DOR: https://csj.isi.org.ir/article_256401.html

DOI: https://doi.org/10.22034/csj.2026.584429.1156

پاییز ۱۴۰۵، دوره ۱۱، شماره ۳، پیاپی ۴۲

دریافت مقاله: ۱۴۰۵/۰۳/۱۲

پذیرش مقاله: ۱۴۰۵/۰۶/۱۷

نوع مقاله: پژوهشی

ارائه چارچوبی مبتنی بر تحلیل گراف و یادگیری ماشین برای پیش‌بینی ارتباط کاربران با نمانام‌ها در شبکه‌های اجتماعی

مرتضی محمدی زنجیره*

گروه مهندسی کامپیوتر دانشکده فنی و مهندسی دانشگاه بین المللی امام خمینی(ره)، قزوین، ایران

پست الکترونیکی: zanjireh@edu.ikii.ac.ir

نسرین نظری

گروه مهندسی کامپیوتر دانشکده فنی و مهندسی دانشگاه بین المللی امام خمینی(ره) ، قزوین، ایران

پست الکترونیکی: nasrinnazari@edu.ikiu.ac.ir

چکیده

در سال‌های اخیر، رشد شبکه‌های اجتماعی برخط و افزایش تعامل کاربران با نشان‌های تجاری، تحلیل رفتار کاربران و پیش‌بینی روابط میان کاربران و نمانام‌ها را به یکی از موضوعات مهم در تحلیل شبکه‌های اجتماعی تبدیل کرده است. در این پژوهش، چارچوبی مبتنی بر نظریه گراف و الگوریتم‌های یادگیری ماشین برای پیش‌بینی ارتباط میان کاربران و نمانام‌های تجاری در شبکه اجتماعی ایکس ارائه شده است. برای انجام این تحقیق، داده‌های مربوط به دنبال‌کنندگان شش نمانام تجاری شامل آدیداس، نایکی، ریبوک، لاکوست، هوکا و جامپمن۲۳ از طریق رابط برنامه‌نویسی کاربردی ایکس جمع‌آوری شد. سپس ویژگی‌هایی شامل مقدار رتبهٔصفحه، کوتاه‌ترین مسیر، تعداد دنبال‌کنندگان، تعداد دنبال‌شوندگان و اشتراک همسایگی استخراج و پس از پیش‌پردازش و نرمال‌سازی، به‌عنوان ورودی مدل‌های یادگیری ماشین استفاده شد. در ادامه، عملکرد سه الگوریتم رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان با استفاده از اعتبارسنجی متقابل ده‌بخشی ارزیابی گردید. نتایج نشان داد که الگوریتم جنگل تصادفی با دقت ۹۷ درصد عملکرد بهتری نسبت به سایر مدل‌ها داشته است، در حالی که دو الگوریتم دیگر به دقت ۹۵ درصد دست یافتند. نتایج این پژوهش نشان می‌دهد که ترکیب ویژگی‌های ساختاری گراف با الگوریتم‌های یادگیری ماشین می‌تواند رویکردی مؤثر برای پیش‌بینی ارتباط کاربران با برندها در شبکه‌های اجتماعی باشد.

کلید واژه‌ها: تحلیل شبکه‌های اجتماعی، یادگیری ماشین، نظریه گراف، پیش‌بینی ارتباط کاربران، شبکه اجتماعی ایکس

۱- مقدمه

در دهه اخیر، شبکه‌های اجتماعی برخط به یکی از مهم‌ترین بسترهای تعامل میان کاربران، سازمان‌ها و نمانام‌های تجاری تبدیل شده‌اند. رشد سریع این شبکه‌ها و افزایش حجم داده‌های تولیدشده توسط کاربران، فرصت‌های گسترده‌ای را برای تحلیل رفتار اجتماعی، شناسایی الگوهای ارتباطی و پیش‌بینی تعاملات کاربران

* نویسندهٔ مسئول

فراهم کرده است. در این میان، شبکه اجتماعی ایکس[1] (توییتر[2]) به دلیل ساختار گراف‌محور، تعداد زیاد کاربران فعال و دسترسی به داده‌های تعاملی، به یکی از منابع اصلی مطالعات حوزه تحلیل شبکه‌های اجتماعی تبدیل شده است [۱, ۲]. تحلیل داده‌های موجود در این شبکه‌ها می‌تواند اطلاعات ارزشمندی درباره علایق کاربران، نحوه شکل‌گیری ارتباطات و میزان تأثیرگذاری نمانام‌ها در اختیار پژوهشگران و کسب‌وکارها قرار دهد.

شبکه‌های اجتماعی معمولاً به‌صورت گراف مدل‌سازی می‌شوند که در آن گره‌ها نمایانگر کاربران و یال‌ها بیانگر روابط یا تعاملات میان آنها هستند. این نوع نمایش امکان استفاده از مفاهیم نظریه گراف را برای تحلیل ساختار شبکه، شناسایی کاربران تأثیرگذار و بررسی روابط اجتماعی فراهم می‌کند [۳]. در سال‌های اخیر، معیارهای مبتنی بر تحلیل گراف مانند درجه[3]، مرکزیت[4]، ضریب خوشه‌بندی[5]، کوتاه‌ترین مسیر[6] و الگوریتم رتبۀصفحه[7] در بسیاری از مطالعات مورد استفاده قرار گرفته‌اند [٤, ٥]. با این حال، این ویژگی‌ها به تنهایی قادر به مدل‌سازی کامل پیچیدگی رفتار کاربران در شبکه‌های اجتماعی نیستند و معمولا قدرت پیش‌بینی محدودی دارند.

در مقابل، الگوریتم‌های یادگیری ماشین به دلیل توانایی در یادگیری الگوهای پنهان داده‌ها، در حوزه تحلیل رفتار کاربران و پیش‌بینی روابط اجتماعی کاربرد گسترده‌ای پیدا کرده‌اند [٦]. روش‌هایی مانند رگرسیون لجستیک[8]، جنگل تصادفی۹ و ماشین بردار پشتیبان۱۰ از جمله الگوریتم‌هایی هستند که در مسائل طبقه‌بندی و پیش‌بینی ارتباطات شبکه‌ای عملکرد قابل قبولی ارائه کرده‌اند [۷, ۸]. با این وجود، بخش قابل‌توجهی از مطالعات پیشین یا تنها بر تحلیل ساختاری شبکه متمرکز بوده‌اند و یا بدون استفاده مؤثر از ویژگی‌های گرافی، صرفاً از مدل‌های یادگیری ماشین استفاده کرده‌اند. بنابراین، همچنان نیاز به رویکردهای ترکیبی وجود دارد که بتوانند همزمان از ساختار شبکه و قدرت یادگیری ماشین استفاده کنند [۹, ۱۰].

در این پژوهش، رویکردی ترکیبی مبتنی بر تحلیل شبکه‌های اجتماعی و الگوریتم‌های یادگیری ماشین برای پیش‌بینی ارتباط میان کاربران و نمانام‌های تجاری ارائه شده است. داده‌های مربوط به دنبال‌کنندگان شش نمانام تجاری شامل آدیداس[11]، نایکی[12]، ریبوک[13]، لاکوست[14]، هوکا[15] و جامپمن۲۳[16] از شبکه اجتماعی ایکس جمع‌آوری شده است. سپس مجموعه‌ای از ویژگی‌های ساختاری و ارتباطی شامل رتبۀصفحه، کوتاه‌ترین مسیر میان گره‌ها، تعداد دنبال‌کنندگان، تعداد دنبال‌شوندگان و میزان اشتراک همسایگی استخراج گردید. پس از پیش‌پردازش و نرمال‌سازی داده‌ها، این ویژگی‌ها به‌عنوان ورودی مدل‌های یادگیری ماشین مورد استفاده قرار گرفتند. در ادامه، عملکرد سه الگوریتم رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان با استفاده از روش اعتبارسنجی متقابل ده‌بخشی ارزیابی شد. نتایج نشان داد که مدل جنگل تصادفی عملکرد بهتری نسبت به سایر مدل‌ها در پیش‌بینی ارتباط کاربران با نمانام‌ها دارد. بر این اساس، پژوهش حاضر بر توسعه یک چارچوب ترکیبی و قابل‌تفسیر برای بهره‌گیری همزمان از ویژگی‌های ساختاری گراف و قابلیت‌های یادگیری مدل‌های طبقه‌بندی در مسئله پیش‌بینی ارتباط کاربر- برند تمرکز دارد.

مشارکت‌های اصلی پژوهش

با توجه به شکاف‌های شناسایی‌شده در مطالعات پیشین، مشارکت‌های اصلی این پژوهش به شرح زیر است

۱. ارائه یک چارچوب ترکیبی برای پیش‌بینی ارتباط

1-X
2-Twitter
3-Degree
4-Centrality
5-Clustering Coefficient
6-Shortest Path
7-PageRank
8-Logistic Regression (LR)
9-Random Forest (RF)
10-Support Vector Machine (SVM)

11-Adidas
12-Nike
13-Reebok
14-Lacoste
15-Hoka
16-Jumpman23

کاربر- نمانام: در این پژوهش، چارچوبی ترکیبی از تحلیل ساختاری شبکه و الگوریتم‌های یادگیری ماشین برای پیش‌بینی ارتباط میان کاربران و نمانام‌های تجاری در شبکه اجتماعی ایکس ارائه شده است. این چارچوب، اطلاعات همبندی شبکه را به ویژگی‌های قابل استفاده برای مدل‌های یادگیری ماشین تبدیل می‌کند.

۲. مدل‌سازی ارتباط کاربر- نمانام در قالب مسئله پیش‌بینی پیوند: در رویکرد پیشنهادی، هر ارتباط کاربر- نمانام به‌صورت یک زوج گره در نظر گرفته شده و وجود یا عدم وجود ارتباط مستقیم میان آن‌ها به‌عنوان یک مسئله طبقه‌بندی دودویی مدل‌سازی شده است.

۳. ترکیب ویژگی‌های محلی و سراسری شبکه: در این پژوهش، ویژگی‌هایی با ماهیت‌های مختلف، شامل اشتراک همسایگی و کوتاه‌ترین مسیر به‌عنوان شاخص‌های ساختار محلی و فاصله همبندی، و رتبهٔ‌صفحه و ویژگی‌های درجه‌ای به‌عنوان شاخص‌های اهمیت و جایگاه گره‌ها، به‌صورت هم‌زمان مورد استفاده قرار گرفته‌اند.

٤. ایجاد نمایش ویژگی برای زوج‌های کاربر- نمانام: برای هر زوج کاربر- نمانام، مجموعه‌ای یکپارچه از ویژگی‌های ساختاری شامل رتبهٔ‌صفحه، تعداد دنبال‌کنندگان، تعداد دنبال‌شوندگان، کوتاه‌ترین مسیر و اشتراک همسایگی استخراج شده و برای آموزش مدل‌های طبقه‌بندی مورد استفاده قرار گرفته است.

۵. ارزیابی مقایسه‌ای مدل‌های یادگیری ماشین: سه الگوریتم رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان تحت یک مجموعه ویژگی و پروتکل ارزیابی یکسان مورد مقایسه قرار گرفته‌اند تا نقش نوع مدل یادگیری در پیش‌بینی ارتباطات کاربر- نمانام بررسی شود.

٦. بررسی مسئله ارتباط کاربر- نمانام در داده‌های واقعی شبکه اجتماعی ایکس: برخلاف مطالعاتی که عمدتاً بر شبکه‌های عمومی یا روابط کاربر- کاربر تمرکز دارند، پژوهش حاضر به‌طور مشخص پیش‌بینی ارتباط کاربران با نمانام‌های تجاری را با استفاده از داده‌های واقعی مربوط به شش برند بررسی می‌کند.

۷. ارائه یک چارچوب قابل‌تفسیر برای تحلیل روابط کاربر- نمانام: استفاده از ویژگی‌های ساختاری قابل تفسیر این امکان را فراهم می‌کند که نقش شاخص‌های همبندی مختلف در تشخیص ارتباط میان کاربران و نمانام‌ها مورد بررسی قرار گیرد.

۸. ارزیابی تجربی چارچوب پیشنهادی بر روی یک شبکه چند نمانامی: استفاده از داده‌های مربوط به شش نمانام تجاری و روابط کاربران، امکان ارزیابی چارچوب پیشنهادی را در یک سناریوی واقعی و چندنمانامی فراهم کرده است.

۹. فراهم کردن مبنایی برای توسعه روش‌های پیشرفته پیش‌بینی پیوند: چارچوب ارائه شده می‌تواند مبنایی برای مقایسه و توسعه روش‌های پیشرفته‌تر، از جمله روش‌های یادگیری نمایش گراف و شبکه‌های عصبی گراف، در مطالعات آینده فراهم کند.

