فصلنامه آینده پژوهی مدیریت

سال سی وششم/ شماره۱٤۱/ تابستان ۱٤۰٤

پیش بینی رفتار بازنشر کاربران با استفاده از یادگیری عمیق از طریق بررسی ارزش محتوایی هر توئیت

حسنیه صفی آریان

دانشجوی دکتری ،دانشکده مدیریت و اقتصاد ،دانشگاه آزاد اسلامی واحد علوم تحقیقات

محمد جعفر تارخ(مسئول مکاتبه)

گروه مهندسی فناوری اطلاعات،دانشکده مهندسی صنایع دانشگاه خواجه نصیرالدین طوسی

mjtarokh@kntu.ac.ir

محمدعلی افشار کاظمی

گروه مدیریت صنایع، دانشکده مدیریت دانشگاه آزاد اسلامی واحد تهران مرکز

تاریخ دریافت:۱۴۰۳/۰۲/۲۳ تاریخ پذیرش:۱۴۰۳/۰۷/۱۵

چکیده:

زمینه : بررسی رخدادها و روند مواجهه کاربران شبکه‌های اجتماعی با پدیده‌های روزمره، بخش مهمی از تحلیل‌های آینده پژوهی را به خود اختصاص داده است. کاربران بعد از مطالعه توئیت‌ها، اقدام به تصمیم‌گیری برای بازنشر توئیت‌های مورد علاقه خود می‌نمایند.

هدف : بررسی‌ها نشان می‌دهد تنها ۱٪ کاربران، ۵۰٪ توئیت‌های توئیتر را ایجاد می‌کنند و ۲۵٪ گردش اطلاعات آن را کنترل می‌نمایند. برطبق این آمار، بازنشر توئیت‌ها در توئیتر، یک ابزار مهم برای انتشار اطلاعات می‌باشد. پیشگویی رفتار بازنشر کاربران یکی از چالش‌های مهم در حوزه میکروبلاگ‌ها به شمار می‌آید.

روش‌ها: روش‌هایی که ارائه شده است مبتنی بر خصیصه‌های آماری مربوط به داده‌های مختلف این میکروبلاگ می‌باشند و کمتر به صورت محتوایی اثرگذاری ، هر توئیت را بر روی کاربر مشخص، مورد اندازه‌گیری قرار داده است. علی رغم تنوع موضوعی ، محتوایی توئیت‌ها و کاربران مختلف، اکثر این روش‌ها با ارائه یک مدل عمومی بر مبنای خصیصه‌های پرتعداد، از دقت برخوردار نیستند و قادر به ارائه پیشگویی در زمان برخط نمی باشند.

یافته‌ها: در این مقاله با معرفی خصیصه‌های مبتنی بر بررسی محتوایی یک توئیت و اندازه‌گیری آن‌ها، تخمین میزان ارزش محتوایی یک توئیت برای کاربر مشخص، صورت می‌گیرد. با تعریف این مسئله به صورت یک مسئله کلاس‌بندی باینری بر اساس خصیصه‌های پیشنهادی، پیشگویی رفتار بازنشر کاربران به صورت انفرادی صورت می‌گیرد.

نتیجه‌گیری: مسئله کلاس بندی با استفاده از جدیدترین دستاوردهای حوزه یادگیری عمیق حل شده است. با کمک سه مجموعه داده جمع آوری شده واقعی، کارایی روش پیشنهادی مورد ارزیابی قرار می‌گیرد. دقت اندازه‌گیری خصیصه‌های پیشنهادی ۰.۸۵.بررسی شده . تاثیرگذار بودن آن‌ها نشان داده می‌شود.

واژگان کلیدی: شبکه های اجتماعی ، کاربر، یادگیری عمیق ، بازنشر، محتوا

۱- مقدمه

در سال های اخیر ،داده های بزرگ نقش حیاتی در سازمان ها ایفا می کنند. در این شرایط بهره گیری از اطلاعات صحیح، دقیق، کامل و به روز در سیستم های اطلاعاتی امری اجتناب ناپذیر محسوب می شود. همچنین با توجه به وجود رقابت گسترده در کسب وکارهای مبتنی بر این حوزه، اتخاذ تصمیم های دقیق، به موقع و تاثیرگذار دارای اهمیت زیادی است.بنابراین داشتن داده های دقیق و ایجاد الگوی درست و به جا ، می تواند یکی از پژوهش های مهم در حوزه آینده پژوهی باشد.

با افزایش نفوذ میکروبلاگ‌ها و تعداد کاربران آن‌ها، مطالعه روی رفتار کاربران در این گونه رسانه‌ها می‌تواند اطلاعات ارزشمندی از باورها و سلایق کاربران را در اختیار پژوهشگران آینده پژوهی قرار دهد. در این گونه سایت‌ها یک کاربر نظرات، اطلاعات و اخبار خود را در قالب محتوا(در سایت توئیتر با عنوان توئیت) محدود (حداکثر280 کلمه) به اشتراک می‌گذارد. پیگیری‌کنندگان این کاربر (کسانی که توئیت‌های این کاربر را پیگیری می‌کنند) می‌توانند این توئیت را مشاهده کرده و با انتخاب گزینه بازنشر این پیغام را به منظور اطلاع از محتوای آن برای پیگیری‌کنندگان خود ارسال نمایند. بازنشر اطلاعات یکی از مهم‌ترین خصوصیت‌های میکروبلاک‌هایی مانند توئیتر به شمار می‌آید. با افزایش کاربردهای میکروبلاگ‌ها در زمینه اطلاع رسانی و آموزش و غیره، بررسی رفتار بازنشر به یک موضوع تحقیقاتی فعال مبدل گردید. این مکانیزم این امکان را فراهم می‌کند تا اطلاعات در قالب یک توئیت در سطح میکروبلاگ‌ها انتشار پیدا کند. بر اساس بررسی‌های انجام شده در ٪۵۰ اطلاعات منتشر شده در میکروبلاگ‌ها توسط ٪۱ از کاربران تهیه می‌شود و ٪۲۵ از گردش اطلاعات را در کنترل خود دارند. با توجه به این آمار، مطالعه روی رفتار بازنشر اطلاعات به عنوان یک مدل انتقال اطلاعات در میکروبلاگ‌ها، حائز اهمیت می‌باشد (Tang &et al.2015. 681–688).

در این مقاله به بررسی یکی از چالش‌های اساسی مرتبط با بررسی رفتار بازنشر کاربران، پرداخته می‌شود. پیشگویی رفتار بازنشر کاربران، به معنی بررسی عوامل اثرگذار در اتخاذ تصمیم برای بازنشر اطلاعات یا عدم بازنشر آن می‌باشد. در بسیاری از کاربردهای نظیر پالایش اطلاعات مربوط به توئیت‌ها، توصیه توئیت‌ها ، رتبه‌بندی توئیت‌ها ، تخمین همه‌گیری یک توئیت و غیره، پیشگویی رفتار بازنشر اطلاعات استفاده می‌گردد. در پژوهش‌های مرتبط با این حوزه بیشتر به صورت یک مسئله کلاس‌بندی باینری بر اساس خصیصه‌های مختلفی مانند ویژگی‌های مرتبط با ساختار شبکه اجتماعی (ارتباطات، تعداد پیگیری‌کنندگان و غیره)، توئیت، نویسنده توئیت و کاربر، رفتار بازنشر کاربران پیشگویی می‌گردد. با توجه به افزایش ارتباطات در میکروبلاگ‌ها و تغییر کارکرد این گونه سایت‌ها به عنوان یک محتوای اطلاعاتی عظیم از اخبار، اطلاعات و نظرات و غیره، رفتار بازنشر بیش از پیش تابع محتوای توئیت‌های مشاهده شده، می‌باشد. در این مقاله با معرفی مفهوم توئیت اثرگذار برای هر کاربر، به تبیین ارزش محتوایی هر توئیت پرداخته می‌شود. روش پیشنهادی برای تعیین ارزش محتوایی توئیت‌ها، شامل معرفی و بکارگیری خصیصه‌هایی است که عمدتا به بررسی محتوایی یک توئیت و میزان اثرگذاری آن بر روی کاربر مورد نظر می‌پردازد، تا بتواند رفتار بازنشر این کاربر را در قبال این توئیت پیشگویی نماید. از جمله خصیصه‌های پیشنهادی برای تعیین ارزش محتوایی یک توئیت، تازگی(غیرمنتظره بودن) توئیت می‌باشد. میزان تازگی یک توئیت یک خصیصه نسبی برای هر کاربر بوده و متناسب با ارتباط موضوعی توئیت با توئیت‌های منتشرشده توسط کاربر می‌باشد. در این مقاله ضمن بیان روشی برای اندازه‌گیری میزان تازگی یک توئیت برای یک کاربر، به بررسی ارتباط این خصیصه با رفتار بازنشر کاربران پرداخته

می‌شود. توئیت‌هایی با محتوای منفی، با توجه به.(Hansen &et al.2011-34–43) از قابلیت بازنشر بیشتری برخوردار می‌باشند. منفی بودن یک توئیت خصیصه دیگری است که برای بازنشر معرفی و روشی برای اندازه‌گیری آن به صورت برخط ارائه می‌شود. شباهت موضوع یک توئیت با علایق کاربر (محتویات توئیت‌های قبلی کاربر) و ارتباط مکانی احتمالی توئیت با کاربر در تصمیم برای بازنشر کاربران بسیار موثر است. در کنار این خصیصه‌ها، خصیصه‌هایی مانند نرخ رشد موضوع یک توئیت و وزن اجتماعی یک توئیت، در این مقاله معرفی و برای اندازه‌گیری آن‌ها روش‌هایی پیشنهاد می‌گردد. به منظور پیشگویی از روش کلاس‌بندی باینری با استفاده از یادگیری به کمک شبکه عصبی عمیق استفاده شده است. یادگیری عمیق تحول شگرفی در افزایش دقت کلاس بند داشته است. برای این کار، مسئله به صورت یک مسئله یادگیری کلاس بند با استفاده از شبکه عصبی عمیق تعریف شده است. ارزیابی روش پیشنهادی به دو صورت شهودی و با آزمایش مجموعه داده صورت می‌گیرد. در ارزیابی شهودی با کمک گروه کاربران داوطلب و ارائه توئیت‌های روزانه، میزان همبستگی بین خصیصه‌های پیشنهادی با رفتار بازنشر با کمک ضریب همبستگی point-biserial اندازه‌گیری می‌شود. همین طور دقت روش‌های اندازه‌گیری خصیصه‌های پیشنهادی با کمک ضریب همبستگی رتبه‌ای کاندال ارزیابی می‌گردد.

