ارائه یک روش جدید آماری- رفتاری برای تشخیص حملات امنیتی به حساب‌های کاربران در شبکه‌های اجتماعی

حسین شیرازی
دانشیار، عضو هیأت علمی دانشکده فرماندهی و کنترل، دانشگاه صنعتی مالک اشتر، تهران، ایران
Shirazi@mut-es.ir

داوود محمدپور زنجانی
مربی، عضو هیأت علمی دانشکده مهندسی کامپیوتر، دانشگاه زنجان، زنجان، ایران
Dmp@znu.ac.ir

سیدمحمدرضا فرشچی
گروه کنترل و بهینه‌سازی، دانشکده ریاضی، دانشگاه فردوسی مشهد، مشهد، ایران
Smr.farshchi@stu-mail.um.ac.ir

چکیده

با محبوب شدن شبکه‌های اجتماعی، حملات امنیتی وجرایم رایانه‌ای به سمت این سایت‌ها، افزایش یافته است. در اغلب موارد، هدف از حمله به حساب‌های کاربران، سرقت اطلاعات شخصی، یا پخش کردن هرزنامه می‌باشد. در یک شبکه‌اجتماعی، حساب‌های جعلی را با بررسی ویژگی‌های دقیق، به‌راحتی می‌توان شناسایی و غیرفعال کرد. درحالت مقابل، ممکن است نفوذگر، حساب قانونی یک کاربر را به تصرف خود درآورده و از آن برای اقدامات مخرب بهره جوید. در این‌صورت تشخیص، و غیرفعال‌کردن این حساب‌ها، برای مدیر شبکه‌اجتماعی به‌راحتی امکان‌پذیر نخواهد بود. تاکنون، روش‌های زیادی برای تشخیص حساب‌های جعلی (سایبیل)، در شبکه‌های‌اجتماعی ارائه شده است، اما طبق بهترین دانش نویسندگان در تشخیص حساب‌های در معرض‌خطر، الگوریتم قابل قبولی وجود ندارد. در این پژوهش ما به ارائه یک روش آماری برای ایجاد مدلی رفتاری از عملکرد کاربران با توجه به تاریخچه فعالیت آنان، پرداخته‌ایم. به‌عبارت دیگر، با استناد به مجموعه ویژگی‌هایی منحصر به‌فرد و با استفاده از یک مدل‌آماری، به کشف تغییرات ناگهانی در پروفایل کاربران پرداخته‌ایم. ازآنجا که تغییرات ناگهانی ممکن است صحیح و از سمت خود کاربر باشد، ما از یک طبقه‌بند و ویژگی‌های مشابهت محتوا در تلفیق با مدل‌آماری پیشنهادی خود استفاده کرده‌ایم. ارائه مدل طبقه‌بندی برای دسته‌بندی رفتارهای کاربر، با توجه به ویژگی‌های ارائه شده در این مقاله، روش بسیار منعطفی را برای شناسایی حساب‌های غیرقانونی فراهم کرده است. بازه‌های استخراج ویژگی یک ساعته، در بررسی جریان‌های نامه‌ای در داده‌های دو شبکه‌اجتماعی معروف فیسبوک و تویتر، دقت بسیار مناسب روش پیشنهادی نمایش داده شده است. علاوه براین، اندازه‌گیری نرخ نمونه غلط صحیح در بررسی بیش از ۵۰ میلیون پست توییت شده، به کمتر از ۵٪ رسیده است. نتایج نشان می‌دهد، روش پیشنهادی، قابلیت تجمیع در کلیه پلت فرم‌های شبکه‌های‌اجتماعی را داراست.

کلمات کلیدی:

شبکه‌های‌اجتماعی، امنیت کاربران، مدل‌آماری، طبقه‌بندی.

مقدمه

شبکه‌های‌اجتماعی برخط، مثل فیسبوک و تویتر، در طول دهه اخیر با محبوبیت شگرفی مواجه شده‌اند. بسیاری از مردم برای تبادل‌اطلاعات با اعضای خانواده، اطلاع از رویدادها، در دسترس بودن برای دوستان و همکلاسی‌های خود و برای به اشتراک‌گذاری اطلاعات از این شبکه‌ها استفاده می‌کنند. در حالت کلی، در اشتراک‌گذاری اطلاعات، عموما اصل بر مبنای صداقت و اعتماد به یکدیگر می‌باشد.. در شبکه‌های‌اجتماعی با استفاده از یک گراف اجتماعی [۱]، می‌توان نحوه تبادل پیام‌ها را مدل‌سازی کرد. در این مدل‌سازی به عنوان مثال، ارسال کننده‌ها، دریافت کننده‌ها یا پست‌های دیواری، تغییر وضعیت کاربر یا حتی، توییت‌ها براحتی قابل تشخیص خواهند بود.

اخیراً، این پایگاه‌ها بدلیل داشتن طیف وسیعی از کاربران، مورد توجه مجرمان سایبری قرار گرفته است. در مطالعه‌ای که در سال ۲۰۱۱ بر روی شبکه‌های‌اجتماعی صورت گرفته است، ۹۵٪ کاربران در طول یک سال عضویت، حداقل یک پیام ناخواسته یا هرزنامه دریافت می‌کنند [۲]. علاوه بر این طیف وسیعی از بدافزارها نیز برای دسترسی به اطلاعات موجود در این پایگاه‌ها، طراحی شده است. مطالعه قبلی نشان داد، از دید امنیتی سه نکته حائز اهمیت برای شبکه‌های‌اجتماعی وجود دارد، که به ترتیب، هرزنامه‌ها، فیشینگ و بدافزارها می‌باشند.

برای کنترل فعالیت‌های رو به گسترش مخرب، در شبکه‌های‌اجتماعی، محققین روش‌های مختلفی را برای کاهش یا مقابله با این فعالیت‌ها ارائه کرده‌اند. کارهای اولیه با ارائه روش‌های [۳]، و بررسی حساب‌های جعلی در شبکه‌های‌اجتماعی شروع گردید. در این تحقیق هدف بر روی حساب‌های کاربری متمرکز خواهد بود که، دارای هدف گسترش یک محتوای مخرب باشند. کارهای پیشنهادی در این زمینه، علیرغم استقبال زیاد در ابتدای پیشنهاد موضوع، بدلیل عدم توانایی در تمایز میان حساب‌های سایبیل، و به‌خطر افتاده، به مرور از زمره تکنیک‌های پرکاربرد در تأمین امنیت شبکه‌های‌اجتماعی، خارج شد.

حساب‌های به خطر افتاده، حساب‌های کاربری صحیح و قانونی است، که توسط یک هکر یا شخص ثالث مورد دستبرد و دسترسی قرار گرفته است. این دسترسی ممکن است توسط اجرای یک اسکریپت ضربدری [۴] بر روی کلاینت و سرور، یا استفاده از تکنیک‌های فیشینگ برای سرقت اطلاعات کاربر صورت گرفته باشد. همچنین اخیرا، چندین مورد استفاده از بات‌ها، و ذخیره‌سازی آنها بر روی سرورهای شبکه‌های‌اجتماعی به منظور سرقت هویت کاربران نیز توسط سیمانتک گزارش شده است [۵].

از طرف دیگر، اگر چه ایجاد یک حساب‌کاربری برای استفاده توسط یک بدافزار، یا هرزنامه نگار، راحت‌ترین روش ممکن می‌باشد، اما حساب‌های کاربری که در خطر قرار گرفته، و مورد دستبرد واقع شده‌اند، برای جرایم سایبری بسیار با ارزش‌تر هستند، چراکه، امکان استفاده از تاریخچه گذشته حساب‌کاربری، و اطلاعات شبکه‌اجتماعی برای هکرها وجود دارد. بعبارت دیگر، نفوذگران بیشتر به دنبال استفاده از حساب‌های قانونی برای انجام عملیات مخرب خود می‌باشند. تنها راه پیشنهادی برای مقابله با دونوع مختلف از حساب‌ها، تمرکز بر روی توزیع و ارسال پیام‌ها و بررسی خود پیام‌ها می‌باشد. برخی از محققین در [۶] از ویژگی‌های مشابه میان پیام‌ها برای ارزیابی توزیع و اقدامات مخرب استفاده کرده‌اند. ممکن است ویژگی‌های مشابه از داخل پیام انتخاب شده [۷] یا بر اساس URL [۸] انتخاب شود. با استفاده از یک تکنیک دسته‌بندی در تلفیق با ویژگی‌های داخلی پیام [۹]، براحتی نمی‌توان پیام‌های ارسالی و دریافتی را گروه‌بندی کرد. چراکه، فرض کنید پیام "تولدت مبارک"، توسط نرم‌افزارهایی مثل Foursquare در طول یک شبکه‌اجتماعی [۱۰]، پیام‌هایی با ویژگی‌های یکسان و قانونی تولید کند. بنابراین روش‌های مبتنی بر URL در [۱۱]، [۱۲] و [۱۳] بعنوان الگورتیم پایه مورد توجه بیشتری قرار گرفته است. این روش‌ها هم در واقعیت، استفاده چندانی ندارند، چراکه، عموماً URLها، از پیام‌ها توسط رمزگذاری براحتی حذف می‌شوند [۱۴]. ما نمونه‌ای از این پیام‌ها را در آزمایشات خود مورد بررسی قرار داده‌ایم.

