ارائه‌ی یک روش داده‌کاوی حجیم بر مبنای متن‌کاوی و آنالیز احساسات به منظور شناسایی رفتار هکرها

ماشاالله فرخی طامه[1]، عبدالرضا رسولی کناری[2]، محبوبه شمسی[3]
1- دانشجوی کارشناسی ارشد نرم افزار، دانشگاه صنعتی قم
2- دکتری نرم افزار، عضو هیئت علمی دانشگاه صنعتی قم
3- دکتری نرم افزار، عضو هیئت علمی دانشگاه صنعتی قم
Mft13802@gmail.com

چکیده
در این پژوهش یک روش داده‌کاوی حجیم بر مبنای متن-کاوی و آنالیز احساسات به منظور شناسایی رفتار هکرها بررسی شده است. در روش پیشنهادی در این پژوهش در مرحله ابتدایی پیش پردازش قرار دارد. دلیل استفاده از پیش پردازش این است که در بین داده‌های موجود متن‌هایی قرار دارند که یا ناقص هستند و یا قابل استفاده نیستند و همچنین بایستی ریشه یابی افعال نیز در این مرحله استفاده گردد. در مرحله دوم کاربران با استفاده از رگرسیون به پنج دسته تقسیم می‌شوند. در مرحله بعد با استفاده از الگوریتم tf-idf وزن کلمات را مشخص و بیان می‌کنند که توزیع کلمات در متن به چه صورت بوده است. در مرحله آخر تحلیل احساسات و تعیین نظر مثبت و یا منفی کاربران بوده است. برای شبیه سازی این روش از نرم‌افزار متلب استفاده شده است. نتایج نهایی این روش بیان می‌کند که میزان خطای مربعات در این روش مقدار 0.32بوده است و میزان دقت به دست آمده در این روش 90.9 درصد اعلام شده است. کارایی این روش در بین روش‌های مقایسه شده دارای مقدار بسیار بهنری است و نتایج خوبی را نمایش می‌دهد.

کلمات کلیدي: داده‌کاوی، متن‌کاوی، آنالیز احساسات، هکرها

1. مقدمه

امنیت سایبری یکی از موضوعات حساس و مهمی است که کل یک جامعه را چه از نظر فردی و چه از نظر گروهی، از نظر صنعتی و غیره تحت‌الشعاع قرار می‌دهد. انجمن اقتصاد مسئله‌ی امنیت

سایبری را هم‌تراز با مسائل اقتصادی و سیاسی قرار داده است. گزارش‌های خبری مربوط به هکرهای اینترنتی که اطلاعات کاربران را تهدید می کنند و از جرائم اینترنتی اهداف بلندپروازانه-ای را دنبال می‌کنند، به وقایع روزمره تبدیل شده است. مطالعاتی که اخیرا در حوزه‌ی امنیت سایبری صورت گرفته است نشان می‌دهد که ارزیابی رفتار و عملکرد هکرها بر مقابله با آنها و امنیت سایبری کمک شایان توجهی خواهد نمود. نمونه‌ای از خلافکاران اینترنتی هکرهایی است که بصورت آنلاین به شبکه-های اجتماعی وارد می‌شوند و ابزارها و تکنیک‌های مدرنی که در هک کردن اطلاعات در اختیار دارند را در بین سایر کاربران شبکه به اشتراک می‌گذارند. بنابراین تهدید اینترنتی هوشمند به عنوان تهدید اینترنتی مربوط به کامپیوترها، شبکه‌ها و فنآوری اطلاعات بصورت گسترده‌ای امروزه مطرح است. بنابراین سازمان‌ها و ارگان‌های مختلف نیازمند رویکردی قدرتمند به منظور آنالیز و تجزیه و تحلیل تهدید اینترنتی هوشمند یا CTI هستند تا بتواند بر توانمندی و قدرت امنیت سایبری خود بیافزایند. آنها به تدریج یاد می گیرند از تجزیه و تحلیل پیشرفته CTI که شامل تجزیه و تحلیل دقیق شبکه های تاریک، پیام های فروم و انجمن های چت اینترنتی است، آگاه باشند. این رویکردها مفیدتر از ابزارها و نرم‌افزارهای سنتی شناسایی حملات مخرب به شبکه است که پس از وقوع یک تهدید عملیاتی می‌شوند[1].

