سمینار کارشناسی ارشد

شناسایی اسپم در نظرات

گرد آورنده :

الهه رحیمی

استاد راهنما :

دکتر محسن کاهانی

شهریور ۹۳

چکیده

اشخاص و سازمان‌ها از نظرات در شبکه‌های اجتماعی به‌صورت گسترده‌ای برای تاثیر بر تصمیمات خریداران، تصمیم‌گیری در انتخابات، بازاریابی و طراحی محصول، استفاده می‌کنند. عدم نظارت بر روی نظرات، باعث می‌شود افراد سودجو با ارسال نظرات نامربوط و یا جعلی، شهرت و اعتبار شرکت را از بین ببرند و یا خدشه‌دار کنند. همچنین می‌توانند یک محصول و یا خدمت بی‌کیفیت را با ارسال نظرات مثبت جعلی و ساختگی ارتقاء داده و بر شهرت آن بیفزایند.

در سال‌های اخیر، نظرات جعلی و اسپم مشکلی است که به‌شدت در حال گسترش و افزایش است. امروزه سایت‌های تجاری، هدف مناسبی برای تولیدکنندگان اسپم برای تولید اسپم هستند. اغلب سایت‌های تجاری قسمتی برای نظرات کاربران دارند و کاربران می‌توانند دیدگاه خود را درباره‌ی محصول انتشار دهند که اطلاعات ارزشمندی برای مشتری‌ها و شرکت سازنده دارد. برای این‌که نظرات و تجربیات واقعی کاربران به درستی منعکس شود، شناسایی نظر جعلی بسیار مهم است. تشخیص نظرات جعلی و اسپم در نظرات به علت تکنیک‌ها و روش‌های جدید برای انجام این کار، بسیار پیچیده است. بنابراین نیاز به بررسی گسترده در روش‌های شناسایی و تشخیص اسپم در نظرات، به‌شدت احساس می‌شود.

این گزارش، شیوه‌های مورد استفاده در تشخیص اسپم در نظرات را مورد بررسی کرده و در انتها مورد ارزیابی قرار می‌دهد.

کلیدواژه: واکاوی نظرات، نظر اسپم ، تشخیص نظر جعلی، برچسب‌گذاری نظرات

فهرست مطالب

فهرست اشکال

فهرست جداول

۱- مقدمه

اینترنت به دلیل خصوصیاتی چون قابلیت استفاده و دسترسی گسترده آن، اشکال کاملاً جدیدی از تعاملات، فعالیت‌ها و سازماندهی‌های اجتماعی را پدید آورده است. با توجه به این که نظارت و کنترلی روی محتوا و داده‌های وارد شده در آن وجود ندارد تحلیل بر روی محتوای آن بسیار زمان‌بر خواهد بود.

بدون اغراق می‌توان گفت که استفاده از نظرات در رسانه‌های اجتماعی در حال افزایش است. اشخاص و سازمان‌ها از نظرات در شبکه‌های اجتماعی به‌صورت گسترده‌ای برای تاثیر بر تصمیمات خریداران، تصمیم‌گیری در انتخابات، بازاریابی و طراحی محصول، استفاده می‌کنند. نظرات برخط به‌صورت افزاینده‌ای توسط افراد و سازمان‌ها استفاده می‌شوند. بنابراین بسیار مهم است که این نظرات خالی از غرض‌ورزی باشند تا مراجعه‌کنندگان به این سایت‌ها بتوانند به نظرات اتکا کنند.نظرات مثبت اغلب به معنی سود بیش‌تر و شهرت برای تجارت و اشخاص است که متاسفانه ابزاری برای بازی دادن سیستم‌ها شده است بدین طریق که با جعل دیدگاه[۱] و یا نظر[۲]، محصولات مشخص، سرویس‌ها، اشخاص و سازمان‌ها را ارتقا داده یا بی‌اعتبار می‌کنند. این نظرات حتی بدون فاش کردن هویت اصلی شخص یا سازمانی است که شخص برای آن کار می‌کند، صورت می‌پذیرد.

تولیدکننده‌ی اسپم[۳] و sybil به کاربرانی که حملات بدخواهانه‌ای دارند، چیزی را درست جلوه می‌دهند[۴]، و یا تبلیغات نامربوط ارسال می‌کنند، گفته می‌شود[۱۲]. تولیدکنندگان اسپم استفاده از اینترنت را از استفاده‌ی عادی به استفاده‌ی بد تغییر دادند برای مثال جعل نظر و یا امتیاز جعلی[۵] [۱۸]، اغراق تاثیر یک محصول، پست توضیحات بدخواهانه علیه رقیب[۱۵] یا ارسال توضیحات[۶] تکراری در فروم‌ها، انتشار مقالات، نظرات نامرتبط، ارسال تبلیغات نامربوط[۱۲].

جاعلین نظرات و دیدگاه، تولیدکننده‌ی اسپم دیدگاه[۷] نامیده می‌شوند و فعالیت و کارشان تولید اسپم[۸] نام دارد[۱و۲].

تولید اسپم بیش‌تر، وسیع‌تر و فریبنده‌تر شده است و درحال حاضر یکی از چالش‌های بزرگ، شناسایی آن است.

۱-۱- انواع اسپم و تولید اسپم[۹]

شناسایی اسپم در زمینه‌های زیادی مطالعه می‌شود. اسپم وب[۱] و اسپم ایمیل[۲] دو نوع از انواع اسپم ه ستند که در مورد آن تحقیقات گسترده‌ای انجام شده است[۳و۴].

---

[1] *fake opinions*
[2] *reviews*
[3] *spammer*
[4] *distorted the true*
[5] *fake scoring*
[6] *comments*
[7] *opinion spammer*
[8] *opinion spamming*
[9] *spamming*

اسپم وب به فعالیت برای فریب‌دادن موتور جستجو برای ارتقاء رتبه[۳]ی صفحات وب گفته می‌شود [۶و ۷] کـه در دسـته‌بندی اسپم محتوا[۴] و اسپم پیوند[۵]، قرار می‌گیرند.

اسپم محتوا کلمات رایج (اما نامربوط) در صفحات وب[۶] است که موتور های جستجو را برای کوئری های جستجو[۷] گمراه می‌کند، این نوع از اسپم در ارسال نظرات زیاد استفاده نمی‌شود[۳].

اسپم پیوند، فراپیوندهای[۸] موجود در نظرات یا لینک‌های تبلیغاتی است که در فروم‌های رسانه‌های اجتماعی متداول می‌با شد. شناسایی آن هم نسبتا آسان است[۳].

اسپم نظرات به نوعی شبیه به اسپم وب است اما اسپم پیوند حتی با استفاده‌ی آن‌هـا در صـفحات وب و نظـرات بـه صـورت کلمات غیر مرتبط (نامربوط)، متفاوت با اسپم وب است.

اسپم ایمیل که همان ایمیل‌های ناخواسته است به صورت مستقیم یا غیر مستقیم و بدون تبعیض، توسط فرستنده‌ای کـه بـه کاربر دیگر هیچ ارتباطی ندارد، ارسال می‌شود [۸]. این اسپم‌ها شامل تبلیغات هستند که به صورت بسیار کم توسـط نظردهنـدگان اسپم استفاده می‌شود. تشخیص آن توسط کاربر آسان است و همین نکته باعث شده تا کم‌تر خطرناک [۵] باشند.

۱-۲- انواع اسپم نظر و دیدگاه

مطالعات در مورد اسپم نظرات از سال ۲۰۰۷ صورت گرفته است. سه نوع از اسپم نظرات[۹] معرفی شدند: [۲]

نوع اول(نظر جعلی[۱۰]): نظرات غیرصادقانه‌ای که درخصوص استفاده از محصولات و یا سرویس‌ها، با غرض‌ورزی پنهان، نوشته شده است. نظرات مثبت برای ارتقاء یک محصـول یـا سـرویس و نظـر منفـیِ کـاذب[۱۱] بـرای لطمـه زدن بـه شـهرت محصـولات و سرویس‌هاست.