روش‌های دیگری در [۱۵] و [۱۶] نیز مورد بررسی قرار گرفته است. این روش‌ها از ویژگی‌های تلفیقی دیگر مثل، سایز نامه، گروه مخاطبین مورد استفاده در ارسال‌ها، استفاده‌های اسکریپتی در متن نامه، یا ...، برای تشخیص و دسته‌بندی عملکرد صحیح، از مخرب استفاده می‌کنند. اگر چه مشکل اولیه با استفاده از ویژگی‌های تلفیقی حل شد، اما دقت این روش‌ها، بدلیل عدم پیروی از الگوی خاص یکسان، و تکیه تنها به استفاده از مدل‌های آماری، به ۸۰٪ نمی‌رسد [۱]. علاوه بر این، مشکل عدم تمایز میان نوع حساب کاربری مورد استفاده همچنان پابرجاست. دلیل اهمیت تمایز میان نوع حساب‌کاربری، بسیار حائز اهمیت است. مدیر شبکه اجتماعی در صورت کشف یک حساب‌کاربری جعلی، براحتی می‌تواند عمل حذف حساب را انجام دهد، از طرف دیگر، حساب به خطر افتاده باید شناسایی شده، و عملیات بازگرداندن حساب به قربانی انجام شود، و عملاً اقدام حذف برای این نوع حساب‌ها، غیر منطقی است [۱۷].

در این پژوهش، ما به ارائه یک روش جدید، در تشخیص نوع حساب مورد حمله قرار گرفته توسط هکرها، یا بدافزارها، پرداخته‌ایم. طبق بهترین دانش‌ما، در این پژوهش برای اولین بار از ترکیب سه ویژگی جدید، برای این تشخیص استفاده کرده‌ایم، که قبلاً در جایی مورد استفاده قرار نگرفته است. اولین ویژگی روش جدید، عدم استفاده از تگ‌های URL در متن پیام‌هاست، بنابراین طیف وسیعی از اسکریپت‌ها و نامه‌های مورد استفاده در شبکه‌های اجتماعی، برای اولین بار در روش پیشنهادی، شناسایی خواهد شد. برخی از این نامه‌ها شامل نامه‌های تماس با ما، کلاهبرداری، لیست مخاطبین و دیگر نامه‌ها که جزو دسته‌های اساسی ایجاد مشکل در [۳] مورد بررسی قرار گرفته‌اند، می‌باشد. دوم، بررسی‌های دقیق انجام شده بر روی طیف وسیع داده‌های ارزیابی نشان می‌دهد، دقت سیستم پیشنهادی بسیار مطلوب بوده و از نرخ نمونه غلط مثبت، بسیار پایینی برخوردار است. سوم، ما در این پژوهش تنها برای یافتن حساب‌های به خطر افتاده، روشی را ارائه کرده‌ایم، چراکه روش‌های کنونی به خوبی بر روی شبکه‌های اجتماعی حساب‌های جعلی را جمع‌آوری می‌کنند. همچنین، پس از شناسایی این نوع حساب‌ها، مدیر شبکه اجتماعی می‌تواند سیاست‌های خود را در قالب‌های متفاوت مثل مدل‌های شناختی اعمال کند.

فرضیه و ایده اصلی در روش پیشنهادی ما، براساس رفتارهای منظم و نامنظم، کاربران خاص در شبکه‌های اجتماعی بنا نهاده شده است. اگر در زمان خاصی یک حساب به خطر افتاده باشد، قطعاً در مدل رفتاری نیز، کاربری مورد حمله قرار گرفته است، که قطعاً در حال استفاده از شبکه اجتماعی هم بوده است. ما فرضیه پیشنهادی، را در ارزیابی‌های خود تأیید کرده‌ایم. برای ایجاد مدل‌رفتاری در زمانی که کاربر حقیقی در حال استفاده از شبکه اجتماعی می‌باشد، ما از یک مدل آماری به نام مدل آماری رفتاری، استفاده می‌کنیم.

طبیعتاً، هر ویژگی پیشنهادی ما در مدل‌آماری رفتاری، مطابق یکی از خصیصه‌های نامه‌ها یا اسکریپت‌های تبادلی است. (بعنوان مثال، زمان ارسال نامه یا زبان نامه‌های ارسالی). مدل‌پیشنهادی یک پروفایل رفتاری برای هرکاربر ایجاد کرده و وابسته به پلتفرم خاصی نمی‌باشد، چنانکه در آزمایشات، قابلیت استفاده مدل را بر روی داده‌های قابل دسترس فیسبوک و توییتر، به نمایش گذاشته‌ایم.

روش پیشنهادی ما، با یک تناقض در مدل رفتاری نرمال، یک کاربر حقیقی، (مشابه کارهای قبلی)، عمل خاصی را انجام نمی‌دهد، چرا که ممکن است رفتار یک کاربر در یک برهه از زمان به‌صورت بدیهی متفاوت از ارسال‌های قبلی (رفتار عمومی) باشد. بنابراین ما دنباله‌ای از آخرین ارسال‌ها یا اسکریپت‌ها را، برای ایجاد یک مجموعه از ویژگی‌ها در تلفیق با مدل رفتاری استفاده خواهیم کرد. این تلفیق در زمانی که نامه‌ها بر روی جمعیت بیشتری از کاربران (بیش از ۱۰)، ارسال گردد، توسط مدل‌آماری رفتاری و با استفاده از پروفایل رفتاری، شناسایی خواهد شد. البته به دلیل ویژگی‌های مناسب انتخاب شده روش پیشنهادی، چنانکه در آزمایشات هم مشخص خواهد شد، با بررسی کمتر از ۹ پیام مشابه در داده‌های توییتر، توانستیم دقت بسیار مناسبی در تشخیص پیام‌های مخرب و هرزنامه‌ها ایجاد کنیم.

به طور خلاصه در روش پیشنهادی، اولاً، ایجاد مجموعه‌ای از پیام‌های مشابه انجام گرفته و ثانیاً، زیرمجموعه مناسبی (تعداد مشخص)، از پیام‌هایی که مدل آماری رفتاری را نقض کرده‌اند، تشکیل خواهد شد. مزیت روش پیشنهادی در عدم تقدم یا تأخر، اجرای هریک از این مراحل است. بعبارت دیگر، می‌توانیم ابتدا مدل آماری رفتاری را بررسی سپس براساس آن اقدام به دسته‌بندی نامه‌ها نماییم، (که در این صورت به ویژگی‌های تشابه نامه، نیاز داریم. بنابراین نسبت به ویژگی‌های کارهای ارائه شده قبلی، می‌توان ویژگی‌های بسیار مناسب‌تری برای دسته‌بندی انتخاب کرد)، یا ابتدا نامه‌ها را دسته‌بندی کرده، سپس برای هر دسته ویژگی آماری رفتاری را ایجاد کنیم. بدیهی است در این مرحله مدلی برای دسته نامه‌ها ساخته می‌شود، و برای مواقعی مناسب خواهد بود، که حجم کمتری از نامه‌ها مورد پردازش قرار می‌گیرد.

روش پیشنهادی ما دارای محیط گرافیکی داخلی است، که توسط مدیران شبکه‌های‌اجتماعی می‌تواند مورد استفاده قرار بگیرد. داده‌های مورد استفاده در ارزیابی‌ها، با همکاری بسیار خوب آقای ویگنا [۱] تهیه گردید. این داده‌ها شامل داده‌های ارسالی فیسبوک از تاریخ ۱۳ اردیبهشت ۱۳۹۱ تا ۱۳ مرداد ۱۳۹۱ و همچنین داده‌های ارسالی تویتر در تاریخ ۱ بهمن ۱۳۹۰ تا ۱ اردیبهشت ۱۳۹۱ بود. داده‌های ارسال شده حدود ۶۰ میلیون و دویست هزار پست توییت شده بود. همچنین داده‌های ارسالی توسط کاربران فیسبوک ۱۱۰ میلیون پست بود. در ارزیابی‌های انجام شده با مدل‌پیشنهادی، ۳۴۳ هزار و ۲۲۹ حساب‌کاربری در تویتر و ۱۱ هزار و ۸۷ حساب‌کاربری در فیسبوک، که مورد حمله واقع شده بود، را شناسایی کردیم.

بعنوان کارهای تحقیقاتی بعدی می‌توان از مدل‌آماری رفتاری جدید برای پیش‌بینی، کاوش رفتار نامه‌ها یا اسکریپت‌های ارسالی، توسط کاربران در شبکه اجتماعی برای تشخیص ناهنجاری یا موضوعات کاربردی دیگر، استفاده کرد.

در ادامه این مقاله، در بخش بعد به بررسی مدل آماری رفتاری پیشنهادی پرداخته‌ایم. سپس روش طبقه‌بندی تلفیقی با ویژگی‌های ارائه شده را توضیح می‌دهیم. در بخش ۴ نیز به الگوریتم دسته‌بندی پرداخته و در بخش ۵ نتایج ارزیابی‌های خود را به نمایش گذاشته‌ایم.

۱- مدل‌آماری- رفتاری

یک مدل آماری رفتاری، تاریخچه فعالیت یک کاربر در شبکه اجتماعی به منظور محاسبه مدل رفتاری وی (مورد توقع)، می‌باشد. برای ساخت مدل رفتاری، باید جریان نامه‌های ارسالی و پست‌های کاربر را در طول بازه‌های زمانی مورد ارزیابی قرار داد. اگرچه ویژگی‌های دیگر نیز، مثل تصاویر، لیست مخاطبین و دوستان نیز می‌تواند اطلاعات بسیار مناسبی در اختیار مدل قرار دهد. طبق گزارش [۱۸] و بررسی‌های دقیق ما، متاسفانه، شبکه‌های‌اجتماعی یک روش هدفمند برای در دسترس قرار دادن این اطلاعات، برای کاربران ثالث، ارائه نکرده‌اند.

