فصلنامه آینده پژوهی مدیریت

سال سی و چهارم/ شماره۱۳۵/ زمستان ۱۴۰۲

اکتشاف کاربران پرنفوذ در بستر رسانه های اجتماعی با استفاده از تکنیک شبکه عصبی مصنوعی در پایتون به منظور بهبود عملکرد بازاریابی اینترنتی

حسین امامی راد[۱]، عباس اسدی[۲]*

۱- دانشجوی کارشناسی ارشد، گروه فناوری اطلاعات، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

۲- استادیار، گروه مدیریت بازاریابی، واحد ورامین- پیشوا، دانشگاه آزاد اسلامی، ورامین، ایران(نویسنده مسئول)

marketingabasasadi@gmail.com

تاریخ دریافت: ۱۴۰۲/۱/۹ **تاریخ پذیرش: ۱۴۰۲/۵/۳۰**

چکیده

در سال‌های اخیر با حضور رسانه‌های اجتماعی، بیشتر از قبل بحث تحلیل رفتار کاربران برحسب داده‌های آن‌ها مطرح‌شده است. در این سال‌ها، سازمان‌ها و شرکت‌های بزرگ علاقه‌مند به سرمایه‌گذاری بر روی‌داده‌ها هستند تا بتوانند مشتریان خود را بهتر درک کنند. یکی از بحث‌های بسیار مهم در حوزه علم داده، تشخیص کاربران پرنفوذ است. این کاربران می‌توانند با تأثیرگذاری بیشتر بر روی سایر کاربران خرید محصولات و سرویس‌های سازمان‌ها را افزایش دهند و هزینه‌های تبلیغات و جذب مشتری را کاهش دهند. هدف از این مسئله بهبود عملکرد تبلیغات و بازاریابی است. برای تشخیص این کاربران از مجموعه داده توییتر استفاده‌شده است و برحسب مدل کریسم ابتدا داده‌ها جمع آوری شده است، بعد داده‌ها تمییز می شود، در نهایت مدلسازی و ارزیابی می شود. در بحث مدلسازی از الگوریتم های یادگیری عمیق استفاده‌شده است و در مرحله ارزیابی این مدل با سایر مدل های یادگیری ماشین که شامل شبکه بیزین، جنگل تصادفی و ماشین بردار پشتیبان مقایسه شده است. با استفاده از روش های ارزیابی موجود در علم داده که شامل معیار حساسیت، معیار تشخیص و معیار دقت و ***F-measure*** است. مدل پیشنهادی از سایر مدل های مرسوم ذکر شده، بهتر بوده است. در نتیجه روش پیشنهادی عملکرد بهتری خواهد داشت.

کلید واژه: شبکه های اجتماعی، کاربران پر نفوذ، یادگیری عمیق، بازاریابی اینترنتی، پایتون

مقدمه

شبکه‌ی جهانی وب بر روی دنیای امروزه به قدری تأثیر داشته است که نمی‌توان دنیا را بدون ارتباطات از طریق شبکه‌های برخط تصور کرد. اجتماعات برخط به عنوان یکی از آخرین کاربردهای ابداعی محبوب، مبتنی بر این شبکه ظهور پیدا کرده‌اند. زندگی روزانه‌ی ما، به عنوان موجوداتی اجتماعی، با زندگی دیگران از قبیل اقوام، دوستان، همکاران و بسیاری دیگر از آشنایان و غریبه‌ها از طریق شبکه‌های اجتماعی در هم تنیده شده است. زندگی دیجیتالی ما در اینترنت از شبکه‌ها و ساختارهای اجتماعی متفاوتی تشکیل شده است. همان طور *(Enrique & Ruiz, 2018)* که بیان نموده است: شبکه‌های کامپیوتری ذاتا شبکه‌های اجتماعی هستند که افراد، سازمان‌ها و دانش آن‌ها را به هم مرتبط می‌سازد. از این دیدگاه می‌توان این استنباط را کرد که اینترنت فقط وسیله‌ای برای پشتیبانی از ارتباطات اجتماعی اعضا نیست، بلکه اینترنت به اعضا امکان ایجاد ارتباطات کاملا مجازی را از طریق عضویت در گروه‌ها و اجتماعات برخط می‌دهد. این امکانات از طریق اطلاعات فراوان رایگان و آماده‌ی استفاده، ارتباطات و فنّاوری‌های ایجاد اجتماعات فراهم شده است. از جمله‌ی این فنّاوری‌ها می‌توان به موارد زیر اشاره نمود:

- فنّاوری‌های ارتباطی مانند نامه‌های الکترونیکی، انجمن‌های برخط تحت وب، صفحات گفتگو، پیام‌های فوری و رسانه‌های اجتماعی
- فنّاوری تبادل و توزیع اطلاعات از صفحات وب، ویکی‌ها، وبلاگ‌ها و ویدئو بلاگ‌ها.
- فنّاوری شبکه‌های اجتماعی سرگرمی و یا با اهداف مشخص مانند **Facebook**، **Linkedin** و **Twitter**.
- درس افزارهای برخط مانند **moodle**، **blackboard** و **sakai**.
- محیط‌های مجازی مانند **Secondlife** و **World Of Warcaft**.

هر پاسخی به یک نامه الکترونیکی، گذاشتن پیوندی به یک صفحه وب، نوشتن بلاگ یا نوشتن نظرات برای ویدوئی در سایت یوتیوب، ردپای دیجیتالی از هر فرد ثبت می‌کند که وی را به صورت ضمنی یا صریح به فرد برخط دیگر مرتبط می‌کند. هر یک از این فعالیت‌های ثبت شده، شبکه‌ای از توجه پیرامون موضوع مورد علاقه، وابستگی‌های مشترک، اجتماعات انسانی و یا اقدامی جمعی را ایجاد می‌نماید.

رشد روزانه‌ی مشارکت‌ها و روابط متقابل برخط، نشان‌دهنده‌ی رشد تعداد شرکت‌کنندگان (اعضا) و افزایش حجم اسناد در این شبکه‌ها است *(Cartwright & F, 1956)* این حجم وسیع از داده‌های برخط اخذ شده و ذخیره شده، فرصتی استثنایی برای دانشمندان علوم اجتماعی و محققان حوزه اینترنت برای مطالعه‌ی نحوه‌ی کار اجتماعات انسانی فراهم آورده است. اکنون محققان می‌توانند از طریق این تعاملات ثبت شده، موضوعاتی مانند علایق و الویت گروه‌ها، چگونگی و چرایی پیدایش یا انقراض اجتماعات برخط و نحوه به اشتراک‌گذاری دانش و اطلاعات میان اعضای گروه‌ها را به دقت بررسی کنند. پاسخ‌گویی به این سوالات و دیگر سوالات مشابه، نه تنها امکان درک نحوه‌ی ارتباطات اجتماعی میان افراد را می‌دهد، بلکه کمک می‌کند که از فنّاوری‌های جدید و شبکه‌های اجتماعی برای فراهم کردن خدماتی مانند حل مسئله با کمک اجتماعات برخط، استفاده شود. نکته بسیار ضروری در این بخش انتشار اطلاعات در سراسر اجتماع با توجه به نیازمندی‌های هر کاربر می‌باشد.

بیان مسئله

اخیرا رسانه های اجتماعی بر زندگی انسان ها در زمینه های مختلف تاثیری بسزا داشته است . به دلیل سرعت انتشار مطالب در این رسانه ها ، کاربران می‌توانند به راحتی در جریان آخرین رویدادها و اخبار باشند. در نتیجه این رسانه ها در دنیای امروزی یک ابزار بسیار مناسب در حوزه بازاریابی دیجیتالی شده اند *(Barbasi & R, 1999)* یکی از چالش های عمده بازاریابی دیجیتالی افزایش اثربخشی و گسترش تبلیغات در سطح شبکه می‌باشد. در این پروژه به دنبال یافتن روشی برای تعیین میزان تاثیرهر محتوا روی کاربران رسانه های اجتماعی به منظور بیشینه کردن انتشار آن می باشیم. برای این کار بایستی کاربران پرنفوذ برحسب محتوای تبلیغ شناسایی شود. منظور از کاربران پرنفوذ، کاربرانی است که در صورت انتشار تبلیغ توسط آنها، گستره تبلیغ در رسانه اجتماعی بیشینه خواهد شد.