نوع دوم (نظرات در مورد یک علامت تجاری منحصر): نظرات درخصوص محصول یا سرویس خاص نیست و فقط در خصـوص علامت تجاری خاص یا تولید کننده‌ی محصول است. با توجه به این که هدفِ نظر، یک محصول خاص نیسـت بـه عنـوان اسـپم در نظر گرفته می‌شوند. به عنوان مثال نظرِ «من از HP متنفرم، من هیچ‌وقت هـیچ کـدام از محصـولات آن را نمی‌خـرم»، بـرای یـک پرینتر خاص از محصولات HP.

نوع سوم(غیرِ نظر[۱۲]): به دو دسته تقسیم می‌شود: (۱) تبلیغات و (۲) متن نامربوط کـه نظـری در آن داده نشـده اسـت. مثـلِ سوالات، پاسخ‌ها و متون تصادفی.

---

[1] *Web spam*
[2] *email spam*
[3] *Rank*
[4] *content spam*
[5] *Link Spam*
[6] *Web pages*
[7] *search queries*
[8] *hyperlinks*
[9] *spam review*
[10] *fake reviews*
[11] *false negative*
[12] *non-review*

نظرات دو نوع ۲و ۳ کم‌یاب هستند و شناسایی آن‌ها با استفاده از یادگیری با ناظر[1] نسبتا ساده است. حتی اگر آن‌ها شناسایی نشوند مشکل بزرگی به‌وجود نخواهد آمد به این دلیل که انسان در حین خواندنِ نظر، آن را کشف می‌کند.

۱-۱-۱- نظرات مضر[2]

نظرات ۲،۳،۴ و ۵ مضر هستند، نظرات ۱ و ۶ هم زیاد مضر نیستند(جدول۱). در حال حاضر، الگوریتم‌های تشخیصِ نظر جعلی در محدوده‌ی نظرات مضر کار می‌کند. الگوریتم‌های تشخیص نظرِ موجود از رتبه‌بندی ویژگی‌ها استفاده می‌کنند. (کیفیت خوب، بد و متوسط برای رتبه‌بندی محصول) این روش هم اگر تولیدکنندگان اسپم زیاد و نظردهی پایین باشد، کار نمی‌کند.

جدول۱- نظرات جعلی در مقابل کیفیت محصول

	Positive fake review	Negative fake review
Good quality product	1	***2***
Average quality product	***3***	*4*
Bad quality product	***5***	**6**

۱-۱-۲- تولید اسپم شخصی[3] و گروهی[4]

نظر جعلی می‌تواند توسط افراد مختلفی نوشته شود. تولیدکننده‌ی اسپم می‌تواند به تنهایی یا به‌صورت گروهی، نظر جعلی بنویسد. اسپم گروهی بسیار خطرناک است زیرا با توجه به تعداد اعضای گروه، به‌راحتی می‌توانند نظر در مورد یک محصول را به‌سرعت تغییر دهند.

۱-۱-۳- انواع داده‌ها و ویژگی‌های نظرات

انواع داده‌هایی که برای شناسایی اسپم در نظرات استفاده می‌شود:

محتوی داده: ویژگی‌های زبان شناختی استفاده می‌شود مانند POS n-gram و ساختارها و معانی که برای تشخیص دروغ استفاده می‌شود.

[1] *supervised learning*
[2] *Harmful Fake Reviews*
[3] *Individual spammer*
[4] *Individual spammer*

متا دیتای نظر: داده‌هایی مانند رتبه‌بندی ستاره‌ای، زمانی که شخص نظر خود را ثبت می‌کند، IP و Mac کامپیوتر فرستنده ثبت می‌شود. هم‌چنین محل جغرافیایی و ترتیب کلیک در سایت بر اساس الگوی رفتاری این افراد و نظراتشان ثبت می‌شود.

مثلاً نظر مثبت در مورد یک هتل که همه‌ی این نظرات از محل‌های اطراف هتل ارسال شده‌اند، قابل اعتماد نیستند یا ارسال چندین نظر از اشخاص مختلف (شناسه‌های مختلف) از یک کامپیوتر.

اطلاعات محصول: اطلاعات در مورد موجودیت، مانند توصیف محصول و میزان فروش محصول، مثلاً محصولی که فروش خوبی نداشته ولی نظرات مثبت درباره‌ی آن زیاد است، قابل باور نیست.

داده‌های عمومی و خصوصی:

داده‌های عمومی که در صفحه‌ی نظرات، قابل مشاهده هستند مثل محتوی نظر و شناسه‌ی فرد و زمانی که نظر ثبت شده است.

داده‌های خصوصی که نشان داده نمی‌شوند مانند آدرس IP و Mac کامپیوتر نظردهنده و اطلاعات کوکی.

هدف اصلی شناسایی اسپم در نظر، تشخیص نظر جعلی، شخص جعل‌کننده‌ی نظر[1] و گروه جعل‌کننده‌ی نظر[2] است. کشف یکی از این سه نوع به شناسایی دو گروه دیگر کمک می‌کند.

در بخش دوم این گزارش به چالش‌ها و انگیزه‌های پیش روی این مبحث اشاره می‌شود و در بخش سوم به کارهای انجام شده در زمینه‌ی شناسایی اسپم در نظرات و مقایسه‌ی روش‌ها و نتایج به‌دست آمده، پرداخته خواهد شد.

[1] *fake reviewer*
[2] *fake reviewer group*

۲- چالش‌ها و انگیزه‌ها

چالش اصلی شناسایی و تشخیص اسپم نظر این است که بر خلاف انواع دیگر اسپم، بسیار سخت است که نظر جعلی را با استفاده از خواندن آن‌ها به‌صورت دستی[1] تشخیص دهیم و این کار را برای طراحی و ارزیابی الگوریتم‌های شناسایی اسپم دیدگاه سخت می‌کند. مثلا ممکن است شخصی نظری در مورد یک رستوران خوب بنویسد و آن را برای یک رستوران بد پست کند تا آن رستوران را ارتقاء دهد. هیچ راهی وجود ندارد که این نظر جعلی را بدون اطلاعاتی که در پس این نوشته وجود دارد، شناسایی کنیم چرا که یک نظر نمی‌تواند در یک زمان هم صادقانه[2] و هم جعلی باشد. روش‌ها و ویژگی‌های بسیار متنوع تشخیص اسپم در نظرات، این مبحث را پیچیده‌تر کرده است.

تحقیقات جاری، از انواع نظرات جعلی ساختگی[3] برای آموزش استفاده می‌کنند. بیش‌ترین نظرات جعلی ساختگی توسط ابزار ماشینی تورک آمازون[4] ساخته می‌شوند. تورکرها[5] ذهنیت جاعلین واقعی نظرات که شغل آن‌ها و کار واقعیشان ارتقاع یا تنزل دادن است، را ندارند. مقایسه‌ی تشخیص نظرات جعلی بر روی داده‌های جعلی ATM و داده‌های زندگی واقعی[6] نشان می‌دهد که با توجه به حالات مختلف ذهن، می‌تواند نتایج متفاوتی در نوشته‌ها و به تبع آن دقت و صحت دسته‌بندی[7] داشته باشد. این مورد یکی از چالش‌های تشخیص و شناسایی اسپم در نظرات است.

متاسفانه کار چندان زیادی در این حوزه صورت نگرفته است به‌خصوص در مقالات فارسی تنها [۲۲] به شناخت و بررسی ابعاد آن پرداخته است.

---

[1] *manually*
[2] *truthful*
[3] *pseudo*
[4] *Amazon Mechanical Turk (AMT)*
[5] *Turkers (AMT authors)*
[6] *real-life*
[7] *classify*

۳- مطالعات انجام شده

ابزار متفاوتی برای کشف اسپم وجود دارد. تکنیک عمده یادگیری با ناظر است. متاسفانه با توجه به فقدان استاندارد برای نظرات جعلی، کارهای فعلی عمدتا بر مدل‌های برچسب‌گذاری نظرات به جعلی و غیرجعلی[1] استوار است.