یک مدل‌آماری رفتاری (به اختصار مدل می نویسیم)، برای کاربر U به‌صورت زیر ساخته می‌شود،

ابتدا جریان ارسال اطلاعات از U در سایت شبکه اجتماعی رهگیری می‌شود. این جریان نامه‌ای، لیست ارسال شده به ترتیب زمانی خواهد بود. مسلما، برای شبکه‌های اجتماعی متفاوت، جریان نامه‌ای به روش‌های مختلفی جمع‌آوری خواهد شد. برای مثال، در تویتر جریان نامه‌ای بیانگر، جدول زمانی عمومی، هر کاربر است. برای یک کاربر فیسبوک، جریان نامه‌ای پستی است، که بر روی دیوار شخصی‌اش، یا بر روی دیوار دوست وی، قرار داده شده است.

برای ساخت یک مدل صحیح و کامل، جریانی که باید مورد پیگیری قرار بگیرد، باید شامل کمترین میزان نامه باشد، چراکه بهینه‌ترین حالت ممکن خواهد بود [۵]. به‌صورت بدیهی، مدل صحیح مدلی است، که با یک تغییر ممکن و صحیح کاربر، تغییری در ویژگی‌های استخراجی صورت ندهد، و در طرف مقابل مدلی غیرقابل قبول است، که فعالیت صحیح کاربر را ناصحیح جلوه دهد. از طرف دیگر در قسمت قبل گفتیم، میزان حداقل نامه‌ها، نباید از مقدار معینی، S، کمتر باشد. چراکه با بررسی‌های کمتر، امکان شناسایی تغییرات صحیح ممکن نخواهد بود. بررسی‌های ما نشان می‌دهد، در حالت کمتر از S=10، ویژگی‌های

استخراجی مدل قابل قبولی را برای یک حساب خاص، ایجاد نخواهد کرد. اگرچه این نقطه ضعفی نخواهد بود، چراکه حساب دارای این محدودیت، احتمالاً، به مدت طولانی غیرقابل استفاده بوده یا با احتمال زیاد، مورد استفاده صحیح بوده است. (اگرچه، احتمال ایجاد حساب جدید نیز در این حالت وجود خواهد داشت.) بحث مفصل‌تری در این باره در بخش‌های بعد ارائه خواهیم کرد.

مدل‌پیشنهادی، با دنبال کردن مسیر جریان نامه‌ای، به استخراج ویژگی‌های مشابه می‌پردازد. بعبارت دقیق‌تر، روتین نوشته شده، به استخراج یک مجموعه ویژگی از هر نامه می پردازد، هر ویژگی استخراج شده به مدل آماری آموزش داده می‌شود. این ویژگی‌ها به مرور تکمیل می‌شوند. بعنوان مثال، زمان ارسال نامه، برنامه‌ای که نامه را تولید کرده است، یا ... ویژگی‌های مورد استفاده مدل‌پیشنهادی را در بخش بعد معرفی خواهیم کرد.

با تولید یک مدل برای یک کاربر خاص، می‌توان با دقت خاصی رفتار بعدی کاربر در تولید نامه، اسکریپت یا پست را پیش‌بینی کرد. علاوه بر این، از تولید مدل به بعد، یک امتیاز برای نامه‌های ارسالی محاسبه می‌کنیم. محاسبه این امتیاز براساس مدل محاسبه شده و ویژگی‌های استخراج شده در مرحله قبل است. هر مدل یک امتیاز (عدد حقیقی)، در بازه [0,1] تولید خواهد کرد. عدد صفر بیانگر عملکرد صحیح (نسبت به ویژگی‌های در نظر گرفته شده)، و عدد یک برای رفتار کاملاً، مخرب به کار خواهد رفت. به همین ترتیب محاسبه مدل برای تمامی حساب‌هایی که مورد پیگیری قرار گرفته است، تکرار خواهد شد.

۱-۱- مدل‌سازی مشخصه‌های نامه‌ها

ما در مدل‌پیشنهادی خود از ویژگی‌های هفت‌گانه زیر برای ایجاد مدل آماری رفتاری استفاده خواهیم کرد.

زمان (در روز). زمان‌های فعال بودن حساب کاربری در طول هر روز، را به عنوان اولین ویژگی ذخیره خواهیم کرد. مطالعات ما و همچنین گزارشات [۱،۲] نشان داده است، حدود ۹۵٪ کاربران در زمان‌هایی که از الگوی خاصی پیروی می‌کنند، اقدام به ارسال پست می‌کنند، زمان‌های زیادی مثل نهار یا خواب، از زمره زمان‌هایی است که عموم کاربران در آن فعالیتی ندارند. قطعاً ارسال‌های بیشتر در ساعاتی که در این ویژگی قرار نگیرد، بیانگر عملکرد غیرقابل قبول در مدل ویژگی زمان خواهد بود.

منبع نامه. منظور از انتخاب این ویژگی بررسی نرم‌افزار ارسال‌کننده نامه است. اکثر شبکه‌های اجتماعی، واسط‌های میل سرورهای تحت وب را در اختیار کاربران خود قرار می‌دهند. علاوه براین، امروزه بسیاری از شبکه‌های‌اجتماعی، استفاده از واسط‌های شخص ثالث را برای ارائه خدمات تحت وب بر روی سیستم عامل های، Android و IOS را در دستور کار خود قرار داده‌اند.

اگرچه از لحاظ امنیتی چنین نرم‌افزاری مستقیماً قادر به ارسال پست به حساب کاربری اشخاص نخواهد بود. اما بیش از نیمی از کاربرانی که از شبکه‌های‌اجتماعی استفاده می‌کنند، به نرم‌افزارهای شخص ثالث اجازه دسترسی به حساب خود را فراهم کرده‌اند، تا بتوانند در تبلت‌ها یا دیگر وسیله‌های همراه خود به چک کردن یا ارسال پست اقدام نمایند [۹].

معروفترین روش دسترسی به حساب‌های شبکه‌های OAUTH در [۱۳] است. این نرم‌افزار توسط فیسبوک و تویتر، مورد استفاده قرار گرفته است. کاربر می‌تواند بدون نگرانی در مورد کلمه عبور خود، حساب خود را بررسی نماید. این ویژگی، به بسیاری از کاربران غیرقانونی، اجازه ارسال نامه‌های ناخواسته یا دسترسی به حساب‌های اشخاص دیگر، را فراهم می‌کند. زمانیکه کاربر برای اولین بار از یک نرم‌افزار خارج از ویژگی رفتاری خود، برای دسترسی به حساب خود استفاده می‌کند، رفتار خارج از مدل ویژگی منبع، اتفاق خواهد افتاد.

زبان نامه (نوشتن). کاربران در شبکه‌های‌اجتماعی در انتخاب زبان نوشتن نامه‌های خود آزاد هستند. برآوردهای ما نشان می‌دهد، عموم کاربران حداکثر از سه نوع زبان در نوشتن نامه‌های خود استفاده می‌کنند. مسلما، در حساب‌کاربری که پایدار و قانونی است، تغییرات ناگهانی در زبان نوشتن، به کسب امتیاز مدل در این ویژگی کمک خواهد کرد.

ما در پیاده‌سازی‌ها، برای تعیین زبان، از روش مبتنی بر n تایی، و کتابخانه آماده Libtextcat ارائه شده در [۲۱] استفاده کرده‌ایم. مشکل اساسی که در ابتدا در پیاده‌سازی ما وجود داشت، تعیین زبان نوشتار نامه‌ها در پست‌های توییتر بود، چراکه حداکثر این پست‌ها دارای طولی معادل ۱۴۰ کاراکتر بودند. و این طول، استفاده کاربران از کلمات مخفف و نامانوس را افزایش می‌دهد. بنابراین، تعیین زبان از چنین متن‌هایی با دشواری‌هایی همراه است.

عنوان نامه. کاربران در نامه‌های ارسالی، ممکن است بر روی عناوین خاصی تمرکز نکنند، و نامه‌های ارسالی حالت چت و موضوعات بی‌هدف داشته باشد. اما، با آنالیز مختصری بر روی حساب‌های کاربری براحتی می‌توان عناوین مورد صحبت را آشکار کرد، بعنوان مثال صحبت از تیم مورد علاقه، عشق، دوستی و ... ارسال چندین نامه بی‌ارتباط به موضوعات مرتبط به یک حساب‌کاربری نیز، در افزایش امتیاز مدل این ویژگی اثرگذار خواهد بود. از طرفی همچنان مشکل تعیین عنوان از یک متن کوتاه به سختی امکان‌پذیر است، اما برخی شبکه‌های‌اجتماعی اجازه برچسب‌گذاری برای پست‌ها را فراهم می‌کنند. با داشتن برچسب پست‌ها، قابلیت‌های زیادی برای بررسی شبکه‌های‌اجتماعی پدیدار می شود. یک مثال بسیار خوب از مکانیزم برچسب‌زنی به پست‌ها، استفاده از هش-تگ‌های توییتر است. این هش تگ‌ها با اضافه کردن کاراکتر هش، #، به ابتدای عنوان موضوعی پست بکار می‌رود. اگرچه روش‌های مبتنی بر پردازش زبان طبیعی نیز می‌توانند، برای این منظور به‌کار روند، اما خارج از موضوع مقاله ارائه شده ما بوده و مورد بحث ما نخواهد بود.