ادامه مقاله به این صورت می‌باشد که در قسمت بعد به مرورکارهای مرتبط به حوزه بازنشر اطلاعات و پیشگویی بازنشر پرداخته می‌شود. در قسمت ۳، ضمن معرفی خصیصه‌های پیشنهادی و روش اندازه‌گیری برای تخمین ارزش محتوایی هر توئیت برای هر کاربر، روش کلاس‌بندی باینری با ناظر با روش شبکه عصبی عمیق ارائه می‌گردد. برای ارزیابی روش پیشنهادی، در قسمت ۴، آزمایش با کمک مجموعه داده (سه مجموعه داده از توئیتر) برای ارزیابی انتخاب شده است.در قسمت ۵، جمع‌بندی مقاله و نتیجه گیری ارائه می‌گردد.

۲- مرورکارهای گذشته:

همواره مسئله ی پیش بینی تصمیم بازنشر در شبکه های اجتماعی به طور گسترده مورد مطالعه واقع شده است. در این بخش، یک مرور کلی از روش های موجود برای پیش بینی تصمیم بازنشر در اجتماعات، گردآوری شده است. در نگارش این بخش سعی شده است که علاوه بر مرور پژوهش‌های صورت گرفته، مزایا و محدودیت های روش های ارائه شده نیز مورد بحث و بررسی قرار گیرد.

درک چگونگی ایجاد یک پست توسط کاربران، و انگیزه های آن‌ها برای بازنشر یک پست، در پیش بینی تصمیم بازنشر یک پست، دارای اهمیت فراوان دارد. در واقع، فهمیدن این که کاربران چه محتواهایی را برای بازنشر انتخاب می کنند، می تواند به توضیح چرایی بازنشر یک پست خاص و پیش‌بینی آن، کمک کند. این انگیزه ها برای تصمیم بازنشر در مطالعه ی انجام شده در (Boyd&et all.2010.2-3) خوبی بررسی‌شده‌اند. آن‌ها دلایل اصلی تصمیم بازنشر توسط کاربران، را مشخص نموده و ده انگیزه ی مختلف را برای بازنشر معرفی کرده‌اند. از جمله این انگیزه‌ها می‌توان به نظر دادن در پست ها، انتشار پست ها به گیرندگان جدید برای آگاهی دادن به اشخاص یا گروه های خاص و ذخیره ی پست ها برای دسترسی آتی شخصی را، اشاره کرد. اگرچه تمرکز مطالعه ی آن‌ها پیش بینی تصمیم بازنشر پست ها نبوده است، اما انگیزه های گفته شدهٔ برای تصمیم بازنشر، می تواند در تعیین ویژگی ها مؤثر برای پیش بینی تصمیم بازنشر پست ها، مورد استفاده قرار گیرند.

مطالعات دیگری برای یافتن عوامل بازنشر توسط سو و همکارانش در (Kwak & et al. 2010. 591-600) انجام شده است. آن‌ها سه عامل نهفته از ویژگی های پست

را، با استفاده از روش تحلیل مؤلفه ی اصلی (PCA) استخراج کرده و تلاش کردند تا آن ها را به ویژگی های واقعی آشکار، ارتباط دهند. آن‌ها سپس یک مدل خطی برای یافتن احتمال بازنشر معرفی کردند. آن‌ها انگیزه ی خود را در انتخاب مدل خطی برای پیش بینی تصمیم بازنشر، ذکر نکرده‌اند و درباره ی مؤثر بودن روش PCA در یافتن عوامل مهم قابلیت بازنشر نیز بحث ننموده‌اند. همچنین آن‌ها تنها آزمایش های خود را بر روی مجموعهٔ محدودی از داده ها که خودشان ارائه دادند، اجرا کردند. آن‌ها نتیجه گرفتند که ویژگی های مبتنی بر محتوا نظیر هَشتَگ و url در قابلیت بازنشر، بسیار مهم هستند. البته این نتیجه‌گیری توسط مطالعات بعدی به چالش کشیده شد. آن‌ها نشان دادند که ویژگی های مربوط به محتوای یک پست به‌تنهایی، دارای اطلاعات کافی برای پیش بینی تصمیم بازنشر پست ها نمب‌باشند. این نکته در (Hwong & et al.2017.480-492) نیز مورد تایید قرار گرفت.

در یک مطالعه ی مشابه (Petrovic & et al. 2011)،یک تحقیق تجربی برای پیش بینی تصمیم بازنشر یک پست انجام شد. در این پژوهش، یک الگوریتم مبتنی بر یادگیری برخط، برای پیش بینی تا حد ممکن سریع، توسعه داده شد. برای مطالعات بیشتر در ارتباط با الگوریتم‌های مبتنی بر یادگیری برخط به دو مقاله (Alshaabi ۲۰۲۰ &et al.)مراجعه نمایید. سپس، برای مرحله آموزش مدل‌های ارائه شده، از زیرمجموعه های مختلف داده که بر اساس زمان و روز تولید شده بودند، استفاده کردند تا از اطلاعات زمانی پست ها برای پیش‌بینی بازنشر یک پست، بهتر بهره ببرند. در این پژوهش، همانند مطالعه ی (Suh &et al.2010. 177-184) ، دلیل استفاده از این مدل به صراحت بیان ن شده است و مدل با استفاده از مجموعه دادگان مختلف، مورد ارزیابی قرار نگرفته است. در این پژوهش، عملکرد روش یادگیری برخط، با پیش بینی‌های انسانی مقایسه شده و ادعا شده است که این روش به خوبیِ پیش بینی های انسانی عمل می کند.

در(Zaman & et al .2010. 17599-601) نیز، پژوهشی در ارتباط با پیش بینی تصمیم بازنشر بر اساس روش تصفیه همکارانه انجام دادند. برخلاف مطالعات دیگر که از ویژگی های مستقیماً استخراج شده از پست ها یا کاربران استفاده می کنند، آن‌ها از بازخوردهای مثبت و منفی غیرمستقیم در مدلشان استفاده کردند. اگر کاربران پیگیری‌کنندهٔ فعال یک پست را بازنشر کنند، به عنوان بازخورد مثبت، و در غیر این صورت به عنوان بازخورد منفی در نظر گرفته می شود. اما یک نقطه ضعف در این پژوهش این است که آن‌ها مدل های خود را بر اساس حداقل یک ساعت داده ی پس از انتشار یک پست آموزش دادند. از سوی دیگر، مطالعات قبلی نشان می دهند که بیشتر از نود درصد پست های بازنشر شده، در اولین ساعت پس از ایجاد اتفاق می افتند. بنابراین، آموزش یک مدل بر اساس چنین بازه ی زمانی طولانی از نظر عملی ارزشمند نیست.

در (Artzi & et al. 2012.602-606)، یک مدل متمایزکننده را برای پیش بینیِ احتمالاتی بازنشر یک پست، توسعه دادند. آن‌ها چندین ویژگی تاریخچه ای و لغوی از متن پست، و برخی ویژگی ها را از کاربران انتشاردهندهٔ پست ها، استخراج کردند. آن‌ها برای پیش بینی از دو طبقه بندی مختلف به نام درخت رگرسیون تجمیعی چندگانه (Wu & et al, 2008) و طبقه بندی بیشین آنتروپی (Luukka. 2011.4600-4607)استفاده کردند. تمرکز اصلی مطالعه ی آن‌ها بر روی ویژگی های محتوایی بود. بنابراین، به عنوان یک محدودیت، کار آن‌ها تنها بر روی پست های انگلیسی انجام شد. همچنین آن‌ها از ویژگی‌هایی نظیر هَشتَگ که برای پیش بینی تصمیم بازنشر پست ها مفید است، استفاده نکردند. در (Zhang& et al ,2015)، از روش تخصیص پنهان دریکله برای پیش‌بینی رفتار بازنشر

کاربران استفاده شده است. این روش در محیط‌های برخط قابلیت عملیاتی ضعیفی دارد.

در (Zhang & et al, 2014) مشابه (Bi & Cho, 2016.459-469)، یک روش مبتنی بر شبکه‌های بِیزی، برای پیش بینی تعداد بازنشرهای یک پست معلوم بر اساس الگوی پخش اولیهٔ آن، ارائه شده است. در این پژوهش، مسئلهٔ پیش بینی تصمیم بازنشر بر اساس مطالعه ی الگوی پخش پست ها انجام شده است. آن‌ها دریافتند که زمان‌های واکنش به پست ها را می توان به خوبی با استفاده از یک توزیع نُرمال لگاریتمی تخمین زد. در این پژوهش، از شبکه بیزی برای مدل‌سازی پویایی بازنشرها در طی زمان، استفاده شده است. برخلاف سایر تحقیقات که عموماً در آن‌ها مدل‌های یادگیری ویژگی‌محور هستند، این تحقیق هیچ ویژگی را به شکل مستقیم از پست‌ها یا کاربر دریافت نمی‌کند. در واقع پیش‌بینی تنها بر پایه‌ی الگوهای انتشار قبلی پست‌ها بنا شده است. آن‌ها ادعا می‌کنند که راهکارشان زمانی خوب کار می‌کند که حداقل ۱۰٪ بازنشرهای یک پست، دیده شوند. این موضوع برای ما خیلی جالب نیست چرا که اولاً روشن نیست چه زمانی ۱۰٪ پست‌ها مشاهده می‌گردد، و ثانیاً در مسئله مطرح شده در این رساله، پیش‌بینی بازنشر یک پست پیش از انتشار و یا مدت کمی پس از انتشار آن‌ها مورد نظر است.