۲- پیشینه پژوهش

تحلیل شبکه‌های اجتماعی در سال‌های اخیر به یکی از مهمترین حوزه‌های پژوهشی در علوم داده، داده‌کاوی[17] و یادگیری ماشین تبدیل شده است. در این حوزه، شبکه‌های اجتماعی معمولاً به صورت گراف مدل‌سازی می‌شوند که در آن گره‌ها بیانگر کاربران و یال‌ها نشان‌دهنده روابط میان آنها هستند. مطالعات اولیه در این زمینه بیشتر بر تحلیل ساختار شبکه و شناسایی گره‌های تأثیرگذار متمرکز بوده‌اند. برای این منظور، معیارهایی مانند درجه، مرکزیت و فاصله‌های گرافی به‌طور گسترده مورد استفاده قرار گرفته‌اند [۲]. پژوهشی در ارتباط با این زمینه نشان داد که معیارهای مرکزیت می‌توانند نقش مهمی در شناسایی کاربران تأثیرگذار و تحلیل جریان اطلاعات در شبکه‌های

17-Data Mining

اجتمـــاعی ایفا کنند [۱۱]. با این حال، محدودیت بنیادین این رویکردها آن است که صرفاً بر ویژگی‌های همبندی ایستا تمرکز دارند و قادر به مدل‌ســـازی رفتار پویا، زمینه‌مند و چندبعدی کاربران نیستند. علاوه‌بر این، این معیارها معمولاً فرض می‌کنند که تمام یال‌های شـــبکه مشاهده‌شده‌اند، در حالی که در شبکه‌های واقعی، داده‌ها اغلب ناقص و پرنوفه هســـتند و داده‌های مفقود[۱۸] نیـــز از الگوهای غیرتصادفی پیروی می‌کنند. [۱۲]

در ادامه، پژوهشـــگران به ســـمت تحلیل ساختارهای گروهی در شبکه‌های اجتماعی و مسئله تشخیص اجتماع[۱۹] حرکـــت کردند. هدف اصلی این روش‌ها، تقسیم شـــبکه به زیرگروه‌هایی بـــا تراکم ارتباطی بالا بـــود. در این زمینه، الگوریتم تشـــخیص اجتمـــاع لوون[۲۰] که توســـط بلوندل[۲۱] و همکاران ارائه شـــد، به‌عنوان یـــکی از روش‌های کارآمد بـــرای شناســـایی اجتماعات در شـــبکه‌های بزرگ معرفی گردید [۱۳]. همچنین، روش‌های مبتنی بر خوشه‌بندی طیفی نیز برای تحلیل ســـاختارهای پیچیده شبکه توسعه یافتند [۱٤] اگرچه این روش‌ها در کشـــف ســـاختارهای ارتباطی موفق عمـــل کرده‌اند، اما چند محدودیت جـــدی دارند: (۱) ماهیـــت توصیفی: این الگوریتم‌ها عمدتاً ســـاختار موجود را توصیـــف می‌کنند و توانایی پیش‌بیـــنی روابط آینده را ندارند؛ (۲) ایســـتایی: فرض می‌کنند شـــبکه ثابت است، در حالی که شبکه‌های اجتماعی واقعی پویا و زمان‌مند هستند؛ (۳) حســـاسیت به پارامتر: عملکرد این روش‌ها به شـــدت وابسته به تنظیم دســـتی پارامترها (مانند تعداد اجتماعات یـــا آســـتانه‌های تراکم) اســـت؛ و (٤) عـــدم تعمیم‌پذیری: الگوریتم‌های تشـــخیص اجتماع معمولاً بر روی یک شبکه خاص آموزش می‌بینند و تعمیم آنها به شبکه‌های جدید یا پویا دشوار است [۱۲].

در کنار روش‌هـــای ســـاختاری، پژوهش‌های دیگری در حـــوزه پیش‌بینی پیوند[۲۲] شـــکل گرفته‌اند که هدف آنها پیش‌بیـــنی احتمـــال ایجاد ارتبـــاط جدید میـــان گره‌ها در آینده اســـت. روش‌های مبتنی بر شباهت ساختاری مانند همســـایه‌های مشـــترک[۲۳]، ضریب جاکارد[۲٤] و شـــاخص آدامیـــک-آدار[۲۵] از جمله رویکردهـــای اولیه در این حوزه بوده‌اند [۱۵]. محدودیت‌های کلیدی این روش‌ها عبارتند از: (۱) تمرکز بر شـــباهت‌های محلی: این شاخص‌ها عمدتاً از اطلاعات همســـایگی درجه یک یا دو بهـــره می‌برند و قادر به مدل‌ســـازی روابط بلندمحدوده یا غیرخطی نیستند؛ (۲) عدم استفاده از اطلاعات زمینه‌ای: این روش‌ها از محتوای تعاملات، ویژگی‌های گره‌ها یا زمینه زمانی بهره نمی‌برند؛ (۳) عملکرد ضعیف در شـــبکه‌های پراکنده: در شبکه‌های بزرگ با چگالی پایین، این شـــاخص‌ها دقت کافی ندارند؛ و (٤) محدود بودن توانایی تمایز الگوهای پیچیده: این روش‌ها بر معیارهای شباهت ساختاری از پیش تعریف‌شده متکی هستند و در مقایســـه با روش‌های یادگیری‌محور، قابلیت محدودتـــری بـــرای یادگیری خـــودکار روابـــط پیچیده و غیرخطی شبکه دارند [۱٦].

باگسترش کاربرد یادگیری ماشین، پژوهش‌های جدیدتر به اســـتفاده از مدل‌های نظارت‌شده برای تحلیل شبکه‌های اجتماعی روی آوردنـــد. الگوریتم‌هـــایی مانند رگرسیون لجســـتیک، درخت تصمیم، جنگل تصادفی و ماشین بردار پشتیبان در مسائل طبقه‌بندی گره‌ها و پیش‌بینی ارتباطات مورد اســـتفاده قرار گرفتنـــد [۷, ۸, ۱٦]. نتایج پژوهش‌ها نشان داد که این روش‌ها نسبت به مدل‌های کلاسیک مبتنی بر شـــباهت، دقت بالاتری ارائه می‌کننـــد [۱۲]. با این حال، این دســـته از روش‌ها نیـــز محدودیت‌های جدی دارند: (۱) وابســـتگی به مهندسی ویژگی: عملکرد این مدل‌ها به شدت وابســـته به کیفیت و جامعیت ویژگی‌های دستی استخراج شـــده اســـت و در صورت عدم طراحی مناسب ویژگی‌ها، دچار افت شـــدید عملکرد می‌شوند [۱۷]؛ (۲) عدم یادگیری خـــودکار نمایش: این الگوریتم‌ها قادر به یادگیری خودکار نمایش‌های غنی از ســـاختار گراف نیستند و باید ویژگی‌ها به صورت دســـتی تعریف شـــوند [۱۸]؛ (۳) تعمیم‌پذیری محدود: بسیاری از این مطالعات بر داده‌های ایســـتا تمرکز

18-Missing
19-Community Detection
20-Louvain
21-Blondel
22-Link Prediction
23-Common Neighbors
24-Jaccard Coefficient
25-Adamic-Adar

داشـــته و تعمیم‌پذیری آن‌ها به شبکه‌های پویا، چندلایه یا ناهمگن محدود است [۱۹]؛ و (٤) ارزیابی نادرست: بسیاری از این پژوهش‌ها از تقسیم‌بندی تصادفی داده‌ها اســـتفاده می‌کنند که در شـــبکه‌های زمان‌مند منجر به نشت اطلاعات آینده به آموزش شـــده و عملکرد را به صورت غیرواقعی بالا نشان می‌دهد [۲۰].

در سال‌های اخیر، ترکیب ویژگی‌های ساختاری شبکه بـــا الگوریتم‌های یادگیری ماشین به یـــکی از رویکردهای مهم در تحلیل شبکه‌های اجتماعی تبدیل شده است. در این رویکرد، ویژگی‌هایی مانند رتبهٔصفحه، همسایگی مشترک، کوتاه‌تریـــن مسیر و شـــاخص‌های مرکزیت از ســـاختار گراف اســـتخراج و به‌عنـــوان ورودی مدل‌هـــای یادگیری ماشین استفاده می‌شوند. مطالعات نشان داده‌اند که ترکیب ویژگی‌های ســـاختاری با روش‌هـــای داده‌کاوی می‌تواند به درک بهتر رفتار کاربـــران و بهبود تحلیل ارتباطات در شـــبکه‌های اجتماعی منجر شـــود [۱۰, ۲۱]. با وجود این، بســـیاری از این پژوهش‌ها از ویژگی‌های دســـتی و ثابت بهـــره می‌برند و قادر بـــه یادگیری خـــودکار نمایش‌های غنی از ســـاختار گراف نیســـتند [۲۲]. همچنیـــن، اغلب این مطالعات بر شـــبکه‌های عمومی (مانند فیسبوک یا توییتر) تمرکز داشته و کمتر به تحلیل تعاملات خاص، مانند ارتباط کاربران با نمانام‌های تجـــاری، پرداخته‌اند. علاوه بر این، در برخی موارد، ویژگی‌های اســـتخراج شده از معیارهای مختلف همبندی ممکن اســـت دارای همبستگی و اطلاعات هم‌پوشان باشند که می‌تواند به افزونگی ویژگی‌ها و کاهش کارایی مدل منجر شود [۲۳].

از ســـوی دیگر، شبکه‌های عصبی گراف به عنوان یکی از روش‌های نوین تحلیل داده‌های گرافی مورد توجه قرار گرفته‌انـــد. این مدل‌ها بـــا بهره‌گیری از ســـاختار ارتباطی گراف، در مسائلی مانند طبقه‌بندی گره‌ها، تشخیص اجتماع و پیش‌بینی ارتباطات عملکرد مناسبی ارائه داده‌اند و کاربرد گســـترده‌ای در تحلیل شـــبکه‌های اجتماعی و سیستم‌های توصیه‌گر یافته‌انـــد [۲٤, ۲۵]. با این حال، چالش‌های مهم و حل‌نشـــده‌ای در به‌کارگیری این مدل‌ها در ســـناریوهای واقعی وجـــود دارد: (۱) مقیاس‌پذیری: آموزش GNNها بر روی گراف‌های بسیار بزرگ با میلیون‌ها گره و یال نیازمند منابع محاسباتی قابل توجهی است و بسیاری از روش‌های موجود مقیاس‌پذیر نیستند [۲٦]؛ (۲) تفسیرپذیری: مدل‌های عمیـــق گرافی اغلب به صورت جعبـــه سیاه عمل می‌کنند و توجیه پیش‌بینی‌های آنها دشوار است، که این موضوع در کاربردهای حســـاس (مانند بازاریابی یا امنیت) مشکل‌ساز اســـت [۲۷] ؛ (۳) کیفیت داده: وجود نوفه، داده‌های ناقص، الگوهای غیرتصادفیِ داده‌های مفقود و عدم تعادل کلاسی می‌توانـــد عملکرد مدل را به شـــدت کاهش دهـــد [۲٦]؛ (٤) تعمیم‌پذیری: بسیاری از مدل‌ها در مواجهه با داده‌های خارج از توزیع آموزش[۲٦] یا گراف‌های ناهمگن دچار افت عملکرد می‌شوند [۲۸]؛ (۵) هموارسازی بیش‌ازحد[۲۷]: در GNNهای عمیق، نمایش گره‌ها پس از چندین لایه به ســـمت همگرایی می‌روند و تمایزپذیری خود را از دست می‌دهند [۲۹]؛ و (٦) ناهم‌دوســـتی[۲۸]: بسیاری از GNNها فرض می‌کنند گره‌های متصل مشـــابه هستند، در حالی که در شبکه‌های اجتماعی واقعی، گره‌های متصل اغلب متفاوت‌اند [۲۸].

همچنین، پژوهش‌هـــای جدید در حوزه پیش‌بینی پیوند نشـــان می‌دهند که روش‌های مبتنی بـــر یادگیری ماشین، یادگیری عمیق و شبکه‌های عصبی گراف به تدریج جایگزین رویکردهای ســـنتی مبتنی بر شباهت ســـاختاری شده‌اند. مطالعات اخیر بیان می‌کنند که اســـتفاده از نمایش‌های یاد گرفته شـــده گره‌ها در کنار ویژگی‌های ســـاختاری شبکه می‌تواند دقـــت پیش‌بینی ارتباطـــات را به‌طور قابل‌توجهی افزایش دهد [۳۰, ۳۱, ۳۲, ۳۳, ۳٤]. با این وجود، شـــواهد اخیر نشـــان می‌دهد که معیارهای ارزیـــابی رایج، به‌ویژه مساحت زیر منحنی[۲۹]، در برخی سناریوهای پیش‌بینی پیوند دارای محدودیت‌هایی هســـتند و ممکن است در شبکه‌های بســـیار نامتوازن عملکرد مدل‌ها را بیش از حد خوش‌بینانه نشـــان دهند [۳۵]. مطالعات نشـــان داده‌اند که (۱) مساحت زیر منحنی در مســـائل بسیار نامتعـــادل، مانند پیش‌بینی

26-Out-of-Distribution (OOD)
27-Over-smoothing
28-Heterophily
29-Area Under the Curve (AUC)

پیوند در شبکه‌های پراکنده، ممکن است توانایی محدودی در ارزیابی کیفیت بازیابی پیوندهای مثبت داشته باشد؛ (۲) در شرایطی که هدف اصلی شناسایی تعداد محدودی از پیوندهای مهم در رتبه‌های بالاست، مساحت زیر منحنی الزاماً کیفیت پیش‌بینی‌های ابتدایی فهرست رتبه‌بندی شده را به‌خوبی منعکس نمی‌کند [۳۵, ۳۶]؛ (۳) معیارهای مختلف ممکن است رتبه‌بندی متفاوتی از الگوریتم‌ها ارائه دهند و در نتیجه، برتری یک روش بر اساس مساحت زیر منحنی لزوماً به معنای برتری آن بر اساس معیارهایی مانند مساحت زیر منحنی دقت- بازخوانی[۳۰]، سود تجمعی تنزیل‌شده نرمال‌شده[۳۱] یا معیارهای رتبه‌بندی نیست [۳۷]؛ و(٤) اتکا به تنها یک معیار ارزیابی می‌تواند تصویری ناقص از عملکرد مدل ارائه دهد. بنابراین، مطالعات اخیر توصیه می‌کنند که برای ارزیابی جامع‌تر روش‌های پیش‌بینی پیوند، از حداقل دو معیار مکمل استفاده شود، به‌ویژه ترکیبی از مساحت زیر منحنی و یکی از معیارهایی مانند مساحت زیر منحنی دقت- بازخوانی یا سود تجمعی تنزیل شده نرمال‌شده [۳٦, ۳۷].