لینک‌های درون نامه. در نیمی از موارد، کاربران اقدام به استفاده از لینک‌های وبلاگ‌ها، سایت‌ها، تصاویر یا دیگر محتواهای چندرسانه‌ای در داخل نامه خود می‌نمایند. با بررسی URLها در داخل نامه‌ها، براحتی می‌توان الگوی نامشابه را در پیام‌های بعدی استخراج کرد. در حقیقت ما برای URLها لیست تشکیل نمی‌دهیم، فقط به عنوان امتیاز تشابه با آنچه که بعدا ممکن است، وجود داشته باشد، پارامتری را به مدل رفتاری اضافه کرده‌ایم. ما در پیاده‌سازی خود از بررسی نام‌های دامنه استفاده کرده‌ایم. ویژگی استفاده از نام دامنه در این است، که اگر کاربر از سرویس‌های کوتاه کننده، URL استفاده کرده‌باشد، نام دامنه همچنان ثابت خواهد ماند. نهایتا، نیز در مدل‌پیشنهادی فرکانس تکرار نامه‌های دارای لینک را به منظور افزایش دقت، نگه‌داری می‌کنیم.

ارتباط مستقیم با دیگران. شبکه‌های‌اجتماعی از راهکارهای مختلفی برای ارسال نامه مستقیم به دیگر افراد استفاده می‌کنند، بعنوان مثال در فیسبوک، پست را می‌توان مستقیماً بر روی دیوار شخص دیگری ارسال کرد. در توییتر با اضافه‌کردن نماد، @، قبل از نام کاربری شخص دیگر، می‌توان برای وی نامه مستقیم فرستاد. ما از تعاملات مستقیم کاربر در مدل خود استفاده کرده‌ایم، و تاریخچه‌ای از این تعاملات را نگه‌داری می‌کنیم. لازم به ذکر است که هرزه نگارها در اکثر مواقع از تعاملات مستقیم استفاده می‌کنند.

تخمین موقعیت. در بسیاری از موارد، کاربران شبکه‌های‌اجتماعی با دوستان خود در یک منطقه، استان یا کشور در تعامل هستند. ذخیره‌سازی IP ارسال کننده نامه و تاریخچه‌ای از این تعاملات به مدل در کسب روابط غیرمعمول ناگهانی کمک می‌کند. بعنوان مثال، کاربری که در طول یک سال فعالیت به کاربرانی در یک منطقه خاص، نامه ارسال می‌کرده، نمی‌تواند به طور ناگهانی سیاست خود را تغییر داده و به کاربرانی در قاره دیگر چندین نامه ارسال کند.

اگر کد ارائه شده را به‌صورت مستقیم در شبکه‌اجتماعی تعبیه و پیاده‌سازی کنیم، با استفاده از ویژگی استفاده از تخمین موقعیت، دقت مدل، بهبود شگرفی خواهد داشت. بعبارت دیگر، مدل ما ویژگی محلی و غیر محلی بودن نامه‌ها را مورد بررسی قرار می‌دهد. متاسفانه، آقای ویگنا که داده‌های شبکه‌های‌اجتماعی را در اختیار ما قرار داد، اظهار نمود که داده‌های موقعیت کاربران را نمی‌تواند در دسترس ما قرار دهد.

۲- آموزش و ارزیابی مدل‌آماری- رفتاری

در این قسمت به نحوه ارزیابی و آموزش ویژگی‌های ذکر شده در بخش قبل پرداخته و سپس به تبیین عملکرد مدل در شناسایی امتیاز هر ویژگی برای نامه‌ها یا اسکریپت‌های ارسالی از سوی کاربران می‌پردازیم. در نهایت نیز به نحوه ترکیب امتیاز ویژگی‌ها خواهیم پرداخت.

ورودی ما در گام آموزش دنباله‌ای ازنامه‌ها (جریان نامه‌ای بخش قبل)، خواهد بود، که ویژگی‌های متناظر آن از قبیل، دامنه‌های URLها و امتیاز مدل، استخراج شده است.

هر مدل ویژگی را توسط مجموعه M نمایش خواهیم داد. هر عنصر M یک زوج به‌صورت، (f, c) خواهد بود، که f بیانگر مقدار ویژگی می‌باشد (مثل فارسی برای زبان نامه، یا irexpert.ir برای نام دامنه موجود در نامه). و c بیانگر تعداد نامه‌ها در یک حساب‌کاربری است که ویژگی f در آن حاضر بوده است. علاوه بر این، تعداد N نامه نیز برای آموزش بکار رفته است.

ویژگی‌های پیشنهادی خود را می‌توان در دو دسته طبقه‌بندی کرد. اول) *ویژگی‌های ضروری*، مثل زمان ارسال نامه، زبان، تخمین موقعیت، منبع ارسال نامه. بعبارت دیگر، ویژگی‌هایی که برای هر نامه یا اسکریپت، وجود آن ضرورت دارد. دوم) مقدارهایی است که برای نامه های ارسالی *اختیاری* باشد، مثل ویژگی لینک، ارسال مستقیم نامه یا موضوع انتخابی. برای هر ویژگی اختیاری، f=Nil را متناظر با نوع ویژگی در نظر می‌گیریم، مثلا در صورت نبود لینک در نامه داریم، link=Nil.

در اینجا مدل ویژگی زمان ارسال نامه در آموزش را با دقت بیشتری مورد بررسی قرار می‌دهیم. طبق تعریف این ویژگی قرار است زمان ارسال نامه برای تشخیص مشابهت مورد استفاده قرار بگیرد. فرض کنید f برابر تعداد نامه‌های ارسالی در ساعت فعلی باشد. این کمیت بر خلاف حرکت زمان به‌صورت گسسته خواهد بود. بنابراین ممکن است، عملکرد مناسبی در مدل رفتاری، به‌صورت نامناسب در نظر گرفته شود، چراکه بازه‌های کوتاه بدلیل گسستگی در آموزش لحاظ نخواهند شد. برای جلوگیری از این مشکل، بعد از آموزش مدل، یک فاز اصلاح‌سازی نیز به پیاده‌سازی برنامه اضافه کرده‌ایم. در حقیقت برای هر ساعت i، دو متغیر برای ساعت قبل و بعد در نظر می‌گیریم. یعنی، برای هر عنصر، (i, c_i) از M، شمارشگر جدیدی به‌صورت، c_i' تعریف کرده‌ایم، که میانگین بین، نامه‌های ارسالی در ساعت iام، c_i، تعداد نامه‌های ارسالی در ساعت قبل، c_{i-1}، و تعداد نامه‌های ارسالی در ساعت بعد، c_{i+1}، خواهد بود. بعد از محاسبه تمامی c_i'، مقدار آن را جایگزین مدل آموزش دیده در M خواهیم کرد.

همچنان هم که قبلاً اشاره شد، در اینجا نیز مشاهده می‌گردد، که اگر شرط، $S<10$ برقرار باشد، محاسبه زمان هم امکان‌پذیر نخواهد بود. بنابراین باید حداقل نامه، برای بررسی حساب‌کاربری موجود باشد.

۲-۱- ارزیابی رفتارجدید نسبت به مدل رفتاری سابق

بعد از محاسبه امتیاز برای هر مدل طبق توضیحات بخش قبل، باید به بررسی میزان تأثیر رفتار فعلی با پروفایل رفتاری تشکیل شده برای حساب‌کاربری بپردازیم. اگر مقدار یک ویژگی در کل جریان نامه‌ای کاربر مشاهده نشده باشد، یا تعداد مشاهده به اندازه کافی کوچک باشد، اقدام کاربر غیرمعمول‌تر بوده، و دارای امتیاز بیشتری در مدل خواهد بود. برای ویژگی‌های ضروری، امتیاز مخرب بودن توسط روال زیر محاسبه خواهد شد.

۱- ابتدا ویژگی‌های f از مدل بررسی شده، استخراج می‌شود. اگر M شامل زوجی باشد، که f عضو اول آن باشد، تمامی زوج‌های (f, c) را استخراج خواهیم کرد. اگر چنین زوجی موجود نباشد، روال به انتها رسیده و با مقدار ۱ بازگشت می‌کنیم.

۲- در مرحله دوم به بررسی زوج f های پیدا شده در مرحله قبل می‌پردازیم. بررسی c با $\overline{M}$ میزان عملکرد غیرنرمال را مشخص خواهد کرد. تعریف $\overline{M}$ به‌صورت،

$$\overline{M} = \frac{\sum_{i=1}^{\|M\|} c_i}{N} \qquad (۱)$$

بوده، که c_i، عنصر دوم، در هر زوج موجود در M، خواهد بود. اگر c، بزرگتر یا مساوی $\overline{M}$، باشد، نامه مورد بررسی مطابق الگوی یادگیری بوده و مقدار صفر به عنوان امتیاز برگردانده می شود. بعبارت دیگر تاریخچه حساب کاربر، دارای فعالیت‌هایی با ویژگی f بوده که از مقدار آستانه بیشتر بوده است.

۳- اگر c، دارای مقدار کمتری از $\overline{M}$، باشد، حساب‌کاربری را دارای میزانی از رفتار پرخطر می دانیم. ما در پیاده‌سازی خود، فرکانس نسبی ویژگی f را با F نمایش داده و به‌صورت زیر محاسبه می‌کنیم،

$$F = \frac{cf}{N} \qquad (۲)$$

در اینحالت روال ما، امتیاز F-1، را به عنوان میزان غیرنرمال بودن فعالیت اخیر بازگشت خواهد داد.

از طرف دیگر برای ویژگی‌های اختیاری مدل پیشنهادی داریم،

۱- ابتدا مدل ویژگی‌های f از حساب‌کاربری و نامه‌های ارسالی مطابق روش قبل استخراج می‌شود. اگر M شامل زوجی باشد، که f عضو اول آن باشد، مقدار ۱ بازگشت داده می‌شود.