۳-۱- تشخیص اسپم با ناظر[2]

در حقیقت، تشخیص اسپم، می‌تواند همان دسته‌بندی مساله به دو کلاس جعلی و غیر جعلی باشد. یادگیری با ناظر در این زمینه کاملا کاربردی است. همان‌طور که پیش از این گفته شد، تشخیص اسپم از طریق خواندن آن بسیار مشکل است. در این خصوص به معرفی چهار روش پرداخته می‌شود.

۳-۱-۱- تشخیص اسپم با استفاده از دسته بندی نظرات برحسب میزان تکرار

در[۲] ابتدا تقسیم‌بندی مساله به دو دسته‌ی اسپم و غیر اسپم[3]، صورت می‌گیرد. بررسی بر روی ۵/۸ میلیون نظر و ۲/۱۴ میلیون نظردهنده‌ی سایت آمازون، تعداد زیادی نظرِ تکراری[4] یا نزدیک به تکراری[5] یافت شده است. منتقدین، نظرات مشابه یا نظرات با کمی تفاوت را برای محصولات متفاوت نوشته‌اند.

سه نوع نظر تکراری یا نزدیک به تکراری وجود دارد.

۱. تکراری از کاربران با شناسه کاربری[6] متفاوت برای محصول مشابه
۲. تکراری از کاربر با شناسه کاربری یکسان برای چندین محصول متفاوت
۳. تکراری از کاربران با شناسه کاربری متفاوت برای چندین محصول متفاوت

نظر تکراری و نزدیک به تکراری از طریق روش 2-gram مبتنی بر مقایسه‌ی محتوا انجام می‌شود. با استفاده از فرمول Jaccard distance یعنی نسبت اشتراک به اجتماع 2-gram دو نظر به یکدیگر به دست می‌آید. دو نظر با امتیاز[7] شباهت حداقل ۹۰ درصد، به‌عنوان نظرات تکراری انتخاب می‌شوند.

---

[1] *non-fake*
[2] *supervised spam detection*
[3] *non-spam*
[4] *duplicate*
[5] *near-duplicate*
[6] *userid*
[7] *score*

برای تشخیص اسپم نوع ۲ و۳ به این صورت عمل می‌شود که چون این دو نوع اسپم توسط انسان قابل شناسایی هستند بنابراین در این روش با استفاده از یادگیری بر اساس نمونه‌های برچسب‌گذاری شده، طبقه‌بندی انجام می‌شود. برای این که تخمین زده شود که چقدر احتمال دارد که نظر اسپم باشد، از رگرسیون منطقی[1] استفاده شده است.

در این مقاله[۲] از سه نوع اطلاعات که به نظر مربوط می‌شود استفاده می‌کند: ۱. محتوای نظر ۲. نظردهنده۳. محصولی که نظر برای آن داده شده است, و برای تشخیص اسپم، نیز از سه مجموعه ویژگی استفاده می‌شود:

۱. ویژگی‌های نظر محور[2]: مربوط به ویژگی‌های هر نظر است. مثلا تعداد دفعاتی که کلمات در یک نظر تکرار شده، n-gramهای متن و تعداد دفعاتی که نام سازنده‌ی مورد نظر تکرار شده است، درصد کلماتِ نظر، طول نظر، تعداد بازخوردهای مفید. در بسیاری از سایت‌ها هر نظر دهنده به جمله‌ای مانند:«آیا این نظر مفید بوده است؟» جواب می‌دهد.

۲. ویژگی‌های نظردهنده محور[3]: مربوط به ویژگی‌های نظردهنده است. ویژگی‌هایی که میانگین رتبه‌ای که به نظردهنده می‌دهند، است. میانه و انحراف معیار در رتبه و نسبت تعداد نظراتی که نظردهنده برای اولین بار برای محصول نظر ارسال کرده است و نسبت تعداد مواردی که آن شخص تنها نظردهنده‌ی محصول است.

۳. ویژگی محصول محور[4]: مربوط به ویژگی‌های محصول است. مانند قیمت محصول، میزان فروش محصول، میانه و انحراف معیارِ رتبه‌ی نظر در مورد محصول.

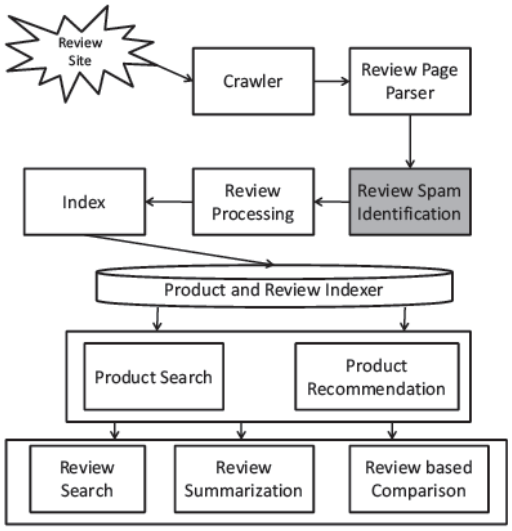
نتایج تجربی به دست آمده، ابتدائی[5] هستند به دو دلیل. ۱) این که به‌طور قطع نمی‌توان گفت سه نوعِ تکرار ذکر شده، نظر جعلی هستند و ۲) بسیاری از نظرات جعلی تکراری نیستند و ممکن است تحت این مدل به عنوان غیر جعلی معرفی شوند.

۳-۱-۲- تشخیص اسپم با استفاده از روشهای یادگیری ماشینی

در این روش[۱۹]، ابتدا چارچوب کاری سیستم استخراج نظرات برای محصول معرفی شده است.(شکل۱)

---

[1] *Logistic Regression*
[2] *Review centric features*
[3] *Reviewer centric features*
[4] *Product centric features*
[5] *tentative*



شکل۱- چارچوب کاری سیستم استخراج نظرات

در این روش فرض شده که بین امتیاز مفید بودنِ نظر و اسپم ارتباط وجود دارد و وقتی امتیاز مفید بودن یک نظر پایین باشد احتمال این‌که نظر اسپم باشد، زیادتر است.

پیش پردازش داده‌ها به این صورت انجام می‌شود:۱- محصولات تکراری با استفاده از تطبیق کامل نام، حذف می‌شوند.۲- نظرات بر اساس امتیازدهی برای مفیدبودن نظر، به سه دسته‌ی بالا[۱]، متوسط[۲] و پایین[۳] تقسیم می‌شوند.۳- استخراج داده‌های محتوای نظر مانند پروفایل نظردهنده، تاریخچه‌ی نظرات و...

در مطالعات اخیر ۳۰ روش جهت شناسایی اسپم، معرفی شده است[۲۰]. از ۱۰ دانشجو بعد از مطالعه‌ی مقالات در خصوص اسپم در نظرات، خواسته شده که نظرات را برچسب‌گذاری کنند. هر نظر توسط قضاوت دو نفر برچسب‌گذاری می‌شود و در صورت تناقض نفر سوم هم روی نظر، قضاوت خواهد کرد. در انتها، بین ۶۰۰۰ نظر، ۱۳۹۸ اسپم به‌دست آمده که با اثبات فرضیه‌ی اولیه، تعداد ۱۲۵۶، ۱۱۲ و ۳۰ اسپم از بین نظرات در محدوده امتیازدهی مفید بودن پایین، متوسط و بالا جمع‌آوری شد.

آزمایشات بر روی چندین روش دسته‌بند شامل SVM، رگرسیون منطقی و مدل بیزین، روی نرم‌افزار یادگیری ماشین Weka صورت گرفته و بهترین نتایج دسته‌بندی با استفاده از روش دسته‌بند نایو بیز[۴] استخراج شده است.

در ادامه با این نتیجه‌گیری که احتمال نوشتن مجدد اسپم توسط تولیدکننده‌ی اسپم زیاد است، از الگوریتم کمک یادگیرنده[۵] برای تشخیص اسپم استفاده کرده است.