علاوه‌بر این، برخی پژوهش‌ها به تحلیل ارتباط کاربران با برندهای تجاری در شبکه‌های اجتماعی پرداخته‌اند. نتایج این مطالعات نشان می‌دهد که داده‌های شبکه‌های اجتماعی می‌توانند اطلاعات ارزشمندی درباره رفتار کاربران، علایق مشتریان و اثربخشی فعالیت‌های بازاریابی دیجیتال فراهم کنند [۳۸, ۳۹] با این حال، بررسی عمیق‌تر این دسته از مطالعات چند شکاف جدی را آشکار می‌سازد: (۱) تمرکز توصیفی: بسیاری از این پژوهش‌ها عمدتاً بر تحلیل و توصیف تعاملات تمرکز داشته‌اند و کمتر به توسعه مدل‌های پیش‌بینی برای ارتباط کاربران و نمانام‌ها پرداخته‌اند [٤۰]؛ (۲) عدم بهره‌گیری از ساختار گراف: برخی مطالعات از مدل‌های یادگیری ماشین یا ویژگی‌های رفتاری استفاده کرده‌اند، بدون آنکه ساختار ارتباطی و وابستگی‌های شبکه‌ای را به‌صورت صریح مدل‌سازی کنند؛ (۳) عدم توجه به زمینه تجاری: اگرچه پژوهش‌های جدیدتر به پیش‌بینی تعامل کاربران با محتوای برندها پرداخته‌اند، ترکیب هم‌زمان ویژگی‌های ساختاری شبکه، اطلاعات کاربر و ویژگی‌های زمینه‌ای تعامل همچنان یک چالش مهم است؛ (٤) ارزیابی نادرست: در مسائل پیش‌بینی پیوند، انتخاب روش تقسیم داده می‌تواند بر نتایج ارزیابی تأثیر قابل توجهی داشته باشد و تقسیم تصادفی در شبکه‌های زمان‌مند ممکن است به برآورد بیش از حد خوش‌بینانه عملکرد منجر شود [٤۱]؛ و (۵) عدم تفسیرپذیری: بسیاری از مدل‌های عمیق گرافی از نظر تفسیرپذیری با چالش مواجه‌اند، هر چند مطالعات جدید تلاش کرده‌اند این مسئله را از طریق روش‌های توضیح‌پذیر و تحلیل اهمیت عوامل مؤثر بر پیش‌بینی کاهش دهند [٤۱, ٤۰].

شکاف‌های پژوهشی و نوآوری مقاله حاضر

با وجود پیشرفت‌های اخیر در حوزه یادگیری ماشین، یادگیری عمیق و شبکه‌های عصبی گراف برای تحلیل شبکه‌های اجتماعی، بررسی انتقادی ادبیات پژوهش چندین شکاف پژوهشی اساسی را آشکار می‌سازد:

۱. عدم ترکیب نظام‌مند ویژگی‌های ساختاری گراف با یادگیری ماشین در زمینه تجاری: هنوز مطالعات محدودی به صورت هم‌زمان از ویژگی‌های ساختاری گراف (مانند مرکزیت، رتبهٔصفحه، همسایگی)، الگوریتم‌های یادگیری ماشین نظارت‌شده (مانند جنگل تصادفی، الگوریتم تقویت گرادیانی افراطی[۳۲])، و زمینه تجاری (تعامل با نمانام‌ها) برای پیش‌بینی ارتباط کاربران با نمانام‌ها استفاده کرده‌اند. در بسیاری از پژوهش‌ها، یا تمرکز صرف بر تحلیل ساختار شبکه بوده و یا مدل‌های یادگیری ماشین بدون بهره‌گیری عمیق از ویژگی‌های گرافی به کار گرفته شده‌اند.

۲. ارزیابی نادرست و تک‌معیاری: بسیاری از مطالعات پیشین تنها از یک معیار ارزیابی (معمولاً مساحت زیر منحنی) استفاده کرده‌اند که شواهد اخیر نشان می‌دهد در مسائل نامتعادل مانند پیش‌بینی پیوند، ناکارآمد و

30-Precision–Recall (AUPR)
31-Normalized Discounted Cumulative Gain (NDCG)
32-Extreme Gradient Boosting (XGBoost)

گمراه‌کننده است. این موضوع منجر به نتیجه‌گیری‌های نادرست درباره برتری الگوریتم‌ها شده است.

۳. عدم توجه به پویایی و زمان‌مندی داده‌ها: اکثر پژوهش‌های موجود بر داده‌های ایستا تمرکز داشته و از تقسیم‌بندی تصادفی استفاده می‌کنند که در داده‌های زمان‌مند منجر به نشت اطلاعات آینده به آموزش می‌شود. این موضوع عملکرد مدل‌ها را به صورت غیرواقعی بالا نشان می‌دهد و تعمیم‌پذیری آن‌ها را کاهش می‌دهد.

٤. عدم تفسیرپذیری مدل‌ها: بسیاری از مدل‌های پیشرفته (به ویژه شبکه عصبی گراف[۳۳]) به صورت جعبه سیاه عمل می‌کنند و بینش عملی درباره عوامل مؤثر بر پیش‌بینی‌ها فراهم نمی‌کنند. این موضوع در کاربردهای تجاری مانند بازاریابی دیجیتال، که نیاز به درک چرایی پیش‌بینی‌ها وجود دارد، مشکل‌ساز است.

۵. عدم تعادل کلاسی و پراکندگی داده‌ها: در شبکه‌های اجتماعی واقعی، تعاملات کاربران با نمانام‌ها بسیار پراکنده است (کلاس مثبت بسیار کمتر از کلاس منفی). بسیاری از مطالعات پیشین این عدم تعادل را به درستی مدیریت نکرده‌اند و از تکنیک‌های مناسب (مانند نمونه‌برداری مجدد، وزن‌دهی کلاس‌ها، یا معیارهای ارزیابی مناسب) بهره نبرده‌اند.

بنابراین، خلأ پژوهشی اصلی که این مقاله به آن می‌پردازد، عدم وجود چارچوبی ترکیبی، تفسیرپذیر و به‌درستی ارزیابی شده است که بتواند به‌صورت هم‌زمان از قابلیت‌های نظریه گراف (ویژگی‌های ساختاری غنی)، یادگیری ماشین (مدل‌های نظارت‌شده پیشرفته)، و زمینه تجاری (تعامل با برندها) برای پیش‌بینی دقیق و قابل تفسیر ارتباطات کاربران با نمانام‌ها در شبکه‌های اجتماعی بهره ببرد.

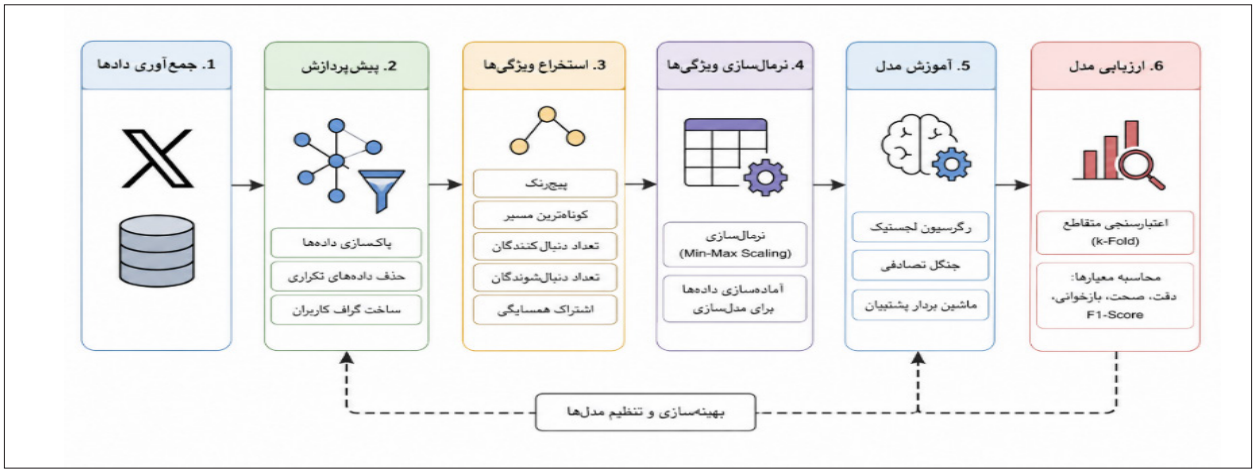
مقاله حاضر با ارائه چنین چارچوبی، چندین سهم پژوهشی کلیدی دارد: (۱) استخراج نظام‌مند ویژگی‌های ساختاری و ارتباطی گراف، شامل رتبهٔصفحه، کوتاه‌ترین مسیر، تعداد دنبال‌کنندگان، تعداد دنبال‌شوندگان و میزان اشتراک همسایگی؛ (۲) ترکیب ویژگی‌های گرافی با الگوریتم‌های یادگیری ماشین نظارت‌شده شامل رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان؛ (۳) ارائه یک چارچوب مقایسه‌ای برای ارزیابی عملکرد الگوریتم‌های مختلف در پیش‌بینی ارتباط کاربران با نمانام‌های تجاری؛ (٤) استفاده از داده‌های واقعی شبکه اجتماعی ایکس مربوط به دنبال‌کنندگان شش نمانام تجاری و ارزیابی چارچوب پیشنهادی در یک سناریوی کاربردی؛ و (۵) تحلیل نقش ویژگی‌های ساختاری و ارتباطی در عملکرد مدل‌ها و شناسایی عوامل مؤثر بر پیش‌بینی ارتباط کاربران با برندها. بدین منظور، خلاصه‌ای از مهمترین پژوهش‌های مرتبط در جدول (۱) ارائه شده است.

جدول ۱. مقایسه مطالعات پیشین با پژوهش حاضر

مرجع	روش/ الگوریتم	ویژگی‌های استفاده شده	نوع داده	محدودیت
فریمن[۳۴] [۱۱]	تحلیل ساختاری گراف	معیارهای مرکزیت	شبکه‌های اجتماعی	عدم توانایی در پیش‌بینی روابط
بلوندل[۳۵] و همکاران [۱۳]	الگوریتم لوون	ساختار اجتماعات	شبکه‌های بزرگ	ماهیت توصیفی و عدم پیش‌بینی ارتباطات
لیبن-ناول[۳۶] و کلاینبرگ[۳۷] [۱۵]	پیش‌بینی پیوند	شباهت ساختاری و همسایگی مشترک	شبکه‌های اجتماعی	عملکرد محدود در روابط غیرخطی
آگاروال۳۸ [۱۲]	داده‌کاوی و تحلیل شبکه	ویژگی‌های ساختاری شبکه	داده‌های شبکه اجتماعی	وابستگی به کیفیت ویژگی‌ها
تانگ[۳۹] و لیو[۴۰] [۱]	تحلیل ترکیبی	ویژگی‌های گراف و داده‌کاوی	شبکه‌های اجتماعی	عدم تمرکز بر برندهای تجاری
گندمی و حیدر [۳۸]	تحلیل داده و یادگیری ماشین	داده‌های رفتاری کاربران	کلان‌داده شبکه‌های اجتماعی	عدم استفاده از ویژگی‌های گراف
پژوهش حاضر	رگرسیون لجستیک، جنگل تصادفی، ماشین بردار پشتیبان	رتبهٔصفحه، کوتاه‌ترین مسیر، اشتراک همسایگی، تعداد دنبال‌کنندگان و دنبال‌شوندگان	شبکه اجتماعی ایکس	

34-Freeman
35-Blondel
36-Liben-Nowell
37-Kleinberg
38-Aggarwal
39-Tang
40-Liu

33-Graph Neural Networks (GNNs)



شکل ۱. معماری کلی پژوهش پیشنهادی

۳- استخراج داده و مدل پیشنهادی

در این پژوهش، یک چارچوب مبتنی بر تحلیل شبکه‌های اجتماعی و الگوریتم‌های یادگیری ماشین برای پیش‌بینی ارتباط کاربران با نمادهای تجاری در شبکه اجتماعی ایکس ارائه شده است. فرایند پیشنهادی شامل جمع‌آوری داده‌ها، پیش‌پردازش، استخراج ویژگی‌های مبتنی بر ساختار گراف، نرمال‌سازی داده‌ها و آموزش مدل‌های یادگیری ماشین است. در این چارچوب، شبکه اجتماعی به صورت یک گراف جهت‌دار مدل‌سازی شده و ویژگی‌های ساختاری کاربران به منظور تحلیل روابط و پیش‌بینی ارتباطات استخراج شده‌اند. در ادامه این فصل، مراحل مختلف، جمع‌آوری داده، پیش‌پردازش، استخراج داده، مدل‌سازی و ارزیابی عملکرد الگوریتم‌ها تشریح می‌شود. به منظور ارائه دیدی کلی از مراحل انجام پژوهش، معماری کلی چارچوب پیشنهادی در شکل (۱) نمایش داده شده است.

شکل (۱) نمای کلی مراحل انجام پژوهش را نشان می‌دهد. در این چارچوب، ابتدا داده‌های مربوط به کاربران و نمادهای تجاری از شبکه اجتماعی ایکس جمع‌آوری شد. سپس داده‌ها طی مرحله پیش‌پردازش پالایش و آماده‌سازی شدند. در ادامه، ویژگی‌های ساختاری و ارتباطی از گراف شبکه استخراج و نرمال‌سازی گردیدند. پس از آن، داده‌های آماده شده به عنوان ورودی مدل‌های یادگیری ماشین شامل رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان مورد استفاده قرار گرفتند. در نهایت، عملکرد مدل‌ها با استفاده از روش اعتبارسنجی متقابل و معیارهای ارزیابی مورد بررسی و مقایسه قرار گرفت.

۳-۱- جمع‌آوری داده‌ها

داده‌های مورد نیاز این پژوهش از شبکه ایکس و با استفاده از رابط برنامه‌نویسی کاربردی این شبکه اجتماعی جمع‌آوری شد. انتخاب این شبکه به دلیل ساختار گراف‌محور ارتباطات، تعداد بالای کاربران فعال و دسترسی به داده‌های تعاملی کاربران صورت گرفت. در این پژوهش، شش نمادهای تجاری شامل آدیداس، نایکی، ریبوک، لاکوست، هوکا و جامپمن ۲۳ انتخاب شدند. دلیل انتخاب این نمادها، فعالیت گسترده در شبکه‌های اجتماعی، تعداد زیاد دنبال‌کنندگان و تعامل بالای کاربران با صفحات رسمی آنها بوده است.