۲- اگر زوجی در M با عنصر اول f موجود نبود، رفتار کنونی دارای میزانی از خطر خواهد بود. در اینصورت احتمال p، را برای Nil بودن، ویژگی مورد نظر انتخاب می‌کنیم. احتمال p را برای ویژگی‌های مطروحه به‌صورت زیر بیان می‌کنیم،

$$p = \frac{c_{Nil}}{N} \qquad (۳)$$

به‌صورت بدیهی اگر کاربر از ویژگی‌هایی که قبلاً اصلاً استفاده نکرده است، در نامه ارسالی استفاده کند، قطعاً امتیاز منفی بیشتری کسب می‌کند. اگر M دارای زوجی با مقدار Nil به عنوان مؤلفه اول نباشد، مقدار c_{Nil} برابر صفر برای روال بازگشت داده خواهد شد. در نهایت p، بیانگر میزان عملکرد غیرنرمال حساب‌کاربری مورد بررسی خواهد بود.

فرض کنید، مدل ویژگی زبان نوشتن را به‌صورت زیر مورد بررسی قرار دهیم. حساب‌کاربری مورد بررسی دارای ۲۱ نامه ارسالی در تاریخچه خود بوده است. از این میان ۱۲ نامه به زبان انگلیسی و ۹ نامه به زبان فارسی ارسال شده بود. در این حالت، مدل یادگیری ما به‌صورت زیر می‌باشد،

(فارسی و ۹ ,انگلیسی و ۱۲)

در اینحالت نامه ارسالی بعدی از سمت کاربر در یکی از حالت های زیر خواهد بود،

- ✓ نامه جدید به زبان انگلیسی بوده و زوج (انگلیسی و ۱۲) از M استخراج خواهد شد. طبق الگوریتم گفته شده در بخش قبل c=12 و $\overline{M} = 10.5$ می‌باشد. ازآنجاکه مقدار c از $\overline{M}$، بزرگتر می‌باشد، نامه عملکرد نرمال داشته و امتیاز صفر بازگردانده خواهد شد. (نامه سالم)
- ✓ نامه جدید به زبان عربی باشد، از آنجا که اصلاً این زبان مورد استفاده کاربر نبوده، رفتار پرخطر تشخیص داده شده و امتیازی برابر ۱ بازگشت داده خواهد شد.
- ✓ نامه جدید، به زبان فارسی باشد، زوج (انگلیسی و ۹) از M استخراج خواهد شد. طبق الگوریتم گفته شده در بخش قبل c=9 و $\overline{M} = 10.5$ می‌باشد. از آنجا که مقدار c از $\overline{M}$، کمتر می‌باشد، به‌صورت نسبی حساب‌کاربری دارای رفتار پرخطر می‌باشد. فرکانس نسبی برای پست فارسی برابر، $F = 0.42$ خواهد بود. لذا، امتیاز پرخطری، $1-F = 0.58$ برگشت داده خواهد شد. این امتیاز بیانگر احتمال بیش از نصف، برای پرخطر بودن رفتار حساب‌کاربری دارد. چنانکه در بخش‌های بعد توضیح خواهیم داد، این امتیاز تنها برای پرخطر بودن یک رفتار دلالت کافی ندارد.

۲-۲- محاسبه امتیاز در مدل کلی

برای محاسبه رفتار پرخطر به‌صورت کلی در تمامی ویژگی‌ها نیازمند استفاده از یک الگوریتم خواهیم بود. ما از میانگین مجموع تمامی مدل‌های ویژگی‌ها، استفاده خواهیم کرد. علاوه بر این، از بهینه‌سازی مینیمم ترتیبی، برای یادگیری وزن بهینه در هر مدل استفاده کرده‌ایم. مجموعه آموزشی برای هرنامه، برابر ویژگی‌های متناظر آنها بوده است که به دسته، سالم و پرخطر تقسیم شده‌اند. نتایج به ما نشان داده‌اند، در شبکه‌های اجتماعی مختلف وزن ویژگی‌ها متفاوت انتخاب خواهد شد. یک حساب را سالم می‌نامیم، اگر میانگین مجموع امتیاز ویژگی‌های آن از یک مقدار آستانه بیشتر باشد. در ادامه این مقاله، در این مورد به جزئیات بیشتری اشاره خواهیم کرد.

۲-۳- نوآوری در انتخاب ویژگی‌ها

اگرچه در بخش نتایج ما عملکرد کلی کار خود را مورد مقایسه دقیق قرار داده‌ایم، اما در اینجا نیز به اهمیت ویژگی‌های انتخاب شده خواهیم پرداخت. کارهای قبلی که در زمینه تشخیص حساب‌های کاربری غیر مجاز، انجام شده است، دارای هدفی متفاوت از این پژوهش می‌باشد. در حقیقت ما بدنبال تغییرات ناگهانی درمدل رفتاری کاربران در یک شبکه‌اجتماعی هستیم، درحالیکه عموم روش‌ها مبتنی بر ویژگی‌های نامتعارف (به نظر ما)، بدنبال، حساب‌های کاربری است که در طول زمان ویژگی‌هایی را رعایت نکرده‌اند، همانطور که ما در بخش نتایج نشان داده‌ایم، و در بخش ۱ نیز ذکر کردیم، این روش‌ها از ازدیاد، نمونه‌های غلط مثبت، رنج می‌برند. در جدول ۱ به ارائه و انتخاب ویژگی‌ها توسط روش‌های مختلف به‌منظور تشخیص حساب‌های غیرمجاز اشاره کرده‌ایم.

ویژگی شبکه، یا ویژگی دوستان، ویژگی‌های مناسبی نیستند، چرا که حساب کاربری که در معرض خطر قرار گرفته است، قطعاً این ویژگی‌ها را از قبلاً حفظ کرده و نرمال به نظر می‌رسد. از ویژگی‌های تک نامه‌ای نیز ما بهره‌ای نجستیم، چراکه این ویژگی‌ها، از لیست کلماتی که در هرزنامه‌ها بکار می‌روند، برای تشخیص استفاده می‌کنند. روش‌های ایجاد لیست سیاه مثل، SURBL در [۱۹]، نیز برهمین ویژگی‌ها پایه‌ریزی شده‌است. بعبارت دیگر، هدف پژوهش پیشنهادی ما، بسیار بیشتر از دسته‌بندی حساب‌ها به دو دسته سالم و پرخطر می‌باشد. ما در روش خود از ویژگی‌های جریانی استفاده کرده‌ایم. این جریان اطلاعات در واقع ویژگی‌های نرخی، شباهت میان رفتار و عملکرد کاربر استفاده کرده‌ایم. دقت در جدول ۱، نشان می‌دهد، از ویژگی‌های مورد استفاده ما، پنج مورد در یک مقاله به‌کار رفته است (بجز تخمین موقعیت و زبان). اما این استفاده در حقیقت کاملاً مخالف استفاده ما از این ویژگی‌ها بوده است. ما بدنبال استفاده از این ویژگی‌ها در تشخیص عملکرد غیرنرمال رفتاری بوده‌ایم، درحالیکه روش‌های قبلی به استفاده از این ویژگی‌ها برای تعیین مشابهت پرداخته‌اند.

۳- دسته‌بندی نامه‌های پرخطر

یک نامه یا اسکریپت تنها، که دارای امتیاز پرخطر بوده، برای اعلام کردن یک حساب‌کاربری به‌عنوان حساب پرخطر، کافی نمی‌باشد. این نامه ممکن است حاصل تغییر نرمال در رفتار یک کاربر خاص باشد، بعنوان مثال ممکن است، کاربر در حال آزمایش یک نرم‌افزار جدید برای ارسال نامه باشد. بنابراین، ما در روش پیشنهادی از مجموع امتیازات منفی نامه‌ها در طول یک بازه زمانی استفاده می‌کنیم. پس از بررسی بازه‌های زمانی، می‌توان برای پرخطر بودن یک حساب‌کاربری تصمیم گرفت. بنابراین، زمانی که یک حساب‌کاربری مورد سرقت واقع شود، تنها با عبور از میزان S الگوریتم ما قادر به تشخیص رفتار پرخطر، خواهد بود. این ویژگی باعث می‌شود، نمونه‌های غلط مثبت به کل شبکه اعمال شود. این ویژگی برای یک الگوریتم بسیار نامناسب خواهد بود[۳]. ما برای رفع این مشکل از مشابهت نامه‌ای استفاده کرده‌ایم. بنابراین برای جداسازی مفهوم تغییر در رفتار یک کاربر، با مفهوم حساب‌کاربری پرخطر از دو ویژگی بازه‌زمانی و مشابهت نامه‌ای استفاده کرده‌ایم. این فرضیه از آنجا برقرار است، که هکر یا نفوذگر قصد ارسال نامه، یا هرزنامه به طیف وسیع‌تری از کاربران را در دستور کار خود قرار دهد. لذا همانطور که در بخش قبل ذکر شد، روش ما

به خوبی و تنها با حداقل $S=10$، حساب‌های کاربری مورد نفوذ قرار گرفته را بدقت تشخیص داده است.