2- تئوری و پیشینه تحقیق

انجمن‌ها یا فروم‌های مربوط به هکرها یک بستر مناسبی را برای به اشتراک‌گذاری ابزارها و قابلیت‌های هک و خرابکاری فرآهم می‌آورد. در تحقیقاتی که در این زمینه صورت گرفته است، مطالعات تنها بر روی مشاهداتی از این انجمن‌ها صورت گرفته است که بصورت واضح و مشخص قابل دیدن است. اما آنچه که در این پژوهش در نظر گرفته شده است، پا را فراتر نهاده و به آنالیز محتوای پیام‌ها، درک احساسات و غیره برای شناسایی بهتر هکرها و رفتار آنها است. به عبارت دیگر رویکرد مورد نظر در این پژوهش این قابلیت را بدست می‌دهد که بتوان ویژگی‌های ویژه‌ی هر یک از هکرها را بدست آورد. براساس ویژگی‌های مفهومی متن‌های منحصر به فرد هر هکر و آنالیز احساسات پست‌های فروم و همچنین ویژگیهای هر انجمن مربوط به هکرها، یک مدل طبقه‌بندی برای پیش‌بینی نقش احتمالی یک هکر در انجمن بدست می‌آید. اما آنچه که به عنوان قدم اول راهکار ارائه شده در این پژوهش مطرح است، شناسایی انجمن‌ها و گروه-های هکر یا انجمن‌های سیاه برای آنالیز احساسات و متن‌کاوی پیام‌های ارسال شده توسط آنها است[1].

اساس نظر روش پیشنهادی بر اساس مطالعه‌ی رهبری انجمن‌ها قرار دارد که از دو منظر قابل بحث است: رهبری در انجمن‌های عملیاتی و نظریه کنترل و رهبری در شبکه‌های خرابکاری اینترنتی. در یک انجمن عملیاتی، توسعه بستگی به پویایی داخلی و همچنین توانایی رهبر دارد. این انجمن عملیاتی دارای یک مجموعه‌ای ویژگی‌ها است نظیر: حل سریع مشکلات موجود برای هکرها، توسعه‌ی مهارت‌های حرفه‌ای هکرها، انتقال شیوه‌های مدرن و جدید برای اعضای انجمن و ارائه‌ی ابزارهایی که توسط یک هکر طراحی و ساخته شده است و اشتراک آن بین سایر اعضای انجمن. اغلب رهبران قادر به حل و فصل اختلافات میان اعضای خود و یا برطرف کردن مشکلات در انجمن خود هستند. بنابراین در رویکرد پیشنهاد شده در این پژوهش با آنالیز متن‌کاوی و آنالیز احساسات رفتار هکرها از نقطه‌نظر رهبری انجمن هکرها مورد بررسی و آنالیز قرار می‌گیرد.

از طرف دیگر رویکرد بعد تئوری کنترل است. بصورتی که یک کاربر می‌تواند با شناسایی هکرها و اعضای انجمن، اطلاعات اشتراکی را دستکاری کند و به این ترتیب رفتار هکرها را کنترل و هدایت نماید. این رویکرد نیز یک راهکار دیگر در شناسایی رفتار و برخورد با هکرها است. این رویکرد نیز در این تحقیق پیشنهادی بر اساس آنالیز احساسات هکرها در انجمن در نظر گرفته شده است[2].

در زمینه تحلیل احساسات کارهای بسیاری انجام شده است. که هر کدام در یک زمینه و با یک خروجی مقبولی بوده‌اند ولی در هیچیک دقت صد در صد وجود ندارد. به همین دلیل تحقیق در این زمینه ادامه دارد. در منبع شماره [3] جداسازی متن می تواند یک جنبه بسیار مفید برای استخراج اطلاعات مفید از اسناد متنی باشد.ایدئولوژی تشریح متنی، همان شیوه ای است که مردم در مورد یک متن خاص فکر می کنند.این فرآیندی است که در آن داده ها به عنوان مثبت یا منفی طبقه بندی شده اند.مقدار زیادی از داده ها (بررسی ها) در وب موجود است که می تواند مورد تجزیه و تحلیل قرار گیرد تا مفید باشد.این روش می تواند به طور خاص برای بازاریابی، کسب و کار، وجوه مفید باشد، زیرا که باعث می شود تا تحلیل موضوع مورد نظر را به سادگی بررسی کنیم.در عصر امروز اینترنت، بسیاری از افراد می توانند با یکدیگر ارتباط برقرار کنند.اینترنت برای ما امکان اتصال و تشخیص افتراء را فراهم کرده است. اینترنت پلت فرم بسیار زیادی را ارائه می دهد که از طریق آن نظرات از افراد مختلف می تواند از طریق انجمن ها، وبلاگ ها و سایت های شبکه اجتماعی گرفته شود.در این مقاله، استفاده از Tweepyو TextBlobبه عنوان یک کتابخانه پایتون برای دسترسی و طبقه بندی توییت ها با استفاده از Naïve Bayes، یک تکنیک یادگیری ماشین پیشنهاد می شود.تکنیک ما به منظور