این که هر گره (کاربر) در یک شبکه اجتماعی به چه میزان از خود، رفتار تاثیرگذار روی سایر گره‌ها نشان می‌دهد، رهبری گره را تعیین می‌نماید. به طور خاص یک کاربر u رهبر است اگر یک فعالیت a باعث می‌شود تعدادی معینی (حدآستانه مشخص) از کاربران همان فعالیت a را بعد از کاربر u انجام دهد. البته ذکر این نکته مهم است که با انجام یک فعالیت و پیروی سایر گره‌ها نمی‌توان در مورد رهبری یک گره نظر قطعی داد. بایستی این رفتار برای فعالیت‌های معینی (حد آستانه مشخص) تکرار شود.

به عبارت دیگر این مسئله، به دنبال این می‌باشد کاربرانی انتخاب نماید در صورت انتشار تبلیغ، مطلوب کاربران پیگیری کننده آن کاربران نیز باشد که این امر باعث افزایش انتشار پست به صورت بازگشتی می‌شود. در واقع هدف مسئله انتخاب k موثرترین کاربر، یافتن کاربرانی است که در صورت انتشار تبلیغ، k بیشینه ترین سود را می‌توانند دریافت کنند.

روش تعیین کاربران تأثیر گذار علاوه بر بررسی ساختار شبکه باید به محتوای داده‌های منتشر شده نیز توجه کرد. یکی از احتمال های موجود در انتشار محتوای تبلیغاتی ، موضوعاتی که توسط کاربران منتشر می شود. عملکرد کاربران پرنفوذ در موضوعات مختلف می تواند متفاوت باشد. برای مثال برخی کاربران پرنفوذ می‌توانند در موضوعات تکنولوژی بسیار تأثیر گذار باشند و برخی دیگر در موضوعات فرهنگی و هنری اثر گذار باشند.

پیشینه پژوهش

یک شبکه ی اجتماعی، ساختاری اجتماعی است که از مجموعه‌ای از عاملها (مانند افراد یا سازمان‌ها) و مجموعه ی پیچیده‌ای از روابط مابین این عامل‌ها، تشکیل شده است. امکانات ارتباطی وسیع و پردامنه ای که اینترنت و شبکه های ارتباطی در دنیای مجازی در اختیار افراد قرار می دهند، فراتر از مرزهای مرسوم جغرافیایی می باشد. شبکه های اجتماعی بی تردید تأثیر مستقیم خود را بر شکل گیری روابط انسانی، تسهیل، تقویت و همچنین در مواردی تضعیف برخی اشکال مالوف سنتی تر باقی گذاشته اند *(Al-Garadi, 2018)* . شبکه ی اجتماعی، ساختاری اجتماعی است که از گره هایی(که عموماً فردی یا سازمانی هستند) تشکیل شده‌است که توسط یک یا چند نوع خاص از وابستگی به هم متصل اند، برای مثال قیمت‌ها، الهامات۱، ایده‌ها و تبادلات مالی، دوست‌ها، خویشاوندی، تجارت، لینک‌های وب، سرایت بیماری‌ها ۲یا مسیرهای هواپیمایی. ساختارهای حاصل اغلب بسیار پیچیده هستند. تحلیل شبکه‌های اجتماعی روابط اجتماعی را با اصطلاحات رأس۳ و یال۴ می‌نگرد *(Eliacik, et al., 2018)* . رأس‌ها بازیگران فردی درون شبکه‌ها هستند و یال‌ها

[1] ***inspiration***
[2] ***epidemic***
[3] ***node***
[4] ***edge***

روابط میان این بازیگران می با انواعنواع زیادی از یال‌ها می‌توانند، میان رأس‌ها وجود داشته باشند. نتایج تحقیقات مختلف بیانگر آن است که می توان از ظرفیت شبکه‌های اجتماعی در بسیاری از سطوح فردی و اجتماعی به منظور شناسایی مسائل و تعیین راه حل آنها، برقراری روابط اجتماعی، اداره امور تشکیلاتی، سیاستگذاری و رهنمون سازی افراد در مسیر دستیابی به اهداف استفاده نمود. به عنوان مثال، نتایج مطالعات در حوزه سیاستگذاری گردشگری نشان می دهد شبکه‌های اجتماعی ۵به واسطه تاثیرگذاری روی متغیرهای رفتاری بر جذب گردشگران خارجی به مقاصد گوناگون تاثیرگذار هستند و می توان از این شبکه ها به منظور شکل‌گیری اعتماد و کاهش ریسک تصمیم‌گیری کاربران در انتخاب یک مقصد خاص گردشگری بهره گرفت *(Probst, et al., 2013)* .

شبکه های اجتماعی بر نوع سرویس دهی شرکت ها تأثیر می گذارد وادغام سه مفهوم روانشناسی اجتماعی ، جامعه شناسی،تئوری گراف می باشد *(Ghosh & Lerman, 2010)*. امکانات ارتباطی وسیع و پردامنه ای که اینترنت و شبکه های ارتباطی در دنیای مجازی ۶در اختیار افراد فراتر از مرزهای مرسوم جغرافیایی فراهم می آورند، بی تردید تأثیر مستقیم خود را بر شکل گیری روابط انسانی، تسهیل ، تقویت۷ و همچنین در مواردی تضعیف برخی اشکال مالوف سنتی تر باقی گذاشته است در عرصه‌ی اشکال رابطه های صمیمانه در صحنه ی روابط اجتماعی نیز شاهد جایگزینی بسیاری از ارتباطات و تعاملات چهره به چهره ۸و مستقیم با انواعی از دیدارها و تبادل پیام های دوستی و معاشرت در دنیای مجازی هستیم *(Liu, 2014)*.

بنابراین شبکه های اجتماعی دارای مزایا و معایبی هستند. بر طبق *(Silva, 2013)* در خصوص اینکه "چه تعداد زیادی از افراد پس از سر زدن به فیسبوک۹ احساس حسودی و در پی آن احساس تنهایی، سرخوردگی و عصبانیت می‌کنند اظهار شگفتی کرده است. در دیدگاه اخیر قدرت هویت سازی حتی در فضای مجازی برای افراد و کاربران وجود دارد. *(Erlandsson and Fredrik, 2016)*

به طورکلی بحث انتشار در شبکه‌های اجتماعی، مطالعهٔ تأثیر خواص گراف شبکه بر رفتار کاربران است. انتشار، یک رفتار مهم در این حوزه به شمار می‌رود و مدل‌های فراوانی در حوزه‌های مختلف برای آن پیشنهاد شده است، مدل بازاریابی و فروش یک کالا *(Yang et al, 2019)* پخش یک خبر در شبکه‌ی وبلاگ‌ها (چن و همکاران ۱۳۹۸ ، ۲۲:۳۲)، انتقال یک بیماری مسری یا ویروس کامپیوتری *(Ishfaq, 2022)* ، انتشار سیاست‌های دولت در خصوص ممنوعیت کشیدن سیگار یا لاتاری ایالتی *(Featherstone et al, 2020)* انتشار گرایش‌های هالیوودی و حتی سرایت بیماری وسواس و غیره *(Featherstone et al, 2020)* از جمله این موارد می‌باشند.

مدل‌سازی در اجتماعات برخط باید به‌وضوح تعاملات بین افراد، تأثیراتی که از درون به گروه‌ها اعمال می‌شود و تأثیراتی که از محیط بیرون اعمال می‌گردد، را نشان دهد *(Lingam et al, 2019)* تأثیرات رسانه‌های سنتی مثل

---

۵ *Social network*

۶ *Virtual world*

۷ *strength*

۸ *Face to face*

۹ *Face book*

تلویزیون یا حتی رسانه‌های جدیدتر مثل شبکه‌های اجتماعی را می‌توان در این گروه‌ها بررسی و تحلیل نمود.

مطالعات تجربی بر روی شبکه‌های اجتماعی از اواخر قرن بیستم شروع شد ولی مدل ریاضی برای این پدیده چندین دهه بعد از شروع مطالعات شبکه‌های اجتماعی در کارهای*(Fan and et al, 2021)* ارائه شده است.