با دو دیدگاه[۱] از مجموعه‌ی ویژگی‌ها (نظر محور و نظردهنده محور) برای یادگیری ماشین شبه ناظر، استفاده شده است.

---

[1] *top*
[2] *middle*
[3] *low*
[4] *Naive Bayes*
[5] *co-training algorithm*

ویژگی‌های مبتنی بر نظر:

- ویژگی‌های محتوا
- ویژگی‌های احساسی
- ویژگی‌های محصول
- ویژگی‌های متا – دیتا

ویژگی‌های مبتنی بر نظردهنده:

- ویژگی‌های پروفایل
- ویژگی‌های رفتاری

در این الگوریتم، دسته‌بندی با استفاده از ویژگی‌های نظر Fr و ویژگی‌های نظر دهنده Fu از روی مجموعه‌ی کوچکی از نظرات برچسب‌گذاری‌شده L بر روی U مجموعه‌ی بزرگی از نظرات برچسب‌گذاری‌نشده، صورت می‌گیرد.

۱. آموزش دیدگاه اول با دسته‌بند Cr با توجه به مجموعه‌ی L که بر روی ویژگی‌هایFr برچسب‌گذاری شده است.

۲. استفاده از Cr برای برچسب‌گذاری نظراتU بر اساس Fr.

۳. انتخاب P به عنوان پیش‌گویی مثبت و N به عنوان پیش‌گویی منفی نظرات از مجموعه‌ی U.

۴. آموزش دیدگاه دوم کلاس‌بند Cr با توجه به مجموعه‌ی L که بر روی ویژگی‌هایFu برچسب‌گذاری شده است.

۵. استفاده ازCu برای برچسب‌گذاری مجموعه‌یU بر اساس ویژگی‌های نظردهندگان.

۶. انتخاب P به عنوان پیش‌گویی مثبت و N به عنوان پیش‌گویی منفی نظر T از مجموعه‌ی نظردهندگان U.

۷. استخراج نظراتی که یک نظردهنده‌ی خاص نوشته است[۲].

۸. جابه‌جایی نظرات نظردهنده‌ی خاص از مجموعه‌ی U به مجموعه‌ی برچسب‌گذاری شده‌ی L با برچسب‌های پیش‌بینی شده.

این روش در F-Score به نتایج خوبی دست یافته است.(جدول ۲)

جدول۲- نتایج به‌دست آمده از چارچوب کاری با ناظر

	Precision	Recall	F-Score
Helpful	0.184	0.911	0.306
All Features(A)	0.517	0.669	0.583

[1] *view*

[2] *Extract the Reviews T'review Authored by T'reviewer*

A-content	0.502	0.662	0.571
A-sentiment	0.507	0.649	0.569
A-product	0.514	0.665	0.58
A-metadata	0.506	0.632	0.562
A-profile	0.516	0.658	0.578
A-behavior	0.541	0.587	0.563
A-review	0.593	0.504	0.545
A-reviewer	0.571	0.531	0.55

۳-۱-۳- تشخیص با استفاده از محاسبات انسانی

در این روش[۱۱] از روش ارزیابی ماشینی و انسانی برای تشخیص اسپم استفاده می‌شود. نظرات از هشت شرکت تجهیزات ورزشی از چهار منبع Amazon.com، bodybuilding.com، GNC.com و supplementreviews.com جمع‌آوری شده است. نظرات بررسی شده بر روی ویژگی‌های محدود بودند (سه مورد). پس از آن نظرات کوچکتر از ۱۵۰ کاراکتر به این دلیل که احتمالا به موضوع مربوط نیستند یا شامل تکرار یا نزدیک به تکرار هستند، حذف می‌شود. پس از آن از نظرات باقی‌مانده ۲۵ نظر از نظرات با امتیاز بالا T^H و ۲۵ نظر با امتیاز پایین T^L از میان ۸ محصول، انتخاب می‌شوند.

پس از آن از با استفاده از MTurk نظرات جعلی امتیاز بالا و پایین ایجاد می‌شود. برای تکراری نبودن نظرات، از سایت تشخیص سرقت ادبی استفاده شده است. همچنین برای سنجش سبک[۱] نگارش و تحلیل احساس محتوای نظر از API که text-processing.com[۲] ارائه داده است، استفاده شده است. مجموعه‌ی یادگیری از نظرات فیلم استفاده شده است.

$$ARI = 4.71(C/W) + 0.5\,(W/S) - 21.43 \quad (۱)$$

c، W و s به ترتیب تعداد حروف[۳]، کلمات[۴] و جملات[۵] موجود در نظر است.

فرض اشتباه این روش این است که تمام نظرات bodybuilding.com را صادقانه در نظر گرفته است. در مقایسه‌ای که خود مقاله بین یافته‌های خود و مقاله‌ی Yoo and Gretzel (2009) انجام داده‌است، ذکر شده که تفاوت اندازه‌ی کلمات در نظرات با امتیاز بالا یا امتیاز پایین تفاوتی با مقاله‌ی Yoo and Gretzel (2009) ندارد و همان نتایج را به‌دست آورده است. پیچیدگی در نظرات جعلی با استفاده از فرمول API که ارائه داده است به مراتب بیش‌تر است. تفاوت بین گستردگی و وسعت کلمات بین آن دو وجود ندارد. این مقاله از ۸۰۰ نظر در مورد دستگاه‌های بدن‌سازی و Yoo and Gretzel (2009) هشتاد و دو نظر بر روی هتل را بررسی کرده‌است.

[1] *style*
[2] *http://text-processing.com/api/sentiment/*
[3] *character*
[4] *word*
[5] *sentence*

در نتایج بررسی محتوا همچنین ذکر شده است که خود ارجاعی در این مقاله هم در نظرات جعلی بیش‌تر ملاحظه شده است. استفاده از نام سازنده در نظر و یا تناوب تکرار نام سازنده در نظر جعلی وجود دارد و جملات حسی مثبت در نظر جعلی بیش‌تر از جملات حسی منفی در نظرات صادقانه است.

کلاس‌بندی نظرات با استفاده از مدل زبان QuickML[۱] دو مدل جداگانه unigram ایجاد می‌شود و این دو همراه با امتیازات حسی و API به عنوان ورودی به SVM داده می‌شود. در انتها از ارزیاب انسانی و مجموعه‌ی جدیدی از ارزیاب Mturk استفاده شده است.

در نتیجه‌گیری پایانی ذکر شده است زمانی که در یک نظر مقایسه بین دو محصول صورت می‌گیرد، دسته‌بند خودکار درست عمل نمی‌کند و هنوز هم دسته‌بندهای انسانی در مقایسه با دسته بندهای خودکار، دقت بالاتر و بیش‌تری دارند.

۳-۱-۴- تشخیص اسپم توسط امتیاز[۲]دهی

از پایگاه داده‌ی آمازون در این روش[۱۸] استفاده شده است[۳]. اطلاعات در چهار مرحله، پیش پردازش می‌شوند:

۱. حذف کاربران ناشناس[۴]
۲. حذف محصول تکراری[۵]
۳. حذف کاربران غیر فعال و محصولات غیر مرسوم[۶]
۴. شفاف‌سازی مترادف‌های نام سازنده[۷]

جدول۳- داده‌های استفاده شده

	Number (before preprocessing)
U: set of users	11,038 (313,120)
O: set of objects	5,693 (32,075)
V: set of reviews	48,894 (404,637)
E: set of ratings	as above

در این مقاله[۱۸] ثابت می‌شود برای محصولات معدود یا محصولات هدف[۸] اسپم ارسال می‌شود. پس از انجام مراحل پیش پردازش، شناسایی اسپم از طریق الگوی محتوای نظر و امتیاز دهی به مدل کردن چهار رفتار متفاوت نظردهندگان اسپم انجام

[1] *http://www.speech.cs.cmu.edu/tools/lm.html*
[2] *rating*
[3] *This dataset, known as Manufactured Products (MProducts)*
[4] *Removal of Anonymous*
[5] *Removal of Duplicate Products*
[6] *Removal of Inactive Users and Unpopular Product*
[7] *Resolution of Brand Name Synonyms*
[8] *targeted products*

می‌شود. محصول هدف، گروه هدف[۱] و انحراف امتیازدهی[۲] و انحراف ابتدایی امتیاز[۳]. تمام امتیازهای رفتارهای متفاوت با هم ترکیب شده و برای هر کاربر امتیازی برای اسپم بودن، تعلق می‌گیرد.