ساده کردن روند تجزیه و تحلیل، خلاصه سازی و طبقه بندی است. در منبع شماره [4] و در سال 2016 استفاده از روش های یادگیری با ناظر مانند svm[1] و knn[2] در نظر کاوی مشتریان بیان شده است. در این روش عقاید مشتریان مخابراتی اردن که در سایت فیس بوک ثبت شده است با این روش تحلیل و تجزیه شده است. زبان مورد تجزیه و تحلیل در این روش انگلیسی بوده است. در منبع شماره [5] استفاده از تحلیل احساسات در یادگیری زبان دوم بیان شده است. برای این کار با استفاده از روش های یادگیری ماشین برای هر کلمه هم معنی و یا تعریف آن بیان می‌شود. دقت در این روش نیز بسیار خوب بوده است. دلیل این بهبود دقت استفاده از آنتالوژی در این روش می-باشد. در منبع شماره [6] نظرات را در ارتباط با یک محصول خاص مورد بررسی قرار داده اند. در ابتدای کار ورود دیتاست و انجام پیش پردازش است. این پیش پردازش همانند منابع قبل است. در مرحله بعد تمامی پاسخ ها به یک موضوع خاص را پیدا می کنیم . این کار با استفاده از پیدا کردن @ در توییت ها صورت می گیرد. در مرحله بعد ریشه یابی این جملات است. سپس این پاسخ ها را به صورت درختی ریشه یابی کرده و ریشه را پیدا می کنیم . در مرحله بعد پیدا کردن توییت هایی که مستقیما برای این کار ارسال نشده اند می باشد. سپس بر اساس پارامترها قابلیت اطمینان و طول و میانگین کلمات وزن دهی صورت می گیرد و نظرات طبقه بندی می شوند. در این روش در مرحله اول دیتاست وارد سیستم شده و عمل پیش پردازش انجام می شود و سپس نام شخصی را که می خواهیم اطلاعات در مورد این شخص را بررسی کنیم که آیا نظرات درباره این فرد مثبت بوده است یا خیر وارد کرده و در مرحله بعد با استفاده از اکانت توییتر وادر شده و به داده های واقعی دسترسی پیدا می کنیم و سپس با استفاده از الگوریتم ناوی بیزی طبقه بندی اطلاعات صورت می گیرد. این روش به دلیلی اینکه از اطلاعات واقعی استفاده می کند درای محدودیت است. در منبع شماره [7] برای تجزیه و تحلیل یادداشت‌های نویسنده از روش‌های یادگیری ماشین استفاده شده است. این نظریات متعلق به کسانی بوده است که در درس‌های آنلاین شرکت کرده اند و نظرات خود را بیان نموده-اند. نظرات بیان شده عبارتند از: خسته کننده و جذاب و ... برای تحلیل نظرات در این منبع از از روش‌های یادگیری با ناظر و بدون ناظر استفاده شده است و در انتها نتایج با هم مقایسه شده‌اند. نتایج به دست آمده نشان داده‌اند که جنگل تصادفی بیشترین دقت را داشته است. در منبع شماره [8] تجزیه و تحلیل متن‌های زبان چینی با استفاده از الگوریتم LSTM و L2 انجام گرفته است. بسیاری از روش‌ها برای تحلیل