در یک تحقیق دیگر (Yang Zhang & et.al, 2015) ، یک مدل وزن دهی شده بر اساس ویژگی ارائه شد که تصمیم بازنشر پست‌ها را بر اساس تعداد بازنشرهای بالقوه، پیش‌بینی می‌کند. برخلاف کارهای دیگر، این کار یک عمل کلاس‌بندی چندگانه است که در آن یک پست، در یکی از چهار کلاس ممکن جای می‌گیرد. این کلاس‌ها عبارتند از کلاس صفر یا پست‌های بازنشر نشده، کلاس یک یا پست‌های با کمتر از ده بار بازنشر، کلاس دو یا پست‌های با کم‌تر از ۱۰۰ بار بازنشر و کلاس سه یعنی پست‌های با بیش از ۱۰۰ بار بازنشر. مدل استخراج ویژگی آن‌ها، مجموعه‌ای از ویژگی‌ها را از خود پست و از کاربر منتشرکننده‌ی پست استخراج می‌کند. برخلاف (Artzi & et al., 2012) تمرکز آن‌ها بیشتر بر روی ویژگی‌های اجتماعی است. آن‌ها از یک جداکنندهٔ ماشین بردار پشتیبان با هسته‌ی RBF استفاده کردند که مدل آن‌ها را قادر می‌سازد مرزهای پیچیده‌ای برای تفکیک کلاس‌ها به وجود آورد. در روش آن‌ها که بر اساس وزن‌دهی بنا شده است، به هر ویژگی یک وزن اختصاص می‌دهد که بر اساس بهره اطلاعات هر ویژگی، محاسبه می‌شود. این روش به هر ویژگی که بهره اطلاعاتی بیشتری داشته باشد، وزن بالاتری اختصاص می‌دهد. این نوع وزن‌دهی به ویژگی‌ها، سبب می‌شود آن ویژگی‌ها سهم بیشتری در عمل کلاس‌بندی داشته باشند. مقادیر وزن بر اساس یک ارزیابی آزمایشی بر روی مجموعه دادگان، به دست آمده است. هرچند گزارش شده است که روش آن‌ها از روش‌های بدون استفاده از وزن‌دهی بهتر عمل می‌کند، ولیکن نویسندگان مدارک کافی ارائه نکرده‌اند تا نشان دهند این روش برای مجموعه دادگان دیگر و با تنظیمات دیگر نیز، بهینه است.

یک تحقیق مشابه برای پیش‌بینی میزان تصمیم بازنشر پست‌ها نیز توسط هونگ و همکاران (Hong & et al. 2011.57-58)انجام گرفته است. آن‌ها این عمل را بر اساس دو مسئله‌ی کلاس‌بندی دودویی و چندگانه تنظیم کردند. هدف در عمل کلاس‌بندی دودویی این است که مشخص کند یک پیام بازنشر خواهد شد یا خیر، و عمل کلاس‌بندی چندگانه سعی می‌کند حجم بازنشرهای یک پست را پس از انتشار آن پیش‌بینی کند. آن‌ها از روش TF-IDF برای پردازش ویژگی‌های محتوا استفاده کردند. ولی مشخص نکردند دقیقاً از چه رده‌بندی استفاده کرده‌اند و کدام ویژگی‌ها بیشتر در عمل کلاس‌بندی مشارکت داشته است. همچنین روش آن‌ها از نظر عمومیت نیز مشکل دارد، زیرا که بر روی مجموعه دادگان‌ی محدودی آزمایش شده است.

بیشتر کار انجام گرفته در خصوص پیش‌بینی تصمیم بازنشر یک پست، روی یک مجموعه دادگان‌ی محدود با تنظیمات محدود انجام گرفته است. بنا بر اطلاعات ما هیچ تحقیقی انجام ن شده است که سهم هر ویژگی را به‌تنهایی در پیش‌بینی تصمیم بازنشر بررسی کند. در این تحقیق ما سهم هر تک ویژگی را در پیش‌بینی تصمیم بازنشر پست‌ها مطالعه می‌کنیم. همچنین چند ویژگی جدید معرفی کرده‌ایم که در کارهای قبلی استفاده نشده‌اند. علاوه بر این یک مجموعه آزمایش جامع بر روی مجموعه دادگان مختلف اجرا کردیم تا یافته‌های خود را استحکام ببخشیم. روش ما و ویژگی‌های مطرح در پژوهش ما در فصول ۳ و ۴، شرح داده شده‌اند.

۳- روش پیشنهادی پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق DLRBP[1]

بررسی رفتار بازنشر کاربران در سطح انفرادی می‌تواند اطلاعات مهمی در مورد انتشار اطلاعات در سطح اجتماعات برخط در اختیار قرار دهد. توئیت tw_i برای کاربر u_i اثرگذار است اگر و فقط اگر منجر به یک رفتار بازنشر (RT_{tw_i}) توسط کاربر u_i گردد. اندازه‌گیری اثرگذاری یک توئیت نیازمند ایجاد مدلی برای پیشگویی رفتار فردی کاربران در مواجهه با توئیت‌های مختلف می‌باشد. امروزه کارکرد شبکه‌های اجتماعی از حالت سرگرمی فراتر رفته است به صورت کاربردهایی نظیر آموزشی، اطلاع رسانی و حل مسائل نیز فعالیت می‌کنند. کاربران در اغلب مواقع به منظور دسترسی به جدیدترین اخبار و اطلاعات در زمینه مورد علاقه خود، اقدام به پیگیری صفحات مرتبط می‌نمایند. با توجه به این کارکردها در این مقاله از روشی مبتنی بر مفهوم ارزش محتوایی هر توئیت، برای بررسی رفتار بازنشرکاربران استفاده می‌گردد. لازم به توضیح است که ارزش محتوایی یک مفهوم نسبی به شمار می‌رود و برای کاربران مختلف متفاوت می‌باشد. بسته به زمینه فعالیت و میزان اطلاعات و تخصص می‌تواند یک توئیت دارای ارزش محتوایی متفاوتی برای کاربران مختلف باشد. برای سنجش ارزش محتوایی هر توئیت با توجه به رفتار بازنشر کاربران در این مقاله روش پیشنهادی مبتنی بر محتوا ارائه می‌گردد.

تعریف ۱ : رفتار بازنشر: در صورتی که یک میکروبلاگ را به صورت یک گراف $G=(V,E)$ در نظر گرفته شود که V مجموعه کاربران و E مجموعه ارتباطات به صورت پیگیری در نظر گرفته شود. رابطه پیگیری بین دو کاربر u_i,u_j به صورت $u_i \rightarrow u_j$ نشاندهنده پیگیری کاربر j توسط کاربر i می‌باشد. رفتار بازنشر به صورت یک سه تایی (u_j,tw_i,t) نشان داده می‌شود که به معنی بازنشر توئیت tw_i متعلق به توسط کاربر u_j، در زمان t می‌باشد. هر کاربر u_i دارای دو مجموعهٔ توئیت‌های بازنشر شده PR_{u_i} و لیست توئیت‌های ایجاد شده P_{u_i} می‌باشد. رفتار بازنشر کاربر u_i در هنگام مشاهده توئیت tw_i به صورت یک مقدار باینری $y(tw_i,u_i)=\{0,1\}$ نشان داده می‌شود. $y(tw_i,u_i)=0$ نشان دهنده عدم بازنشر توئیت tw_i توسط کاربر u_i می‌باشد و $y(tw_i,u_i)=1$ نشان دهنده بازنشر توئیت tw_i توسط کاربر u_i می‌باشد.

تعریف ۲: مسئله پیشگویی رفتاربازنشر: برای هر توئیت $tw_i \in P$ (P مجموعه توئیت‌ها) و $t \in T$ (T زمان) و هر کاربر $u_i \in U$ (U مجموعه کاربران) تابعی داریم به صورت $\delta_t(tw,u_i)$ که مقدار این تابع ۱ است اگر و فقط اگر یک رفتار (رخداد) بازنشر توئیت tw_i توسط کاربر u_i در زمان $t'>t$ رخ دهد. در غیر این صورت مقدار این تابع صفر است. مسئله پیشگویی رفتار بازنشر کاربر در مواجهه با یک توئیت، پیشگویی مقدار این تابع برای $\forall tw \in P$ و $\forall u_i \in U$ می باشد.

[1] Deep Learning Based Retweet Prediction

۳-۱اندازه‌گیری ارزش محتوایی یک توئیت

برای ارزش محتوایی در منابع گوناگون تعاریف غیررسمی متعددی ارائه شده است. بر اساس زمینه کاربردی تعریف مورد استفاده برای ارزش محتوایی با توجه به کاربرد، برگرفته از ارزش خبری استفاده شده (Gupta& et al. 2014. 228-243) می‌باشد. در این مقاله ارزش محتوایی یک توئیت به میزان توجهی که یک توئیت در مخاطب خود ایجاد می‌کند، اطلاق می‌گردد. این توجه در قالب انجام رخداد بازنشر، پاسخ و یا ذکرکردن توسط کاربر نشان داده می‌شود. همان طور که گفته شد روش پیشنهادی برای اندازه‌گیری آن، مبتنی بر اندازه‌گیری مبتنی بر محتوا و مبتنی بر آمار توئیت می‌باشد، که در ادامه به بررسی آنها پرداخته می‌شود.

در ارتباط با بررسی محتوای یک توئیت از جهت سنجش ارزش محتوایی کارهای بسیاری در زمینه‌های خبررسانی صورت گرفته است. مجموعه خصوصیات‌های فراوانی برای بررسی محتوا ارائه شده است که شامل یک سری از خصوصیات مربوط به پیغام (طول، سوالی بودن یا نبودن و غیره)، کاربر نویسنده (تعداد پیگیری‌کنندگان، سال ثبت نام و غیره)، موضوع پیغام می‌باشند. این دسته از روش‌ها با بررسی لغوی کلمات و تکنیک‌های آماری به دسته بندی، رتبه‌بندی موضوعی و شباهت سنجی‌های مورد نیاز می‌پردازند. در روش پیشنهادی خصوصیات‌های جدیدی برای اندازه‌گیری برخطِ ارزش اطلاعات بر اساس محتوا بر اساس کارکرد مسئله در نظر گرفته شده است. تازگی یک توئیت یک خصوصیت مهم در ارزش محتوایی آن می‌باشد. بنا به تعریف این یک توئیت tw برای کاربر u_i تازگی دارد اگر مرتبط با علایق u_i باشد اما برای او ناشناخته باشد. برای محاسبه اطلاعات غیرمنتظره در بازیابی صفحات وب در (Wu & et al., 2011)و برای خبرهای غیرمنتظره در (Alshaabi & et al., 2020)روش‌های ذکر شده است. مجموعه $P_{u_i} = \{tw_1, tw_2, ..., tw_n\}$ توئیت‌های ایجاد شده توسط کاربر u_i می‌باشد. همان طور که در تعریف تازگی یک توئیت اعلام شد، اولا بایستی مرتبط بودن توئیت را علایق کاربر پیدا کرد. به منظور تعیین میزان ارتباط یک توئیت با موضوعات مورد علاقه کاربر در اغلب کارهای انجام شده از روش TF-IDF و کسینوس زاویه بین بردار کلمات tw_i و P_{u_i} استفاده می‌گردد. این روش با توجه به متنوع بودن کلمات بکارگرفته شده، دارای دقت پایینی می‌باشد. به همین منظور با استفاده از روش مدل سازی موضوع Latent Dirichlet Allocation، می‌توان موضوعات مختلف در متون را دسته بندی کرد. هر موضوع شامل مجموعه‌ای از کلمات $M_{topic\#} = \{w_1, w_2, ..., w_L\}$، می‌باشد که احتمال وقوع آن‌ها در موضوع مربوطه مشخص شده است. برای افزایش دقت سنجش ارتباط بین توئیت tw_i و مجموعه توئیت‌های ایجاد شده توسط کاربر u_i ($P_{u_i} = \{tw_1, tw_2, ..., tw_n\}$)، با استفاده از رابطه ۱، شباهت موضوعی اندازه‌گیری می‌گردد.