به منظور استخراج ساختار ارتباطی کاربران، داده‌ها در دو سطح جمع‌آوری شدند. در سطح نخست، دنبال‌کنندگان مستقیم هر نماد استخراج شدند. سپس در سطح دوم، دنبال‌کنندگان کاربران سطح اول نیز جمع‌آوری گردیدند تا نمای گسترده‌تری از ساختار شبکه و روابط میان کاربران شکل گیرد. در نتیجه، شبکه نهایی شامل همسایگی‌های مرتبه اول و دوم نمادهای مورد بررسی است. داده‌های

استخراج‌شده در قالب جیسون[41]، به عنوان یکی از قالب‌های استاندارد تبادل داده در سرویس‌های وب، ذخیره شدند. همچنین به منظور حفظ حریم خصوصی کاربران، در تمامی مراحل پردازش تنها از شناسه‌های عددی کاربران استفاده شد. پس از تکمیل فرایند جمع‌آوری و پالایش داده‌ها، شبکه نهایی شامل حدود ۷ هزار گره و بیش از ۳۵ هزار یال ارتباطی بود.

جدول ۲. مشخصات مجموعه‌داده

مشخصه	مقدار
تعداد نمانام‌های مورد بررسی	۶
تعداد گره‌ها	حدود ۷۰۰۰
تعداد یال‌ها	بیش از ۳۵۰۰۰
نوع شبکه	جهت‌دار
سطح جمع‌آوری داده	دنبال‌کنندگان سطح اول و دوم
نوع مسئله	طبقه‌بندی دودویی
کلاس مثبت	وجود ارتباط
کلاس منفی	عدم وجود ارتباط

همان‌گونه که در جدول (۲) مشاهده می‌شود، مجموعه‌داده مورد استفاده از نظر تعداد گره‌ها و یال‌ها دارای ابعاد مناسبی برای تحلیل ساختار شبکه‌های اجتماعی است. این داده‌ها مبنای استخراج ویژگی‌های ساختاری، تشکیل مجموعه داده نهایی و آموزش مدل‌های یادگیری ماشین در پژوهش حاضر قرار گرفتند.

۳-۲- تشکیل نمونه‌ها و برچسب‌گذاری روابط

پس از جمع‌آوری و تشکیل گراف اولیه، به منظور تبدیل مسئله پیش‌بینی ارتباط به یک مسئله طبقه‌بندی دودویی، نمونه‌های آموزشی بر اساس جفت‌های گره مبدأ و مقصد تشکیل شدند. در این فرایند، هر جفت گره که در شبکه مشاهده‌شده دارای یک رابطه دنبال‌کردن بود، به‌عنوان نمونه مثبت در نظر گرفته شد و برچسب ۱ دریافت کرد. بر اساس شبکه نهایی، حدود ۳۵,۰۰۰ نمونه مثبت از روابط موجود میان کاربران و نمانام‌ها استخراج شد.

41-JavaScript Object Notation - JSON

در مقابل، جفت‌گره‌هایی که در گراف مشاهده شده دارای رابطه مستقیم نبودند، به عنوان نمونه‌های منفی در نظر گرفته شدند. به منظور جلوگیری از افزایش بیش از حد حجم نمونه‌های منفی و ایجاد عدم توازن شدید میان دو کلاس، نمونه‌های منفی با استفاده از روش نمونه‌برداری تصادفی از جفت‌گره‌های فاقد ارتباط ایجاد شدند. جفت‌های منفی از میان جفت‌های ممکنِ گره‌های موجود در شبکه که در مجموعه یال‌های مشاهده‌شده دارای رابطه مستقیم نبودند، به صورت تصادفی انتخاب شدند. پیش از نمونه‌برداری، تمامی یال‌های مشاهده شده از فضای کاندیدا حذف شدند تا هیچ نمونه منفی دارای رابطه واقعی ثبت شده نباشد و از ایجاد برچسب‌گذاری نادرست جلوگیری شود. در مجموع، حدود ۳۵,۰۰۰ نمونه منفی تولید شد.

بنابراین، مجموعه داده نهایی شامل حدود ۷۰,۰۰۰ نمونه بود که تقریباً ۵۰ درصد آنها مربوط به کلاس مثبت و ۵۰ درصد مربوط به کلاس منفی بودند. به این ترتیب، مسئله مورد بررسی به صورت یک مسئله طبقه‌بندی دودویی تعریف شد که در آن، هدف مدل تعیین وجود یا عدم وجود رابطه میان دو گره بود.

جدول ۳. توزیع نمونه‌ها و نحوه برچسب‌گذاری

مشخصه	مقدار
تعداد نمونه‌های مثبت	حدود ۳۵,۰۰۰
تعداد نمونه‌های منفی	حدود ۳۵,۰۰۰
تعداد کل نمونه‌ها	حدود ۷۰,۰۰۰
برچسب نمونه مثبت	۱ (وجود رابطه)
برچسب نمونه منفی	۰ (عدم وجود رابطه)
نسبت کلاس مثبت	حدود ۵۰ درصد
نسبت کلاس منفی	حدود ۵۰ درصد
روش تولید نمونه منفی	نمونه‌برداری تصادفی از جفت‌گره‌های فاقد رابطه

در این تعریف، نمونه مثبت نشان‌دهنده یک رابطه مشاهده شده در شبکه و نمونه منفی نشان‌دهنده یک جفت گره بدون رابطه مشاهده‌شده در داده‌های جمع‌آوری شده است. این نحوه تشکیل نمونه‌ها امکان استفاده از الگوریتم‌های یادگیری ماشین برای یادگیری الگوهای ساختاری مرتبط با وجود یا عدم وجود ارتباط را فراهم می‌کند.

۳-۳- پیش‌پردازش داده‌ها

داده‌های استخراج‌شده از شبکه‌های اجتماعی معمولاً شامل داده‌های ناقص، نوفه‌دار و تکراری هستند و به‌صورت مستقیم قابل استفاده در مدل‌های یادگیری ماشین نیستند. علاوه‌بر این، محدودیت‌های مربوط به رابط برنامه‌نویسی کاربردی بُن‌سازه ایکس و ماهیت پویای ارتباطات کاربران می‌تواند منجر به ایجاد داده‌های ناقص یا نامعتبر شود. به همین دلیل، پیش از مرحله استخراج ویژگی‌ها، عملیات پیش‌پردازش بر روی داده‌ها انجام شد. در این مرحله، داده‌های ناقص، نوفه‌دار و تکراری حذف شدند و ساختار ارتباطی کاربران به صورت یک گراف جهت‌دار مدل‌سازی شد؛ به‌طوری‌که گره‌ها نمایانگر کاربران و یال‌ها نشان‌دهنده روابط دنبال‌کردن میان آنها هستند. در ادامه، ارتباطات نامعتبر و گره‌های فاقد اطلاعات کافی از شبکه حذف شدند تا ساختار نهایی برای تحلیل آماده شود. در نهایت، داده‌های پالایش‌شده برای استخراج ویژگی‌های ساختاری و ورود به مدل‌های یادگیری ماشین مورد استفاده قرار گرفتند.

۴-۳- استخراج ویژگی‌ها

پس از پیش‌پردازش داده‌ها و مدل‌سازی شبکه به‌صورت گراف جهت‌دار، مرحله استخراج ویژگی‌ها با هدف تبدیل ساختار غیرخطی و پیچیده ارتباطات شبکه به یک نمایش برداری[42] قابل استفاده در الگوریتم‌های یادگیری ماشین انجام شد. این مرحله نقش اساسی در انتقال اطلاعات همبندی شبکه به فضای ویژگی ایفا می‌کند. در این پژوهش، مجموعه‌ای از ویژگی‌های مبتنی بر تحلیل ساختار گراف و نظریه شبکه‌های پیچیده برای توصیف روابط بین کاربران و نمانام‌ها استخراج گردید. این ویژگی‌ها به‌گونه‌ای انتخاب شده‌اند که بتوانند هم اطلاعات اهمیت محلی گره‌ها و هم ویژگی‌های ساختاری جهانی شبکه را پوشش دهند. در این راستا، معیار رتبهٔ‌صفحه به‌عنوان یکی از شاخص‌های مبتنی بر اهمیت ساختاری گره‌ها مورد استفاده قرار گرفت. این معیار با در نظر گرفتن اهمیت گره‌های مجاور، میزان تأثیرگذاری هر گره را در ساختار کلی شبکه مدل‌سازی می‌کند. به عبارت دیگر، گره‌هایی که از سوی گره‌های مهمتر ارجاع بیشتری دریافت می‌کنند، دارای اهمیت بالاتری در شبکه خواهند بود. برای سنجش فاصله ساختاری بین گره‌ها، طول کوتاه‌ترین مسیر در گراف محاسبه شد که به عنوان معیاری از نزدیکی همبندی میان کاربران و نمانام‌ها در نظر گرفته می‌شود. این ویژگی امکان مدل‌سازی ارتباطات غیرمستقیم در شبکه را فراهم می‌سازد.

علاوه‌بر این، ویژگی‌های مبتنی بر درجه گره شامل تعداد دنبال‌کنندگان و تعداد دنبال‌شوندگان استخراج گردید که بیانگر سطح نفوذ، میزان تعامل و جایگاه هر گره در شبکه است. این ویژگی‌ها اطلاعات مهمی از توزیع قدرت ارتباطی در شبکه ارائه می‌دهند. در سطح محلی، میزان شباهت ساختاری بین گره‌ها از طریق معیار اشتراک همسایگی محاسبه شد که مطابق رابطه (۱) تعریف می‌شود

$$CN(u,v) = |\Gamma(u) \cap \Gamma(v)| \qquad \text{رابطه (۱)}$$

در این رابطه، $\Gamma(u)$ مجموعه همسایگان گره (u) را نشان می‌دهد و (CN(u,v) تعداد همسایگان مشترک بین دو گره را بیان می‌کند که معیاری از شباهت ساختاری محلی در شبکه محسوب می‌شود. در نهایت، کلیه ویژگی‌های استخراج‌شده در قالب یک ماتریس ویژگی چندبعدی[43] سازمان‌دهی شده و به عنوان ورودی مدل‌های یادگیری ماشین جهت انجام وظیفه طبقه‌بندی مورد استفاده قرار گرفتند. جدول (۳) نمونه خروجی ویژگی‌های استخراج شده را نشان می‌دهد.

جدول (۳) نمونه‌ای از ماتریس ویژگی‌های استخراج‌شده از شبکه را نشان می‌دهد که در آن هر نمونه بیانگر یک رابطه میان گره مبدأ و گره مقصد است. این ماتریس شامل ویژگی‌های ساختاری و مبتنی بر همسایگی است که برای توصیف تعاملات میان کاربران و نمانام‌ها استخراج

42-Feature Representation

43-Feature Matrix

جدول (۴). نمونه خروجی از ویژگی‌های استخراج شده

گره مبدا	گره مقصد	وضعیت ارتباط	رتبهٔصفحه مبدا	رتبهٔصفحه مقصد	کوتاه‌ترین مسیر	دنبال کنندگان مبدا	دنبال شونده‌گان مبدا	دنبال کنندگان مقصد	دنبال شونده‌گان مقصد	اشتراک دنبال کنندگان مقصد	اشتراک دنبال شونده‌گان مقصد
۳۰۰۱۱۴۶۳۴	۱۲۷۴۵۶۳۲۴۸	۱	۰,۰۰۰۰۲۶	۰,۰۰۰۰۱۳	-۱	۰	۳	۰	۳	۰	۰
۳۰۰۱۱۴۶۳۴	۲۷۱۷۴۵۶۰۰۷	۰	۰,۰۰۰۰۲۳	۰,۰۰۰۰۱۴	-۱	۰	۳	۰	۱۵	۰	۰
۳۰۰۱۱۴۶۳۴	۱۵۲۰۳۳۹۷۱۸۴۹۱۱۷۲۸۶۵	۱	۰,۰۰۰۰۲۸	۰,۰۰۰۰۲۴	۴	۰	۳	۲	۰	۰	۰
۳۰۰۱۱۴۶۳۴	۳۸۴۶۴۱۹۵۳۷	۱	۰,۰۰۰۰۲۷	۰,۰۰۰۰۲۹	-۱	۰	۳	۰	۲۰	۰	۰
۳۰۰۱۱۴۶۳۴	۱۹۵۰۹۶۰۱۹	۱	۰,۰۰۰۰۲۱	۰,۰۰۰۰۲۶	۷	۰	۳	۵	۴	۰	۰
۳۰۰۱۱۴۶۳۴	۲۵۶۲۹۲۶۴۴	۱	۰,۰۰۰۰۲۲	۰,۰۰۰۰۱۹	-۱	۰	۰	۰	۱	۰	۵
۳۰۰۱۱۴۶۳۴	۱۵۶۳۱۳۲۷۲۳۳۸۴۱۷۲۵۴۵	۱	۰,۰۰۰۰۱۳	۰,۰۰۰۰۱۶	-۱	۰	۳	۰	۱۸	۰	۱
۳۰۰۱۱۴۶۳۴	۱۱۷۵۵۸۲۳۵۷۷۲۳۷۹۹۵۵۳	۰	۰,۰۰۰۰۱۶	۰,۰۰۰۰۴۲	-۱	۰	۳	۱۶	۱۵	۰	۰
۳۰۰۱۱۴۶۳۴	۱۰۸۹۱۷۴۲۲۱۹۸۷۶۸۴۳۵۲	۱	۰,۰۰۰۰۱۸	۰,۰۰۰۰۵۰	۵	۰	۳	۵	۰	۰	۲
۳۰۰۱۱۴۶۳۴	۳۸۵۴۶۴۵۷۱۷	۱	۰,۰۰۰۰۲۲	۰,۰۰۰۰۲۰	۳	۰	۳	۸	۱۳	۰	۰

شده‌اند. در این چارچوب، ویژگی‌هایی نظیر رتبهٔصفحه، تعداد دنبال‌کنندگان و دنبال‌شوندگان، طول کوتاه‌ترین مسیر و میزان اشتراک همسایگی برای هر جفت گره محاسبه شده‌اند. این ویژگی‌ها به منظور نمایش اهمیت ساختاری، فاصله همبندی و شباهت محلی در شبکه مورد استفاده قرار گرفته‌اند. همچنین مقدار «۱-» در ویژگی کوتاه‌ترین مسیر نشان‌دهنده عدم وجود مسیر ارتباطی بین دو گره در گراف است.