ارزیابی نظردهندگان و نظرات توسط انسان انجام می‌گیرد و برای سادگی کار از نرم‌افزار[۴] استفاده می‌شود. توسط این برچسب‌گذاری مدل رگرسیون آموزش می‌بیند و به نظردهندگان نمره[۵] می‌دهد.

این روش کارآیی بهتری نسبت به روش‌های مبتنی بر امتیاز مفید بودن نظر، دارد.

۳-۲- تشخیص اسپم بدون ناظر[۶]

با توجه به مشکل بودن برچسب‌زنی داده‌ها برای آموزش، یادگیری با ناظر به‌تنهایی برای شناسایی و تشخیص نظر جعلی، کافی نبوده و کار شناسایی بسیار سخت است. در این بخش نیز به معرفی سه روش بدون ناظر پرداخته می‌شود.

۳-۲-۱- تشخیص اسپم با استفاده از انتشار در شبکه نظرات

در این روش[۱۰] تمرکز بیش‌تر بر روی متن نظر و تحلیل رفتاری است یعنی سرعت تشخیص در این روش با رشد شبکه به‌صورت خطی افزایش پیدا می‌کند.

بر خلاف کارهای قبلی که بر روش‌های اکتشافی یا بر شواهد رفتاری تمرکز کرده بودند در این روش از شبکه‌ی نظردهندگان و اطلاعات مربوط به محصولات، اطلاعات قوی و خوبی به‌دست آورده است.

از چارچوب کاری که از الگوریتم مبتنی بر انتشار برای دسته‌بندی تاثیر شبکه است و خلاصه‌سازی و تحلیل نتایج را فراهم می‌کند، استفاده می‌کند. چارچوب کاری معرفی شده FRAUDEAGLE نام دارد.

مجموعه داده‌های نظرات بر خط[۷] شامل مجموعه‌ای از کاربران(مشتری یا نظردهنده و...) و مجموعه‌ای از محصولات(هتل‌ها، رستوران‌ها و...) است.هر نظر توسط یک کاربر و درخصوص یک محصول است که امتیازدهی ستاره‌ای[۸] آن بین یک تا پنج است. در این شبکه کاربران گره[۹]هایی هستند که به گره‌های محصول متصل هستند. و این ارتباطات از طریق ارسال نظر ایجاد می‌شود.

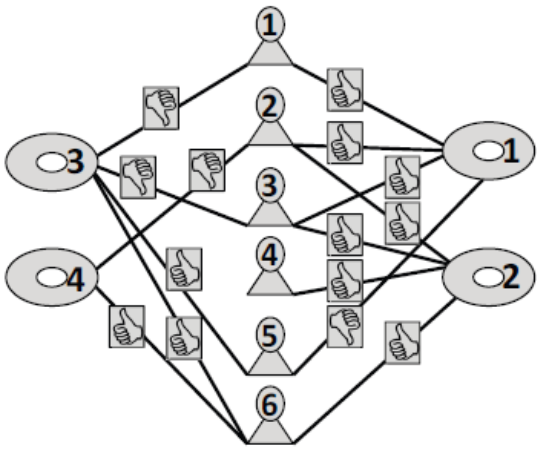
دو کلاس برای هر نوع در نظر گرفته شده است: محصولات کیفیت خوب یا بد دارند، کاربران نیز صادق[۱] و کلاهبردار[۲] هستند و نظرات نیز ممکن است واقعی یا دروغین باشد. پس یک محصول خوب (بد) است، اگر در اغلب موارد نظرات مثبت(منفی) از کاربر

---

[1] *Product Group*
[2] *general rating deviation*
[3] *early rating deviation*
[4] *Review spam detection software*
[5] *Score*
[6] *Unsupervised Spam Detection*
[7] *online*
[8] *star-rating*
[9] *node*

صادق (کلاهبردار) داشته باشد. به صورت مشابه یک کابر صادق(کلاهبردار) است اگر به‌طور عمده نظرات مثبت(منفی) برای محصولات خوب و نظرات منفی(مثبت) برای محصولات بد ارسال کند. بنابراین می‌توان در حالت ایده‌آل تمام نظرات کاربر درست‌کار را حقیقی و به‌همین صورت نظرات کاربر کلاهبردار را جعلی دانست، اگر چه چنین کاربری برای پنهان کردن اعمال خودش نظرات غیر جعلی هم دارد.

ابتدا باید دانست که تشخیص اسپم به عنوان یک مساله‌ی دسته‌بندی بسیار سخت است، در این‌جا این دسته‌بندی بر روی شبکه‌ی دو قسمتی[۳] نظرات صورت می‌گیرد. بنابراین کنترل تاثیر شبکه، این دسته‌بندی را بهبود می‌دهد.(شکل۲)



شکل۲- شبکه‌ی نظرات شامل ۶ کاربر و ۴ محصول، حس نظر بر روی یال‌ها با شست بالا+ و شست پایین - نمایش داده شده.

الگوریتم امتیازدهی با استفاده علامت‌دهی به شبکه انجام می‌شود. این روش ارزیابی را بر روی سه نوع داده انجام می‌دهد. نظرات: جعلی یا صادقانه, کاربران: درست‌کار و کلاه‌بردار، محصولات: کیفیت خوب یا کیفیت بد.

چارچوب کاری معرفی شده برای انواع نظرات کاربردی است و چون بدون ناظر است به هیچ پیش‌زمینه و اطلاعی در مورد دسته‌بندی نیاز ندارد و پیچیدگی آن با پیچیدگی زمان اجرا[۴]، مقیاس‌پذیر است.

این چارچوب کاری علاوه بر تشخیص اسپم، در حملات گروهی اسپم نیز موفق بوده است.

---

[1] *honest*
[2] *fraud*
[3] *bipartite*
[4] *run-time*

۳-۲-۲- تشخیص اسپم به‌روش مکاشفه‌ای

در این روش[۱۳] تشخیص اسپم به‌عنوان یک مساله‌ی داده کاوی برای کشف قواعد غیرمنتظره[۱]ی کلاس وابستگی[۲]، فرض شده است. این تکنیک مشکلات کلاس را طبق استقلال دامنه[۳]، حل می‌کند.

قواعد کلاس وابستگی، انواع خاص قواعد وابستگی هستند که کلاسی ثابت از صفات را دارند. این داده‌ها شامل مجموعه‌ای از رکوردهایی است که توسط صفات[۴] متعارف یا معمولی $A = \{A_1, \cdots, A_n\}$ و کلاس صفات $C = \{c_1, \cdots, c_m\}$ از مقادیر گسسته‌ی m، تعریف می‌شوند. که m کلاس برچسب[۵] نامیده می‌شود.

یک قاعده‌ی CAR این‌گونه نمایش داده می‌شود، $X \rightarrow c_i$ که X مجموعه شرایطی از صفاتی است که در A وجود دارد و c_i کلاسِ برچسب C است. قاعده با احتمال شرطی[۶] $\Pr(c_i \mid X)$ (که اطمینان[۷] نامیده می‌شود) و احتمال توأم[۸] (مشترک) $\Pr(X, c_i)$ (که پشتیبان[۹] نامیده می‌شود) محاسبه می‌شود.