[1] Support vector machine
[2] K- nearest neighbor

احساسات در زبان انگلیسی نتایج بسیار خوبی را به دست آورده است اما در زبان‌هایی به غیر از انگلیسی این نتایج خوب نبوده‌اند لذا، تحقیقات در این زمینه همچنان ادامه دارد. در تحلیل متن‌ها با زبان‌هایی به غیر از انگلیسی ممکن است برخی واژه‌ها شناسایی نشوند و یا کلا دقت سیستم پایین باشد. این الگوریتم دقت بسیار بالایی را در نتایج نشان داده است. در منبع شماره [9] استفاده از رسانه های اجتماعی به بخشی جدایی ناپذیر از روال روزانه در جامعه مدرن تبدیل شده است. پورتال های رسانه های اجتماعی شامل سیستم های عمومی قدرتمند است که مردم می توانند آزادانه نظرات و احساسات خود را در مورد موضوعات مختلف با جمعیت های زیادی به اشتراک بگذارند. در منبع شماره [10] برای نظر کاوی ده هزار توییت سایت توییتر از ابزار knime استفاده شده است. knime ابزاری بسیار قدرتمند در زمینه داده کاوی است که بصری سازی زیادی دارد و استفاده از آن بسیار راحت و آسان است. روش استفاده شده همان روش‌های یادگیری ماشین هستند. در این تحقیق فاکتروهای اساسی در شناسایی و تجزیه و تحلیل هکرها مورد بررسی قرار می‌گیرد. این رویکرد بر اساس آنالیز متنی، استخراج متن از انجمن‌ها و تجزیه و تحلیل احساسات قرار دارد. هر پیام انجمن به یک موضوع بحث مرتبط است و توسط کاربر نوشته شده است. بنابراین پیام‌های فردی را تبدیل کرده و فاکتورهای اصلی در متن‌کاوی را برای هر کاربر دسته‌بندی می‌نمائیم. بر اساس ارزیابی پیام‌های فردی، اقدامات زیر بر اساس انجمن را تعریف می کنیم: تعداد موضوعات درگیر، متوسط طول پیام، تعداد کل پیام‌ها، مدت زمان. سپس با بکارگیری الگوی متن‌کاوی، یک مجموعه از واژگان هکرها ایجاد می‌شود و ارتباط محتوای پیغام هر کدام را با این مجموعه مورد آنالیز قرار می‌دهیم. سپس، آنالیز احساسات را بر روی پیام اجرا می-کنیم و آن را برای ایجاد طبقه‌بندی هکرها بر مبنای نقشی که دارند، اعمال می‌کنیم. هر یک از این ویژگی‌ها به عنوان یک فرضیه در مدل در نظر گرفته می‌شود[2].

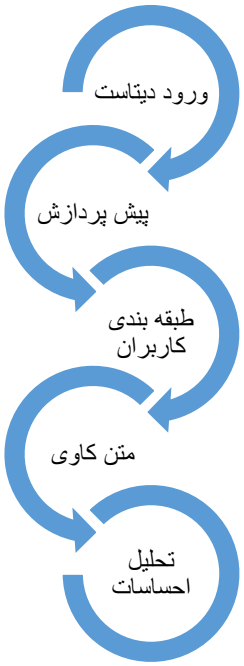
3- مواد و روش ها

می‌توانیم گام‌های اجرایی برای روش پیشنهادی در مدل آزمایشگاهی را بصورت زیر در نظر می‌گیریم:

1. یک انجمن سیاه یا یک انجمن مربوط به هکرها در نظر گرفته می‌شود.
2. پیام‌های فردی ارسال شده در این انجمن یا فروم را بصورت مجزا در نظر می‌گیریم.
3. ارتباطات پیام را از طریق ارائه‌ی الگوی متن‌کاوی بدست می‌آوریم.

4. مجموعه‌ی واژگان مربوط به هکرها را شکل می‌دهیم.
5. آنالیز و تجزیه و تحلیل احساسات را بر روی پیام‌های ارسال شده توسط رهبر انجمن انجام می‌دهیم.
6. رفتار هکرها در یک انجمن را بدست آورده و پیش‌بینی می‌کنیم.

از نرم افزار متلب به منظور تجزیه و تحلیل رویکرد پیشنهادی استفاده می‌شود. رویکرد متن‌کاوی بر مبنای الگوی شناسایی کلمات کلیدی، کلمات پرکاربرد، خلاصه‌سازی و غیره در نرم افزار متلب اجرا می‌شود. مجموعه‌ی واژگان هکرها بر اساس آنالیز پیام‌های ارسال شده در فروم بصورت یک متن در نظر گرفته شده و توسط نرم افزار متلب استخراج و در یک مجموعه ذخیره می‌شود. سپس بر اساس آنالیز احساسات بر روی پیام‌های ارسال شده توسط رهبر انجمن، رفتار هکرها از نظر رفتاری که بصورت معمول در برابر مشکلات بوجود می‌آید، بدست خواهد آمد.