۳-۵- نرمال‌سازی ویژگی‌ها

پس از استخراج ویژگی‌ها و تشکیل ماتریس ویژگی، مقادیر به‌دست آمده دارای مقیاس‌های متفاوتی بودند که می‌توانستند باعث ایجاد سوگیری در عملکرد الگوریتم‌های یادگیری ماشین شوند. به‌عنوان مثال، برخی ویژگی‌ها مانند تعداد دنبال‌کنندگان دارای مقادیر بزرگتر و برخی ویژگی‌ها مانند رتبهٔصفحه دارای مقادیر بسیار کوچک بودند. این اختلاف مقیاس می‌تواند منجر به کاهش دقت مدل و ناپایداری در فرایند آموزش شود. به منظور رفع این مشکل، از روش نرمال‌سازی مینیمال- ماکسیمال برای هم‌مقیاس‌سازی ویژگی‌ها استفاده شد. در این روش، مقادیر هر ویژگی به بازه [۰, ۱] تبدیل می‌شوند تا تمامی ویژگی‌ها در یک مقیاس یکسان قرار گیرند و اثرگذاری آنها در فرایند یادگیری متعادل شود. فرمول نرمال‌سازی به‌صورت رابطه (۲) تعریف می‌شود:

رابطه (۲)

$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

در این رابطه، (x) مقدار اولیه ویژگی، X_{min} و X_{max} به‌ترتیب حداقل و حداکثر مقدار آن ویژگی در مجموعه داده هستند و x' مقدار نرمال‌شده را نشان می‌دهد. پس از اعمال این تبدیل، تمامی ویژگی‌ها در بازه یکسان قرار گرفته و داده‌های نهایی برای ورود به مدل‌های یادگیری ماشین آماده شدند.

۳-۶- مدل‌های یادگیری ماشین

پس از استخراج و نرمال‌سازی ویژگی‌ها، سه الگوریتم یادگیری ماشین شامل رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان برای پیش‌بینی وجود یا عدم وجود ارتباط میان گره‌ها در شبکه اجتماعی مورد استفاده قرار گرفتند. هدف این مدل‌ها، یادگیری الگوهای نهفته در

ویژگی‌های ساختاری و تعیین ارتباط میان کاربران و برندهای تجاری است. در رگرسیون لجستیک، احتمال تعلق هر نمونه به کلاس مثبت با استفاده از تابع سیگموید[44] مدل‌سازی می‌شود. این روش به دلیل سادگی، قابلیت تفسیر بالا و کارایی مناسب در مسائل طبقه‌بندی دودویی به‌عنوان مدل پایه مورد استفاده قرار گرفت.

جنگل تصادفی به‌عنوان یک روش تجمیعی، با ترکیب مجموعه‌ای از درخت‌های تصمیم، قادر به مدل‌سازی روابط غیرخطی میان ویژگی‌ها بوده و در برابر داده‌های نوفه عملکرد پایدارتری ارائه می‌دهد. ماشین بردار پشتیبان نیز با یافتن ابرصفحه بهینه و بیشینه‌سازی حاشیه بین کلاس‌ها، برای داده‌های با ابعاد بالا و ساختار پیچیده مناسب بوده و دقت بالایی در تفکیک کلاس‌ها ارائه می‌کند. در نهایت، عملکرد مدل‌ها با استفاده از معیارهای ارزیابی مناسب مورد سنجش و مقایسه قرار گرفت تا بهترین مدل برای مسئله پیش‌بینی ارتباط انتخاب شود.

۳-۷- تنظیم پارامترهای مدل‌ها

به منظور اطمینان از قابلیت بازتولید نتایج و ایجاد شرایط یکسان برای مقایسه مدل‌ها، پارامترهای اصلی الگوریتم‌های یادگیری ماشین در این پژوهش مشخص و تنظیم شدند. برای رگرسیون لجستیک، مقدار پارامتر C برابر با ۱ و حل‌کننده LBFGS با حداکثر ۱۰۰۰ تکرار مورد استفاده قرار گرفت. در مدل جنگل تصادفی، تعداد درخت‌ها برابر با ۲۰۰، حداکثر عمق درخت‌ها برابر با ۲۰، معیار تقسیم جینی[45]، حداقل تعداد نمونه برای تقسیم یک گره برابر با ۲ و حداقل تعداد نمونه در هر برگ برابر با ۱ در نظر گرفته شد. همچنین، تعداد ویژگی‌های مورد استفاده در هر تقسیم بر اساس ریشه دوم تعداد کل ویژگی‌ها[46] تعیین شد.

برای ماشین بردار پشتیبان، از تابع هسته پایه شعاعی[47] استفاده شد و مقدار پارامتر C برابر با ۱ و پارامتر γ برابر با مقیاس خودکار[48] تنظیم گردید. حداکثر تعداد تکرارهای الگوریتم نیز ۱۰۰۰ در نظر گرفته شد. تنظیمات استفاده‌شده برای هر سه مدل در جدول ۵ خلاصه شده‌اند.

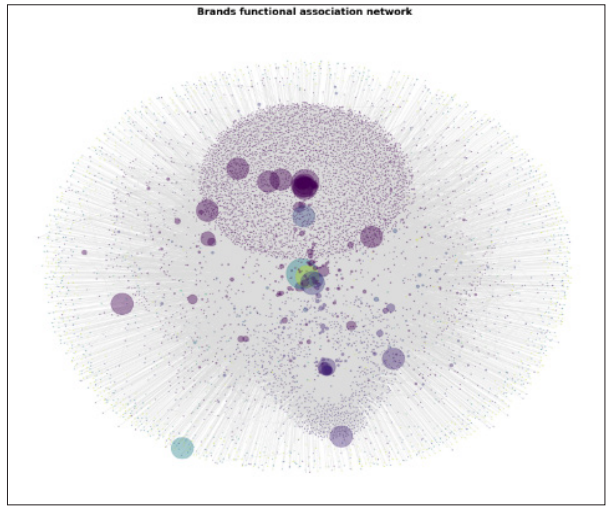
جدول ۵. پارامترهای تنظیم‌شده مدل‌های یادگیری ماشین

مدل	پارامتر	مقدار
رگرسیون لجستیک	حل‌کننده[49]	LBFGS
	ضریب منظم‌سازی C	۱,۰
	حداکثر تعداد تکرار	۱۰۰۰
جنگل تصادفی	تعداد درخت‌ها	۲۰۰
	حداکثر عمق درخت	۲۰
	معیار تقسیم	جینی
	حداقل تعداد نمونه برای تقسیم	۲
	حداقل تعداد نمونه در برگ	۱
	تعداد ویژگی‌ها در هر تقسیم	ریشه دوم تعداد کل ویژگی‌ها
ماشین بردار پشتیبان	نوع هسته	تابع هسته پایه شعاعی
	پارامتر C	۱,۰
	پارامتر γ	مقیاس خودکار
	حداکثر تعداد تکرار	۱۰۰۰

۳-۶- معیارهای ارزیابی

برای ارزیابی عملکرد مدل‌های یادگیری ماشین در مسئله پیش‌بینی ارتباط میان گره‌ها در شبکه اجتماعی، از معیارهای استاندارد شامل دقت[50]، صحت[51] و معیار F1[52] استفاده شد. این معیارها امکان سنجش میزان کارایی مدل‌ها در تشخیص صحیح روابط و ارزیابی کیفیت پیش‌بینی را فراهم می‌کنند. دقت به‌عنوان یک معیار کلی، نسبت پیش‌بینی‌های صحیح به کل نمونه‌ها را نشان می‌دهد و بیانگر عملکرد کلی مدل در طبقه‌بندی ارتباط‌ها است. این معیار دید جامعی از میزان درستی پیش‌بینی‌ها ارائه می‌دهد، اما به‌تنهایی قادر به بررسی دقیق خطاهای نوع اول و دوم نیست. صحت نشان‌دهنده نسبت پیش‌بینی‌های مثبت صحیح به کل نمونه‌هایی است که به‌عنوان مثبت پیش‌بینی شده‌اند و میزان خطای مثبت کاذب را مشخص

44-Sigmoid
45-Gini
46-Sqrt
47-Radial Basis Function (RBF)
48-Scale
49-Solver
50-Precision
51-Recall
52-F1-score



شکل ۲. نمایش گراف ارتباط بین کاربران و برندها

می‌کند. این معیار برای تحلیل میزان اطمینان مدل در پیش‌بینی ارتباط‌های صحیح اهمیت دارد. معیار F1 نیز به‌عنوان میانگین هارمونیک دقت و صحت، برای ایجاد تعادل بین این دو معیار مورد استفاده قرار می‌گیرد. این معیار به‌ویژه در شرایطی که داده‌ها دارای عدم توازن باشند، نقش مهمی در ارزیابی عملکرد مدل ایفا می‌کند. در نهایت، از این معیارها برای مقایسه عملکرد مدل‌های مختلف و انتخاب بهترین مدل جهت پیش‌بینی ارتباط در شبکه اجتماعی استفاده شد.

۴- نتایج و بحث

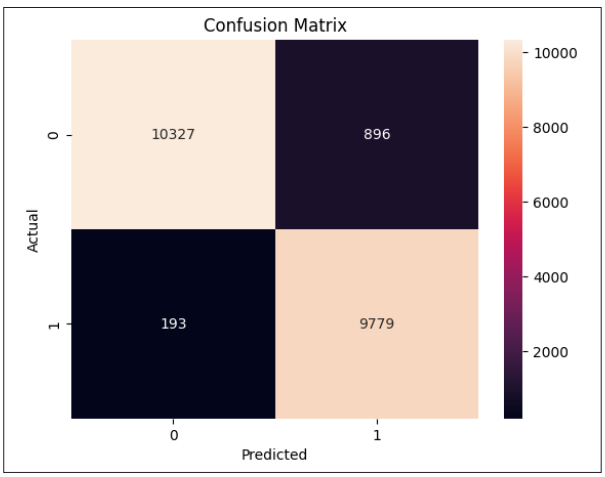
در این پژوهش، داده‌های استخراج شده از شبکه اجتماعی ایکس برای تحلیل ارتباط میان کاربران و نمانام‌های تجاری مورد استفاده قرار گرفت. با توجه به حجم بالای داده، برای نمایش بصری تنها بخشی از شبکه به‌عنوان نمونه ترسیم شد تا ساختار کلی ارتباطات قابل مشاهده باشد. این نمایش، دید کلی از توزیع گره‌ها و یال‌ها در شبکه ارائه می‌دهد.

۴-۱- نتایج مدل رگرسیون لجستیک

برای ارزیابی مدل‌های یادگیری ماشین، از روش اعتبارسنجی متقابل ده‌تایی استفاده شد. در این روش، داده‌ها به ده بخش تقسیم شدند و در هر مرحله، یک بخش به‌عنوان داده آزمون و نه بخش دیگر برای آموزش مدل مورد استفاده قرار گرفت. این فرایند باعث کاهش سوگیری در ارزیابی و افزایش پایداری نتایج شد. مدل رگرسیون لجستیک با استفاده از داده‌های اعتبارسنجی متقابل ده‌تایی آموزش داده شد و نتایج آن در جدول زیر ارائه شده است

جدول (۶). نتایج مدل رگرسیون لجستیک

	دقت	صحت	معیار F1
۰	۰/۹۸	۰/۹۲	۰/۹۵
۱	۰/۹۲	۰/۹۸	۰/۹۵



شکل (۳). ماتریس درهم‌ریختگی مدل رگرسیون لجستیک

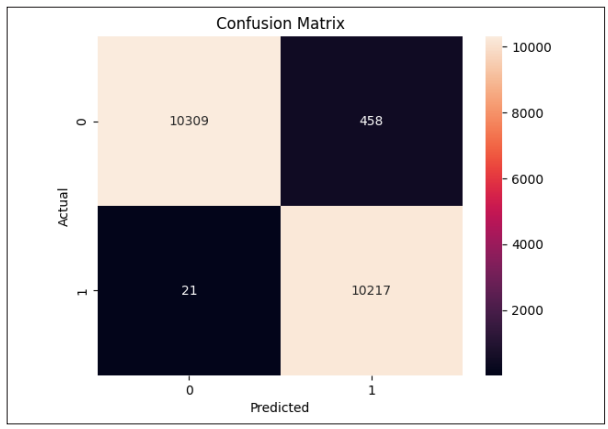
نتایج مدل رگرسیون لجستیک نشان می‌دهد که ویژگی‌های استخراج شده از ساختار گراف توانایی مناسبی در تفکیک ارتباطات میان کاربران و نمانام‌ها دارند. مقدار بالای معیار F1 در هر دو کلاس بیانگر تعادل مناسب میان دقت و صحت مدل است. همچنین، مقادیر به‌دست آمده نشان می‌دهد که روابط موجود در شبکه اجتماعی تا حد زیادی قابل یادگیری و پیش‌بینی هستند.

۴-۲- نتایج مدل جنگل تصادفی

در ادامه، مدل جنگل تصادفی با همان تنظیمات اعتبارسنجی اجرا شد. این مدل با بهره‌گیری از ساختار تجمیعی، توانایی مناسبی در یادگیری روابط غیرخطی میان ویژگی‌ها دارد.