افراد اثر گذار به تحقیقاتی در حوزه رهبران مرتبط می شود. افراد پر نفوذ به کاربرانی گفته می شود که می تواند نقش حیاتی و مهم را در انتشار اطلاعات ایفا کنند. این کاربران بر سایر کاربران نفوذ دارند. تحقیقات صورت گرفته در این حوزه شامل ، ساختار شبکه ، اطلاعات مشترک و ویژگی کاربران می باشد. در اولین روش از معیارها و استانداردهایی برای اندازه گیره قدرت گره های شبکه استفاده کردند.***page Rank*** و سایر الگوریتم ها در این مرحله مورد ارزیابی قرار گرفت. این روش آسان بود اما دقت کافی نداشت*(Eirinaki et al, 2012)* . در دومین روش تمرکز بر روی اطلاعات مشترکی که بین کاربران رد و بدل می شود قرار گرفت. این روش از روش اول بسیار مناسب تر و دقت بیشتری داشت. اما بر روی شبکه های بزرگ قابلیت پیاده سازی نداشت. سومین روش بر پایه رفتار کاربران ، فعالیت ها و سایر فاکتورها قرار گرفت. اگر چه این روش ها ذهنی تر می باشد اما غیر قابل استفاده برای پیدا کردن کاربران اثر گذار است.

در *(Fiorllio and Diana, 2020)* معیار ساده‌ای برای مؤثر بودن کاربران ارائه شده است، این معیارمبتنی بر درجه ورودی (تعداد دوستان) است، اهمیت کاربران با افزایش میزان درجه آن‌ها تشخیص داده می‌شود. در *(Amati and Giambattista, 2019)* اهمیت و شهرت همسایه‌های یک کاربر را مشخص کرده است و معیار بازخورد نامیده شده است، هرچه همسایگان یک کاربر مهم‌تر باشند، آن کاربر به مراتب از اهمیت بیشتری برخوردار است. در*(Kostkova et al, 2017)*تأثیر کاربران بر یکدیگر را برحسب تعداد دوستا ری کننده‌ها و محتوای تولید شده در شبکه ۱۶۹ یت ر بررسی کرد. در *(Kostkova et al, 2017)* تأثیر کاربران بر یکدیگر را برحسب سه معیار: درجه ورودی (تعداد دنبال کننده‌ها)، دوباره بازنشر، ذکر شدن (چند بار اسم یک کاربر ذکر می‌شود) در شبکه اجتماعی توییت ر بررسی کرد. در *(Eskandanian et al, 2019)* روشی را برای تعیین کاربران تأثیرگذار برحسب فعالیت کاربران در شبکه اجتماعی فیس بوک پیشنهاد دادند. از سه معیار مرکزیت داده، تعداد گروه‌هایی که به آن‌ها متعلق هستند و به روز رسانی را مورد توجه قرار دادند. در*(Said et al 2018)* معیارهای مرکزیت (هاب، درجه ورودی) در شبکه مشتریان یک شرکت مخابراتی در نظر گرفت و تأثیر کاربران بر یکدیگر را مورد سنجش قرار داده است. در *(Said et al 2018)* سه ویژگی از وبلاگ‌ها که شامل ساختار شبکه، محتوای تولید شده و فعالیت‌های کاربران را مورد تحلیل قراردادند و با استفاده از یک شبکه عصبی مصنوعی یک روشی را برای تأثیرگذاری وبلاگ نویسان ارائه دادند و وبلاگ نویس های تأثیرگذار را روی وبلاگ ریتچ اکتشاف کردند. در *(Mahmoudi et al, 2018)* مدلی را به نام ***BARA*** پیشنهاد دادند تا رهبران را روی شبکه وبلاگ ***myspace*** اکتشاف کنند. آن‌ها پارامترهایی چون محتوا، ارتباط بین کاربران، ویژگی‌های خواننده و ویژگی‌های نویسنده را مدنظر قرار دادند. در *(Alwan et al, 2020)* مدلی را به نام ***cssm*** ارائه دادند این مدل کاربران تأثیرگذار را برحسب سه پارامتر تأثیرگذار فعالیت گره، تعامل گره و موقعیت اجتماعی گره بر روی شبکه‌های اجتماعی ***Twitter, Sina Weibo*** تشخیص داد*(Azcorra et al, 2018)*

مرکزیت درجه*(Jain and Sinha, 2023)* ساده ترین متریک برای تعیین گره های تأثیرگذار است. با این وجود ، غالباً اتفاق می افتد که گره هایی با درجات کوچکتر ، تأثیر بسیار بیشتری نسبت به آنهایی که دارای درجه های

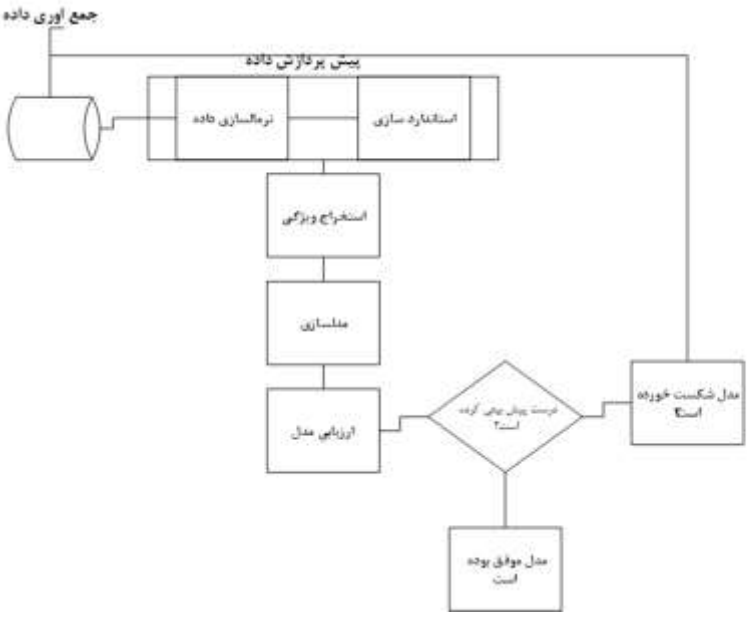
بالاتر هستند ، داشته باشند. در واقع ، گره کم درجه ممکن است در یک مکان استراتژیک شبکه قرار داشته باشد به گونه ای که انتشار اطلاعات بیشتر از برخی گره ها با درجه بالاتر را تسهیل می کند *(Shazad et al, 2020)* که اندازه گیری مبتنی بر نزدیکی یک گره به هسته گراف را معرفی کرد. آنها از روش تجزیه ***k-shell*** استفاده کردند ، جایی که گره های نزدیک به هسته به عنوان گره های تأثیرگذار شناخته می شوند .*(Deng, 20221)* گره هایی را به عنوان گره های پرنفوذ معرفی کردند که همسایگان آن ها به هسته گراف نزدیکتر بودند. نویسندگان *(Deng, 20221)* نشان دادند که اهمیت پیوندها بین گره و همسایگان آن نیز در تأثیرپذیری گره نقش مهمی دارد. آنها یک الگوریتم رتبه بندی را بر اساس درجه ، ***k-shell*** و اهمیت لینک پیشنهاد دادند.

با توجه به درجه همسایگان هر گره و بر اساس اندازه گیری ***h-index*** ، یک متریک مرکزیت در *(Kao and yo-ping, 2015)* پیشنهاد شده است ، که بسیار مناسب تر از درجه تعیین شده برای گره ها ی پرنفوذ است. علاوه بر این ، ***h-index***مجموعه پیرامون هر گره را با ترکیب ***h-indices*** هر گره دقت بیشتری را در سنترالیتی ایجاد می کند*(Zhang, sun and liu, 2011)*. در *(Gomasta et al, 2022)* متریک ***h-index*** گسترش یافت . در*(Mohammadi and Saraee, 2018)* نشان داده شد، که گره هایی با درجه های پایین که به گره های درجه بالا متصل هستند می‌توانند کاندیداهای خوبی برای گره های تأثیرگذار باشند. در روش مبتنی بر مرکزیت آنتروپی *(Mohammadi and Saraee, 2018)* ، محدوده نفوذ یک گره بر اساس توزیع درجه همسایگان مرتبه اول و دوم آن تعیین می شود. رویکرد درخت پوشای تصادفی در *(Wu and Wang, 2014)* برای تعیین گره های تأثیرگذار بر اساس الگوهای اتصال محلی و جهانی شبکه پیشنهاد شده است. با توجه به فاصله بین گره ها ، روشی در *(Wu and Wang, 2014)* برای رتبه بندی گره های تأثیرگذار بر اساس بعد محلی فازی ارائه شده است. در *(Al-Otaibi et al, 2022)*، تأثیر حذف گره در متوسط کوتاهترین مسیرهای شبکه برای پیشنهاد یک اندازه گیری مرکزیت ارزیابی شده است. در *(Wang et al, 2023)* ، با استفاده از تئوری مشهود بر اساس اطلاعات توپولوژیکی گره ها از قبیل درجه ، فاصله و همبستگی تخمین زده شد. روش دیگری در *(Novak, Anucis and Mozetic, 2018)* برای شناسایی تأثیرپذیری گره ها براساس درجه ، فاصله و نزدیکی آنها معرفی شد. برای تعیین تأثیرپذیری بر اساس اقدامات مرکزیت گره ، روش وزن گیری آنتروپی در *(Huang et al, 2019)* ارائه شده است.