شکل 1-دیاگرام روش پیشنهادی

در ابتدا نیاز است تا مجموعه داده ای وجود داشته باشد تا به عنوان ورودی، به سیستم پیشنهادی، داده شود. تحقیق پیش رو از مجموعه های داده توییتر موجود در http://www.sananalytics.com/lab/twitter-sentiment استفاده می‌کند که این مجموعه داده دارای 5513 توییت طبقه بندی شده همراه با نظرات افراد می‌باشد. این داده بعد از عملیات نرمال سازی و یک فرمت مشخص و قابل فهم در محیط MATLAB، وارد می‌شود. سپس

عملیات استخراج ویژگی همراه با طبقه بندی انجام خواهد شد. در ابتدا داده‌ها به صورت حذف داده های پرت صورت می‌گیرد که با هدف کاهش خانه‌های خالی یا فیلدهای خالی از داده است. سپس متن‌های همه کاربران در یک فایل تجمیع می‌شود. این فایل به نام normalized_dataset ذخیره شده است.
سپس کاربران به 8 دسته تقسیم می‌شوند که شامل موارد ذیل هستند:
1)کاربران خبره یا EXM[1]، 2)کاربران پیشرفته یا AM[2]، 3)کاربران متوسط یا IM[3]، 4)کاربران عضو یا Me[4]، 5)کاربران تازه کار یا Be[5]، 6)کاربران تازه وارد یا NW[6]، 7)کاربران معلق یا SU[7] و 8)کاربران ممنوع یا Ba[8]. این طبقه بندی و استخراج ویژگی به طور همزمان با مدل رگرسیون لجستیک صورت می‌گیرد. متغیرهایی که برای رگرسیون لجستیک مورد استفاده قرار می‌گیرد به شرح زیر است:

جدول 1-متغیرهای رگرسیون لجستیک

متغیر	شرح	مقدار
SupportVectors		[7x4 double]
alpha	فاصله اطمینان	[7 x 1 double]
bias	میزان اریب بودن داده‌ها	0.1309
Kernel function	تابع هسته	@linear_kernel

سپس درون این عملیات، وزن دهی و وزن شناسی لغات و جملات مبتنی بر TF-IDF انجام می‌شود. از جمله متغیرهایی که مورد بررسی واقع می‌شود، لغات، کلمات، جملات و به صورت کلی، تک تک واژه‌های موجود به عنوان نظر می‌باشند. همچنین ارزیابی ویژگی های حاصل از طبقه‌بندی در زمان یادگیری دوباره و آزمون با مدل رگرسیون لجستیک، متغیر مهم دیگری تلقی می‌شود. در نظر گرفتن معیارهایی چون دقت، حساسیت و نرخ ویژگی‌ها برای نتایج حاصل نیز دارای ضرورتی خاص در ارائه نتایج صحیح می‌باشد. در زمان عملیات استخراج ویژگی‌ها، مرحله کاهش ابعاد داده‌ای و انتخاب ویژگی‌ها نیز، وجود دارد که همراه با عملیات طبقه بندی، به صورت دو کلاس مجزا، با مدل رگرسیون لجستیک، این عملیات، انجام خواهد گرفت.
برای تحلیل این روش از برخی پارامترها استفاده می‌شود که در زیر بیان شده‌اند:
1)میانگین خطای مربعات(MSE)

[1] Expert member
[2] Advanced member
[3] intermediate member
[4] Member
[5] Beginner member
[6] New born member
[7] Suspend member
[8] Banned member

2) ماتریس سردرگمی[1]: این ماتریس یک ماتریس مربعی N*N می‌باشد که منظور از N همان تعداد کلاس‌های دسته می‌باشد.

3) دقت[2]: این میزان برابر است با مقدار تشخیص درست داده‌ها که با استفاده از فرمول زیر محاسبه می‌شود:
Accuracy = (Correct_Results/No_Test_Data)*180

4) حساسیت[3]: به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به درستی به عنوان منفی علامت‌گذاری می‌کند. این مقدار با استفاده از فرمول زیر محاسبه می‌گردد:
Sensitivity = (False_Results/No_Test_Data)*180

یافته ها:
نتایج پارامتر MSE، ماتریس سردرگمی، میزان دقت و میزان حساسیت در جدول های زیر آورده شده است.