جدول (۷). نتایج مدل جنگل تصادفی

	دقت	صحت	معیار F1
۰	۰/۹۸	۰/۹۶	۰/۹۷
۱	۰/۹۶	۰/۹۸	۰/۹۷



شکل (۴). ماتریس درهم‌ریختگی مدل جنگل تصادفی

مدل جنگل تصادفی بهترین عملکرد را در میان مدل‌های مورد بررسی ارائه داد. این موضوع نشان می‌دهد که وجود روابط غیرخطی میان ویژگی‌های ساختاری شبکه نقش مهمی در پیش‌بینی ارتباط کاربران با برندها دارد. ساختار تجمیعی این الگوریتم باعث کاهش خطای طبقه‌بندی و افزایش پایداری مدل شده است.

۳-۴- نتایج مدل ماشین بردار پشتیبان

در این بخش، مدل ماشین بردار پشتیبان برای طبقه‌بندی روابط میان گره‌ها مورد استفاده قرار گرفت.

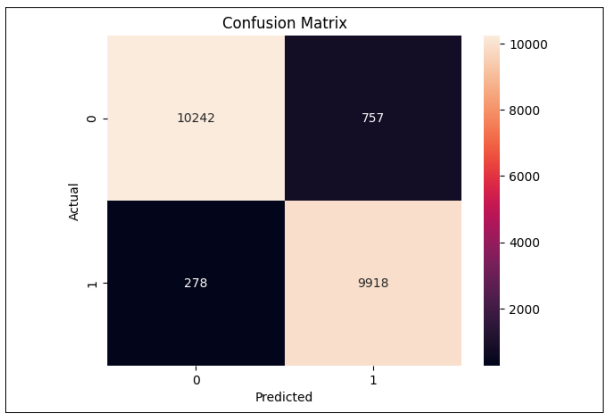
نتایج مدل ماشین بردار پشتیبان نشان می‌دهد که این الگوریتم نیز توانایی مناسبی در تفکیک کلاس‌ها دارد. با این حال، عملکرد آن نسبت به جنگل تصادفی اندکی ضعیف‌تر است که می‌تواند ناشی از پیچیدگی و غیرخطی بودن روابط موجود در شبکه اجتماعی باشد.

۴-۴- بحث و مقایسه نتایج

با مقایسه نتایج سه مدل مورد بررسی شامل رگرسیون لجستیک، جنگل تصادفی و ماشین بردار پشتیبان، مشاهده می‌شود که مدل جنگل تصادفی با دقت کلی ۹۷ درصد

جدول ۸. نتایج مدل ماشین بردار پشتیبان

	دقت	صحت	معیار F1
۰	۰/۹۷	۰/۹۳	۰/۹۵
۱	۰/۹۳	۰/۹۷	۰/۹۵



شکل ۵. ماتریس درهم‌ریختگی مدل ماشین بردار پشتیبان

عملکرد بهتری نسبت به سایر مدل‌ها داشته است. مدل‌های رگرسیون لجستیک و ماشین بردار پشتیبان نیز هر دو با دقت ۹۵ درصد نتایج قابل قبولی ارائه کرده‌اند.

برتری مدل جنگل تصادفی را می‌توان به توانایی این الگوریتم در مدل‌سازی روابط غیرخطی و پیچیده میان ویژگی‌های شبکه نسبت داد. یکی دیگر از دلایل عملکرد بهتر جنگل تصادفی، توانایی این مدل در شناسایی اثرات متقابل و غیرخطی میان ویژگی‌های ساختاری شبکه است. در حالی که رگرسیون لجستیک عمدتاً روابط خطی میان ویژگی‌ها و متغیر هدف را مدل‌سازی می‌کند، جنگل تصادفی با استفاده از مجموعه‌ای از درخت‌های تصمیم می‌تواند ترکیب‌های مختلفی از ویژگی‌هایی مانند رتبهٔصفحه، تعداد دنبال‌کنندگان، تعداد دنبال‌شوندگان، کوتاه‌ترین مسیر و اشتراک همسایگی را به صورت هم‌زمان در نظر بگیرد. این ویژگی به‌ویژه در شبکه‌های اجتماعی که روابط میان ویژگی‌های همبندی معمولاً پیچیده و غیرخطی هستند، می‌تواند به تفکیک بهتر نمونه‌های مثبت و منفی منجر شود.

در شبکه‌های اجتماعی، ارتباط میان کاربران معمولاً دارای ساختارهای پیچیده و الگوهای غیرخطی است؛ به همین دلیل، مدل جنگل تصادفی با استفاده از مجموعه‌ای از درخت‌های تصمیم قادر است تعامل میان ویژگی‌های مختلف را بهتر یاد بگیرد. علاوه بر این، این الگوریتم

نسبت به داده‌های نوفه‌دار و تغییرات موجود در شبکه مقاومت بیشتری دارد و می‌تواند تعمیم‌پذیری مناسب‌تری نسبت به سایر مدل‌ها ارائه دهد. نتایج این پژوهش نشان می‌دهد که ترکیب ویژگی‌های ساختاری استخراج شده از شبکه اجتماعی با الگوریتم‌های یادگیری ماشین می‌تواند در پیش‌بینی ارتباط میان کاربران و برندهای تجاری عملکرد مؤثری داشته باشد. همچنین، استفاده همزمان از مفاهیم نظریه گراف و مدل‌های یادگیری ماشین نسبت به روش‌های صرفاً ساختاری، دقت بالاتری در تحلیل رفتار کاربران و روابط شبکه‌ای فراهم می‌کند.

به منظور بررسی آماری تفاوت عملکرد مدل‌ها و اطمینان از معنادار بودن برتری جنگل تصادفی، آزمون ویلکاکسون[53] به‌صورت جفتی بین مدل جنگل تصادفی و هر یک از دو مدل رگرسیون لجستیک و ماشین بردار پشتیبان انجام شد. با توجه به ماهیت جفت‌شده نتایج حاصل از اعتبارسنجی متقابل، مقادیر معیار F1 حاصل از ده بخشی اعتبارسنجی برای هر مدل به‌صورت جفتی مورد مقایسه قرار گرفتند. نتایج آزمون در جدول (۹) ارائه شده است.

جدول ۹. نتایج آزمون آماری ویلکاکسون برای مقایسه مدل‌ها

مقایسه مدل‌ها	مقدار p	نتیجه
جنگل تصادفی- رگرسیون لجستیک	۰,۰۰۱ >	تفاوت معنادار
جنگل تصادفی- ماشین بردار پشتیبان	۰,۰۰۱ >	تفاوت معنادار

همان‌گونه که در جدول (۹) مشاهده می‌شود، تفاوت عملکرد مدل جنگل تصادفی با هر دو مدل رگرسیون لجستیک و ماشین بردار پشتیبان از نظر آماری معنادار است. مقدار p در هر دو مقایسه کمتر از ۰/۰۰۱ و در نتیجه کمتر از سطح معناداری ۰/۰۵ است؛ بنابراین، فرض صفر مبنی بر نبود تفاوت معنادار بین عملکرد مدل‌ها رد می‌شود. این نتیجه نشان می‌دهد که برتری مشاهده شده برای جنگل تصادفی صرفاً ناشی از تغییرات تصادفی در فرایند ارزیابی نبوده و از نظر آماری نیز پشتیبانی می‌شود.

نتایج این پژوهش نشان می‌دهد که ترکیب ویژگی‌های ساختاری استخراج شده از شبکه اجتماعی با الگوریتم‌های یادگیری ماشین می‌تواند در پیش‌بینی ارتباط میان کاربران و نشان‌های تجاری عملکرد مؤثری داشته باشد. همچنین، استفاده همزمان از مفاهیم نظریه گراف و مدل‌های یادگیری ماشین نسبت به روش‌های صرفاً ساختاری، دقت بالاتری در تحلیل رفتار کاربران و روابط شبکه‌ای فراهم می‌کند.

۴-۴-۱- تحلیل اهمیت ویژگی‌ها و عوامل مؤثر بر پیش‌بینی

به منظور بررسی نقش ویژگی‌های استخراج‌شده در پیش‌بینی ارتباط میان کاربران و برندهای تجاری، اهمیت نسبی ویژگی‌های مورد استفاده در مدل جنگل تصادفی مورد بررسی قرار گرفت. از آنجا که جنگل تصادفی بر اساس کاهش ناخالصی درخت‌های تشکیل‌دهنده خود می‌تواند سهم نسبی ویژگی‌ها را در فرایند تصمیم‌گیری محاسبه کند، از مقادیر اهمیت ویژگی‌های این مدل برای شناسایی عوامل مؤثر بر پیش‌بینی استفاده شد. اهمیت هر ویژگی بر اساس میانگین کاهش ناخالصی[54] در درخت‌های تشکیل‌دهنده جنگل تصادفی محاسبه و سپس به‌صورت نسبی گزارش شد. نتایج این تحلیل در جدول (۹) ارائه شده است.

جدول ۱۰. اهمیت نسبی ویژگی‌ها در مدل جنگل تصادفی

رتبه	ویژگی	اهمیت نسبی
۱	اشتراک همسایگی	۰/۲۴
۲	رتبهٔ‌صفحه گره مقصد	۰/۱۹
۳	دنبال‌کنندگان گره مقصد	۰/۱۵
۴	دنبال‌شوندگان گره مبدأ	۰/۱۱
۵	رتبهٔ‌صفحه گره مبدأ	۰/۱۰
۶	دنبال‌کنندگان گره مبدأ	۰/۰۸
۷	دنبال‌شوندگان گره مقصد	۰/۰۷
۸	کوتاه‌ترین مسیر	۰/۰۶
جمع		۱/۰۰

همان‌گونه که در جدول (۹) مشاهده می‌شود، ویژگی اشتراک همسایگی با اهمیت نسبی ۰/۲۴ بیشترین نقش را در پیش‌بینی وجود ارتباط میان گره‌ها داشته است. این نتیجه

53-Wilcoxon Signed-Rank Test

54-Mean Decrease in Impurity

نشـــان می‌دهد که شباهت ســـاختاری و وجود همسایگان مشـــترک می‌تواند شـــاخص مهمی برای شناسایی روابط میان کاربران و نمانام‌ها باشد. پس از آن، رتبهٔصفحه گره مقصد و تعداد دنبال‌کنندگان گره مقصد به ترتیب با مقادیر ۰/۱۹ و ۰/۱۵ در رتبه‌هـــای بعـــدی قـــرار گرفته‌اند. اهمیت بالاتر این ویژگی‌ها نشـــان می‌دهد که جایگاه ســـاختاری و میزان نفوذ گره مقصد در شـــبکه می‌تواند در پیش‌بینی ارتباط نقش قابل‌توجهی داشته باشد.

همچنین، ویژگی‌های مربوط به گره مبدأ، شـــامل تعداد دنبال‌شـــوندگان، رتبهٔصفحه و تعداد دنبال‌کنندگان، سهم قابل‌توجـــهی در تصمیم‌گیـــری مدل داشـــته‌اند. در مقابل، ویـــژگی کوتاه‌ترین مسیر بـــا اهمیت نســـبی ۰/۰۶، نقش محدودتری در مقایسه با سایر ویژگی‌ها نشان داده است. این یافته بیانگر آن است که در مجموعه‌داده مورد بررسی، شـــاخص‌های مربوط به اهمیت گره و شباهت محلی شبکه نسبت به صرف فاصله همبندی، اطلاعات متمایزکننده‌تری برای پیش‌بینی ارتباط فراهم می‌کنند.

در مجمـــوع، تحلیل اهمیت ویژگی‌ها نشـــان می‌دهد که ویژگی‌هـــای ســـاختاری و محلی گراف، به‌ویژه اشـــتراک همســـایگی و شـــاخص‌های اهمیت گره، نقـــش مهمی در پیش‌بینی ارتباط کاربران بـــا نمانام‌های تجاری دارند. این نتایـــج همچنین می‌تواند یـــکی از دلایل عملکرد بهتر جنگل تصادفی نســـبت به رگرسیون لجســـتیک و ماشین بردار پشتیبان باشـــد، زیرا این مدل قادر است اثرات غیرخطی و تعامل میـــان ویژگی‌های مختلف را در فرایند تصمیم‌گیری لحـــاظ کند. بنابرایـــن، اســـتفاده هم‌زمـــان از ویژگی‌های ساختاری گراف و الگوریتم‌های یادگیری ماشین می‌تواند علاوه بر بهبود عملکرد پیش‌بینی، بینش قابل‌توجهی درباره عوامل مؤثر بر ارتباط کاربران با نمانام‌ها فراهم کند.

۴-۴-۲- مقایسه با روش‌های نوین پیش‌بینی پیوند

با توجه به توســـعه روش‌های مبتنی بر یادگیری نمایش گراف و شبکه‌های عصبی گراف در سال‌های اخیر، روش‌هایی ماننـــد GAT، GCN، GraphSAGE، DeepWalk، Node2Vec و SEAL در مســـائل مختلف پیش‌بینی پیوند مورد توجه قرار گرفته‌انـــد. این روش‌ها بـــا بهره‌گیری از ســـاختار گراف، اطلاعات همســـایگی و نمایش‌های برداری یا معماری‌های یادگیری عمیق، قادر به اســـتخراج الگوهای پیچیده موجود در شبکه هســـتند. به منظور تبیین جایگاه رویکرد حاضر در میان روش‌های رایج پیش‌بینی پیوند، ویژگی‌های اصلی این روش‌هـــا و تفاوت آنها با رویکـــرد حاضر به‌صورت مقایســـه‌ای بررسی شده‌اند. خلاصه این مقایسه در جدول (۱۱) ارائه شده است.