روش پژوهش

در این بخش بر طبق روش کریسم که در شکل زیر نشان داده شده است به ارائه روش پیشنهادی پرداخته خواهد شد. فازهای اجرایی این روش به ترتیب ، جمع آوری کردن داده ، تمیز سازی داده که این مرحله شامل چندین زیر بخش است ، مدلسازی داده که بر طبق روش یادگیری عمیق انجام خواهد شد و در نهایت به ارزیابی روش پیشنهادی که آن را ***DIUS*** [۱۰] نام گذاری شده است با فرمول های رایج در یادگیری ماشین پرداخته خواهد شد.

[۱۰] ***Discovering influential users on the social media***



شکل ۱: مدل اجرای تحقیق

۱. جمع آوری داده‌ها: در این بخش به جمع آوری داده پرداخته خواهد شد. این مرحله در سازمان و پروژه های مختلف متفاوت است . در برخی سازمان داده‌ها به صورت لحظه ایی بر روی سرور نشسته و در پایگاه های داده جمع آوری می شود. در این پژوهش این مرحله به صورت یک فایل ***json*** بوده است که شرکت توییتر در اختیار پژوهشگران حوزه داده قرار داده است.

۲. آماده سازی داده: یکی از مراحل بسیار مهم در حوزه تحلیل داده، آماده سازی داده می باشد. در این بخش به تمیز کردن، از بین بردن داده نویزی، اکتشاف و جایگزینی داده از دست رفته و حذف داده‌های پرت پرداخته می شود. با پیش پردازش داده می توان مدلسازی درست تر و نتایج دقیق تری را در مراحل بعد بدست آورد.

در این بخش یکسری از ویژگی ها توسط کاربر وارد نشده است . سناریو استفاده‌شده در این پژوهش***KNN*** است)

(Guo et al, 2003). که در ادامه به صورت مبسوط توضیح داده خواهد شد.

مراحل ***KNN*** به شرح زیر است:

- داده‌ها را بارگذاری می‌کنیم.
- مقدار ***K*** را تعیین می‌کنیم که همان تعداد نزدیک‌ترین همسایه‌ها هستند.
- برای هر نمونه داده فاصله‌ی میان نمونه داده‌ی جدید را با نمونه داده‌های موجود محاسبه می‌کنیم.
- فاصله و شاخص هر نمونه را به یک فهرست وارد می‌کنیم.
- کل لیست را براساس فاصله‌ی نمونه داده‌ها، از کمترین به بیشترین فاصله، مرتب می‌کنیم.
- ***K***تا از اولین نمونه‌های فهرست مرتب‌شده را به‌عنوان ***K*** نزدیک‌ترین همسایه انتخاب می‌کنیم.

۸

- برچسب این K نمونه را بررسی می‌کنیم.
- اگر مسئله رگرسیون باشد، میانگین برچسب‌های این K نمونه داده برچسب نمونه داده جدید خواهد بود.
- درصورتی‌که مسئله طبقه‌بندی باشد، نمونه‌ی جدید هم همان برچسب K همسایه را خواهد داشت.
- تشخیص و حذف داده‌های تکراری و اضافه
- مدیریت و تصمیم گیری درباره داده‌های خارج از محدوده
- شناسایی داده‌های دارای اختلال
- تبدیل داده‌های پیوسته به گسسته
- تصمیم گیری درباره داده‌های ناقص و ناسازگار
- تحلیل همبستگی داده‌ها
- ایجاد فیلد محاسباتی جدید
- فشرده سازی داده‌ها
- نرمال سازی داده‌ها

۳. در این بخش به استخراج ویژگی های مورد نظر برای طراحی مدل، پرداخته می شود. ویژگی ها در اصل مولفه های اصلی برای طراحی مدل است. این مرحله از جهت انتخاب ویژگی های موثر در انتخاب کاربران پرنفوذ بسیار مهم هست. روش های مختلفی برای مهندسی ویژگی ها وجود دارد یکی از روش های مرسوم برای تشخیص ویژگی ها استفاده از روش ***PCA*** است.

مهندسی ویژگی، استخراج و انتخاب ویژگی

انتخاب ویژگی های مناسب باعث بهبود عملکرد یادگیری می شود. این ویژگی ها شامل موارد است:

- ارزش محتوایی یک توئیت

در قالب انجام رخداد بازنشر، پاسخ و یا ذکرکردن توسط کاربر نشان داده می‌شود

- موضوع یک توییت از روش ***TF-IDF*** استفاده‌شده است *(Aizawa, 2003)*.

بنا به تعریف ، یک توئیت tw_i برای کاربر U_i تازگی دارد اگر مرتبط با علایق U_i باشد اما برای او ناشناخته باشد*(Jain and Sinha, 2022)*

$$W_{ki} = TFIDF(t_k, d_i) = tf(t_k, d_i) * idf(t_k, d_i)$$

- احساسات یک توییت

–ابزار پردازش زبان طبیعی پایتون ***NLTK***

–روش ***Naïve Bayes***

- وزن توییت

از ویژگی های مورد نظر ، تعداد دوستان ، تعداد لایک ، تعداد به اشتراک گذاری است که در مدل پیشنهادی این موارد با تابع ***COUNT*** در پایتون محاسبه شده است.

٤. مدل سازی

برای مدلسازی ، می توان از الگوریتم های یادگیری ماشین استفاده کرد. این الگوریتم ها به سه دسته با نظارت ، بدون نظارت و نیمه نظارت تقسیم شده است.

۹

در این پژوهش ، برای مدلسازی از یادگیری عمیق استفاده می شود. یادگیری عمیق زیرمجموعه یادگیری ماشین است و به سه دسته یادگیری با نظارت ، بدون نظارت و نیمه نظارت تقسیم می شود. الگوریتم استفاده‌شده در این پژوهش یادگیری پیش خور عمیق

است. که در ادامه به صورت مبسوط توضیح داده شده است.

*شبکه عصبی پیش خور عمیق (**DFFNN**)*[11]

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. شکل ۳-۲ شبکه عصبی پیش خور عمیق را نشان می دهد.

پس انتشار شبیه به محاسبه قانون دلتا برای یک شبکه پیشخور چندلایه است. بنابراین، مانند قاعده دلتا، پس انتشار به سه چیز نیاز دارد:

- **مجموعه داده**
- **پارامترهای یک شبکه عصبی پیشخور**
- **تابع خطا**

آموزش یک شبکه عصبی با گرادیان کاهشی نیازمند محاسبه گرادیان خطای $E(X,\theta)$ نسبت به وزن‌های W_{ij}^{k} و بایاس‌های b_i^k است. بر اساس نرخ یادگیری α، هر تکرار گرادیان کاهشی وزن‌ها و بایاس‌ها (که به صورت تجمعی با θ مشخص شده‌اند) را بر اساس رابطه زیر به‌روز می‌کند:

$$\Theta^{t+1} = \Theta^{t} - \alpha \frac{(\delta E(X, \Theta^{t}))}{\delta \Theta}$$

که در آن، Θ^t پارامترهای شبکه عصبی را برای تکرار t در گرادیان کاهشی مشخص می‌کند.

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.

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.

۵. ارزیابی مدل:

در بخش ارزیابی مدل از روش های مرسوم تست کردن مدل در علم داده استفاده می شود. یکی از روش های ۱۰ رسوم در این علم استفاده از روش های ماتریس ***Confusion*** است.