جدول 2-پارامتر MSE

پارامتر	مقدار
MSE داده‌های آزمایشی	0.1648
MSE داده‌های آموزشی	0.1563
MSE در کل داده‌ها	0.3210

جدول 3-ماتریس سردرگمی

	مثبت پیش بینی شده	منفی پیش بینی شده
مثبت واقعی	25	25
منفی واقعی	24	25

جدول4-میزان دقت

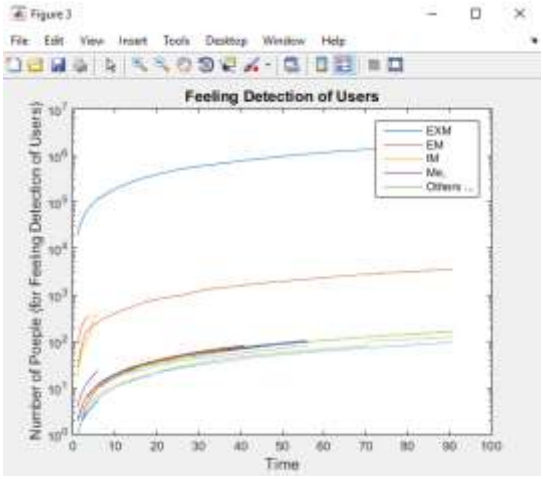
پارامتر	مقدار
دقت	90.9091

جدول 5-میزان حساسیت

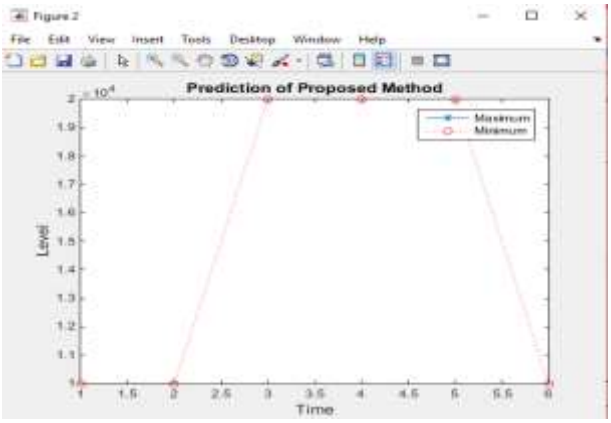
پارامتر	مقدار
حساسیت	89.0909

نمودار میزان پیش‌بینی بر اساس زمان و تعداد افراد را در شکل 2 مشاهده می شود. همانطور که در شکل 2 مشاهده می‌شود دسته بندی‌های مختلف افراد را بر حسب زمان تشخیص نمایش می-دهد. حداقل و حداکثر زمان تشخیص در شکل 3 مشاهده می شود و در انتها نمودار زمانی فاکتورهای تاثیرگذار ر در شکل 4 نمایش داده شده است.

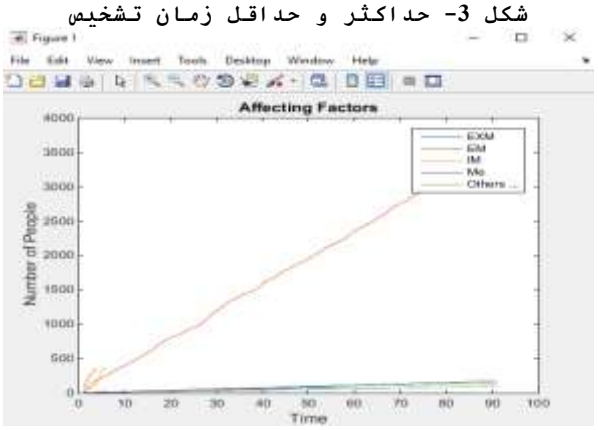
[1] Confusion matrix
[2] accuracy
[3] sensitivity



شکل 2-پیش بینی زمان و تعداد افراد



شکل 3- حداکثر و حداقل زمان تشخیص



شکل 4-نمودار زمانی فاکتورهای تاثیرگذار

مسئله تحلیل احساسات هکرها بر اساس روش‌های زیر در منابع دیگر حل شده نتایجی را به دست آورده‌اند. بر این اساس در این قسمت دقت به دست آمده در هر یک از روش‌های فوق با روش پیشنهادی مقایسه می شوند که نتایج در جدول 6 آمده است.

