‌شود: نرخ مثبت کاذب، نرخ تشخیص و F- measure.

۵-۱- اعتبارسنجی ویژگی‌ها:

برای تایید اعتبار اینکه نتایج روش پیشنهادی عملکرد بهتری داشته است، یک آزمایش به منظور مقایسه عملکرد مجموعه ویژگی‌ها، یکبار با ویژگی‌های پیشین و بار دیگر با ویژگی‌های جدید پیاده‌سازی می‌شود.

جدول (۲) نشان می‌دهد که برای هر یک از چهار الگوریتم طبقه‌بندی، با اضافه کردن ویژگی‌های به تازگی طراحی شده، نرخ تشخیص (DR) تا ۱۰٪ افزایش می‌یابد، در حالی که نرخ مثبت کاذب (FPR) در همان حد پایین، حفظ می‌شود. این نتایج، بهبود عملکرد تشخیص را با توجه به ویژگی‌های جدید، تایید می‌کند.

جدول (۲) : مقایسه معیارهای ارزیابی ویژگی‌های موجود با اضافه‌کردن ویژگی‌های جدید

	ویژگی‌های پیشین			ویژگی‌های اضافه شده		
طبقه‌بندی	FPR	DR	F-Measure	FPR	DR	F-Measure
جنگل تصادفی	۰.۰۰۴	۰.۸۴۸	۰.۹۰	۰.۰۰۳	۰.۹۸۵	۰.۹۷۸
درخت تصمیم	۰.۰۰۸	۰.۸۴۰	۰.۸۷۶	۰.۰۰۴	۰.۹۶۱	۰.۹۶۲
شبکه بیزین	۰.۰۱	۰.۸۳۸	۰.۸۳۳	۰.۰۲۱	۰.۹۲۶	۰.۸۶۶
تزئین	۰.۰۱۲	۰.۸۵۴	۰.۸۸۴	۰.۰۰۴	۰.۹۷۷	۰.۹۶۷

۵-۲- مقایسه عملکرد

در این آزمایش به مقایسه عملکرد این کار با پنج روش موجود پرداخته می‌شود که در شکل (۳)، نشان داده شده است:

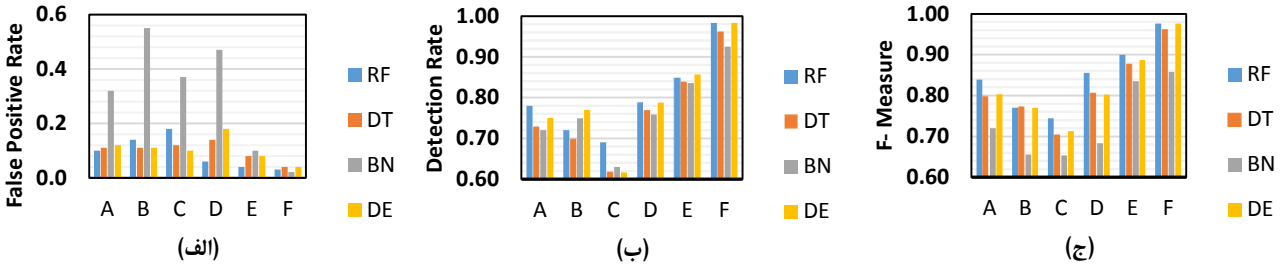
برای بررسی این موضوع، کار [11] به عنوان A، [10] به عنوان B، [14] به عنوان C، [16] به عنوان D، [8] به عنوان E و روش جدید در این مقاله به عنوان F، در نظر گرفته می‌شود. این ارزیابی با استفاده از چهار الگوریتم طبقه‌بندی مختلف یادگیری ماشین، انجام می‌گیرد: جنگل‌تصادفی (RF)، درخت تصمیم(DT)، شبکه بیزین(BN) و تزئین(DE).

همانطور که در شکل (۳) دیده می‌شود، روش جدید از تمامی کارهای موجود، بهتر عمل کرده است. به طور خاص، شکل (۳)- (الف)، نشان می‌دهد که نرخ‌های مثبت کاذب در این کار طبق هر چهار الگوریتم طبقه‌بندی یادگیری ماشین، کمترین مقدار را دارد و می تواند همواره زیر ۲٪ حفظ شود. همچنین شکل (۳)-(ب)، نشانگر این است که نرخ‌های تشخیص در این کار طبق هر چهار الگوریتم طبقه‌بندی نیز بالاترین هستند. ودر نهایت شکل (۳)- (ج) نشان می‌دهد که مقادیر F-measure از روش جدید، بالاترین امتیاز را به میزان ۰.۹۷۸ دارد. نتایج بالا تایید می‌کند که مجموعه ویژگی‌های جدید، به طور موثرتر هرزنامه‌نویسان توییتر را تشخیص می‌دهد.

۵- نتیجه‌گیری

در این مقاله، یک روش جدید مبتنی بر گراف اجتماعی جهت شناسایی هرزنامه‌نویسان اجتماعی در دنیای واقعی، طراحی و ارزیابی گردید. تحقیقاتی که تاکنون بر روی شناسایی هرزنامه‌نویسان صورت گرفته‌اند، عمدتاً بر اساس مدل‌های طبقه‌بندی شده بر روی ویژگی‌های مبتنی بر پروفایل و مبتنی بر محتوا می‌باشند. اما مشکل عمده روش‌های پیشین ضعف در تشخیص هرزنامه‌نویس‌ها و کاربران عادی به خاطر تقلید هرزنامه‌نویس‌ها از رفتار و نوع متون ارسالی کاربران عادی است. هدف این تحقیق به طور کلی، بررسی روش‌ها و توسعه ابزارهای موثر برای شناسایی و فیلترکردن خودکار هرزنامه‌نویسانی است که سیستم‌های اجتماعی را مورد هدف قرار داده‌اند.