هر رکورد داده‌ای در این روش بدین شکل است که: هر نظر همراه با مجموعه صفاتی، ذخیره و ثبت می‌شود. به عنوان مثال شناسه‌ی نظردهنده[۱۰]، شناسه‌ی سازنده[۱۱]، شناسه‌ی محصول[۱۲]، و یک کلاس. که این کلاس حس نظردهنده بر روی محصول، مثبت، منفی، یا بی‌طرف[۱۳] را بر پایه‌ی امتیاز نظر را نشان می‌دهد. در بسیاری از سایت‌ها مانند سایت amazon.com امتیازدهی بین یک تا پنج است، یک برای کم‌ترین و پنج برای بیش‌ترین امتیاز. امتیاز ۴ و ۵ مثبت، ۳ بی‌طرف، و ۱و ۲ منفی در نظر گرفته می‌شوند. یکی از قواعد کشف شده، می‌تواند این باشد که نظردهنده تمام امتیازهای مثبت خود را به سازنده‌ی خاصی از محصولات می‌دهد.

چهار نوع از قواعد غیر منتظره که معرفی شده است:

- اطمینان غیرمنتظره[۱۴]: با استفاده از این سنجه می‌توان نظردهنده‌هایی را یافت که همه‌ی امتیازهای بالا را به محصولات یک سازنده خاص می‌دهد اما بقیه‌ی نظردهندگان عموما در مورد این سازنده، نظرات منفی دارند.
- پشتیبانی غیرمنتظره[۱۵]: با استفاده از این سنجه می‌توان نظردهندگانی را یافت که چندین نظر برای یک محصول ارسال کرده‌اند درصورتی بقیه‌ی نظردهندگان فقط یک نظر نوشته‌اند.

---

[1] *Unexpected*
[2] *class association rules (CAR)*
[3] *domain independence*
[4] *Attribute*
[5] *class labels*
[6] *conditional probability*
[7] *confidence*
[8] *joint probability*
[9] *support*
[10] *reviewer-id*
[11] *brand-id*
[12] *product-id*
[13] *neutral*
[14] *Confidence unexpectedness*
[15] *Support unexpectedness*

- توزیع صفات غیرمنتظره[1]: با استفاده از این سنجه می‌توان نظرات مثبت بسیاری برای محصولات یک سازنده‌ی خاص که توسط یک نظردهنده نوشته شده است را یافت درصورتی که تعداد زیادی نظردهنده درخصوص محصولات این سازنده، نظر دارند.
- صفات غیرمنتظره[2]: با استفاده از این سنجه نظردهندگانی را که فقط نظر مثبت برای سازنده‌ی خاص ارسال کردند و فقط نظر منفی برای دیگر سازنده‌ها دادند را می‌توان یافت.

مزیت استفاده از این روش و این سنجه‌ها بر روی قواعد CAR تعریف شده است. این در حوزه‌های دیگر نیز می‌تواند برای یافتن الگوهای غیرمنتظره استفاده شود. ضعف آن این است که برخی رفتارهای غیرمعمولی نمی‌تواند شناسایی شود، مانند رفتارهایی که به زمان مرتبط هستند به این علت که کلاس وابستگی به زمان مرتبط نیست.

باید توجه داشت که مطالعات بر روی رفتارها مبتنی بر داده‌های عمومی[3] بر روی صفحات نظر میزبان سایت‌ها است. همان‌طور که پیش از این گفته شد، میزبان این سایت‌ها داده‌های دیگری درخصوص نظردهنده و فعالیت‌های او در سایت، ذخیره می‌کند. این داده‌ها برای عموم قابل مشاهده نیستند اما می‌توانند برای تشخیص اسپم، حتی بیش‌تر از داده‌های عمومی، مفید واقع شوند. برای مثال اگر چندین شناسه‌ی کاربری با آدرس IP یکسان نظرات مثبت در مورد یک محصول ارسال کنند، این کاربر مشکوک به تولیدکننده‌ی اسپم است.اگر تمام نظرات مثبت در مورد یک هتل همه از محدوده‌ی اطراف هتل باشد، باز هم نظرات مشکوک هستند. برخی از سایت‌های نظرات، داده‌های داخلی[4] برای تشخیص اسپم و تولیدکنندگان اسپم، استفاده می‌کنند.

۳-۲-۳- تشخیص اسپم با استفاده از گراف نظرات

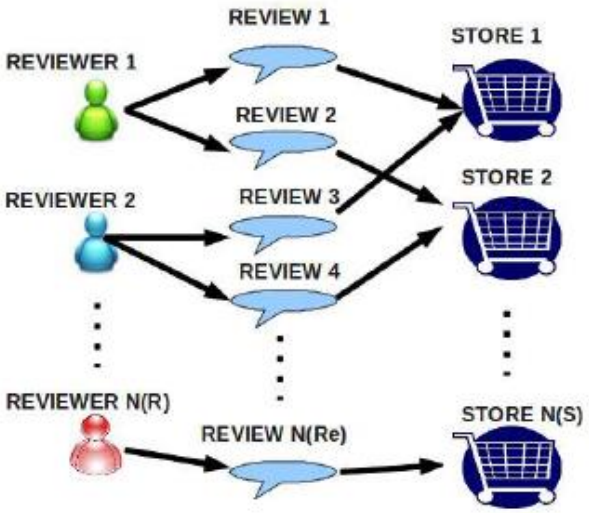
روش مبتنی بر گراف[۱۴] برای شناسایی اسپم در نظرات مغازه و فروشگاه معرفی شده است. این مطالعه بر روی نظرات سایت resellerratings.com انجام شده است. این وب سایت یک آدرس URL جداگانه برای پروفایل هر کاربر دارد که شامل متادیتاهایی مانند شناسه‌ی کاربر، تمام نظرات کاربر و امتیازات و زمان ارسال نظرات است. در هر صفحه‌ی فروشگاه، اطلاعاتی در مورد میانگین امتیاز همه‌ی نظردهندگان با نظرات آن‌ها ثبت شده است. اطلاعات در تاریخ ۶ اکتبر سال ۲۰۱۰ از سایت اخذ شده است پس از آن کاربران و نظراتشان از سایت حذف شد. نظرات جمع‌آوری شده ۳۴۳۶۰۳ بودند که توسط ۴۰۸۴۷۰ نظردهنده بر روی ۱۴۵۶۱ فروشگاه نظر داده بودند.

بر اساس ایده‌های تشخیص اسپم برای محصول، می‌توانیم نظراتی که شخصی برای یک محصول ارسال کرده است، مشکوک فرض کنیم، اما نظراتی که برای چندین محصول از یک کاربر ارسال می‌شود را معمولی و عادی بدانیم. همچنین نظرات نزدیک به

[1] *Attribute distribution unexpectedness*
[2] *Attribute unexpectedness*
[3] *public*
[4] *internal data*

تکراری از یک کاربر برای چندین فروشگاه، به این علت که برای محصولات متفاوت ارسال شده است، می‌تواند عادی باشد. بنابراین ویژگی‌ها یا اثر[1] معرفی شده در تشخیص اسپم در نظرات فروشگاه مناسب نیست. بنابراین به چارچوب کاری مکمل نیاز است.

این مقاله از گراف ناهمگن با سه نوع از نود، نود نظرات، نظردهندگان، و فروشگاه‌ها برای به‌دست آوردن ارتباط بین آن‌ها و مدل کردن اثر تولید اسپم، استفاده کرده است. نود نظردهنده با نوشتن نظر، به آن متصل می‌شود. نود نظر به فروشگاه زمانی متصل می‌شود که درمورد فروشگاه باشد. فروشگاه از طریق نظرِ نظردهنده روی فروشگاه به نظردهنده متصل می‌شود. هر نود به مجموعه‌ای از ویژگی‌ها نیز متصل است. به عنوان مثال نود فروشگاه ویژگی‌هایی در مورد میانگین امتیازها، تعداد نظرات و غیره را شامل می‌شود.