جدول 6-نتایج دقت در تحلیل احساسات

روش	دقت
SVM	72.90
KNN	65.43
CART	65.91
رگرسیون چند جمله ای	80.57
روش پیشنهادی	90.9091

دلیل بهبود روش پیشنهادی را می‌توان در دلایل زیر جستجو کرد. درخت های تصمیم بسیار احتمال Overfit شدن را دارند. درخت تصمیم CART نیز چنین است. جهت جلوگیری از Overfit شدن درخت تصمیم CART میتوان از یک شرطِ توقف استفاده کرد. این شرطِ توقف به الگوریتم می‌گوید که دیگر ادامه‌ی درخت را متوقف کند. این کار باعث می‌شود که درخت CART دیگر ریشه‌سازی را متوقف کند و درخت را بیش از یک حدِ آستانه پیچیده نکند(همان طور که می‌دانید پیچیدگی یکی از دلایل Overfit شدن مدل طبقه‌بند بود). یکی از این روش‌ها استفاده از تعداد مشخص نمونه در زیر درخت خاص است به گونه‌ای که اگر تعداد نمونه‌ها در یک زیر درختْ از یک حد آستانه کمتر شد، دیگر درخت ریشه‌سازی را ادامه نمی‌دهد. الگوریتم k-نزدیکترین همسایگی یکی از ساده‌ترین الگوریتم های طبقه‌بندی است. اما با وجود سادگی، نتایج آن به وضوح قابل رقابت با دیگر الگوریتم ها است. این الگوریتم اغلب به دلیل سهولت تفسیر نتایج و زمان محاسبه پایین مورد استفاده قرار می‌گیرد. لذا نتایج این روش از روش CART بهتر می‌باشد.

از جمله معایب روش svm را می‌توان به شرح زیر بیان کرد:

- این نوع الگوریتم ها، محدودیت های ذاتی دارند مثلا هنوز مشخص نشده است که به ازای یک تابع نگاشت ، پارامترها را چگونه باید تعیین کرد.
- ماشینهای مبتنی بر بردار پشتیبان به محاسبات پیچیده و زمان بر نیاز دارند و به دلیل پیچیدگی محاسباتی، حافظه زیادی نیز مصرف می کنند.
- داده های گسسته و غیر عددی هم با این روش سازگار نیستند و باید تبدیل شوند.

4-نتایج و بحث

یکی از مهم‌ترین کارها در زندگی هر انسانی مشورت کردن است. مشورت با دیگران در تمام اموری که به تصمیم‌گیری مربوط است انجام می‌گیرد. هم اکنون در عصر ارتباطات نیاز به ارتباط مستقیم برای افراد وجود دارد که بتوانند از اینترنت استفاده کرده و از نظرات و احساسات افراد در زمینه‌ای که میخواهند استفاده کنند. نظرات کاربران و مشتریان در زمینه-های مختلف با ورود وب 2.0 در اینترنت ثبت شد. در این راستا و پس از به وجود آمدن این تکنولوژی افراد توانستند به طور فزاینده‌ای با هم تعامل داشته باشند و در بیان نظرات شرکت کنند. این کاربران می‌توانند دیدگاه‌ها و نظرات خود را با استفاده از انجمن‌ها و وبلاگ‌ها و یا دیگر مکان‌ها اجتماعی به اشتراک بگذارند. این نظرات می‌توانند برای صاحبان صنایع و محصولات بسیار مفید باشد. همین مسئله باعث شده است تا این نظرات در بین مصرف کنندگان محبوب شده و کاربران فعال می-توانند انها را پیگیری کنند.