مقایســـه روش‌هـــای فوق نشـــان می‌دهـــد که بخش قابل‌توجهی از روش‌های نوین پیش‌بینی پیوند بر یادگیری خودکار نمایش گراف یا استفاده از معماری‌های عمیق برای استخراج الگوهای پیچیده شبکه متمرکز هستند. برای مثال، روش‌هـــای Node2Vec و DeepWalk تلاش می‌کنند نمایش برداری مناســـبی از گره‌ها ایجاد کنند، در حالی که Graph-GCN، SAGE و GAT با اســـتفاده از معماری‌های مبتنی بر شـــبکه‌های عصبی گراف، اطلاعات ساختاری و همسایگی گره‌ها را به‌صورت لایه‌ای یاد می‌گیرند. روش SEAL نیز با تمرکز بر زیرگراف پیرامون جفت گره‌های مورد بررسی، اطلاعات محلی مرتبط با وجود پیوند را استخراج می‌کند.

در مقابل، رویکرد حاضر بـــه جای اتکا به نمایش‌های بـــرداری یا معماری‌هـــای عمیق گـــراف، از مجموعه‌ای از ویژگی‌هـــای ســـاختاری قابل تفسیر، شـــامل رتبهٔصفحه، تعـــداد دنبال‌کنندگان، تعـــداد دنبال‌شـــوندگان، کوتاه‌ترین مسیر و اشتراک همســـایگی استفاده می‌کند. این ویژگی‌ها مســـتقیماً بیانگر خصوصیـــات همبندی گره‌هـــا و روابط میان آنها هستند و امکان تحلیل نقش هر ویژگی در فرایند پیش‌بینی را فراهـــم می‌کنند. همان‌طور که در بخش تحلیل اهمیت ویژگی‌ها نشان داده شد، ویژگی اشتراک همسایگی و شـــاخص‌های اهمیت گره از مهمتریـــن عوامل مؤثر در پیش‌بینی روابط در مجموعه‌داده مورد بررسی بوده‌اند.

از دیـــدگاه تفسیرپذیـــری، رویکـــرد حاضـــر دارای ویژگی‌هایی است که آن را از روش‌های مبتنی بر یادگیری عمیق متمایز می‌کند. روش‌های مبتنی بر شبکه‌های عصبی

جدول ۱۱: مقایسه رویکرد پیشنهادی با روش‌های نوین پیش‌بینی پیوند

روش	نوع رویکرد	نحوه استخراج اطلاعات گراف	مزیت اصلی	تفاوت با رویکرد حاضر
Node2Vec	یادگیری نمایش گراف	تولید embedding با جست‌وجوی تصادفی	یادگیری نمایش انعطاف‌پذیر گره‌ها	استفاده از نمایش برداری به جای ویژگی‌های ساختاری مستقیم
DeepWalk	یادگیری نمایش گراف	پیمایش تصادفی و یادگیری embedding	استخراج الگوهای ساختاری شبکه	وابستگی بیشتر به embedding گره‌ها
GraphSAGE	یادگیری عمیق گراف	تجمیع اطلاعات همسایگان	قابلیت یادگیری نمایش برای گره‌های دیده نشده	استفاده از معماری عمیق و aggregation
GCN	شبکه عصبی گراف	انتشار و تجمیع اطلاعات همسایگان	یادگیری روابط غیرخطی گراف	استفاده از شبکه عصبی گراف به جای ویژگی‌های مهندسی‌شده
GAT	شبکه عصبی گراف مبتنی بر توجه	وزن‌دهی متفاوت به همسایگان	شناسایی اهمیت نسبی همسایگان	استفاده از سازوکار attention
SEAL	پیش‌بینی لینک مبتنی بر زیرگراف	استخراج زیرگراف پیرامون جفت گره‌ها	استفاده مستقیم از ساختار محلی پیرامون پیوند	استفاده از ساختار زیرگراف پیرامون جفت گره‌های نامزد
رویکرد حاضر	یادگیری ماشین مبتنی بر ویژگی	استخراج ویژگی‌های ساختاری و همبندی	سادگی، تفسیرپذیری و قابلیت تحلیل اهمیت ویژگی‌ها	-

گراف معمولاً به تنظیم معماری، تعداد لایه‌ها، ابعاد نمایش برداری، نرخ یادگیری و سایر اَبَرپارامترها نیاز دارند، در حالی که ویژگی‌های مورد استفاده در پژوهش حاضر از مفاهیم شناخته شده نظریه گراف استخراج شده و تفسیر آنها نسبتاً مستقیم است. بنابراین، هدف پژوهش حاضر ارائه یک معماری جدید یادگیری عمیق برای پیش‌بینی پیوند نیست، بلکه بررسی میزان توانایی ویژگی‌های ساختاری قابل تفسیر در ترکیب با الگوریتم‌های یادگیری ماشین برای تشخیص روابط میان کاربران و نمانام‌های تجاری است.

در نتیجه، مقایسه حاضر نشان می‌دهد که رویکرد حاضر از نظر طراحی با روش‌های مبتنی بر ابعاد نمایش برداری و یادگیری عمیق تفاوت دارد. در حالی که روش‌های نوین بر یادگیری خودکار نمایش و الگوهای پیچیده گراف تمرکز می‌کنند، روش حاضر بر استخراج ویژگی‌های ساختاری قابل تفسیر و استفاده از آنها در الگوریتم‌های یادگیری ماشین تأکید دارد. از این منظر، رویکرد حاضر می‌تواند به عنوان یک چارچوب مبتنی بر ویژگی‌های ساختاری برای پیش‌بینی ارتباط در شبکه‌های اجتماعی در نظر گرفته شود که علاوه‌بر عملکرد پیش‌بینی، امکان تحلیل عوامل مؤثر بر تصمیم‌گیری مدل را نیز فراهم می‌کند.

۴-۵- تحلیل زمان اجرای مدل‌ها

به منظور بررسی هزینه محاسباتی مدل‌های مورد استفاده و مقایسه کارایی آنها از نظر زمان اجرا، زمان اجرای رگرسیون لجستیک، ماشین بردار پشتیبان و جنگل تصادفی اندازه‌گیری و مقایسه شد. برای اطمینان از قابلیت مقایسه، هر سه مدل تحت شرایط یکسان و با استفاده از مجموعه ویژگی‌ها و داده‌های یکسان اجرا شدند. زمان اجرای هر مدل از مرحله آموزش تا تکمیل فرایند پیش‌بینی اندازه‌گیری شد. نتایج این تحلیل در جدول (۱۲) ارائه شده است.

جدول ۱۲: مقایسه زمان اجرای مدل‌های یادگیری ماشین

مدل	زمان اجرا (ثانیه)
رگرسیون لجستیک	۱۲/۴
ماشین بردار پشتیبان	۴۸/۷
جنگل تصادفی	۳۱/۶

همان‌گونه که در جدول (۱۲) مشاهده می‌شود، رگرسیون لجستیک با زمان اجرای ۱۲/٤ ثانیه کمترین هزینه محاسباتی را در میان مدل‌های مورد بررسی داشته است. این نتیجه را می‌توان به ساختار نسبتاً ساده این مدل و تعداد محدود پارامترهای آن نسبت داد. در مقابل،

ماشین بردار پشتیبان با زمان اجرای ٤٨/٧ ثانیه بیشترین زمان محاسباتی را به خود اختصاص داده است. این زمان بیشتر می‌تواند ناشی از فرایند محاسباتی پیچیده‌تر این الگوریتم در تعیین مرز تصمیم و پردازش نمونه‌ها باشد.

مدل جنگل تصادفی با زمان اجرای ٣١/٦ ثانیه در جایگاه میانی قرار گرفته است. اگرچه زمان اجرای آن بیشتر از رگرسیون لجستیک است، اما نسبت به ماشین بردار پشتیبان زمان محاسباتی کمتری نیاز دارد. با توجه به این که جنگل تصادفی در این پژوهش بهترین عملکرد پیش‌بینی را با دقت ۰/۹۷ و معیار F1 برابر با ۰/۹۷ ارائه کرده است، می‌توان گفت که این مدل در کنار عملکرد پیش‌بینی بالاتر، هزینه محاسباتی قابل قبولی نیز دارد.

در مجموع، نتایج نشان می‌دهد که انتخاب مدل مناسب تنها بر اساس زمان اجرای کمتر انجام نمی‌شود، بلکه باید میان عملکرد پیش‌بینی و هزینه محاسباتی تعادل برقرار شود. در این پژوهش، جنگل تصادفی با وجود زمان اجرای بیشتر نسبت به رگرسیون لجستیک، به دلیل عملکرد پیش‌بینی بهتر و زمان اجرای قابل قبول، گزینه مناسب‌تری برای مسئله پیش‌بینی ارتباط میان کاربران و نمانام‌های تجاری محسوب می‌شود.

۴-۶-تحلیل حساسیت

به منظور بررسی پایداری عملکرد مدل و میزان تأثیر عوامل مختلف بر نتایج پیش‌بینی، تحلیل حساسیت برای مدل جنگل تصادفی انجام شد. در این تحلیل، سه عامل شامل مجموعه ویژگی‌های مورد استفاده، حجم داده‌های آموزشی و تعداد فولدهای[55] اعتبارسنجی مورد بررسی قرار گرفتند. هدف از این تحلیل، تعیین میزان وابستگی عملکرد مدل به انتخاب ویژگی‌ها، اندازه مجموعه داده و تنظیمات فرایند ارزیابی بود.

۴-۶-۱- تأثیر حذف ویژگی‌ها

به منظور بررسی نقش هر یک از ویژگی‌های ساختاری، هر بار یکی از ویژگی‌ها از مجموعه ویژگی‌های ورودی حذف و مدل جنگل تصادفی مجدداً ارزیابی شد. سایر شرایط آزمایش ثابت نگه داشته شدند. نتایج در جدول (۱۳) ارائه شده است.

همان‌گونه که در جدول (۱۳) مشاهده می‌شود، حذف اشتراک همسایگی بیشترین کاهش عملکرد را ایجاد کرده و معیار F1 از ۰/۹۷ به ۰/۹٤ کاهش یافته است. این نتیجه با تحلیل اهمیت ویژگی‌ها در بخش قبل نیز همخوانی دارد؛ زیرا اشتراک همسایگی با اهمیت نسبی ۰/۲٤ مهمترین ویژگی مدل بوده است. حذف رتبهٔصفحه نیز کاهش محدودی در عملکرد ایجاد کرده است، در حالی که حذف کوتاه‌ترین مسیر تغییر محسوسی در معیار F1 ایجاد نکرده است. این یافته نشان می‌دهد که ویژگی‌های مربوط به ساختار محلی و اهمیت گره‌ها در مجموعه‌داده مورد بررسی، اطلاعات متمایزکننده‌تری برای پیش‌بینی پیوندها فراهم می‌کنند.

۴-۶-۲- تأثیر حجم داده‌ها

برای بررسی میزان حساسیت مدل نسبت به حجم داده‌های مورد استفاده، مدل جنگل تصادفی با استفاده از بخش‌های مختلفی از مجموعه داده آموزش و ارزیابی شد. سایر تنظیمات مدل ثابت نگه داشته شدند. نتایج در جدول (١٤) ارائه شده است.

جدول ۱۴. تأثیر حجم داده بر عملکرد مدل

درصد داده مورد استفاده	دقت	صحت	بازخوانی	معیار F1
%۶۰	۰/۹۴	۰/۹۳	۰/۹۵	۰/۹۴
%۷۰	۰/۹۵	۰/۹۴	۰/۹۶	۰/۹۵
%۸۰	۰/۹۷	۰/۹۶	۰/۹۸	۰/۹۷
%۹۰	۰/۹۷	۰/۹۶	۰/۹۸	۰/۹۷

نتایج نشان می‌دهد که با افزایش حجم داده از ٦٠ به ٨٠ درصد، عملکرد مدل بهبود می‌یابد؛ در حالی که افزایش حجم داده از ٨٠ به ٩٠ درصد تغییر محسوسی در معیارهای ارزیابی ایجاد نمی‌کند. این موضوع نشان می‌دهد که مدل پس از استفاده از بخش عمده داده‌های آموزشی به سطح نسبتاً پایداری از عملکرد می‌رسد.

55-Fold

جدول ۱۳.تحلیل حساسیت نسبت به حذف ویژگی‌ها

مجموعه ویژگی‌ها	دقت	صحت	بازخوانی	معیار F1
همه ویژگی‌ها	۰/۹۷	۰/۹۶	۰/۹۸	۰/۹۷
بدون اشتراک همسایگی	۰/۹۴	۰/۹۴	۰/۹۵	۰/۹۴
بدون رتبهٔصفحه	۰/۹۵	۰/۹۴	۰/۹۶	۰/۹۵
بدون تعداد دنبال‌کنندگان	۰/۹۶	۰/۹۵	۰/۹۷	۰/۹۶
بدون تعداد دنبال‌شوندگان	۰/۹۶	۰/۹۵	۰/۹۷	۰/۹۶
بدون کوتاه‌ترین مسیر	۰/۹۷	۰/۹۶	۰/۹۷	۰/۹۷

۴-۶-۳- تأثیر تعداد فولدهای اعتبارسنجی

در ادامه، به منظور بررسی تأثیر تعداد فولدهای اعتبارسنجی بر نتایج، مدل با تعداد فولدهای مختلف شامل ۵، ۱۰ و ۱۵ اجرا شد. نتایج این آزمایش در جدول (۱۵) ارائه شده است.

جدول ۱۵: تأثیر تعداد فولدهای اعتبارسنجی بر عملکرد مدل

تعداد فولد	دقت	صحت	بازخوانی	معیار F1
۵	۰/۹۶	۰/۹۵	۰/۹۷	۰/۹۶
۱۰	۰/۹۷	۰/۹۶	۰/۹۸	۰/۹۷
۱۵	۰/۹۷	۰/۹۶	۰/۹۸	۰/۹۷

همان‌گونه که در جدول (۱۵) مشاهده می‌شود، افزایش تعداد فولد از ۵ به ۱۰ موجب بهبود جزئی عملکرد مدل شد، به‌طوری‌که معیار F1 از ۰/۹۶ به ۰/۹۷ افزایش یافت. با افزایش تعداد فولد از ۱۰ به ۱۵، معیارهای ارزیابی ثابت باقی ماندند. بنابراین، انتخاب اعتبارسنجی متقابل ده‌تایی ضمن فراهم کردن ارزیابی پایدار، از نظر هزینه محاسباتی نیز انتخاب مناسبی است.