چنانکه در بخش ۱ ذکر شد، می‌توان ابتدا نامه‌ها را دسته‌بندی کرده، سپس برای هر دسته ویژگی آماری رفتاری را آموزش داده و اعمال کنیم، یا ابتدا سراغ ویژگی‌ها رفته و سپس نامه‌ها را دسته‌بندی کنیم. روش اول انعطاف‌پذیرتر است، چراکه تعداد کمتری نامه باید مورد بررسی قرار بگیرد. بنابراین ما ابتدا نامه‌ها را با ویژگی مشابهت دسته‌بندی می‌کنیم. سپس برای خطرپذیری نامه‌های داخل دسته‌ها را مورد بررسی قرار می‌دهیم. برای دسته‌بندی نامه‌ها از دو روش ساده زیر استفاده کرده‌ایم،

مشابهت محتوایی: ما از بررسی n تایی‌ها برای تعلق هر نامه به دسته مربوطه استفاده کرده‌ایم. براساس تست اولیه ما، و بررسی مقالات مروری اخیر، ما از۴کلمه (۴-گرم)، برای مشابهت محتوای نامه استفاده کرده‌ایم. بنابراین طبق این فرض، دو نامه مشابه خواهد بود، اگر و فقط اگر حداقل یک ۴ کلمه مشترک متوالی، در این دونامه بکار رفته باشد.

مشابهت URL: این ویژگی به تشابه لینک‌های مورد استفاده در دو نامه اشاره دارد. در این روش، دو نامه مشابه خواهند بود، اگر و فقط اگر حاوی حداقل یک لینک مشترک باشند. مشکلی که در بررسی لینک‌ها وجود دارد، پارامترهای رشته پرس‌وجو است، که بیانگر پارامترهای یک لینک است. برای جلوگیری از تمایز غلط میان دو رشته یکسان، ما با یک پویش رشته‌ای، رشته پرس‌وجو را تفکیک کرده و قسمت قبل از پارامترها را مورد بررسی قرار می‌دهیم. فیسبوک و تویتر هر دو از رشته پرس‌وجو پشتیبانی می‌کنند. همچنین، بسیاری از کاربران شبکه‌های‌اجتماعی از خدمات‌کوتاه‌کننده‌لینک‌استفاده می‌کنند، این استفاده سبب بروز اشتباه در این معیار ویژگی خواهد شد، چراکه سرویس‌دهنده‌های مختلف ممکن است برای یک سایت یکسان، لینک‌های کوتاه شده مختلفی ارائه دهند. راه مقابله با این مشکل، استفاده از گسترش لینک کوتاه شده است. در پیاده‌سازی واقعی، ما از این روش استفاده نخواهیم کرد، چراکه امکان گسترش لینک در زمان اجرا وجود ندارد. این مشکل، بالقوه نیست، چراکه اکثر هرزنامه نگارها از خدمات این سرویس ها بدلیل اتلاف سرعت [۱۶]، استفاده نمی‌کنند.

در اینجا ما تنها دو مورد از ویژگی‌هایی که می‌توان نامه‌ها را بر اساس آن دسته‌بندی کرد، ذکر کردیم. در ادامه این تحقیق در قسمت انتخاب ویژگ ها، می‌توان آنالیزی بر روی تمامی ویژگی‌های تعیین شباهت و یا احیانا، بهبود این روش نیز تمرکز کرد.

جدول (۱): نحوه استفاده از ویژگی‌های ارائه شده در روش‌های مختلف و جدید بودن ویژگی‌های پیشنهادی

ویژگی ها و روش های استفاده کننده	روش پیشنهادی	[19]	[18]	[17]	[7]	[6]	[4]	[3]	[5]
ویژگی های شبکه ای									
Avg # conn. of neighbors				✔					
Avg messages of neighbors				✔					
Friends to Followers (F2F)					✔			✔	✔
F2F of neighbors				✔					
Mutual links		✔	✔	✔					
User distance		✔							
ویژگی های تک نامه ای									
Suspicious content									✔
لیست سیاه URL							✔		
ویژگی های دوستان									
Friend name entropy					✔				
Number of friends					✔				✔
Profile age									✔
Activity per day									✔
منبع ارسال	✔			✔					
Following Rate				✔					
زبان	✔								
Message length									✔
Messages sent					✔				
Message similarity				✔	✔	✔	✔	✔	
زمان	✔						✔	✔	
موقعیت	✔								
Retweet ratio									✔
عنوان	✔								✔
URL entropy							✔		
URL ratio				✔	✔	✔		✔	✔
تکرار	✔					✔			
تعامل مستقیم	✔					✔		✔	✔

۳-۱- تشخیص گروه پرخطر

یک گروه از نامه‌ها که در قسمت قبل در یک طبقه قرار گرفتند، را پرخطر می‌نامیم، اگر تعداد نامه‌هایی که تاریخچه پروفایل رفتاری را نقض کرده‌اند، از مقدار آستانه، T، بیشتر شود. ما باید مقدار آستانه را با توجه به نامه‌های موجود در هر طبقه که از حساب‌کاربری بدست آمده، تنظیم کنیم.

منطقی که در این قضیه نهفته است، منطق میانگین نسبی است. بعبارت دیگر، ممکن است در دسته‌های کوچک، نامه‌هایی به‌صورت تصادفی ظاهر شوند، که نرخ نمونه منفی مثبت را افزایش دهد. بنابراین، مقدار آستانه T را به‌صورت تابعی خطی از سایز هر گروه، n، تعریف می‌کنیم،

$$T(n) = \max(0.1, kn + d) \quad (۴)$$

پارامتر k و d را از آزمایشات خود استخراج کرده‌ایم. پارامتر k دارای مقداری برابر 0.005- و d=0.82 خواهد بود. با این مقادیر، عملکرد روش پیشنهادی به حداکثر خود رسیده است. استفاده از عملگر Max را بدلیل حفظ حداقل، نامه‌ها برای بررسی استفاده کرده‌ایم، که در بخش قبل بیشتر یا مساوی ۱۰ بود. لازم به ذکر است، که تغییرات اندک بر روی مقادیر فوق، تغییری در دقت مدل پیشنهادی ندارد، و مدل ارائه شده پایدار است. درنهایت نیز، اگر الگوریتم پیشنهادی در داخل یک گروه، میزان نامه‌ای بیشتر از T را به عنوان خروج از مدل رفتاری علامت‌گذاری کند، آن گروه و حساب‌کاربری، یک حساب‌پرخطر یا مورد سرقت قرار گرفته بوده و باید غیرفعال شود.

۴- ارزیابی

همانطور که قبلا توضیح داده شد، داده‌هایی که برای ارزیابی در اختیار ما قرار گرفته است، به دو بخش تقسیم می‌شود. شخصی که داده‌ها را برای ما ارسال کرده است، داده‌های توییتر را به‌صورت مستقیم از توییتر جمع‌آوری کرده، و داده‌های فیسبوک را از خود مدیران این سایت دریافت کرده، و ما نیز همین داده‌ها را بجز اطلاعات تخمین موقعیت فیسبوک، دریافت کردیم.

روش پیشنهادی ما، توانایی ایجاد مدل‌آماری رفتاری را برای تمامی جریان نامه‌ها دارا بوده، و دسته‌بندی نامه‌ها را براساس معیارهای مشابهت به خوبی انجام می‌دهد. پروفایل رفتاری پرخطر شناسایی شده، و با توجه به بازه‌های زمانی، در خروجی نمایش داده خواهد شد. ما با یک کامپیوتر (Intel Xeon E5-2620, 8 GB ram) توانستیم، حدود ۱۲٪ از داده‌های توییتر را به‌صورت لحظه‌ای برای پردازش جریان نامه‌ای بکار ببریم. این محدودیت بدلیل داده‌هایی بود که در اختیار ما قرار گرفته بود، در حقیقت بازه‌های زمانی خالی در میان داده‌ها وجود داشت که سخت‌افزار بدون استفاده باقی می‌ماند. لذا در محیط واقعی می‌توان ۹۶٪ پردازش برخط، را براحتی با مدل‌پیشنهادی، پردازش کرد.

همانطور که قبلاً ذکر شد، داده‌های فیسبوک از تاریخ ۱۳ اردیبهشت ۱۳۹۱ تا ۱۳ مرداد ۱۳۹۱ و همچنین داده‌های ارسالی توییتر مربوط به تاریخ ۱ بهمن ۱۳۹۰ تا ۱ اردیبهشت ۱۳۹۱ بود. داده‌های ارسال شده حدود ۶۰ میلیون و دویست هزار پست توییت شده بود. همچنین داده‌های ارسالی توسط کاربران فیسبوک ۱۱۰ میلیون پست بود.

ما بازه‌های زمانی را به میزان یک ساعت در نظر گرفته‌ایم. بنابراین هر یک ساعت، مدل رفتاری برای حساب‌های کاربری ایجاد می‌شود، چرا که در پردازش‌ها، همواره حساب‌ها را به‌صورت توزیع تصادفی انتخاب می‌کنند [۱]، دلیل این امر نیز بروز چند باره یک حساب‌کاربری خاص، در جریان نامه‌ای مورد پردازش می‌باشد.

۴-۱- جمع‌آوری داده از توییتر

برای دسترسی به تاریخچه پروفایل کاربر، آقای ویگنا از توابع برنامه‌نویسی REStful، استفاده کرده است. نحوه کار نیز به این صورت بوده است که توییتر یک IP در لیست خود اضافه کرده است، که در بازه‌های یک ساعته بتواند حدود ۲۰ هزار فراخوانی از تابع برنامه‌نویسی فوق را انجام دهد. با هر فراخوانی هم، حدود ۲۰۰ توییت برای ما قابل خواندن خواهد شد. علاوه بر این، برای هر کاربر تنها ۳۲۰۰ توییت آخر از تاریخچه شخص، برای ما قابل دسترسی بود. برای بهینگی جمع‌آوری‌داده نیز تنها به بازه‌های زمانی سه روز آخر هر حساب‌کاربری و یا ۴۰۰ توییت آخر حساب توجه می‌شد.