، هدف از این پژوهش انتخاب k پست اثر گذار جهت بیشینه شدن انتشار است. این موضوع بسیار مهم است که پیش بینی درستی در انتخاب k پست اثر گذار ارائه گردد و همینطور دقت این پیش بینی اندازه گیری شود.

[11] deep feedforward neural network

روش هـای ارزیـابی ، نقش بســیـار مهمی را بـه ویژه هنگامی که با یک مجموعه داده نامتعادل روبه رو هستیم ایفا می کند.

برای ارزیابی از معیارهای مطرح برای نشان دادن دقت پیشگویی استفاده می‌شود. این معیارها برای مسائل پیشگویی از دسته مسائل کلاس‌بندی دو–دویی در نظر گرفته شده اند.

- مثبت صحیح[۱۲] (**TP**): یک پست بازنشر شده، به درستی به عنوان پست بازنشر شده، تشخیص داده می‌شود.
- مثبت کاذب[۱۳] (**FP**): یک پست بازنشر نشده، به اشتباه به عنوان پست بازنشر شده، تشخیص داده می‌شود.
- منفی صحیح[۱۴] (**TN**): پست بازنشر نشده، به درستی به عنوان پست بازنشر نشده، تشخیص داده می‌شود.
- منفی کاذب[۱۵] (**FN**): پست بازنشر شده، به اشتباه به عنوان پست بازنشر نشده، تشخیص داده می‌شود.

۱–میزان حساسیت[۱۶](نرخ مثبت صحیح و بازیابی مثبت)

معیار میزان حساسیت، بعضاً با عنوان نرخ مثبت صحیح[۱۷] و یا بازیابی[۱۸]، شناخته می‌شود. این یک نمونه مثبت صحیح به شمار می‌آید. در این رابطه، TP و FN به ترتیب، مثبت صحیح و منفی کاذب، می‌باشند.

$$\boldsymbol{Sensitivity = \frac{TP}{TP + FN} = TPR = +Recall}$$

۲–تشخیص[۱۹](نرخ منفی کاذب ،بازیابی منفی)

این معیار در بعضی از مراجع به عنوان بازیابی منفی[۲۰]، شناخته می‌شود. اگر یک نمونه پست، در داده واقعی بازنشر نشده است و روش پیشنهادی نیز تصمیم عدم بازنشر را درست پیش‌بینی نماید، این نمونه به صورت مثبت کاذب در نظر گرفته می‌شود. در این رابطه، TN و FP به ترتیب، منفی صحیح و مثبت کاذب، می‌باشند.

$$\textbf{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}} = \boldsymbol{TNR = -Recall}$$

۳–دقت[۲۱]

این معیار نشان می‌دهد که پیش‌بینی‌های تصمیم بازنشر، تا چه حد با عملکرد کاربران در داده واقعی منطبق می‌باشد. ، عبارت دیگر چند مورد از پیش‌بینی تصمیم بازنشر، در ۱۱ مل نیز بازنشر شده‌اند.

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

F– measure–٤

این معیار، میانگین موزون دقت و بازیابی می‌باشد. β پارامتری است که توازن را بین این دو معیار فراهم می‌نماید. 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

---

۱۲ True Positive
۱۳ False Positive
۱۴ True Negative
۱۵ False Negative
۱۶ Sensitivity
۱۷ True Positive Rate
۱۸ Recall
۱۹ Specificity
۲۰ Recall-
۲۱ Precision

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.

در این بخش به ارزیابی مدل پیشنهادی برحسب مجموعه دادگان توییتر (از مجموعه دادگان سال ۲۰۲۰ توییتر استفاده‌شده) پرداخته می شود که چگونه مجموعه داده‌ها به یک پایگاه داده رابطه ای منتقل شده است و همچنین استراتژی های تقسیم بندی داده که در این کار برای آموزش و تست در نظر گرفته شده است معرفی می گردد. در آخر عملکرد روش پیشنهادی با سایر روش ها مقایسه می شود و در نهایت ش پیشنهادی نشان داده خواهد شد.

۱۷۵

مجموعه دادگان

سه مجموعه دادگان مختلف با استفاده از twitter streaming API[۲۲] ایجاد می شود. از مجموعه موضوعات روز[۲۳] به سه موضوع تکنولوژی ، سیاسی و سلامتی پرداخته شده است. این مجموعه دادگان در جدول ۱ نشان داده شده است.

جدول ۱ مجموعه دادگان توییتر

شماره	مجموعه دادگان	توضیح
مجموعه دادگان یک	تکنولوژی	رونمایی از ایفون ۱۱
مجموعه دادگان دو	سیاسی	انتخابات امریکا
مجموعه دادگان سه	سلامتی	انتشار ویروس کرونا

جزئیات بیشتری در ارتباط با مجموعه داده و مباحث آماری آن ، در جدول ۲مطرح شده است.

جدول ۲ جزئیات مجموعه داده

$$F_1 = (\beta^2 + 1)\frac{precision.recall}{\beta^2 precision + recall}(0 \le \beta \le \infty)$$

شماره	مجموعه دادگان	تعداد پست‌ها	تعداد کاربران	تعداد بازنشرها
مجموعه دادگان یک	رونمایی از ایفون ۱۱	۱۰۲۵۸۵۲	۶۷۸۵۲۰	۸۵۰۲۳۰

[۲۲] https://api.twitter.com/

[۲۳] *Hot topic*

مجموعه دادگان دو	انتخابات امریکا	۲۵۴۵۶۰۲	۳۲۵۲۸۰	۶۵۰۲۸۵
مجموعه دادگان سه	انتشار ویروس کرونا	۸۵۹۶۳۵۸	۴۸۷۵۲۶	۹۴۳۵۶۸

ساختار توییتر

کاربران با افزودن # هشتگ ، پست ها را بر اساس موضوع دسته بندی می کنند و به کاربران در جستجوی موضوع کمک می کنند و این می تواند در هر کجای توییت ((در ابتدا ، وسط یا انتهای آن)) رخ دهد. وقتی کلمات هشتگ محبوب می شوند ، به آنها موضوعات پرطرفدار می گویند. برای ارسال پیام خود به یک کاربر خاص ، کافی است نام کاربر در آن پست ذکر شود و سپس آن ها در صفحه **Mentions** خود ، **Tweet** را مشاهده می کنند.

Tweet @ [account]

به این ترتیب ، خالق این توییت می تواند سایر کاربران را به پست اضافه کند. هنگامی که کاربر صفحه خود را باز می کند ، می تواند تمام پستهایی را که در آنها ذکر شده است مشاهده کند. استفاده دیگری که قابل ذکر است شامل پخش مجدد پست های افراد دیگر یا (بازنشر) است. کاربران می‌توانند از گزینه ((دکمه)) بازنشر موجود در زیر پست استفاده کنند یا می‌توانند نام کاربری RT @ را در ابتدای پست ذکر کنند. بازنشر ها مفید هستند زیرا به آنها امکان می دهد جریان اطلاعات را در توییتر پیگیری کنند. بازنشر توییت در اصل یک یال دوجهته در گراف شبکه اجتماعی محسوب می شود.

[additional text] RT @[account] : [original tweet]

ساختار پیام توییتر

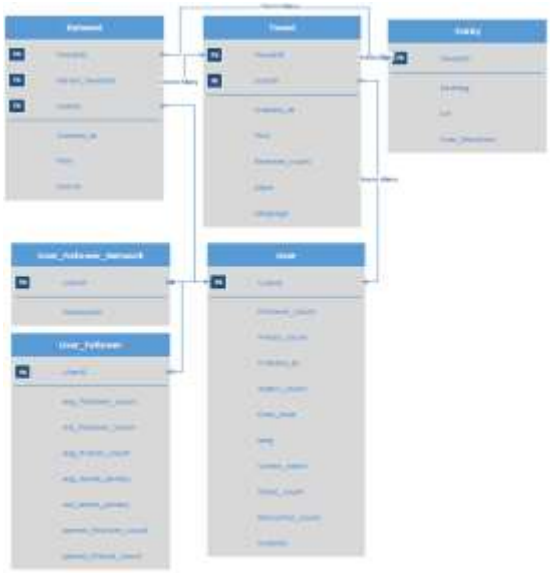
هر توییت دارای یک بدنه اصلی است که حاوی ویژگی های منفرد مانند، (شناسه، متن، منبع، در پاسخ به شناسه وضعیت) و ویژگی های پیچیده مانند (کاربر، نهاد[۲۴]، جغرافیا، مکان) است که حاوی صفات بیشتری در داخل آنها است . در این پژوهش سه قسمت متمایز از هرتوییت که شامل توییت ، کاربر و موجودیت در لیست JSON می باشد ، مورد بررسی قرار می گیرد. در این پژوهش، برای سهولت پیاده‌سازی پرس‌وجوهای متعدد، از پایگاه دادۀ رابطه‌ای برای نگهداری خروجی JSONها استفاده‌شده است.