گراف‌ها برای حل مسائل زیادی در ریاضیات و علوم کامپیوتر استفاده می‌شوند. در این مقاله همانطور که نشان داده شد، با در نظر گرفتن رفتار کاربران در گراف شبکه اجتماعی به صورت سلسله مراتبی، از ساختار گراف‌ها بهره گرفته شده که کمک شایانی به شناسایی این کاربران نامطلوب می‌کند و از طرفی مانع تقلید هرزنامه‌نویس‌ها از رفتار کاربران عادی می‌شود. بنابراین، دو ویژگی موثر بر اساس شباهت کاربران با همسایه‌های خود، با استفاده از الگوریتم ساختاری SRank و همچنین نرخ همسایگی مشترک محاسبه شده و ثابت شد که این روش بسیار موثرتر از روش‌های پیشین است. درنهایت، با توجه به ارزیابی صورت گرفته، نشان داده شد که با پایین ماندن حداقل نرخ مثبت کاذب، سرعت کشف و شناسایی هرزنامه‌نویس‌ها همچنان با استفاده از ویژگی‌های جدید، بالاتر از کارهای موجود است.



کل (۳) : مقایسه عملکرد این مقاله با روش‌های موجود.
(الف) نرخ مثبت کاذب. (ب) نرخ تشخیص. (ج) F- measure.

مراجع

[1] S. Abu-Nimeh, T. Chen, and O. Alzubi, "*Malicious and spam posts in online social networks*." IEEE Computer Society, 44(9), Sep. 2011.

[2] K. Thomas, and D. M. Nicol, "*The koobface botnet and the rise of social malware*," In Proceedings of the 5th International Conference on Malicious and Unwanted Software, Oct. 2010.

[3] G. Yan, G. Chen, S. Eidenbenz, and N. Li, "*Malware propagation in online social networks: nature, dynamics, and defense implication*," In Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, Mar. 2011.

[4] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "*Who is tweeting on Twitter: human, bot, or cyborg?*" In Proceedings of the 26th Annual Computer Security Applications Conference(ACM), Dec. 2010.

[5] F. Benevenuto, G. Magno, T.Rodrigues and Y. Almeida, "*Detect spammers on Twitter*," In Proceedings of the 7th Annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference(CEAS), July 2010.

[6] Z. Chen, J. Yang and J. H. Wang, *"A Cascading Framework for Uncovering Spammers in Social Networks"*, Networking Conference, 2014 IFIP-ieeexplore.ieee.org.

[7] X. Zheng, Z. Zeng and Z. Chen. *"Detecting spammers on social networks.*" Neurocomputing, 2015 – Elsevier. 159: pp. 27-34.

[8] C. Yang, R. Harkreader, G. Gu, *"Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers*", Information Forensics and Security, IEEE Journals & Magazines, Vol. 8, Issue: 8, PP. 1280 - 1293, 2013.

[9] J. Song, S. Lee, and J. Kim, "*Spam Filtering in Twitter using SenderReceiver Relationship*," In Proceedings of International Symposium on Recent Advances in Intrusion Detection (RAID), Sept. 2011.

[10] G. Stringhini, C. Kruegel, and G. Vigna, "*Detecting spammers on social networks*," In Proceedings of the 26th Annual Computer Security Applications Conference(ACM), Dec. 2010.

[11] K. Lee, J. Caverlee, and S. Webb, "*Uncovering social spammers: social honey pots + machine learning*," In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, July 2010.

[12] The definition of Spam on Twitter Help Center, https://support.twitter.Comlgroups/31-twitter-basics /topics/114-guidelines-best -ractices/articles /18311-the-twitter-rules.

[13] H. Khosravi-Farsani , M. Nematbakhsh, and G. Lausen, *"SRank: Shortest paths as distance between nodes of a graph with application to RDF clustering,"* Journal of Information Science, vol. 39, pp. 198-210 , 2013.

[14] A. H. Wang, *"Security and Cryptography (SECRYPT), Don't Follow Me: Spam Detection in Twitter,"* Proceedings of the 2010 International Conference, PP. 1-10, 26-28 July 2010, IEEE.

[15] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonalves, *"Detecting spammers and content promoters in online video social networks,"* in Proc. ACM SIGIR Conf. (SIGIR), Boston, MA, USA, 2009.

[16] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, *"Detecting spammers on Twitter,"* in Proc. Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), Redmond, WA, USA, 2010.

[17] H. Gao, J. Hu, C.Wilson, Z. Li, Y. Chen, and B. Zhao, "*Detecting and characterizing social spam campaigns*," in Proc. ACM SIGCOMMIMC (IMC'10), Melbourne, Australia, 2010.

[18] A. Gupta, R. Kaushal, *"Improving Spam Detection in Online Social Networks,"* Cognitive Computing and Information Processing (CCIP), International Conference on 2015 , ieeexplore.ieee.org.

[19] Weka, the machine learning tool, http://www.cs.waikato.ac.nz/m1/weka

زیرنویس‌ها

[1] Online Social Networks
[2] Spammers
[3] Machine learning
[4] Short Rank
[5] False Positive Rate
[6] Spam
[7] Follower
[8] Following
[9] Tweet
[10] Detection rate
[11] Neighbor-Based
[12] Automation-Based
[13] Decision Tree
[14] Random Forest
[15] Bayes Net
[16] Decorate
[17] Waikato Environment for Knowledge Analysis