$$topic - sim(tw_i, P_{u_i}) = \frac{topic_{tw_i} . topic_{P_{u_i}}}{\| topic_{tw_i} \| . \| topic_{P_{u_i}} \|} \quad (۱)$$

با استفاده از روش مطرح LDA می‌توان موضوع مربوط به هر توئیت را در قالب یک موضوع شامل دسته ایی از کلمات با احتمال رخداد در این موضوع بدست آورد. در نتیجه با استفاده از این روش موضوع توئیت مورد نظر را پیدا کرده و دسته کلمات مرتبط با این موضوع مشخص می‌گردد. در صورتی توئیت tw_i ابتدا با یک روش استاندارد کلمات ساکن آن حذف شود و مجموعه $tw_i = \{w_1, w_2, ..., w_m\}$، مجموعه کلمات استفاده شده در توئیت tw_i، و مجموعه $M_{topic\#} = \{w_1, w_2, ..., w_L\}$، مجموعه کلمات مرتبط با

موضوع tw_i باشد، رابطه دو میزان تازگی توئیت را مشخص می‌نماید.

در رابطه فوق $M_{topic\#} - \{M_{topic\#} \cap tw_i\}$ مجموعه کلمات مرتبط با موضوع می‌باشد که در توئیت وجود ندارد. همان طور که در رابطه ۲ مشخص است، هر چه کلمات مرتبط با موضوع شباهت بیشتری با موضوعات مورد بحث توسط کاربر را داشته باشد، احتمال تازگی این توئیت کاهش پیدا می‌کند. تازگی توئیت بنا به تعریف با شباهت آن با توئیت‌های ایجاد شده توسط کاربر نسبت مستقیم دارد.

علاوه بر تازگی توئیت، ارتباط یک توئیت با کاربر می‌تواند نکته کلیدی در ارزش محتوایی این توئیت توسط کاربر باشد. در این مقاله علاوه بر محاسبه شباهت لغوی بین توئیت tw_i و توئیت‌های بکارگرفته شده توسط کاربر، شباهت جغرافیایی در نظر گرفته شده است. به طور معمول میزان اثرگذاری توئیت‌هایی که به مباحث منطقه‌ای می‌پردازند، بیش از تویئت‌های سایر مناطق است. به همین جهت در این مقاله از مجموعه داده maxmind شامل ۴ میلیون اسامی شهرها و مناطق مختلف به همراه اطلاعات تکمیلی از کشور، موقعیت و غیره برای یافتن کلمات مربوط به شهرها و مناطق جغرافیایی استفاده می‌گردد. برای یافتن کلمات مربوط به شهرها از توئیت tw_i تمام کلمات موجود در توئیت به جز کلمات اضافی مورد استفاده قرار می‌گیرد (حتی هشتگ‌ها با حذف نماد #). درصورتی که Loc_{tw_i} مجموعه شهرها به علاوه کشور و منطقه استفاده شده در توئیت باشند و Loc_{u_i} مجموعه شهرها، کشور و منطقه‌های استفاده شده توسط کاربر u_i در کل توئیتهای ایجاد شده باشند، رابطه ۳ نشان دهنده میزان ارتباط جغرافیایی بین توئیت tw_i و کاربر u_i می‌باشد.

$$\left| \left(Loc_{tw_i} \bigcap_{u_i \in U \,\&\, tw_i \in P} Loc_{u_i} \right) \right| \qquad (۳)$$

در(Kapanipathi & et all. 2011.6891-02) و (Gou & et al, 2010. 313-322) نشان داده شده است که اخبار و اطلاعات منطقی قابلیت انتشار بیشتری نسبت به رخدادهای مثبت دارند. به همین منظور به منظور تحلیل محتوایی یک توئیت برای اندازه‌گیری ارزش محتوایی، در این مقاله منفی بودن توئیت مشخص میگردد.

$$Novelty = \frac{topic - sim(tw_i, P_{u_i})}{sim(M_{topic\#} - \{M_{topic\#} \cap tw_i\}, P_{u_i})} \qquad (۲)$$

در روش بکارگرفته شده برای اندازه‌گیری مقدار منفی بودن توئیت سعی شده است از روش ساده برای اجرا برخط استفاده گردد. برای این‌کار از مجموعه داده توئیتر دانشگاه میشیگان[۲] استفاده شده است. در این مجموعه داده 1,578,627 توئیت برچسب‌گذاری شده[۳] (برچسب ۱ برای توئیت‌های مثبت و ۰ برای توئیت‌های منفی) گردآوری شده است. با استفاده از ابزار پردازش زبان طبیعی پایتون NLTK[4] و روش naïve bayes، تحلیل احساسات[۵] توئیت‌ها صورت گرفته است. برای بدست‌آوردن جداکننده از ۱/۱۰ داده‌ها به صورت تست و بقیه برای آموزش استفاده شده است. به عنوان پیش‌پردازش دو مرحله نشانه‌گذاری[۶] و نرمال‌سازی[۷] روی مجموعه داده صورت‌می‌پذیرد. مخفف‌سازی‌ها[۸] و ایکن‌های احساسی[۹] (OMG, ASAP,WTF...) در مرحله نشانه‌گذاری به صورت یک نشانه[۱۰] مجزا درنظرگرفته می‌شود. در مرحله نرمال‌سازی کلمات باحروف بزرگ به حروف کوچک تبدیل می‌شوند(برای مثال I LOVE

[۲] Sentiment Analysis competition on Kaggle
[۳] labled
[4] (NLTK (Natural Language Tool Kit
[۵] Sentiment analysis
[۶] tokenization
[۷] normalization
[۸] abbreviations
[۹] Emoticons
[۱۰] token

$$Topic\_growth = \left[\frac{1}{24}\left(\frac{\sum_{w_i \in M_{topic\#}} \#tW_{w_i}^{t} - \sum_{w_i \in M_{topic\#}} \#tw_{w_i}^{t-24}}{\sum_{w_i \in M_{topic\#}} \#tw_{w_i}^{t-24}}\right)\right] \times 100 \quad (۴)$$

it!!!!) و تکرار حروف در توئیت‌ها(برای مثال I am happyyyy!!) حذف می‌گردند.

پارامتر دیگری که مقدار ارزش محتوایی یک توئیت را تعیین می‌نماید، رشد موضوع توئیت است. موضوعات توئیت برگرفته از وقایع خارج از محیط میکروبلاگ می‌توانند بر رفتار بازنشر توئیت تاثیر مستقیمی بگذارد. در این مقاله با استفاده از روش LDA که پیشتر بدان اشاره شد، موضوعات مربوط به یک توئیت استخراج می‌گردد. درصد رشد موضوع توئیت در بازه زمانی (۲۴ ساعت) با استفاده از رابطه ۵ محاسبه می‌گردد. در رابطه ۴، $M_{topic\#}$ مجموعه کلمات مرتبط با موضوع توئیت tw_i و $\#tw_{w_i}^{t-24}$تعداد توئیت‌های ۲۴ ساعت گذشته است که حاوی کلمات $w_i \in M_{topic\#}$ به ازای تمامی توئیت‌های $\forall tw_i \in P$ میباشند.

$$weight(tw_i, t) = \left[\sum_{u_j \in U_{tw_i}^{RT}} log\left(\frac{\#f(u_j)}{\#E(u_j)}\right) + log\left(\frac{\#f(Au_{tw_i})}{\#E(Au_{tw_i})}\right)\right] \times (t - t_{tw_i})^{-\beta} \quad (۵)$$

برای تکمیل تعیین ارزش محتوایی توئیت، از خصیصهٔ مربوط به توئیت نیز استفاده می‌گردد. این خصیصه به صورت مکمل بخش محتوا به گونه ایی انتخاب شده است که بتوان به صورت برخط و کارامد از آن استفاده کرد. خصیصه‌های زیادی برای پیشگویی رفتار بازنشر در مقالات متعددی استفاده شده اند، با این حال در انتخاب این خصیصه‌ها دو جنبه قابلیت عملیاتی و مرتبط بودن به ارزش محتوایی لحاظ شده است. با توجه به آمار بازنشرهای یک پست، خصیصه‌ای معرفی می‌گردد که در این مقاله وزن یک توئیت گفته می‌شود. وزن توئیت با توجه به اعتبار نویسنده توئیت و کاربرانی که این توئیت را بازنشر کرده‌اند، نسبت مستقیم دارند. اگر $U_{tw_i}^{RT} = \{u_1, u_2, ..., u_R\}$ مجموعه کاربرانی است که تا زمان t، توئیت tw_i را بازنشر کرده اند، رابطه ۵ وزن توئیت tw_i را تعیین می‌نماید.