شکل۳- گراف نظر

بر این اساس سه مفهوم تعریف می‌شود. قابلیت اعتماد[2] نظردهنده، صداقت[3] نظر و قابلیت اطمینان فروشگاه. نظردهنده زمانی قابل اعتماد[4] است که نظرات صادقانه ارسال کرده باشد. فروشگاه زمانی قابل اطمینان است که نظرات مثبت از نظردهندگان قابل اعتماد داشته باشد و نظر زمانی صادقانه است که توسط بقیه‌ی نظرات صادقانه، پشتیبانی شود. بنابراین اگر نظردهنده برای فروشگاه نظر بدی ارسال کند، صداقت نظر تحت تاثیر قابلیت اعتماد نظردهنده که با فروشگاه برخورد داشته است، پایین می‌آید. با توجه به چگونگی نظرِ نظردهندگان با دیگر نظرات در مورد فروشگاه یکسان، قابلیت اعتماد نظردهنده تغییر می‌کند.

فرضیات مورد نظر مقاله برای ایجاد اسپم بدین‌صورت ذکر شده است:

- ایجاد منفعت و سود، اتصال آن‌ها به فروشگاه بابت بدنام کردن فروشگاه‌های دیگر و ارتقاء خودشان است

[1] *clues*
[2] *trustiness*
[3] *honesty*
[4] *trustworthy*

- تولیدکنندگان اسپم معمولا توسط فروشگاه‌های بی‌کیفیت استخدام می‌شوند تا نظر جعلی بنویسند. فروشگاهی با شهرت خوب و مشتریان ثابت، نیازی به تولیدکننده‌ی اسپم ندارند. حتی اگر این فروشگاه‌ها تولیدکنندگان اسپم را برای نوشتن نظر جعلی مثبت استخدام کنند، این مورد مضر نیست. بنابراین فروشگاه‌هایی که قابلیت اطمینان پایین‌تری دارند در تولید اسپم نقش بیش‌تری دارند.

- اسپم‌های مضر همیشه از حقیقت انحراف دارند. پس یا نظرات مثبت برای فروشگاه‌های بد و یا نظرات منفی در خصوص فروشگاه‌های خوب.

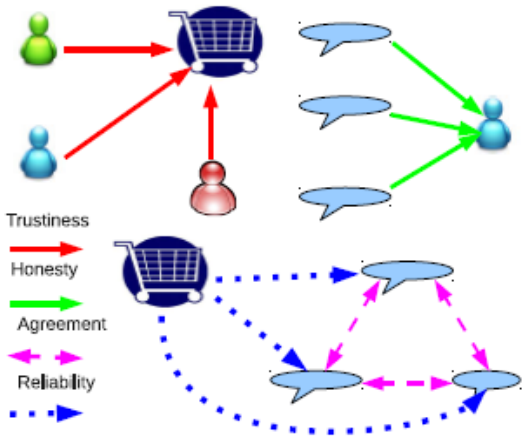
- تمام نظرات منحرف اسپم نیستند. مردم احساس و تجارب متفاوتی نسبت به یک سرویس مشابه دارند.

بر اساس مطالب ذکر شده، مشاهدات بدین صورت است:

۱. می‌توان درستی نظر را بر اساس قابلیت اطمینان فروشگاهی که به آن پست شده و سازش[1] نظر با دیگر نظراتی که در مورد فروشگاه یکسان پست شده است، قضاوت کرد.

۲. اگر امتیاز همه‌ی نظرات یک نظردهنده صادقانه باشد، می‌توان استنباط کرد قابلیت اعتماد ، دارد. به این علت که قابلیت اعتماد کسی بالاست که تمام نظرات آن صداقت بالا داشته باشد.

۳. فروشگاه زمانی قابلیت اطمینان بالا دارد که نظردهندگان قابل اعتماد با نظرات مثبت داشته باشد و زمانی قابلیت اطمینان کم دارد که نظردهندگان قابل اعتماد با نظرات منفی دارد. شکل۴ موراد ذکر شده در خصوص قابلیت اعتماد در نظر، نظردهنده و فروشگاه رانمایش می‌دهد.



شکل۴ - تاثیر انواع نودها

[1] *agreement*

روش محاسباتی تکرار شونده‌ای برای محاسبه‌ی سه مقدار جهت رتبه‌بندی[1] نظردهنده، نظر و فروشگاه استفاده می‌شود.

ارزیابی توسط قضاوت انسانی و مقایسه با امتیاز[2] فروشگاه‌ها که از از BBB[3] اخذ شده است. BBB موسسه‌ای در آمریکا است که گزارشاتی را در مورد قابلیت اطمینان کسب و کار جمع‌آوری کرده و در خصوص کسب و کار یا مصرف‌کننده به عموم هشدار می‌دهد. دقت[4] ارزیاب‌ها در تشخیص اسپم ۴۹درصد اعلام شده است. با وجود این که در تشخیص اسپم از روش ماهرانه‌ای استفاده شده است اما دلیل دقت نه‌چندان خوب این روش، قضاوت ارزیاب‌ها در ارزیابی بیان شده است.

۳-۳- تشخیص اسپم گروهی

در [۱۶و۱۷] گفته شده تا زمان ارائه‌ی مقاله، به موضوع تشخیص اسپم گروهی، پرداخته نشده است. الگوریتمی که برای تشخیص استفاده می‌شود در دو مرحله انجام می‌شود:

۱. استخراج الگوی تکرار شونده[5]: در ابتدا داده‌های نظرات را برای تهیه‌ی مجموعه‌ای از تراکنش‌ها، استخراج می‌کنیم. هر تراکنش یک محصول منحصر و (شناسه) نظردهندگانی که به محصول نظر دادند را نمایش می‌دهد. پس از آن استخراج الگوی تکرار شونده، اجرا می‌شود. هر الگو، گروهی از نظردهندگان است که روی محصول، نظر ارسال کرده‌اند. در انتها گروه‌هایی که کاندیدای اسپم گروهی هستند، به‌دست می‌آیند. تشخیص و شناسایی گروهی از نظردهندگان که سعی در ارتقا یا بی‌اعتبار کردن یک محصول را دارند از طریق جمع‌آوری رفتار آن‌ها بسیار سخت است، به همین دلیل از این شیوه استفاده می‌شود. استخراج الگوی تکرار شونده می‌تواند افرادی که با یکدیگر بر روی چندین محصول اسپم ارسال می‌کنند را هم بیابد.

۲. رتبه‌بندی گروه مبتنی بر شاخص‌های گروه اسپم[6]: ممکن است تمام گروه‌هایی که در مرحله‌ی قبلی به‌دست آمدند، گروه‌های تولیدکننده‌ی اسپم نباشند. بسیاری از نظردهندگان در حین استخراج الگو تصادفا با یکدیگر در یک گروه قرار گرفته‌اند. پس از مجموعه‌ای از شاخص‌ها جهت اخذ رفتارهای نوع متفاوت گروهی غیر عادی و عضو انفرادی، استفاده می‌شود.

پنجره‌ی زمانی[7]: گروه تولیدکننده‌ی اسپم معمولا با هم کار می‌کنند و در یک بازه‌ی زمانی کوتاه، نظرات جعلی برای یک محصول خاص ارسال می‌کنند.

انحراف گروه[1]: زمانی‌که اعضای گروه برای تولید اسپم، با هم کار می‌کنند امتیاز بسیار بالا یا بسیار کم برای محصول ارسال می‌کنند. معمولا گروه اسپم در امتیازدهی به محصولات از امتیاز نظردهندگان عادی، به‌صورت قابل توجهی انحراف دارند.

---

[1] *Rank*
[2] *Score*
[3] *Better Business Bureaus*
[4] *Precision*
[5] *Frequent pattern mining*
[6] *Rank groups based on a set of group spam indicators*
[7] *Time Window*

شباهت محتوای نظر گروهی[۲]: اعضای گروه تولیدکننده‌ی اسپم ممکن‌است یکدیگر را بشناسند و نظر یکدیگر را در مورد محصول کپی‌برداری کنند، بنابراین محصولاتی که مورد حمله‌ی گروه اسپم قرار گرفته است، محتوای نظر شبیه به یکدیگر داشته باشند.