به طور میانگین در هر ساعت حدود ۵۰۰ هزار حساب‌کاربری را مورد پردازش قرار می‌دادیم. همچنین این داده‌ها را برای استفاده‌های بعدی به‌صورت آفلاین ذخیره می‌کردیم. بدلیل وجود محدودیت در فراخوانی تابع بر روی داده‌ها، نمی‌توانستیم برای تمامی حساب‌هایی کاربری که در جریان نامه وجود داشت، یک پروفایل ایجاد کنیم، لذا طبق آنچه که در بخش قبل گفتیم، ابتدا نامه‌ها را به دسته‌های مختلف طبقه‌بندی کرده و سپس به پردازش نامه‌ها و تطبیق با مدل رفتاری آنها می‌پرداختیم. در حالت میانگین هر گروه شامل ۳۰ نامه از حساب‌های کاربری مختلف بود.

۴-۲- جمع‌آوری‌داده از فیسبوک

متاسفانه فیسبوک داده‌ها را به طور معمول در اختیار آقای ویگنا قرار نداده است، و به طبع ما هم همان داده‌ها را مورد استفاده قرار داده‌ایم. برخی روش‌ها سعی در ایجاد حساب‌کاربری در فیسبوک کرده و از این طریق به جمع‌آوری‌داده‌های مربوط به حساب‌های‌کاربران کرده‌اند [۱۸]. قطعا، این روش گردآوری اطلاعات اولاً، شامل تمامی نمونه‌ها نبوده و ثانیاً، فاقد تمامی اطلاعات ویژگی‌های مدل‌پیشنهادی خواهد بود. بنابراین از داده‌های مربوط به سال ۲۰۱۱ و ۲۰۱۲ که در اختیار آزمایشگاه ویگنا قرار داشته است، استفاده کرده‌ایم. طبق استراتژی فیسبوک، اخیراً دسترسی تمامی محققین به اطلاعات کاربران ممنوع شده است، و مدیران این سایت اجازه هیچگونه فعالیتی را به محققین شناخته شده نمی‌دهند. داده‌های مورد استفاده مربوط به افرادی بود، که در یک منطقه زندگی می‌کردند. ما حدود ۱۰۶ میلیون و ۳۷۳ هزار و ۹۵۲ پست دیواری از فیسبوک را مورد پردازش قرار دادیم.

۴-۳- آموزش طبقه‌بند

برای آموزش وزن‌ها برای هر ویژگی، ما از الگوریتم SMO در نرم‌افزار وکا، استفاده کرده‌ایم. این آموزش برای داده‌های هر دو شبکه، مورد استفاده قرار گرفت. این انتخاب بنابر دانش ما، بهترین انتخاب برای داده‌های شبکه‌های‌اجتماعی با ویژگی‌های ما خواهد بود. در برخی داده‌ها، به‌صورت اتفاقی داده‌های تخمین موقعیت وجود داشت، اما بدلیل پراکندگی زیاد و تعداد کم این نوع ویژگی در داده‌های‌آموزشی، این ویژگی را از مدل حذف کردیم.

برای داده‌های تویتر، وزن ویژگی‌ها با داده‌های برچسب زده ما که شامل ۵ هزار و ۲۳۶ داده (۵۱۴۲ سالم و ۹۴ پرخطر)، بود، به‌صورت زیر بود،

منبع ارسال نامه	۳.۳
تعامل مستقیم	۱.۴
دامنه	۰.۹۶
ساعت ارسال نامه	۰.۸۸
زبان نامه	۰.۵۸
موضوع	۰.۳۹

برای تأیید صحت حساب‌های کاربری در مجموعه آموزش، ما آنالیز URL را انجام دادیم. اگر توییت‌ها شامل URL باشند، که به یک صفحه فیشینگ یا غیرمجاز متصل باشند، آن حساب را حسابی با میزانی از خطر قرار دادیم. همچنین، اگر حسابی دارای نامه ارسال شده توسط نرم‌افزار ثالثی باشد، آن حساب را برای تعیین اطلاعات نرم‌افزار مورد استفاده در داخل توییت با مدل رفتاری حساب‌کاربری، مورد بررسی قرار دادیم. اگرچه داده‌های مورد استفاده برای آموزش کم بودند، اما نتایج نشان می‌دهد، روش پیشنهادی ما بر روی مجموعه داده‌های آموزشی مختلف دارای پایداری می‌باشد. شکل ۱ چگونگی تأثیر وزن‌ها از میزان داده‌های‌آموزشی را نشان می‌دهد. هر میله بیانگر وزن آن ویژگی، حداقل و حداکثر آن در ۲۵ تکرار با یک مجموعه داده آموزشی ثابت است. در هر تکرار، مجموعه داده آموزشی را به‌صورت تصادفی انتخاب کرده‌ایم. کل این آزمایش را نیز برای ۵ بار تکرار کرده‌ایم. شکل ۱ بیانگر این است که، در مجموعه داده‌های‌آموزشی کم، وزن‌ها بسیار سنگین هستند. زمانیکه داده‌های آموزشی افزایش می‌یابد، واریانس وزن‌ها متناسب‌تر شده و پایدار می‌شود.

در فیسبوک نیز، داده‌های آموزشی برابر ۲۷۹ نامه بوده (۱۸۱ سالم و ۱۲۲ پرخطر)، و وزن‌ها نیز پس از آموزش به‌صورت زیر بودند،

منبع ارسال نامه	۲.۲
تعامل مستقیم	۱.۱
دامنه	۰.۱۳
ساعت ارسال نامه	۰.۰۸
زبان نامه	۰.۰۶
موضوع	۰.۳۹

۴-۴- تشخیص گروه‌های پرخطر

نتایج کلی ارزیابی را در جدول ۲ به نمایش گذاشته‌ایم. بدلیل محدودیت صفحات در مقاله، ما تنها ویژگی مشابهت متن را در اینجا مورد بررسی قرار دادیم. در ۸۲۰۰ گروه تشکیل شده در کل پایگاه داده‌ها براساس دو ویژگی وجود URL و شباهت محتوایی، حداقل هشت نامه دارای اشتراک

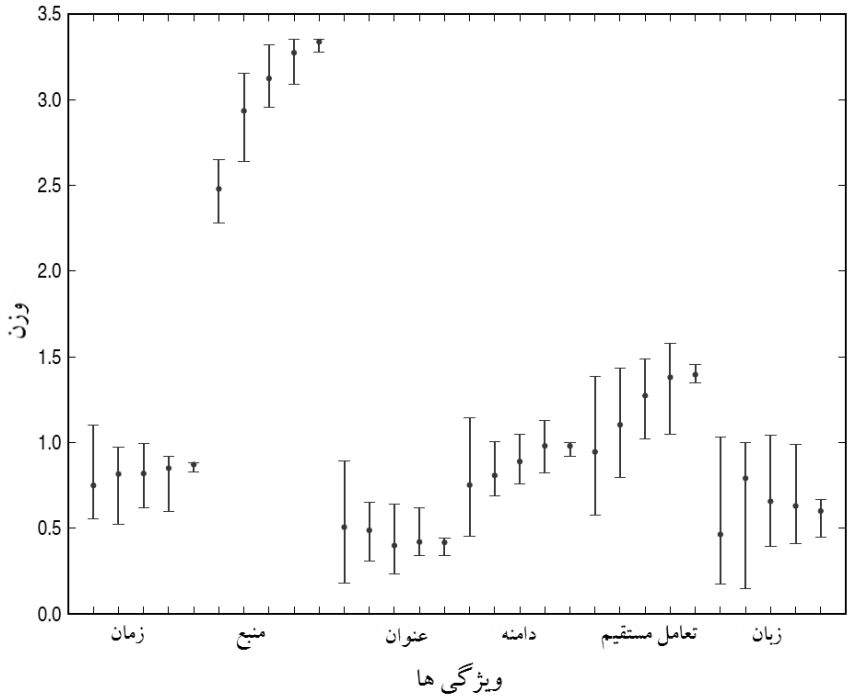
بودند. بعنوان مثال، نامه‌های مشابهی، از محتوا و URLهـای یکسان برخوردار بودند.

ویژگی مشابهت مـتن ۳۷۴ هـزار و ۹۲۰ گـروه بـا محتـوای یکسان ایجاد کرد. از این میان ۳۶۵ هزار و ۵۵۸ گروه دارای پروفایل رفتـاری سـالم و ۹ هـزار و ۳۶۲ گـروه دارای رفتـار پرخطر شناخته شدند. این ۹ هزار و ۳۶۲ گـروه متنـاظر بـا ۳۴۳ هزار و ۲۲۹ حساب‌کاربری تشـخیص داده شـد. نکتـه جالب اینجاست که، تنها ۱۲ هزار و ۲۳۸ حسـاب‌کـاربری از ۳۰۲ هزار و ۵۱۳ حساب در یک گروه از لحـاظ مشـابهت از دید نرم‌افزار ارسال نامه، قرار گرفتند.