چالشی که در این زمینه وجود دارد این است که این نظرات حجم بسیاری را در اینترنت اشغال کرده و هر روز بر این حجم افزاده می‌شود. لذا، با توجه به این حجم داده، استخراج مطالب مفید مرتب‌سازی آنها بسیار کاری پر هزینه و وقتگیر است. در دهه اخیر تمایل زیادی برای شناسایی و استخراج خودکار قطبیت نظرات از اینترنت به وجود آمده است. کاربران مایل هستند تا نظرات مثبت و یا منفی را از بین نظرات موجود استخراج کنند.
در این زمینه چالش‌های زیادی وجود دارد که در اینجا به صورت تیتروار مشاهده می‌شود:
1-در بیشتر تحقیقات صورت گرفته فقط قطبیت نظرات در نظر گرفته شده است ولی با انجام تحلیل‌های بیشتر می‌توان از روی لحن بیانات شدت قطبیت را نیز تشخیص داد.
2-طول مورد استفاده در متونی که برای نقد و یا تایید یک محصول بیان می‌شود بسیار کوتاه است.
3-برای بیان متون در باره محصولات از یک فرمت استاندارد استفاده نشده است.
4-اکثر متون وارد شده برای بیان احساسات به زبان فارسی هنوز دارای دقت تشخیص پایین هستند.
5-برای نقد و نوشتن نظرات راجع به محصولات از زبان‌های مختلف استفاده شده است.
تجزیه و تحلیل احساسات نوعی داده‌کاوی است که اهمیت نظرات مردم را از طریق پردازش زبان طبیعی (NLP)، زبان شناسی محاسباتی و تجزیه و تحلیل متن را محاسبه کرده و برای استخراج و تجزیه و تحلیل اطلاعات ذهنی از وب و رسانه های اجتماعی و منابع مشابه استفاده می‌کند. داده‌های تجزیه و تحلیل داده‌ها احساسات یا واکنش عمومی را نسبت به محصولات،

افراد یا ایده های خاصی نشان می‌دهد و قطعیت محتوا اطلاعات را نشان می‌دهد.
برای تحلیل احساسات عموما از سه روش استفاده می‌کنند که عبارتند از:
1-روش‌های بانظارت
2-روش‌های بدون نظارت
3-روش‌های نیمه نظارتی
در روش پیشنهادی در این پژوهش در مرحله ابتدایی پیش پردازش قرار دارد. دلیل استفاده از پیش پردازش این است که در بین داده های موجود متن‌هایی قرار دارند که یا ناقص هستند و یا قابل استفاده نیستند و همچنین بایستی ریشه یابی افعال نیز در این مرحله استفاده گردد. در مرحله دوم کاربران با استفاده از رگرسیون به پنج دسته تقسیم می‌شوند. در مرحله بعد با استفاده از الگوریتم tf-idf وزن کلمات را مشخص و بیان می‌کنند که توزیع کلمات در متن به چه صورت بوده است. در مرحله آخر تحلیل احساسات و تعیین نظر مثبت و یا منفی کاربران بوده است. برای شبیه سازی این روش از نرم افزار متلب استفاده شده است. نتایج نهایی این روش بیان می‌کند که میزان خطای مربعات در این روش مقدار 0.32بوده است و میزان دقت به دست آمده در این روش 90.9 درصد اعلام شده است. کارایی این روش در بین روش‌های مقایسه شده دارای مقدار بسیار بهنری است و نتایج خوبی را نمایش می‌دهد.

منابع

[1] Mahmood, A.M.; Siponen, M.; Straub, D.; Rao, H.R.; and Raghu, T.S. Moving toward black hat research in information systems security: An editorial introduction to the special issue. MIS Quarterly, 34(3), 2-22. (2010).
[2] Hackett R. Facebook Awards Server-crushing Hacker with Its Biggest Ever Bounty, (2017). Available: http://fortune.com/2017/01/19/facebook-hacker-bug-bounty/. [Accessed: 28-Apr-2017].
[3] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of a ffective computing: from unimodal analysis to multimodal fusion, Inf. Fus. 37 (2017) 98–125.
[4] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, ACL, (2017), pp. 873–883.
[5] E. Cambria, A. Hussain, Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis, Springer, Cham, Switzerland, 2015.
[6] Walaa Medhat a,*, Ahmed Hassan b, Hoda Korashy," Sentiment analysis algorithms and applications: A survey", Production and hosting by Elsevier B.V. on behalf of Ain Shams University. http://dx.doi.org/10.1016/j.asej.2014.04.011
[7] ناصر یعقوبی و حمید ناصری و "تحلیل احساسات در شبکه اجتماعی توییتر با متن کاوی" و مجله انفورماتیک مشهد و جلد 25 و ص 56-67 و 1391
[8] S.L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artif. Intell. Rev. 48 (4) (2017)499–527.
[9] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, A Practical Guide to Sentiment Analysis, Springer, Cham, Switzerland, 2017.