شباهت محتوای عضو[۳]: در صورتی‌که اعضای گروه یکدیگر را نشناسند، هر عضو فقط نظر قبلی خودش را کپی می‌کند و یا تغییر می‌دهد. اگر تعداد زیادی از اعضای گروه این کار بیش‌تر شبیه به کار تولیدکنندگان اسپم گروهی خواهد شد.

قاب زمان ابتدائی[۴]: یکی از فعالیت‌های مخرب این گروه، ارسال نظر بلافاصله بعد ازعرضه‌ی محصول است. این عملیات تاثیر زیادی در کنترل احساس مشتریان روی محصول دارد.

نسبت اندازه‌ی گروه[۵]: نسبت اندازه‌ی گروه و تعداد نظرات روی محصول نیز شاخص خوبی برای شناسایی تولید اسپم است. در بدترین حالت تمام تظرات روی یک محصول فقط نظر اعضای گروه تولیدکننده‌ی اسپم باشد که بسیار مخرب است.

اندازه‌ی گروه[۶]: اگر اندازه‌ی گروه بزرگ باشد، احتمال اعضا به‌صورت اتفاقی در یک گروه قرار بگیرند، کوچک است. بنابراین گروه بزرگ‌تر تاثیر مخرب کم‌تری خواهد داشت.

تعداد پشتیبانی[۷]: تعداد محصولاتی که گروه تولیدکننده‌ی اسپم روی آن کار می‌کنند. بالا بودن این تعداد هشداردهنده است.

پس از آن مدل رابطه‌ای به نام GSRank[۸] معرفی شده است که روابط بین گروه‌ها، افراد منحصر از گروه و محصولات را استخراج می‌کند. الگوریتم تکرار شونده برای حل این مساله پیشنهاد شده است. مجموعه‌ای از گروه تولیدکننده‌ی اسپم، به‌صورت دستی، برچسب‌گذاری می‌شوند و برای ارزیابی مدل ارائه شده، استفاده می‌شود.

با توجه به آستانه[۹]ی تکراری که در استخراج الگو استفاده شد، اگر گروه، اغلب اوقات با هم کار نمی‌کردند(۳ مرتبه و یا بیش‌تر)، توسط این روش شناسایی نمی‌شدند و این یکی از ضعف‌های این روش است.

---

[1] *Group Deviation (GD)*
[2] *Group Content Similarity (GCS)*
[3] *Member Content Similarity (MCS)*
[4] *Early Time Frame (ETF)*
[5] *Ratio of Group Size (RGS)*
[6] *Group Size (GS)*
[7] *Support count*
[8] *Group Spam Rank*
[9] threshold

۳-۴- ارزیابی روش‌ها

نظرات به عنوان ملاک عمل برای انتخاب بسیار حائز اهمیت است. این مهم، امروزه توسط افرادی بـا انگیـزه‌ی شخصـی و یـا سازمانی، دست‌خوش تغییر و حمله قرار می‌گیرند. یکی از نیازهای فعلی در زمینه‌ی سایت‌های نظر، خالی شدن نظرات سـاختگی و جعلی از این سایت‌ها است. روش‌هایی از سال ۲۰۰۷ تا کنون در این زمینه ارائه شده است. روش‌هایی مبتنی بر محتوای نظر، و یـا مبتنی بر اطلاعات نظردهنده، تفکیک نظرات صادقانه از نظرات جعلی و اسپم از طریق یادگیری با ناظر یا بدون نظارت انسان.

در این گزارش تکنیک‌ها و روش‌های تشخیص اسپم در نظرات معرفی شدند و در هر بخش مختصری در خصوص تکنیک‌های به‌کار رفته، ارائه شد. در انتها ارزیابی روش‌های ذکر شده را از نظر دقت بررسی می‌شود.

جدول ۴- ارزیابی روش‌های بررسی شده[۲۱]

ردیف	روش	Precision (دقت)	ویژگی(های) مورد استفاده
	تشخیص با ناظر		
۱	تشخیص اسپم به روش دسته‌بندی نظرات به تکرای و نزدیک به‌تکراری [۲]	٪۸۵	مبتنی بر نظر
۲	تشخیص اسپم به روش آموزش ماسین بادگیری [۱۹]	٪۵۱/۷	مبتنی بر نظر و نظردهنده
۳	تشخیص با استفاده از محاسبات انسانی[۱۱]	٪۷۴/۴-٪۴۳/۶* [1]	مبتنی بر نظر
۴	تشخیص با استفاده از تاثیر شبکه‌ای[۱۰]	-	مبتنی بر نظر و نظردهنده
	تشخیص بدون ناظر		
۵	تشخیص اسپم به‌روش مکاشفه‌ای[۱۳]	-	مبتنی بر نظر و نظردهنده
۶	تشخیص اسپم به روش امتیازدهی[۱۸]	٪۷۸	مبتنی بر نظر و نظردهنده
۷	تشخیص اسپم با استفاده از گراف نظرات[۱۴]	٪۴۹	مبتنی بر نظر و نظردهنده

فهرست مراجع

[1] Jindal, Nitin, and Bing Liu. "Review spam detection." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[1] *for Low-rated review and high-rated review*

[2] Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.

[3] Bing Liu, Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012

[4] Castillo, Carlos, et al. "A reference collection for web spam." ACM Sigir Forum. Vol. 40. No. 2. ACM, 2006.

[5] Jindal, Nitin, and Bing Liu. "Analyzing and detecting review spam." Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007.

[6] Z. Gyongyi & H. Garcia-Molina. Web Spam Taxonomy. Technical Report, Stanford University, 2004.

[7] Nikita Spirin and Jiawei Han. Survey on Web Spam Detection: Principles and Algorithms Department of Computer Science University of Illinois at Urbana- Champaign Urbana, IL 61801, USA

[8] Gordon Cormack and Thomas Lynam. Spam corpus creation for TREC. In Proceedings of Second Conference on Email and Anti-Spam,CEAS'2005, 2005.

[9] Patil et al., Review on Brand Spam Detection Using Feature Selection, International Journal of Advanced Research in Computer Science and Software Engineering 3(9), September - 2013, pp. 744-748

[10] Akoglu, Leman, Rishi Chandy, and Christos Faloutsos. "Opinion Fraud Detection in Online Reviews by Network Effects." ICWSM. 2013.

[11] Harris, C. "Detecting deceptive opinion spam using human computation." Workshops at AAAI on AI. 2012.

[12] Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter//Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS). 2010.

[13] Jindal, Nitin, Bing Liu, and Ee-Peng Lim. "Finding unusual review patterns using unexpected rules." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[14] Wang, Guan, et al. "Review graph based online store review spammer detection." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

[15] Gao H, Hu J, Wilson C, et al. Detecting and characterizing social spam campaigns//Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010: 35-47.

[16] Mukherjee, Arjun, et al. "Detecting group review spam." Proceedings of the 20th international conference companion on World wide web. ACM, 2011.

[17] Mukherjee, Arjun, Bing Liu, and Natalie Glance. "Spotting fake reviewer groups in consumer reviews." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[18] Lim E P, Nguyen V A, Jindal N, et al. Detecting product review spammers using rating behaviors//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 939-948.

[19] Li, Fangtao, et al. "Learning to identify review spam." IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. No. 3. 2011.

[20] http://consumerist.com/2010/04/how-you-spot-fake-onlinereviews.html

[21] Dixit, Snehal, and A. J. Agrawal. "Survey on Review Spam Detection." International Journal of Computer & Communication Technology ISSN (PRINT): 0975-7449.2013

[۲۲] اعراب شیبانی امیر، پویان علی اکبر، دستاوردهای اخیر در واکاوی نظرات با تاکید بر روش‌های اسپم شناسی، اولـین همـایش ملی فناوری اطلاعات و شبکه‌های کامپیوتری دانشگاه پیام نور، بهمن ماه ۱۳۹۱