در رابطه 5، $\#f(u_j)$ و $\#E(u_j)$ به ترتیب تعداد کاربران پیگیری‌کننده‌های کاربر u_j و تعداد پیگیری شونده‌ها توسط کاربر u_j می‌باشد. t زمان محاسبه و Au_{tw_i} نویسنده توئیت tw_i را نشان می‌دهد. به علت وجود رابطه پیگیری متقابل، نمی توان فقط با استناد به $\#f(Au_{tw_i})$ اعتبار نویسنده یا کاربر را تعیین کرد. هر چه تعداد پیگیری کنندگان کاربر u_j نسبت به تعداد پیگیری شونده‌ها توسط کاربر u_j بیشتر باشد، اعتبار کاربر u_jبیشتر می‌گردد ($\frac{\#f(Au_{tw_i})}{\#E(Au_{tw_i})}$). یکی از جنبه‌هایی مهم و تعیین کننده برای ارزش محتوایی عمر توئیت است. با توجه به مجموعه داده‌های ارزیابی این مقاله میزان بازنشر یک توئیت با توجه به مدت زمانی که از ایجاد آن می‌گذرد بسیار شبیه به قانون قدرت[۱۱] می‌باشد. β توان زمان گذشته شده از ایجاد توئیت tw_i می‌باشد که مطابق قانون قدرت به وزن توئیت اعمال شده است. هر چه از زمان ایجاد یک توئیت بگذرد، وزن و ارزش محتوایی آن توئیت کاسته می‌شود.

[۱۱] Power law

۳-۲ یادگیری اثرگذاری با استفاده از روشی مبتنی بر یادگیری عمیق

در این مقاله با استفاده از معماری‌های مختلف مدل یادگیری عمیق برای توسعه یک سیستم پیشگو برای پیش بینی رفتار بازنشر کاربران ارائه شده است. در مرحله آموزش، یک الگوریتم رمزنگار خودکار عمیق با استفاده از مشاهدات شبکه ای نرمال برای ایجاد پارامترهای اولیه آموزش داده می‌شود. پارامترهای اولیه شامل وزن و بایاس بوده و این الگوریتم یک بازنمایی عمیق از مشاهدات نرمال را می‌آموزد. این پارامترها به عنوان یک مرحله اولیه برای آموزش یک شبکه عصبی پیش خور عمیق استاندارد (DFFNN) برای یادگیری ضرایب ویژگی‌های موجود و پیش بینی رفتارهای بعدی کاربران در مواجهه با یک پست مورد استفاده قرار می‌گیرد. در مرحله آزمایش، شبکه عصبی پیش خور عمیق استاندارد برای یادگیری ضرایب و تشکیل دسته بند مورد استفاده قرار می‌گیرد. گره‌های مختلف پنهان در این تکنیک به صورت حرفه ای بازنمایی ویژگی‌های عمیق را یاد گرفته و مهمترین ویژگی‌ها را با تبدیل ابعاد بالای داده به ابعاد پایین مبتنی بر کاهش لایه پنهان دریافت می‌نمایند. جزئیات روش پیشنهادی به تفضیل در ادامه تحقیق شرح داده می‌شود.

الف- شبکه عصبی پیش خور عمیق (DFFNN)

به صورت معمول، شبکه عصبی پیش خور عمیق به عنوان یک شبکه عصبی هوشمند در نظر گرفته می‌شود که یک لایه ورودی، بیش از یک لایه پنهان و یک لایه خروجی با اتصالات مستقیم بدون چرخه مابین آنها دارا می‌باشد.هر لایه پنهان از گره ها، ویژگی‌های انتزاعی مبتنی بر خروجی سطح قبلی را نمایش می‌دهد که به صورت خودکار در چندین لایه برای تولید خروجی تعیین و جمع آوری شده است. برای آموزش این روش، یک الگوریتم پس انتشار نزولی با گرادیان تصادفی استفاده شده است. در این الگوریتم یادگیری عمیق، داده‌های ورودی درون یک لایه ورودی قرار گرفته و سپس به لایه پنهان، پخش می‌شود، خروجی از آن که یک انتقال غیر خطی از داده‌ها است، به لایه خروجی عبور داده می‌شود. تابع ضرر یا پس انتشارِ خطا که تفاوت مابین خروجی پیش بینی شده و خروجی واقعی است، برای ارزیابی عملکرد مدل مورد استفاده قرار می‌گیرد و مقدار آن به صورت پس انتشار از طریق لایه‌های پنهان برای به روز رسانی وزن‌ها پخش می‌گردد. تابعِ ضرر یا تابعِ هزینه در واقع میزان خطا در هر بار اجرای شبکه‌ی عصبی را برای داده‌های آموزشی نمایش می‌دهد. محاسبه تابع ضرر بر اساس یک نمونه داده یا دسته ای کوچک از داده‌های آموزشی به جای کل داده‌ها مورد محاسبه قرار می‌گیرد، با وزن‌های به روز رسانی شده بعد از هر نمونه به منظور سازگار شدن با مدل مورد پردازش قرار می‌گیرد. فرایند یادگیری نظارت شده در این الگوریتم بستگی به تصادفی بودن مقدار دهی اولیه پارامترهای شبکه عصبی دارد که تمایل به قرار دادن مدل در یک راه حل حداقل محلی با تنظیمات ضعیف دارد. برای داشتن ویژگی همگرایی بهتر و بهبود نتایج یادگیری نظارت شده، تکنیک‌های پیش یادگیری بدون نظارت، بویژه شبکه‌های خود رمز نگار، می‌تواند برای ایجاد پارامترهای اولیه مورد استفاده قرار گیرد.

ب- خود رمزنگار عمیق (DAE)

DAE یک الگوریتم شبکه عصبی پیشخور برای یادگیری موثرکدینگ با استفاده از یک تکنیک بدون نظارت می‌باشد. این الگوریتم یک مجموعه داده (x) را از طریق یادگیری تقریبی یک تابع تشخیص نمایش می‌دهد، در جایی که خروجی ($\hat{x}$) با ورودی (x) یکسان می‌باشد، یعنی $x \rightarrow \hat{x}$. ساختار شماتیک آن شامل بردارهای ($x^{(i)}$) در لایه ورودی و بیش از یک لایه پنهان با یک تابع فعال ساز غیر خطی می‌باشد. لایه‌های پنهان برای یادگیری یک نمایش فشرده از داده ورودی توسط نرون‌های کمتر از لایه ورودی مورد استفاده قرار می‌گیرد. در نتیجه، مهمترین ویژگی‌ها را فرا

می‌گیرد، ابعاد را کاهش داده و چکیده ای از داده‌های ورودی را نمایش می‌دهد. در نهایت، لایه خروجی ($x^{(i)}$) به عنوان یک تقریب از نمایش داده ورودی، نمایش داده می‌شود. ساده ترین معماری یک خودرمزنگار شامل یک لایه ورودی، لایه پنهان و لایه خروجی است. فرض شده است که داده یادگیری ($x^{(i)}$، n نمونه دارد، به طوری که هر $x^{(i)}$ ($i \in (1, \ldots, n)$) ابعاد بسیاری داشته و یک بردار ویژگی یک بعدی (d^0) وجود دارد، تابع فعال ساز Tanh مورد استفاده قرار گرفته و به صورت معادله (۶) محاسبه می‌شود:

$$T(t)=\frac{1-e^{-2t}}{1+e^{-2t}} \qquad (۶)$$

الگوریتم خودرمزنگار دارای دو قسمت اصلی رمزنگار و رمزگشا می‌باشد. برای نگاشت بردار ورودی ($x^{(i)}$) به یک لایه پنهان که با ($z^{(i)}$) نمایش داده می‌شود، یک نگاشت قطعی که فرایند رمزگذاری (f_θ) نامیده می‌شود، استفاده شده است و ابعاد ($x^{(i)}$) برای فراهم کردن تعداد درستی کدها به صورت معادله (۷) کاهش یافته است:

$$f_\theta (x^{(i)}) = T (Wx^{(i)} + b) \quad (۷)$$

که W ماتریس وزنی اندازه $d^0 \times d^h$ می باشد، d^h تعداد نرون‌ها در لایه پنهان ($d^h < d^0$), b بردار بایاس، T هم تابع فعال ساز Tanh و θ هم پارامترهای نگاشت $[W, b]$ می‌باشد. برای بازسازی ورودی به صورت یک تقریب ($x^{\wedge(i)}$)، نتیجه نمایش لایه پنهان نگاشت شده و فرایند رمزگشایی با نگاشت قطعی ($g_{\theta'}$)به صورت معادله (۸) محاسبه می‌شود:

$$g_{\theta'}(x^{(i)}) = T(w'z^{(i)} + b) \quad (۸)$$

که w' ماتریس وزنی $d^h \times d^0$ ، b' بردار بایاس و θ' هم پارامترهای نگاشت $[W, b]$ می‌باشد.

ورودی در یک نمایش فشرده برای تناسب با لایه پنهان تشکیل می‌شود، سپس داده ای که به صورت ورودی استفاده شده است، برای بازسازی داده اصلی مورد استفاده قرار می‌گیرد. فرایند یادگیری، خطای باز سازی (به عنوان مثال تمایز بین داده اصلی و بازسازی با ابعاد کم آن) را به حداقل می‌رساند و برای یک نمونه یا دسته کوچک s به صورت رابطه‌های (۹) و (۱۰) محاسبه شده است :

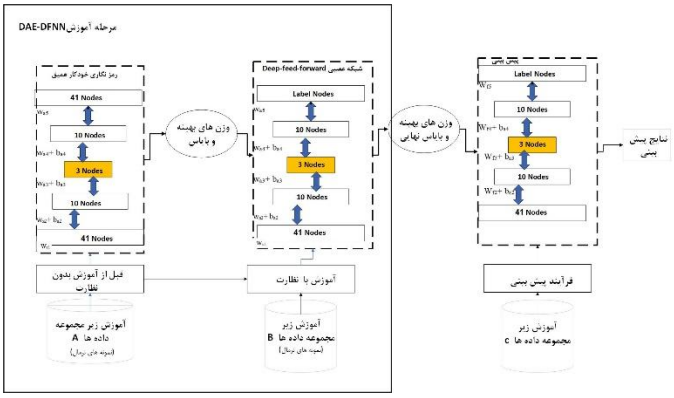
$$E(x, \hat{x}) = \tfrac{1}{2} \sum_i^s || x(i) - \hat{x}(i) ||^2 \quad (۹)$$

$$\theta=\{w,b\}=\operatorname{argmin}_\theta E(x, \hat{x}) \quad (۱۰)$$

معماری DAE شامل سه لایه پنهان، یک رمزنگار، تنگه (که شامل گره‌های کمتری از لایه‌های قبلی بوده و برای نمایش داده ورودی با کاهش ابعاد غیر خطی استفاده شده است و در آن تعداد گره‌ها تعداد ابعاد را نشان می‌دهد) و یک رمزگشا می‌باشد. این مدل به صورت بالقوه با استفاده از تکنیک آنالیز مولفه اصلی غیر خطی (PCA) برای کاهش ابعاد مورد استفاده قرار می‌گیرد. به منظور پیشگویی رفتار بازنشر کاربران در شبکه‌های اجتماعی، ما یک روش پیش بینی پیشنهاد می‌دهیم که شامل فازهای یادگیری و تست می‌باشد و در ادامه توضیح داده می‌شود.