در مجموع، تحلیل حساسیت نشان می‌دهد که عملکرد مدل جنگل تصادفی نسبت به تغییرات در مجموعه ویژگی‌ها، حجم داده و تعداد فولد اعتبارسنجی از پایداری مناسبی برخوردار است. در میان ویژگی‌های مورد بررسی، اشتراک همسایگی بیشترین تأثیر را بر عملکرد مدل داشته است، در حالی که افزایش حجم داده از ۸۰ به ۹۰ درصد و افزایش تعداد فولد از ۱۰ به ۱۵، تغییر محسوسی در معیارهای ارزیابی ایجاد نکردند. این نتایج از پایداری نتایج اصلی پژوهش و انتخاب تنظیمات مورد استفاده پشتیبانی می‌کنند.

۵- نتیجه‌گیری

در این پژوهش، مسئله پیش‌بینی ارتباط کاربران با نمانام‌های تجاری در شبکه‌های اجتماعی با استفاده از ویژگی‌های ساختاری گراف و الگوریتم‌های یادگیری ماشین بررسی شد. داده‌های کاربران و ارتباطات آنها از شبکه اجتماعی ایکس استخراج و پس از پیش‌پردازش، به‌صورت یک گراف جهت‌دار مدل‌سازی شدند. سپس ویژگی‌هایی شامل رتبهٔصفحه، کوتاه‌ترین مسیر، تعداد دنبال‌کنندگان، تعداد دنبال‌شوندگان و اشتراک همسایگی از ساختار شبکه استخراج و برای پیش‌بینی ارتباط میان کاربران و نمانام‌ها مورد استفاده قرار گرفتند.

از دیدگاه علمی، یافته‌های پژوهش نشان می‌دهد که ویژگی‌های ساختاری و همبندی شبکه می‌توانند اطلاعات متمایزی برای شناسایی روابط بالقوه میان گره‌ها فراهم کنند. به‌ویژه، اهمیت بالاتر ویژگی اشتراک همسایگی نشان داد که ساختار محلی شبکه و میزان اشتراک در همسایگان می‌تواند نقش مهمی در تشخیص روابط میان کاربران و نمانام‌ها داشته باشد. همچنین، عملکرد بهتر جنگل تصادفی نسبت به رگرسیون لجستیک و ماشین بردار پشتیبان نشان می‌دهد که توانایی مدل‌سازی روابط غیرخطی و تعامل میان ویژگی‌های ساختاری می‌تواند برای مسئله پیش‌بینی لینک اهمیت داشته باشد. بنابراین، یافته اصلی پژوهش صرفاً دستیابی به عملکرد پیش‌بینی مناسب نیست، بلکه نشان می‌دهد ویژگی‌های ساختاری قابل تفسیر، در ترکیب با الگوریتم‌های یادگیری ماشین، می‌توانند اطلاعات مؤثری برای تحلیل روابط شبکه‌های اجتماعی فراهم کنند.

از دیدگاه کاربردی، چارچوب ارائه‌شده می‌تواند برای شناسایی روابط بالقوه میان کاربران و نمانام‌های تجاری و پشتیبانی از تحلیل ساختار شبکه‌های اجتماعی مورد استفاده قرار گیرد. شناسایی چنین روابطی می‌تواند در تحلیل تعاملات کاربران، شناسایی کاربران بالقوه مرتبط با نمانام‌ها و پشتیبانی از تصمیم‌گیری‌های مرتبط با بازاریابی شبکه‌های اجتماعی مفید باشد. همچنین، استفاده از ویژگی‌های ساختاری قابل تفسیر این امکان را فراهم

می‌کنـد که عوامل مؤثـر بر پیش‌بیـنی روابط به صورت مستقیم مورد بررسی قرار گیرند.

محدودیت‌ها و پیشنهادهای پژوهش‌های آینده

با وجـود یافته‌های حاصل از ایـن پژوهش، داده‌های مورد اسـتفاده به شـش نمانام تجاری در شبکه اجتماعی ایکـس محدود بوده و بنابراین بررسی قابلیت تعمیم نتایج به سـایر نمانام‌ها و شـبکه‌های اجتماعی نیازمند ارزیابی بیشتر است. همچنین، به دلیل محدودیت‌های دسترسی به داده‌های شـبکه اجتماعی و تمرکز پژوهش بر ویژگی‌های سـاختاری، اطلاعات محتوایی و معنایی تعاملات کاربران به‌صورت مستقیم در مدل مورد استفاده قرار نگرفته است. علاوه‌بر این، اگرچه پایداری مدل نسـبت به برخی عوامل آزمایشی بررسی شـد، تغییرات زمانی روابط کاربران در طول دوره‌های مختلف به‌صورت مستقل مدل‌سازی نشده است.

در پژوهش‌هـای آینده می‌توان چارچوب پیشـنهادی را بر روی مجموعه گسـترده‌تری از کاربـران، نمانام‌ها و شـبکه‌های اجتمـاعی ارزیـابی کرد. همچنیـن، ترکیب ویژگی‌های سـاختاری با ویژگی‌های زمـانی، محتوایی و معنایی می‌توانـد اطلاعات مکملی بـرای پیش‌بینی روابط فراهم کند. بررسی روش‌های پیشرفته مبتنی بر شبکه‌های عصـبی گرافی ماننـد GAT، GCN و GraphSAGE در کنار ویژگی‌هـای مهندسی شـده نیز می‌تواند امکان مقایسـه جامع‌تـر روش‌های مبتنی بر ویژگی و روش‌های یادگیری عمیـق را فراهم کنـد. در نهایت، بـررسی تغییرات زمانی روابط و توسـعه مدل‌های قادر به پیش‌بینی شـکل‌گیری پیوندهای جدیـد در دوره‌های زمانی آینده می‌تواند مسیر مناسبی برای ادامه این پژوهش باشد.

منابع

1. Tang, L., & Liu, H. 2010. Community Detection and Mining in Social Media, vol. 2, Synthesis Lectures on Data Mining and Knowledge Discovery, pp. 1-137.
2. Wasserman, S., & Faust, K. 1994. Social Network Analysis: Methods and Applications, Cambridge University Press.
3. Chakraborty, T., Dalmia, A., Mukherjee, A. & Ganguly, N. 2018. Metrics for Community Analysis: A Survey, ACM Computing Surveys, vol. 50, no. 4, p. Article 64.
4. Tabassum, S., Pereira, F. S. F., Fernandes, S., & Gama, J. 2018. Social Network Analysis: An Overview, WIREs Data Mining and Knowledge Discovery, vol. 8, no. 5, p. e1256.
5. Page, L., Brin, S., Motwani, R., & Winograd, T. 1999. The PageRank Citation Ranking: Bringing Order to the Web, Stanford InfoLab Technical Report.
6. Kotsiantis, S. B. 2007. Supervised Machine Learning: A Review of Classification Techniques, Informatica, vol. 31, no. 3, pp. 249-268, 2007.
7. Cortes, C., & Vapnik, V. 1995. Support-Vector Networks, Machine Learning, vol. 2, no. 3, pp. 273-297.
8. Breiman, L. 2001. Random Forests, Machine Learning, vol. 45, no. 1, pp. 5-32.
9. Fan, W. 2012. Graph Pattern Matching Revised for Social Network Analysis, in Proceedings of the 15th International Conference on Database Theory (ICDT).
10. Jin, D., You, X., Li, W., He, D., Cui, P., & Zhang, W. 2019. Community Detection in Social Networks: Approaches and Applications, in Proceedings of the AAAI Conference on Artificial Intelligence.
11. Freeman, L. C. 1979. Centrality in social networks: Conceptual clarification, Social Networks, vol. 1, no. 3, p. 215–239.
12. Aggarwal, C. C. 2011. Social Network Data Analytics, Springer.
13. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. 2008. Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, p. P10008.
14. Ng, A. Y., Jordan, M. I., & Weiss, Y. 2002. On Spectral Clustering: Analysis and an Algorithm, in Advances in Neural Information Processing Systems (NIPS.
15. Liben-Nowell, D., & Kleinberg, J. 2007. The link-prediction problem for social networks, Journal of the American Society for Information Science and Technology, vol. 58, no. 7, p. 1019–1031.
16. Kapoor, P., Kaushal, S., Kumar, H. & Kanwar, K. 2024. A Survey on Feature Extraction and Learning Techniques for Link Prediction in Homogeneous and Heterogeneous Complex Networks, Artificial Intelligence Review, vol. 57.
17. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. 2013. Applied Logistic Regression, 3rd Edition ed., Wiley.
18. Mumin, D., Shi, L.-L., & Liu, L. 2022. An efficient algorithm for link prediction based on local information: Considering the effect of node degree, Concurrency and Computation: Practice and Experience, vol. 34, no. 7, p. e6289.
19. Qin, M., & Yeung, D.-Y. 2023. Temporal Link Prediction: A Unified Framework, Taxonomy, and Review, ACM

Computing Surveys, vol. 56, no. 4, pp. 1- 40, 2023.
20. Zare, G., Jafari Navimipour, N., Hosseinzadeh, M., & Sahafi, A. 2024. Network link prediction via deep learning method: A comparative analysis with traditional methods, Engineering Science and Technology, an International Journal, vol. 56, p. 101782.
21. Wu, H., Song, C., Ge, Y., & Ge, T. 2022. Link Prediction on Complex Networks: An Experimental Survey, Data Science and Engineering, vol. 7, p. 253–278.
22. Tran, K. L., Le, H. A., Nguyen T. H., & Nguyen, D. T. 2022. Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam, Data, vol. 7, no. 11, p. 160.
23. Makarov, I., Kiselev, D., Nikitinsky, N., & Subelj, L. 2021. Survey on graph embeddings and their applications to machine learning problems on graphs, PeerJ Comput Sci, vol. 4, no. 7, p. e357.
24. Bayrak, A. E., & Polat, F. 2019. Effective feature reduction for link prediction in location-based social networks, Journal of Information Science, vol. 45, no. 5, p. 665–679.
25. Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. 2024. Graph Neural Networks, Nature Reviews Methods Primers, vol. 4, no. 1, 2024.
26. Khemani, B., Patil, S., Kotecha, K., & Tanwar, S. 2024. A Review of Graph Neural Networks: Concepts, Architectures, Techniques, Challenges, Datasets, Applications, and Future Directions, Journal of Big Data, vol. 11, no. 1.
27. Ju, W., Yi, S., Wang, Y., Xiao, Z., Mao, Z., Li, H., Gu, Y., Qin, Y., Yin, N., Wang, S., Liu, X., Yu, P. S., & Zhang, M. 2026. A Survey of Graph Neural Networks in Real World: Imbalance, Noise, Privacy and OOD Challenges, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 3, p. 3036–3055.
28. Li, X., Wang, J., & Yan, Z. 2025. “Can Graph Neural Networks be Adequately Explained? A Survey,” ACM Computing Surveys, vol. 57, no. 5, pp. 1-36.
29. Zheng, X., Wang, Y., Liu, Y., Li, M., Zhang, M., Jin, D., Yu, P. S., & Pan, S. 2026. Graph Neural Networks for Graphs with Heterophily: A Survey, IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 38, no. 7, p. 4385–4404, 2026.
30. Rusch, T. K., Bronstein, M. M., & Mishra, S. 2023. A Survey on Oversmoothing in Graph Neural Networks, arXiv preprint arXiv:2303.10993.
31. Wu, H., Song, C., Ge, Y., & Ge, T. 2022. Link Prediction on Complex Networks: An Experimental Survey, Data Science and Engineering, vol. 7, no. 3, p. 253–278.
32. Kaczyńska, P., Sienkiewicz, J., & Ślęzak, D. 2026. Accumulated Local Effects and Graph Neural Networks for Link Prediction, Scientific Reports, vol. 16, no. 8574.
33. Liang, M., & Zhang, H. 2026. AGNN: Topology-Aware Graph Neural Network Framework for Link Prediction, Scientific Reports, vol. 16.
34. Xiong, J., Zareie, A., & Sakellariou, R. 2026. A Survey of Link Prediction in Temporal Networks, SN Computer Science, vol. 7, no. 100.
35. Yang, Y., Lichtenwalter, R. N., & Chawla, N. V. 2015. Evaluating Link Prediction Methods, Knowledge and Information Systems, vol. 45, no. 3, p. 751–782.
36. Zhou, T. 2023. Discriminating abilities of threshold-free evaluation metrics in link prediction, Physica A: Statistical Mechanics and its Applications, vol. 615, p. 128529.
37. Bi, Y., Jiao, X., Lee, Y.-L., & Zhou, T. 2024. Inconsistency among evaluation metrics in link prediction, PNAS Nexus (Oxford University Press / National Academy of Sciences), vol. 3, no. 11, p. 498.
38. Gandomi, A., & Haider, M. 2015. Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, vol. 35, no. 2, p. 137–144.
39. He, W., & Zha, S. 2014. Social media competitive analysis and text mining: A case study in the pizza industry, International Journal of Information Management, vol. 34, no. 3, p. 464–472.
40. Ma, T., Hu, Y., Lu, Y., & Bhattacharyya, S. 2025. Customer engagement prediction on social media: A graph neural network method, Information Systems Research (INFORMS) , vol. 36, no. 3, p. 1887–1897.
41. De Bruin, G. J., Veenman, C. J., Herik, H. J. V. D. & Takes , F. W. 2021. Supervised temporal link prediction in large-scale real-world networks, Social Network Analysis and Mining (Springer), vol. 11, no. 1, p. 80.