ایجاد پایگاه داده رابطه ایی [۲۵]

در بخش قبل داده‌هایی که شامل توییت و بازتوییت بوده است ، جمع آوری شده است. و از توییت ها ،فیلدهایی جهت ساختن یک پایگاه داده رابطه ایی استخراج شده است. پایگاه داده رابطه ای مجموعه ای از آرایه های دو بعدی است که جداول یا رابطه نامیده می شوند. شش جدول جهت تعریف پایگاه داده استفاده‌شده است، که شامل Tweet, Retweet, User, User_ Follower, User_Follower_Network, Entity می باشد. نمودار پایکاه داده در شکل۲ نشان داده شده است.

[۲۴] **Entity**

[۲۵] Relational Data Base



شکل ۲پایگاه داده رابطه ایی

استراتژی تقسیم مجموعه داده

استراتژی های متفاوتی برای این موضوع مورد بررسی قرار می گیرد که در این پژوهش از استراتژی اعتبارسنجی متقابل که در ادامه به صورت جزیی توضیح داده می شود ، مورد استفاده قرار می گیرد.

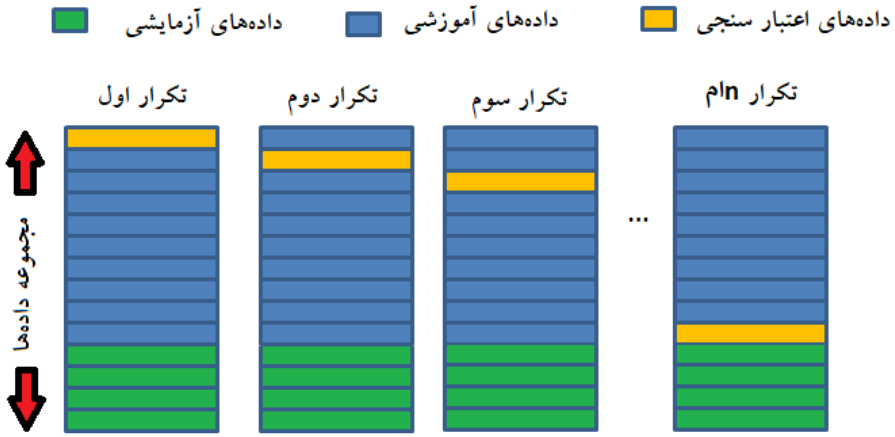
اغلب در مدل‌سازی، بخصوص در مبحث ((آموزش ماشین[۲۶])) احتیاج به برآورد پارامترهای مدل می باشد. در صورتی که تعداد پارامترها زیاد باشد، پیچیدگی مدل زیاد شده و ممکن است محاسبات به سادگی قابل انجام نباشند. اعتبارسنجی متقابل[۲۷] یکی از راه‌هایی است که می‌توان تعداد پارامترها (متغیرهای) مدل را بصورت بهینه تعیین کرد.

.هدف در اعتبارسنجی متقابل، دستیابی به مدلی است که تعداد پارامترهای آن بهینه باشد. یعنی پیدا کردن مدلی است که دچار بیش‌برازش نباشد. برای دستیابی به این هدف در ((آموزش ماشین)) معمولا داده‌ها را به دو قسمت تفکیک می‌کنند.

- قسمت داده‌های آموزشی[۲۸] :از این بخش از داده‌ها به منظور ایجاد مدل و برآورد پارامترهای آن استفاده می‌شود.
- قسمت داده‌های آزمایشی [۲۹]:این قسمت از داده‌ها برای بررسی کارایی مدل استفاده می‌شود. اهمیت این بخش از داده‌ها در این نکته است که این مشاهدات شامل مقدارهای متغیرهای مستقل ($\boldsymbol{x}$)ها و پاسخی (y) هستند که در مدل به کار نرفته ولی امکان مقایسه مقدار پیش‌بینی شده ($\boldsymbol{y}^{\wedge}$) را با مقدار واقعی به ما می‌دهند .البته توجه داریم که این

26 Machine learning
27 Cross validation
28 Training set
29 Test set

داده‌ها مدل را تحت تأثیر قرار نداده‌اند، پس در تعیین پارامترهای مدل نقشی نداشته و فقط برای ارزیابی مدل به کار می‌روند. شکل ۳ به درک ماهیت داده‌های تست در فرآیند اعتبارسنجی متقابل کمک کند.



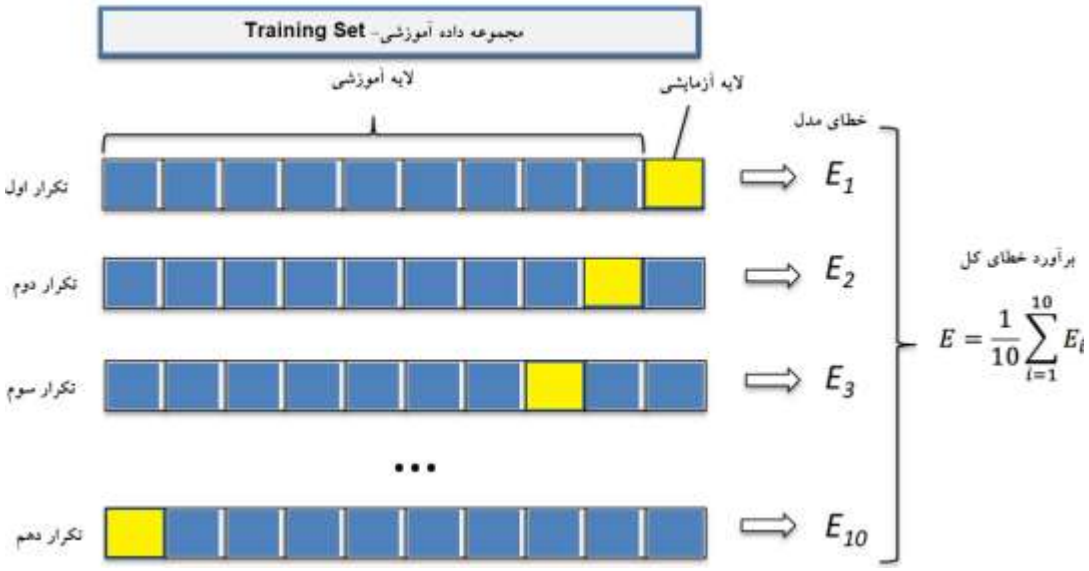
شکل ۳ فرآیند اعتبار سنجی متقابل

از روش اعتبار سنجی K-fold اعتبار سنجی متقابل استفاده‌شده است.

- **اعتبارسنجی** k-Fold

اگر مجموعه داده‌های آموزشی را به طور تصادفی به k زیرنمونه یا لایه [۳۰] با حجم یکسان تفکیک کنیم، می‌توان در هر مرحله از فرایند اعتبار سنجی متقابل ، تعداد k-1 از این لایه‌ها را به عنوان مجموعه داده آموزشی و یکی را به عنوان مجموعه داده اعتبارسنجی در نظر گرفت. شکل ۴-۵ مراحل روش k-Fold را به خوبی نشان می‌دهد. مشخص است که با انتخاب k=10 ، تعداد تکرارهای فرآیند اعتبار سنجی متقابل برابر با ۱۰ خواهد بود و دستیابی به مدل مناسب به سرعت امکان‌پذیر می‌شود.