ج- فازهای یادگیری و تست سیستم پیش بینی رفتار بازنشر مبتنی بر یادگیری عمیق

شبکه عصبی پیش خور عمیق (DFFNN) و خود رمزنگار عمیق (DAE) که در بخش‌های قبلی مورد بحث قرار گرفت، مکانیزم پایه برای ارائه پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق می‌باشند. ساختار شبکه DAE-DFFFN در شکل ۱ نمایش داده شده است.



شکل ۱- معماری پیشنهادی مدل DAE-DFFFN

در فاز یادگیری، مجموعه داده یادگیری نرمال فاقد برچسب A، مجموعه داده یادگیری دارای برچسب B، که در آن ACB، یک خود رمزنگار عمیق (DAE) با یک لایه تنگه تنها با رکوردهای نرمال (A) بدون هر گونه بردار غیر عادی برای یادگیری، آموزش داده می‌شود و مهمترین ویژگی‌ها را برای نماش الگوهای نرمال پوشش می‌دهد. این آموزش با تمام داده ها، جایی که ورودی به شبکه $(x^{(i)})$ از سه لایه پنهان عبور می‌کند، شامل یک تنگه برای بازسازی $(x^{\wedge(i)})$، انجام می‌گرد در جایی که W_e، W_n و W_f وزن‌های شبکه عصبی پیش خور عمیق (DFFNN)، خود رمزنگار عمیق (DAE) و مدل پیش بینی نهایی هستند که در شکل ۱ توصیف شده است.

در مرحله رمزگذاری، لایه ورودی در اولین لایه پنهان با استفاده از معادلات (۶) و (۷) مورد پردازش قرار می‌گیرد. در مرحله تنگه، یک تبدیل غیر خطی با ابعاد پایین از ویژگی ورودی برای استخراج ویژگی‌های موثر در تصمیم بازنشر کاربران اجرا می‌شود. سپس در مرحله رمزگشایی، آخرین لایه پنهان در ویژگی تنگه برای تقریب تکرار ورودی با استفاده از معادلات (۶) و (۸) مورد استفاده قرار می‌گیرد. و پس انتشار شیب تصادفی نزولی برای کاهش تابع ضرر یا تابع هزینه یعنی خطای میانگین مربعات مابین $(x^{(i)})$ و $(x^{\wedge(i)})$ با استفاده از معادلات (۹) و (۱۰)، مورد استفاده قرار می‌گیرد. فرایند کلیدی برای یادگیری بدون نظارت پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق پیشنهادی در شکل ۲ ارائه شده است.

شکل ۲ – فاز یادگیری بدون نظارت برای پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق

Input: training dataset (A) with n umber of samples (n) of $(x^{(i)})$, where $i \in (1, ..., n)$.
Output: parameters $\theta = \{W, b\}$
Begin
  Initialize $\{W, b\}$;
    repeat
      For each record $(x^{(i)})$, do
        compute the activation ($z^{(i)}$) in/at hidden layer and
        give output ($\hat{x}^{(i)}$) to outhe tput layer.
        compute the training error ($E(x^{(i)}, \hat{x}^{(i)})$.
        Back-propagate E and update parameters
        $\theta = \{W, b\}$;
    End
until converged
end

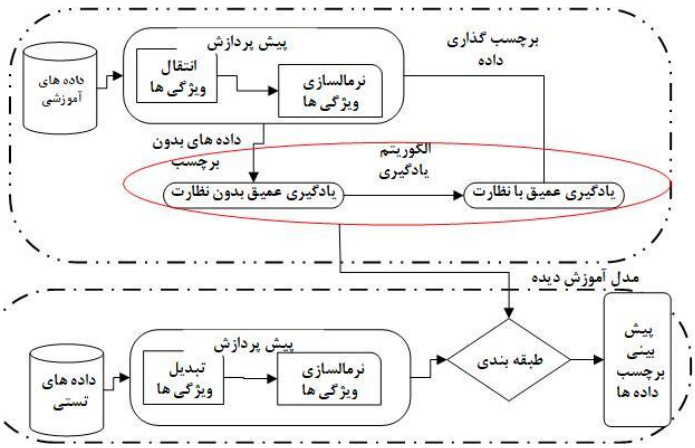
سپس مدل آموزشی برای نقطه شروع و مقدار دهی اولیه پارامترهای وزن و بایاس برای یادگیری برای شبکه عمیق یادگیری نظارتی و انجام تنظیمات مدل شبکه ای با استفاده از مجموعه داده یادگیری برچسب دار $(B((x^{(i)}), (y^{(i)})))$ مورد استفاده قرار می‌گیرد. همان مراحل که قبلا بیان شد، برای یادگیری و اعتبارسنجی روش پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق پیشنهادی روی مجموعه داده B که شامل داده‌های مربوط به رفتار بازنشر و عدم بازنشر برای تست دقت روش پیشنهادی می‌باشد، نیز دنبال می‌شود. با جزئیات بیشتر، مدل شبکه ای مبتنی بر مکانیزم پس انتشار شیب تصادفی نزولی برای به حداقل رساندن تابع ضرر با خطای میانگین مربعات محاسبه شده از تفاوت مابین مقدار خروجی هدف $(y^{(i)})$ و خروجی پیش بینی شده $g_{\theta'}(x(i))$ آموزش داده می‌شود که $g_{\theta'}$ تابع فرضی است که حاصل یک خروجی تخمینی است.

در فاز تست، پس از اینکه پارامترها به صورت خودکار در فاز یادگیری، آموزش داده شدند، نمونه مجموعه داده جدید $(C \subset_{\neq} [A, B])$ مبتنی بر مدل شبکه ساخته شده نهایی مورد آزمایش قرار می‌گیرد. هر رکورد ورودی $(x^{\wedge(i)})$ به لایه ورودی با مقدار اولیه وزنی و بایاس $\theta_f = (W_f, b_f)$ تطبیق

داده شده و سپس داده ورودی از طریق لایه‌های پنهان مورد پردازش قرار می‌گیرد. سرانجام، لایه خروجی کلاس داده ورودی را به عنوان موارد بازنشر یا عدم بازنشر مبتنی بر مقدار تخمینی تابع ضرر هر کلاس پیش بینی می‌نماید.

د- چارچوب پیشنهادی برای سیستم مورد نظر

این مقاله یک مدل پیش بینی رفتار بازنشر کارآمد را برای مطالعه رفتارهای کاربران در مواجهه با توئیت‌های مختلف پیشنهاد می‌دهد، که در شکل ۳ نشان داده شده است. همانگونه که در شکل ذیل مشاهده می‌شود، معماری شامل مراحل یادگیری و تست می‌باشد. پیش پردازش داده ها، که شامل تغییر و تبدیل ویژگی‌ها و نرمال سازی می‌باشد، اولین مرحله در مکانیزم پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق پیشنهادی می‌باشد که اطلاعات مهم را از داده‌های در مقیاس بزرگ در محیط شبکه اجتماعی مورد بررسی قرار داده و انتخاب می‌نماید.



شکل ۳- معماری پیشنهادی برای پیش بینی رفتار بازنشر کاربران مبتنی بر یادگیری عمیق

- **تبدیل ویژگی :** از آنجایی که مدل پیشنهادی فقط ویژگی‌های عددی را قبول می‌کند، هر مقدار ویژگی سمبولیک به یک مقدار عددی تغییر می‌یابد. به عنوان مثال، مجموعه داده جمع آوری شده تعدادی ویژگی سمبولیک همانند منفی بودن توئیت وجود دارد که شامل توئیت منفی، مثبت و یا خنثی می باشد که به مقادیر ۲،۱ و ۳ نگاشت می‌شوند.
- **نرمال سازی ویژگی :** ازآنجایی که یادگیری عمیق به وزن بستگی دارد، مقیاس‌های متفاوت ویژگی می‌تواند داده را به لایه‌های خاصی بایاس نماید که سبب می‌شود وزن‌های مشخصی سریع تر از سایر وزن‌ها به روز شوند. در نتیجه لازم است اینکار با نرمال سازی آماری تصحیح شود بدین صورت که امتیاز z برای هر مقدار ویژگی $(v^{(i)})$ به صورت رابطه (۱۱) محاسبه شود.

$$Z^{(i)} = \frac{v(i) - \mu}{\sigma} \qquad (۱۱)$$

که μ میانگین مقدار n برای ویژگی داده شده ((2,3,4,..) $v^{(i)}$ $(i \in 1,$) و σ انحراف استاندارد می‌باشد. از آنجایی که داده‌های شبکه شامل فضایی با ابعاد بالا می‌باشند، ضروری است که کاهش ابعاد داده برای بهبود منابع محاسباتی برای طراحی روش پیش بینی رفتار بازنشر کاربران به صورت مقیاس پذیر سبک وزن انجام گیرد. در نتیجه، مدل DAE-DFFFN بایستی از ابعاد بالا به ابعاد پایین تر با استفاده از لایه کاهش یافته مرکزی، کاهش یابد. با جزئیات بیشتر، یک تابع غیر خطی در مدل وجود دارد که تعداد زیادی از ویژگی‌ها را به مجموعه ویژگی‌های پایین تر در لایه پنهان کاهش یافته رمزگذاری می‌نماید. بنابراین کاهش ویژگی بدون نیاز به دانش بشری انجام می‌گیرد. هدف از کاهش ویژگی مدل DAE-DFFFN محدود کردن ساختار مبهم در توزیع ورودی و یافتن نمایش بخوبی طراحی شده از لحاظ یادگیری سطح بالا، فیلتر با اهمیت و ویژگی کاهش یافته می‌باشد. بنابراین یک فرایند یادگیری بدون نظارت برای مشاهدات رفتار بازنشر کاربران بکار برده می‌شود تا تخمین اولیه پارامترهای وزن و بایاس یک ورودی داده شده از DFFNN استاندارد برای کاهش زمان پردازش ساخت این مدل مشخص شود. همچنین پارامترها مجددا با استفاده از داده‌های برچسب دار (نرمال و مخرب) در روش

یادگیری عمیق نظارت شده، با مدل یادگیری نهایی ارزیابی شده مبتنی بر نمونه داده‌های جدید که در طی فاز تست بدست می‌آیند، تنظیم می‌شوند.