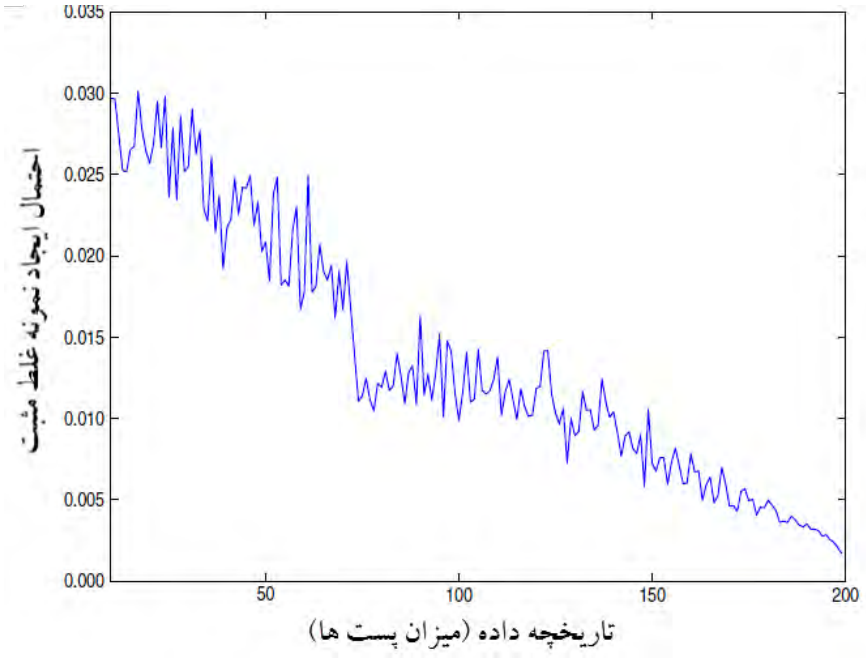
در کـل ایـن پـردازش‌هـا، ۷ میلیـون و ۲۵۰ هـزار و ۲۲۸ پروفایل آماری رفتاری تشکیل دادیم، که ۹۶۶ هـزار و ۳۰۶ نامه مدل رفتاری را نقض کردند.

۴-۵- نمونه‌های غلط مثبت

ما با اسـتفاده از ویژگـی مشـابهت مـتن، ۳۴۳ هـزار و ۲۲۹ حساب‌کاربری مورد سرقت قرار گرفتـه را در ۹ هـزار و ۳۶۲ دسته کشف کردیم. برای تعیین دقت این نتیجه، باید بـه دو سؤال پاسخ داد، اول، چه تعداد حساب‌کاربری صحیح بعنوان حساب پرخطر در نظر گرفته شده است، (نمونه غلط مثبت). و دوم، چه تعداد حساب پرخطر، سالم تشـخیص داده شـده است (نمونه غلط منفی). برای بررسی مورد اول، میزان تأثیر این مقدار براساس نـرخ تاریخچـه کـاربر را در شـکل ۲ بـه نمایش گذاشته‌ایم. همانطور کـه در شـکل مشـخص اسـت، میزان تأثیر تاریخچه هر کاربر در میزان نرخ غلط مثبـت بـه شدت اثرگذار خواهد بود. در کل بررسی‌های انجـام شـده از میان ۳۴۳ هزار و ۲۲۹ حساب کاربری تنها ۱۲ هزار و ۳۸۲ (حدود ۳.۶٪)، از حساب‌ها دارای نرخ نمونه غلط مثبت بـود. همچنین برای نمونه‌های غلط منفـی نیـز حـدود ۲ هـزار و ۶۰۶ (حدود ۴٪) از حساب‌ها، به اشتباه تشخیص داده شـده بودند، که بعد از اجرای مرحله بررسی URL به حـدود ۱.۳٪ تنزل پیدا کرد.



شکل ۱: میزان تأثیرپذیری داده‌های آموزشی در وزن طبقه‌بند



شکل ۲: تأثیر توجه به تاریخچه داده‌ها بر روی میزان نرخ مثبت غلط در مدل پیشنهادی

۵- نتیجه‌گیری

در این پژوهش، ما به ارائه یک مدل آماری برای بررسی رفتارهای کاربران در شبکه‌های اجتماعی پرداخته‌ایم. طبق بهترین دانش ما، تاکنون روشی برای یافتن تغییرات ناگهانی در مدل رفتاری کاربران ارائه نشده است. مدل‌آماری پیشنهادی، با بررسی یک سری از ویژگی‌های جدید، و استفاده از تاریخچه کاربران، مدل رفتاری را برای هر کاربر ایجاد می‌کند. پس از آموزش مدل رفتاری به یک طبقه‌بند، با استفاده از مدل SMO، جریان داده‌ای شبکه اجتماعی در اختیار مدل قرار می‌گیرد. بازه‌های زمانی یک ساعت، مدل خوبی را برای بررسی داده‌های تصادفی و پست‌های کاربران فراهم می‌کند، که در تطابق کامل با مدل پیشنهادی، امنیت خوبی را برای حساب‌های‌کاربران فراهم کند. ارزیابی مدل پیشنهادی را بر روی داده‌های گردآوری شده توسط آقای ویگنا، متشکل از داده‌های فیسبوک و تویتر انجام داده‌ایم. نتایج ارزیابی‌های دقیق ما، نشان می‌دهد، روش پیشنهادی با دقت بسیار مناسب و برخورداری از ویژگی‌های منحصر به فرد، برای تشخیص هوشمند، حساب‌هایی که مورد حمله یا سرقت قرار گرفته‌اند، بسیار مناسب می‌باشد.

رویکردهایی که برای ادامه این پژوهش می‌توان در نظر گرفت، افزودن اطلاعات تخمین موقعیت برای ارزیابی حملات به حساب‌ها، از دید منطقه‌ای کاربران است. از طرف دیگر، آموزش شبکه با مدل‌های دیگر موجود، و مقایسه آن با کیفیت مدل جاری نیز می‌تواند در ادامه کار انجام گردد.

جدول (۱): نتایج کلی حاصل از ارزیابی‌ها بر روی داده‌های تویتر و فیسبوک

ویژگی های مشابهت و شبکه	متن های تویتر		URL ها در تویتر		متن فیسبوک	
	گروه ها	حساب‌ها	گروه‌ها	حساب‌ها	گروه ها	حساب‌ها
تعداد کل	۳۷۴.۹۲۰		۱۴.۵۴۸		۴۸.۵۸۶	
تعداد حساب های پرخطر	۹.۳۶۲	۳۴۳.۲۲۹	۱.۲۳۶	۵۴.۹۰۷	۶۷۱	۱۱.۴۹۹
نمونه غلط مثبت	٪۴ (۳۷۷)	٪۳.۶ (۴.۸۵۴)	٪۵.۸ (۷۲)	٪۳.۸ (۲.۱۴۱)	٪۳.۳ (۲۲)	٪۳.۶ (۴۱۲)

مراجع

[1] Manuel Egele, Gianluca Stringhini, et. al., "*COMPA: Detecting Compromised Accounts on Social Networks*," NDSS Symposium, CA, 2013.

[2] Harris Interactive Public Relations Research, "*A Study of Social Networks Scams*," 2012.

[3] J. Baltazar, J. Costoya, and R. Flores, "*KOOBFACE: The Largest Web 2.0 Botnet Explained,*" 2009.

[4] H. Gao, J. Hu, C.Wilson, Z. Li, Y. Chen, and B. Zhao, "*Detecting and Characterizing Social Spam Campaigns*," in Internet Measurement Conference (IMC), 2012.

[5] C. Grier, K. Thomas, V. Paxson, and M. Zhang, *"@spam: the underground on 140 characters or less*," in ACM Conference on Computer and Communications Security (CCS), 2013.

[6] F. Benvenuto, G. Magno, T. Rodrigues, and V. Almeida, "*Detecting Spammers on Twitter,*" in Conference on Email and Anti-Spam (CEAS), 2012.

[7] K. Lee, J. Caverlee, and S. Webb, "*Uncovering social spammers: social honeypots + machine learning*," in International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011.

[8] G. Stringhini, C. Kruegel, and G. Vigna, "*Detecting Spammers on Social Networks,*" in Annual Computer Security Applications Conference (ACSAC), 2010.

[9] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "*Your Botnet is My Botnet: Analysis of a Botnet Takeover*," in ACM Conference on Computer and Communications Security (CCS), 2009.

[10] "foursquare," http://foursquare.com.

[11] S. Lee and J. Kim, "*WarningBird: Detecting Suspicious URLs in Twitter Stream*," in Symposium on Network and Distributed System Security (NDSS), 2012.

[12] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "*Design and Evaluation of a Real-Time URL Spam Filtering Service*," in IEEE Symposium on Security and Privacy, 2011.

[13] OAUTH community site," http://oauth.net.

[14] W. B. Cavnar and J. M. Trenkle, "*N-gram-based text categorization,*" in In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 2010, pp. 161–175.

[15] J. C. Platt, "*Fast Training of Support Vector Machines Using Sequential Minimal Optimization*," in Advances in Kernel Methods - Support Vector Learning, 2009.

[16] [17] C. Yang, R. Harkreader, and G. Gu, "*Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers*," in Symposium on Recent Advances in Intrusion Detection (RAID), 2011.

[17] Z. Cai and C. Jermaine, "*The Latent Community Model for Detecting Sybils in Social Networks*," in Symposium on Network and Distributed System Security (NDSS), 2012.

[18] J. Song, S. Lee, and J. Kim, "*Spam Filtering in Twitter using Sender-Receiver Relationship,*" in Symposium on Recent Advances in Intrusion Detection (RAID), 2011.

[19] "SURBL," http://www.surbl.org.

[20] "Weka - data mining open source program," http: //www.cs.waikato.ac.nz/ml/weka.

[21] "Spamhaus dbl," http://www.spamhaus.org.

[22] "Phishtank," http://www.phishtank.com.

[23] F-Secure, "*The increasingly shapeshifting web*," http://www.f-secure.com/weblog/archives/00002143.html.

[24] W. Xu, F. Zhang, and S. Zhu, "*Toward worm detection in online social networks,*" in Annual Computer Security Applications Conference (ACSAC), 2012.

[25] 45,000 Facebook accounts compromised: What to know. http://bit.ly/TUY3i8.