[۳۰] Fold



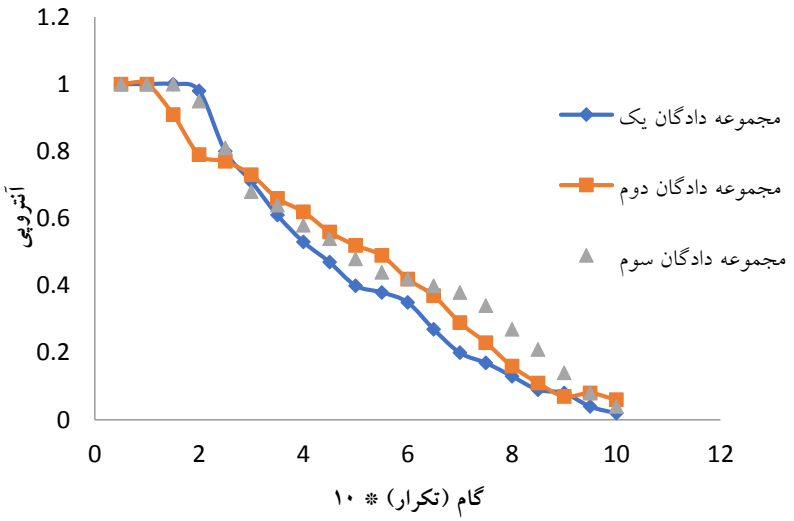
شکل۴اعتبارسنجی K-fold

٤-٨-٢ انتخاب الگوریتم و یادگیری

در این بخش بر طبق مدل یادگیری پیشنهادی در فصل سوم و براساس استراتژی اعتبارسنجی متقابل ***K-fold*** به یادگیری پرداخته می شود.

برای نشان دادن فرآیند یادگیری در روش پیشنهادی، از روند تغییرات آنتروپی، می‌توان استفاده کرد. نمودار روند تغییرات آنتروپی محیط در هر سه مجموعه دادگان با رنگ‌های مختلف در شکل **5** 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 پژوهش مقدار دو صدم برای آن در نظر گرفته شده است.

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 **5** 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 **5** نمودار روند تغییرات آنتروپی در هر سه مجموعه دادگان ارزیابی بعد از طی گام‌های یادگیری

٤-۱۰-۲ارزیابی عملکردها

همانطور که در بخش های قبل ارزیابی عملکردها معرفی گردید ، در این بخش به ارزیابی روش پیشنهادی با سه روش دیگر یعنی شبکه بیزین، جنگل تصادفی و ماشین بردار پشتیبان[۳۱] علت انتخاب این سه روش برای مقایسه با روش پیشنهادی (***DIUS***) این بوده که این روش ها با استناد به مقالات مختلف بسیار مهم و مرسوم بوده است.

در جدول ۳ مقادیر دقیق حساسیت روش‌های مختلف ارزیابی شده را نشان می‌دهد. در بررسی نتایج، نکته حائز اهمیت دقت بالای روش ***DIUS*** پیشنهاد شده در مقایسه با دیگر روش‌هاست. ***DIUS*** در برابر روشی مانند جنگل تصادفی بسیار کند است ، با این وجود، دقت بالای این روش علی‌رغم کندی آن می‌تواند قابلیت‌های این روش را نشان دهد. دقت بالای روش ***DIUS*** به ماهیت این روش برمی‌گردد. در میان روش‌های دیگر، روش شبکه بیزین عملکرد مناسبی داشته است. علت این برتری در قدرت طبقه‌بندی این روش می‌باشد.

الف) مقایسه نرخ حساسیت روش‌های پیشنهادی با بقیه روش‌ها

[۳۱] Support vector machine

جدول ۳مقایسه نرخ حساسیت روش‌های پیشنهادی با بقیه روش‌ها

معیار حساسیت	میانگین	مجموعه دادگان اول	مجموعه دادگان دوم	مجموعه دادگان سوم
DIUS	***0.83***	۰,۸۵	۰,۸۱	۰,۸۳
شبکه بیزین	***0.75***	۰,۷۳	۰,۷۵	۰,۷۶
جنگل تصادفی	***0.53***	۰,۵۳	۰,۵۱	۰,۵۴
ماشین بردار پشتیبان ***SVM***	***0.70***	۰,۶۰	***0.62***	۰,۶۴

ب)مقایسه معیار تشخیص روش‌های پیشنهادی با بقیه روش‌ها

در جدول ۴، 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. ***DIUS*** بهترین عملکرد را داشته است. در معیار تشخیص نیز، روش شبکه بیزین به صورت میانگین، بهترین عملکرد را بعد از ***DIUS*** به خود اختصاص داده است. روش ***DIUS*** توانسته است، موارد عدم بازنشر را بهتر از روش‌های دیگر، تشخیص دهد. در این بخش، روش جنگل تصادفی بدترین عملکرد را به خود اختصاص داده است.

جدول ۴مقایسه معیار تشخیص روش‌های پیشنهادی با بقیه روش‌ها

معیار تشخیص	*میانگین*	*مجموعه دادگان اول*	*مجموعه دادگان دوم*	*مجموعه دادگان سوم*
DIUS	***0.88***	***0.91***	***0.86***	0.88
شبکه بیزین	***0.73***	***0.66***	***0.63***	0.91
جنگل تصادفی	***0.52***	***0.56***	***0.51***	0.48
ماشین بردار پشتیبان ***SVM***	***0.69***	***0.60***	***0.71***	0.75

ج) مقایسه نرخ دقت روش‌های پیشنهادی با بقیه روش‌ها

در ادامه ارزیابی روش‌های پیشنهادی برای پیش‌بینی تصمیم بازنشر کاربر در مواجهه با یک پست، به بررسی معیار نرخ خطای روش‌های پیشنهادی در مقایسه با روش‌های دیگر می‌پردازیم. 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 ۵ آورده شده است.

جدول۵ مقایسه نرخ دقت روش‌های پیشنهادی با بقیه روش‌ها

معیار دقت	میانگین	مجموعه دادگان اول	مجموعه دادگان دوم	مجموعه دادگان سوم
DIUS	***0.80***	۰,۷۶	۰,۸۱	***۰,۸۲***
شبکه بیزین	***0.56***	۰,۶۱	۰,۵۱	***۰,۵۵***
جنگل تصادفی	***0.47***	۰,۴۳	۰,۴۸	***۰,۵۱***
ماشین بردار پشتیبان ***SVM***	***0.73***	۰,۷۲	۰,۷۸	***۰,۶۹***

روش ***DIUS*** 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.

د)مقایسه معیار *F- measure* روش‌های پیشنهادی با بقیه روش‌ها

همان‌طور که پیش‌تر اشاره شد، معیار ***F-measure*** بکارگرفته شده در این پژوهش، میانگین موزون دقت و بازیابی (حساسیت) می‌باشد (***β***=۱). همان‌طور که اشاره شد، میزان این معیار، ارتباط مستقیمی به دقت و بازیابی روش‌ها دارد. هر دو روش پیشنهادی، بهترین عملکرد را داشته‌اند. نکته جالب در این جدول، پایین بودن مقدار ***F- measure*** برای روش جنگل تصادفی است. با وجود دقت نسبتاً خوب این روش، با توجه به ضعف شدید این روش در بازیابی، بدترین عملکرد را در این معیار دارد. شبکه بیزین ، در نقطه مقابل، علی‌رغم، دقت پایین، در بخش حساسیت، عملکرد مناسبی داشت. در نتیجه مقدار ***F-measure*** برای این روش چندان پایین نیست.

جدول ۶ مقایسه معیار F measure روش‌های پیشنهادی با بقیه روش‌ها

معیار F measure	میانگین	مجموعه دادگان اول	مجموعه دادگان دوم	مجموعه دادگان سوم
DIUS	***2.62***	***2.7***	***2.52***	2.65

شبکه بیزین	*2.18*	*1.96*	*2.12*	2.45
جنگل تصادفی	*1.59*	*1.59*	*2.61*	1.56
ماشین بردار پشتیبان ***SVM***	*1.83*	*1.72*	*1.65*	2.11

بحث و نتیجه گیری

در این پژوهش به یکی از مسائل چالش بر انگیز در حوزه علم داده پرداخته شده است. تشخیص کاربران پرنفوذ در رسانه های اجتماعی بسیار مهم و برای سازمان‌ها حیاتی است. این مسئله می تواند در خیلی از جنبه های سازمان پر اهمیت باشد. در این پایان نامه برای تشخیص کاربران پرنفوذ ویژگی های بسیار مهمی از داده‌ها مطرح شده است و با استفاده از معماری کریسپ به تشخیص و تحلیل آن ها پرداخته شده است. مرحله اول از این روش جمع آوری داده است که از ***API*** توییتر استفاده‌شده است. مرحله دوم ، پیش پردازش داده است که در این بحش داده‌ها به دو بخش ساختاریافته و غیر ساختاریافته تقسیم می شود در مرحله سوم به استخراج ویژگی های مناسب پرداخته می شود و در نهایت در مرحله آخر ارزیابی این روش صورت می گیرد.