۴- ارزیابی

برای ارزیابی چندین سناریو در نظر گرفته شده است. سه مجموعه داده برای ارزیابی روش پیش بینی رفتار بازنشر به کار گرفته شده است که با روش‌های مختلف به یادگیری خصیصه‌های موثر در ارزش محتوایی یک توئیت پرداخته شده است.

٤-۱ مجموعه داده

برای جمع‌آوری داده‌های توئیتر چندین ابزار کاربردی متفاوت، وجود دارد. ۱) برنامه کاربردی جستجوی توئیتر، ۲) برنامه کاربردی جریان توئیتر، ۳) ابزار آتش‌نشانی توئیتر. برنامه کاربردی جریان توئیتر، امکان دسترسی به پست‌های عمومی نزدیک به زمان آزمایش را فراهم می‌آورد. برای استفاده از این برنامه کاربردی، ابتدا لازم است، ثبت نام صورت گیرد و در ادامه، مشخصات جریان مورد نظر، تعیین گردد. ضوابطی نظیر کلمه کلیدی، اسامی کاربران، مکان و یا اسم محل، می‌تواند توسط کاربر تعیین می‌شود و این برنامه‌های کاربردی بر اساس این مشخصات، اقدام به جمع‌آوری پست‌ها، در توئیتر نمایند. ایراد اصلی این ابزار برای جمع‌آوری دادگان درخواستی این است که این ابزار فقط نمونه‌هایی از فعالیت‌های رخداده در توئیتر را فراهم می‌کند. بر اساس توضیحات ارائه شده توسط توئیتر، به کمک این ابزار می‌توان یک تا چهل درصد پست‌های عمومی اخیر را جمع‌آوری کرد. ابزار آتش‌نشانی توئیتر، جمع‌آوری صد در صد پست‌ها را تضمین می‌نماید. متأسفانه این ابزار رایگان نبوده و با توجه به محدودیت‌های تحریم، امکان استفاده از آن در این مقاله میسر نشد. با توجه به این توضیحات، در این مقاله از خدمت رایگان برنامه کاربردی جریان توئیتر، استفاده شده است.

برای ارزیابی روش پیشنهادی از سه مجموعه داده از رسانه اجتماعی توئیتر استفاده شده است. برای این منظور برای انتخاب داده شرایط زیر لحاظ شده است. ۱) تمام کاربران بایستی بیش از ۳۰ در لیست پیگیره کننده و پیگیری شوندشان موجود باشد. ۲) کاربرانی که حداقل ۲۰ پست در هفته ایجاد یا بازنشر می‌نمایند. سه مجموعه داده انتخاب شده، هر کدام دارای یک ماه توئیت‌های عمومی توئیتر می‌باشند که شرایط مطرح شده را دارا باشند. این سه مجموعه داده از مجموعه توئیت‌های عمومی در سال ۲۰۲۰ انتخاب شده است. دلیل انتخاب سه مجموعه داده این است که باتوجه به ماهیت روش پیشنهادی که بیشتر بر روی محتوا تمرکز دارد. همان‌طور که در جدول ۱ مشاهده می‌نمایید این سه مجموعه دادگان، دارای اندازه‌های مختلف بوده و برای ارزیابی مطابق مفروضات مطرح در مستند پیشنهاد مقاله، مناسب می‌باشند. این سه مجموعه، موضوعات مختلف را در ارتباط با پست‌ها پوشش می‌دهند. مجموعه اول، مربوط به کاربران و پست‌های مرتبط با حوزه فناوری می‌باشد. در مجموعه دوم، داده‌های یک زمینه سیاسی و اجتماعی جمع‌آوری شده است، و در مجموعه سوم، پست‌های سرگرمی و روزمره در نظر گرفته شده‌اند.

جدول ۱- مشخصات مجموعه دادگان ارزیابی

شماره	مجموعه دادگان	تعداد پست‌ها	تعداد کاربران	تعداد بازنشرها
۱	*Steve Jobs Death*	19834	11155	158133
۲	*US Election*	253680	57267	741064
۳	*:)*[12]	13956	1162	25578

[12] این مجموعه داده مربوط به «نشانک لبخند» می‌باشد که در اغلب پست‌ها بسیار رایج است.

برای ارزیابی از معیارهای مطرح برای نشان دادن دقت پیشگویی استفاده می‌شود. این معیارها برای مسائل پیشگویی از دسته مسائل کلاس‌بندی باینری در نظر گرفته شده اند(Browne, 2000.108-132).

۱) **تشخیص**[۱۳]:این معیار در بعضی از مراجع به عنوان بازیابی منفی[۱۴]، شناخته می‌شود. اگر یک نمونه پست، در داده واقعی بازنشر ن شده است و روش پیشنهادی نیز تصمیم عدم بازنشر را درست پیش‌بینی نماید، این نمونه به صورت مثبت کاذب در نظر گرفته می‌شود. در رابطه (۱۲)، طریقه محاسبه این معیار نشان داده شده است. در این رابطه، TP و FN به ترتیب، مثبت صحیح و منفی کاذب، می‌باشند.

$$Specificity = \frac{TN}{TN + FP} \qquad (۱۲)$$

۲) **دقت**[۱۵]: این معیار نشان می‌دهد که پیش‌بینی‌های تصمیم بازنشر، تا چه حد با عملکرد کاربران در داده واقعی منطبق می‌باشد. به عبارت دیگر چند مورد از پیش‌بینی تصمیم بازنشر، در عمل نیز بازنشر شده‌اند. طریقه محاسبه این معیار، در رابطه (13) نشان داده شده است.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \qquad (۱۳)$$

۳) **F measure**: این معیار، میانگین موزون دقت و بازیابی می‌باشد. β پارامتری است که توازن را بین این دو معیار فراهم می‌نماید. همان‌طور که در رابطه (۱۶) مشاهده می‌شود، در صورتی که $\beta = 1$ باشد، F_1 برابر با میانگین موزون دقت و بازیابی می‌باشد. در صورتی که $\beta > 1$ باشد این معیار، مبتنی بر بازیابی خواهد شد. به همین صورت اگر $\beta < 1$ باشد، این معیار مبتنی بر دقت خواهد بود. در ارزیابی روش پیشنهادی از میانگین موزون دقت و بازیابی(ضریب مساوی یک)، استفاده می‌شود.

$$Precison = \frac{TP}{TP + FP} \qquad (۱۴)$$

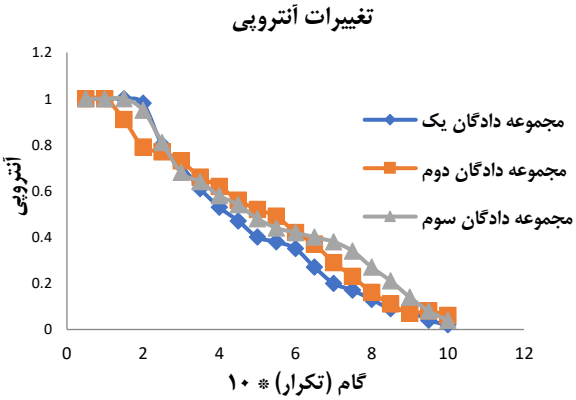
$$Recall = \frac{TP}{TP + FN} \qquad (۱۵)$$

$$Fmeasure = \frac{2 * (Precison * recall)}{Precison + recall} \qquad (۱۶)$$

نتایج ارزیابی: برای نشان دادن فرآیند یادگیری در روش پیشنهادی، از روند تغییرات آنتروپی، می‌توان استفاده کرد (Thathachar & Sastry, 2004.139-176).نمودار روند تغییرات آنتروپی محیط در هر سه مجموعه دادگان با رنگ‌های مختلف در شکل ۴ نشان داده شده است. نمودار مربوط به هر یک از مجموعه دادگان یک، دو و سه به ترتیب با رنگ‌های آبی، قرمز و خاکستری مشخص شده است. در این نمودار، محور افقی، گام‌های مختلف یادگیری با ضریب ده می‌باشد. محور عمودی این نمودار مربوط به آنتروپی محیط در این سه مجموعه دادگان است. این نمودار صد گام اول یادگیری هر سه مجموعه دادگان را نشان می‌دهد. میزان آنتروپی در حالت اولیه برابر با یک می‌باشد و با یادگیری روش مقدار آن کاسته می‌شود. برای شرط همگرایی یک حد آستانه آنتروپی تعیین می‌گردد که در این مقاله مقدار دو صدم برای آن در نظر گرفته شده است. همان طور که در شکل ۴ مشاهده می‌شود، روند تغییرات آنتروپی روش پیشنهادی در هر سه مجموعه دادگان نمایش داده شده است. در روش‌های یادگیری به کمک اتوماتای یادگیرنده، آنتروپی به عنوان ابزاری برای سنجش کارایی روش به‌کار گرفته می‌شود. هر چه محیط تصادفی‌تر باشد، مقدار آن افزایش می‌یابد. به صورت عکس در صورتی

[۱۳] Specificity
[۱۴] Recall-
[۱۵] Precision

که محیطی قادر باشد که عمل بهینه را بیابد، آنتروپی اتوماتای یادگیرنده و به تبعیت آن آنتروپی محیط کاهش می‌یابد. میزان آنتروپی در حالت بی نظمی مطلق برابر با یک می‌باشد و هر چه به سمت یادگیری پیش برویم، از میزان آن کاسته شده و به صفر نزدیک می‌گردد. این روند در شکل ۴ نیز وجود دارد. در هر سه مجموعه داده، میزان آنتروپی روش پیشنهادی در ابتدا برابر با یک می‌باشد و بعد از حدود صد گام این میزان در هر سه مجموعه دادگان به زیر «۰/۱» رسیده است.



شکل ۴–نمودار روند تغییرات آنتروپی در هر سه مجموعه دادگان ارزیابی بعد از طی گام‌های یادگیری

در ادامه بخش ارزیابی، نتایج ارزیابی روش پیشنهادی و مقایسه آن با دیگر روش‌ها ارائه می‌گردد. در این مقاله برای پیش‌بینی تصمیم بازنشر کاربران روش **DLRBP** پیشنهاد شده است. نتایج ارزیابی‌ها نشان‌دهنده کیفیت روش‌های پیشنهادی در مقایسه با دیگر روش‌ها می‌باشد.