رفتارشناسی کاربران در شبکه‌های اجتماعی یکی از جذاب‌ترین بحث‌های حوزه فناوری اطلاعات در دهه اخیر می‌باشد. به عنوان پیشنهاد برای محققین آینده، می‌توان روی رفتارشناسی انتشار در شبکه‌های اجتماعی مطالعات تکمیلی صورت پذیرد.

- علایق کاربران با گذشت زمان تغییر می‌کند، این تغییرات یا بستگی به برهه‌های خاص زمانی دارد، مانند اوایل سال جدید و یا بر اثر تغییر طبع کاربر با گذر زمان ایجاد می‌شود، با در نظر گرفتن این پویایی در شبکه‌های اجتماعی می‌توان روش پیشنهادی را انعطاف پذیرتر نمود. با استفاده از روش‌های پیوندی می‌توان ویژگی‌های مثبت روش پیشنهادی را با سایر روش‌های مطرح شده در این زمینه ترکیب کرد و عملکردروش پیشنهادی را ارتقاء داد.
- یکی از زمینه‌های مناسب برای محققین آینده ، « ارائه سازوکاری برای انتخاب محتوای چندرسانه‌ای اثرگذار در اجتماعات برخط» می‌باشد. امروزه درگاه‌های اشتراک‌گذاری ویدئو نظیر یوتیوب و آپارات، روزانه میلیون‌ها بازدید کننده دارد. چرخه عظیم مالی این درگاه‌ها از طریق تبلیغات میسر شده است. بازدید هر چه بیشتر ویدئوها در این سامانه‌ها موجب افزایش درآمد شده است. بر اساس تحقیقات انجام شده میزان اثرگذاری محتواهای چندرسانه‌ای بسیار بیشتر از محتواهای متنی می‌باشد که استفاده از این نوع محتوا برای تبلیغات را مناسب‌تر کرده است. امروزه با پیشرفت‌های صورت گرفته، امکاناتی نظیر یادداشت‌گذاری، برچسب‌گذاری، پسندیدن، اشتراک‌گذاری و سایر موارد در روی ویدئوها بوجود آمده است. این اطلاعات می‌تواند برای انتخاب محتواهای چندرسانه‌ای اثرگذار بسیار مفید باشد. به عنوان پیشنهاد می‌توان با اندازه‌گیری مدت زمان مشاهده، بررسی موضوع، هشتگ و یادداشت‌های یک ویدیو، مدلی از کاربران ایجاد کرد. با استفاده از این مدل‌سازی،

بـازخورد کاربران در مواجهه با ویدیوهای دیگر قـابـل پیش‌بینی می‌باشـــد. این کار به افزایش بازدید و درآمد سامانه‌های اشتراک گذاری ویدیو و افزایش رضایتمندی کاربران کمک می‌کند.

- اســـتفـاده از الگوریتم های ادغامی با نظارت و بدون نظارت به گونه ایی که احتمال گیر کردن روش در مینیم هـای محلی کـاهش پیدا کند و دقت پیش بینی بالاتر رود.

منابع

1- Enrique Tica, G.I. and Pineda Ruiz, D.A., 2018. El marketing digital en las redes sociales facebook, linkedin y youtube y su influencia en la fidelización de los clientes de la empresa Atanasovski corredores de seguros.

2-Cartwright, D., & Harary, F. (1956). Structural balance: a generalization of Heider's theory. Psychological Review, 63, 277–292.

3- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks Science, 286, 509–512.

.

4-Eirinaki, Magdalini, Sumit Pal Singh Monga, and Shreedhar Sundaram. "Identification of influential social networkers." **International Journal of Web Based Communities** *8.2 (2012): 136-158.*

5-Fiorillo, Diana, et al. "Identification of influential user locations for smart meter installation to reconstruct the urban demand pattern." **Journal of Water Resources Planning and Management** *146.8 (2020): 04020070.*

6-Amati, Giambattista, et al. "Influential users in Twitter: detection and evolution analysis." **Multimedia Tools and Applications** *78 (2019): 3395-3407.*

7-Kostkova, Patty, et al. "Who is spreading rumours about vaccines? Influential user impact modelling in social networks." **Proceedings of the 2017 international conference on digital health.** *2017.*

8-Eskandanian, Farzad, Nasim Sonboli, and Bamshad Mobasher. "Power of the few: Analyzing the impact of influential users in collaborative recommender systems." **Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization.** *2019.*

9-Said, Anwar, et al. "Proactive caching at the edge leveraging influential user detection in cellular D2D networks." **Future Internet** *10.10 (2018): 93.*

10-Mahmoudi, Amin, Mohd Ridzwan Yaakub, and Azuraliza Abu Bakar. "New time-based model to identify the influential users in online social networks." **Data Technologies and Applications** *52.2 (2018): 278-290.*

11-Jain, Somya, and Adwitiya Sinha. "TriBeC: identifying influential users on social networks with upstream and downstream network centrality." **International Journal of General Systems** *(2023): 1-22.*

12-Shazad, Babar, et al. "Finding Temporal Influential Users in Social Media Using Association Rule Learning." **Intelligent Automation & Soft Computing** *26.1 (2020)*

13-Deng, Yujing, et al. "Energy consumption characteristics and influential use behaviors in university dormitory buildings in China's hot summer-cold winter climate region." **Journal of Building Engineering** *33 (2021): 101870.*

14-Deng, Yujing, et al. "Energy consumption characteristics and influential use behaviors in university dormitory buildings in China's hot summer-cold winter climate region." **Journal of Building Engineering** *33 (2021): 101870.*

15-Kao, Li-Jen, and Yo-Ping Huang. "Mining influential users in social network." **2015 IEEE** *International*

19-Wu, Xinmiao, and Jianmin Wang. "Micro-blog in China: identify influential users and automatically classify posts on Sina micro-blog." **Journal of Ambient Intelligence and Humanized Computing** *5 (2014):*

20-Novak, Petra Kralj, Luisa De Amicis, and Igor Mozetič. "Impact investing market on Twitter: influential users and communities." **Applied network science** *3 (2018): 1-20.*

21-Jain, Somya, and Adwitiya Sinha. "Discovering influential users in social network using weighted cumulative centrality." **Concurrency and Computation: Practice and Experience** *34.1 (2022): e6521.*

22-Aizawa, Akiko. "An information-theoretic perspective of tf–idf measures." **Information Processing & Management** *39.1 (2003): 45-65.*

Conference on Systems, Man, and Cybernetics. IEEE, 2015.

16-Guo, Gongde, et al. "KNN model-based approach in classification." On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, 2003.

17-Mohammadi, Azadeh, and Mohamad Saraee. "Finding influential users for different time bounds in social networks using multi-objective optimization." **Swarm and evolutionary computation** *40 (2018): 158-165.*

18-Wu, Xinmiao, and Jianmin Wang. "Micro-blog in China: identify influential users and automatically classify posts on Sina micro-blog." **Journal of Ambient Intelligence and Humanized Computing** *5 (2014): 51-63.*

Exploring Influential Users on Social Media Using Artificial Neural Network Techniques in Python for Enhancing Online Marketing Performance

In recent years, with the advent of social media, the discussion of analyzing user behavior based on their data has gained more attention than ever. During these years, organizations and large companies are interested in investing in data to better understand their customers. One of the crucial topics in the field of data science is the identification of influential users. These users can increase product and service purchases and reduce advertising and customer acquisition costs by exerting more influence on other users. The aim of this issue is to improve advertising and marketing performance. To identify these users, Twitter data has been used, and, according to the crisp model, data is first collected, then preprocessed, and finally modeled and evaluated. Deep learning algorithms have been used for modeling, and in the evaluation phase, this model has been compared to other machine learning models, including Bayesian networks, random forests, and support vector machines. Using data science evaluation methods, including sensitivity, specificity, accuracy, and F-measure, the proposed model has outperformed traditional models. Therefore, the proposed method is expected to have better performance.

Key Words: social networks, Influential users, Deep learning, Internet marketing, Python