در جدول ۲، روش‌های پیشنهادی با سه روش دیگر از نظر معیار تشخیص مقایسه شده است. در این ارزیابی، میزان پیش‌بینی درست تصمیمات عدم بازنشر، در معیار تشخیص بسیار موثر است. همان‌طور که در جدول ۲، مشاهده می‌شود، DLRBP بهترین عملکرد را داشته است. در معیار تشخیص نیز، روش لجستیک رگرسیون به صورت میانگین، بهترین عملکرد را بعد از DLRBP به خود اختصاص داده است. روش DLRBP توانسته است، موارد عدم بازنشر را بهتر از روش‌های دیگر، تشخیص دهد. در این بخش، روش درخت تصمیم، بدترین عملکرد را به خود اختصاص داده است.

جدول ۲– مقایسه معیار تشخیص روش‌های پیشنهادی با بقیه روش‌ها

معیار تشخیص	میانگین	مجموعه دادگان اول	مجموعه دادگان دوم	مجموعه دادگان سوم
DLRBP	0.834835449	0.6707317	0.969697	0.86407767
لجستیک رگرسیون	0.68581518	0.6593407	0.4953271	0.902777778
درخت تصمیم	0.54729571	0.703125	0.365591398	0.57317073
ماشین بردار پشتیبان SVM	0.65835223	0.4591837	0.7857143	0.73015873

معیار دقت، یکی از پرکاربردترین معیارها برای ارزیابی روش‌های پیش‌بینی و کلاس‌بندی است. هر چه نسبت میزان پیش‌بینی درست تصمیم بازنشر، به پیش‌بینی‌های نادرست، بیشتر باشد، دقت روش مربوطه بالاتر خواهد بود. جزئیاتی بیشتری از نتایج مقایسه روش‌های پیشنهادی با دیگر روش‌ها از نظر معیار دقت، در جدول ۳، آورده شده است.

جدول۳– مقایسه نرخ دقت روش‌های پیشنهادی با بقیه روش‌ها

معیار دقت	میانگین	مجموعه دادگان اول	مجموعه دادگان دوم	مجموعه دادگان سوم
DLRBP	0.852342394	0.7428571	0.9615385	0.852631579
لجستیک رگرسیون	0.675483757	0.6436782	0.4857143	0.897058824
درخت تصمیم	0.568715506	0.712121212	0.391752577	0.60227273
ماشین بردار پشتیبان SVM	0.687730423	0.4952381	0.7887324	0.779220779

همان طور که در جدول ۳ مشاهده می‌شود، روش DLRBP در این معیار، بهترین عملکرد را داشته است. بعد از این روش ماشین بردار پشتیان بهترین دقت را در بین روش‌های دیگر داشته است. یکی از دلایل برتری روش‌های پیشنهادی بر اساس معیار دقت در ماهیت این روش‌ها می‌باشد. در اجتماعات برخط ممکن است کاربران دچار تغییر سلایق شوند. برای نمونه کاربری که قبلا پست‌های سیاسی را بازنشر می‌کرد و علایق سیاسی داشت دچار تغییر رفتار شده و اقدام به بازنشر پست‌ها با موضوعات دیگر نماید. روش پیشنهادی در مقابل این گونه وضعیت‌ها کاملا انعطاف‌پذیر بوده و مدل جدیدی بر اساس داده‌های جدید ساخته می‌شود. این برتری روش پیشنهادی در دقت پیش‌بینی‌های انجام شده بسیار موثر است. از طرف دیگر در روش پیشنهادی، برای هر کاربر بر اساس معیارهای موثر در تصمیم بازنشر یک مدل طبقه‌بندی ساخته می‌شود، در حالی که در روش‌های دیگر به صورت دسته‌ای برای همه کاربران یک مدل واحد استفاده می‌شود. این امکان در روش پیشنهادی باعث برتری این روش بر روش‌های دیگر پیش‌بینی تصمیم بازنشر شده است.

در جدول ۴، معیار F-measure روش‌های مختلف مقایسه شده است. همان‌طور که پیش‌تر اشاره شد، معیار F-measure بکارگرفته شده در این مقاله، میانگین موزون دقت و بازیابی (حساسیت) می‌باشد (β=۱). همان‌طور که اشاره شد، میزان این معیار، ارتباط مستقیمی به دقت و بازیابی روش‌ها دارد.روش پیشنهادی، بهترین عملکرد را داشته‌اند.

جدول ۴- مقایسه معیار F- measure روش‌های پیشنهادی با بقیه روش‌ها

معیار F measure	میانگین	مجموعه دادگان اول	مجموعه دادگان دوم	مجموعه دادگان سوم
DLRBP	2.47	2.6	2.34	2.47
لجستیک رگرسیون	2.248	1.90	2.18	2.65
درخت تصمیم	0.58	0.59	0.59	0.55
SVM	1.78	1.71	1.44	2.19

نتیجه گیری:

این مقاله با مطالعه روی رفتار بازنشر کاربران، به پیشگویی رفتار بازنشر یک توئیت توسط کاربر می‌پردازد. به همین جهت یک توئیت بر اساس ارزش محتوایی آن، برای یک کاربر اثرگذار و یا غیراثرگذار میباشد. با معرفی ارزش محتوایی، روشی مبتنی بر محتوا برای اندازه‌گیری آن ارائه شده است که شامل خصیصه‌های گوناگونی است که در تعیین ارزش محتوایی موثر هستند. برای اندازه‌گیری تک تک این خصیصه‌ها با بررسی توئیت روش‌هایی ارائه شده است. راه حلی برای مسئله پیشگویی رفتار بازنشر به صورت یک مسئله کلاس‌بندی باینری با ناظر و با استفاده از روش یادگیری عمیق ارائه گردید. برای ارزیابی از روش آزمایشات با استفاده از مجموعه داده (سه مجموعه داده واقعی) در نظر گرفته شد که نتایج نشان دهنده ارتباط این خصیصه‌ها با رفتاربازنشر می‌باشد. در آزمایش‌های متعدد کارایی روش پیشنهادی مورد ارزیابی قرار گرفته است.

منابع:

Alshaabi, T., Dewhurst, D. R., Minot, J. R., Arnold, M. V., Adams, J. L., Danforth, C. M., & Dodds, P. S. (2020). The growing echo chamber of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. arXiv preprint arXiv:2003.03667.

Artzi, Y., Pantel, P., & Gamon, M. (2012, June). Predicting responses to microblog posts. In proceedings of the 2012 conference of the north American chapter of the Association for Computational Linguistics: human language technologies (pp. 602-606).

Bi, B., & Cho, J. (2016, April). Modeling a retweet network via an adaptive bayesian approach. In proceedings of the 25th International conference on world wide web (pp. 459-469).

Browne, M. W. (2000). Cross-validation methods. Journal of mathematical psychology, 44(1), 108-132.

Danah, B., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS-43. IEEE: Kauai, HI, 6.

Foust, A., Popovic, M., Zecevic, D., & McCormick, D. A. (2010). Action potentials initiate in the axon initial segment and propagate through axon collaterals reliably in cerebellar Purkinje neurons. Journal of Neuroscience, 30(20), 6891-6902.

Gou, L., Zhang, X., Chen, H. H., Kim, J. H., & Giles, C. L. (2010, June). Social network document ranking. In Proceedings of the 10th annual joint conference on Digital libraries (pp. 313-322).

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014, November). Tweetcred: Real-time credibility assessment of content on twitter. In International conference on social informatics (pp. 228-243). Springer, Cham.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., & Etter, M. (2011). Good friends, bad news-affect and virality in twitter. In Future information technology (pp. 34-43). Springer, Berlin, Heidelberg.

Hong, L., Dan, O., & Davison, B. D. (2011, March). Predicting popular messages in twitter. In Proceedings of the 20th international conference companion on World wide web (pp. 57-58).

Hwong, Y. L., Oliver, C., Van Kranendonk, M., Sammut, C., & Seroussi, Y. (2017). What makes you tick? The psychology of social media engagement in space science communication. Computers in Human Behavior, 68, 480-492.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600).

Li, M., Zhang, T., Chen, Y., & Smola,

approach. Knowledge-Based Systems, 89, 681-688.

Thathachar, M. A. L., & Sastry, P. S. (2004). Learning automata for pattern classification. In Networks of Learning Automata (pp. 139-176). Springer, Boston, MA.

Wu, S., Tan, C., Kleinberg, J., & Macy, M. W. (2011, July). Does bad news go away faster?. In Fifth International AAAI Conference on Weblogs and Social Media.

Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in twitter. In Workshop on computational social science and the wisdom of crowds, nips (Vol. 104, No. 45, pp. 17599-601). Citeseer.

Zhang, Q., Gong, Y., Guo, Y., & Huang, X. (2015, February). Retweet behavior prediction using hierarchical dirichlet process. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

A. J. (2014). Efficient mini-batch training for stochastic optimization. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2623330.2623612

Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications, 38(4), 4600-4607.

Petrovic, S., Osborne, M., & Lavrenko, V. (2011, July). Rt to win! predicting message propagation in twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In 2010 IEEE second international conference on social computing (pp. 177-184). IEEE.

Tang, X., Miao, Q., Quan, Y., Tang, J., & Deng, K. (2015). Predicting individual retweet behavior by user similarity: A multi-task learning

The Srediction of Users Sharing Behavior by Deep Learning Method via Analysis the Content of Tweet

Abstract

Recently, Social media is an important way for the future of research that is why users follow events and trends on it. When users study their interesting tweets, there is likelihood of them sharing the tweets. Research shows that only 1% users can create 50% tweets and 25% users share the tweets. According to it, sharing tweet is an important tool for diffusion of information and news. The prediction of user's sharing behavior is one of the most important challenges. The many of methods pay attention to statistical features also these methods do not focus on the content of tweets. Although there are many different content of issues , many of methods are not accurate as well as can not predict fast. This paper defines the content of feature and measure them to predict the value of tweet for users .This method is a binary classification based on preposed features to predict user's sharing behavior individually. The classification problem is solved by deep learning . The proposed method is evaluated by three datasets.The accuracy of features is 85% as well as it is effective.

Key words: Social media , User , Deep learning , Share , Content