Rasanch \ A Quarterly Journal of Media Studies and Research
http//qjmn.farhang.gov.ir DOI: 10.22034/bmsp.2023.376274.1813
Research paper Received: 11/12/2022
Vol. 35, No. 2, Summer 2024, Issue 4 Accepted: 25/09/2023

Improving User Relationship Prediction in Twitter Metadata Using Aggregate Classification

Mohammad Chakerolhoseini Firouzabad: Doctoral student of Computer Engineering, Faculty of Technology and Engineering, Islamic Azad University, Neyshabur branch, Neyshabur, Iran.
email: m.chaker71@iau-neishabour.ac.ir

Reza Ghaemi: Assistant Professor, Department of Computer Engineering, Faculty of Technology and Engineering, Islamic Azad University (corresponding author), Quchan Branch, Quchan, Iran.
email: r.ghaemi@iauq.ac.ir

In today's world, social networks that have become a part of people's daily life, including Twitter, Telegram, Instagram, etc., are increasing and expanding day by day. Therefore, the number of their users is also increasing and as a result, a large amount of data is being exchanged and stored in these network; and this huge amount of data has turned social networks, especially Twitter, into big data. It is very important to manage, organize and prune these big data, as well as to predict the behavior of social network users.
One of the most important and effective methods for predicting user relationships in social networks is classification techniques, which in most of the applications and researches in the background of the research, are still based on criteria such as 'accuracy; and accuracy of prediction. have weakness In this article, in order to predict the user relationship in Twitter social networks, the cumulative classification method based on voting, which has two basic steps, has been used. In the first step, by using basic classification algorithms including nearest neighbor, decision tree, random forest and simple Bayesian, the outputs of each classification are obtained. In the second step, the final output of cumulative classification is calculated using the voting method. The results of the experiments on the dataset of the Twitter social network and based on the criteria of accuracy, correctness and coverage, argue that the proposed cumulative classification method based on voting has more favorable results than It has other algorithms

Keywords: Big Data, Social Network, Prediction of User Relationship, Cumulative Classification, Twitter.

How to cite this paper: Chakeralhoseini Firouzabad, M., & Ghaemi, R. (2024). Improving User Relationship Prediction in Twitter Metadata Using Aggregate Classification. *Rasaneh*, 35(2), 107-131 . **[In Persian]**

رسانه / فصلنامه مطالعاتی و تحقیقاتی وسایل ارتباط جمعی
Rasanch
نوع مقاله: پژوهشی
DOI: 10.22034/bmsp.2023.376274.1813
http//qjmn.farhang.gov.ir
Vol. 35, No. 2, Summer 2024

بهبود پیش‌بینی علاقهٔ کاربران در کلان‌داده توییتر با استفاده از طبقه‌بند تجمعی

نوشتهٔ
محمد چاکرالحسینی فیروزآباد*
رضا قائمی**

تاریخ دریافت: ۱۴۰۱/۰۹/۲۰
تاریخ پذیرش: ۱۴۰۲/۰۷/۰۳

چکیده

در دنیای امروزی، شبکه‌های اجتماعی که بخشی از زندگی روزمره انسان‌ها شده‌اند، از جمله توییتر، تلگرام، اینستاگرام و غیره، روزبه‌روز در حال افزایش و گسترش هستند. لذا تعداد کاربران آن‌ها نیز در حال افزایش است و در نتیجه، حجم دادهٔ زیادی در این شبکه‌ها در حال تبادل و ذخیره‌سازی است که این حجم عظیم داده، شبکه‌های اجتماعی به‌خصوص توییتر را تبدیل به کلان‌داده کرده است. مدیریت، سامان‌دهی و هرس‌کردن این کلان‌داده‌ها و همچنین، پیش‌بینی رفتار کاربران شبکه‌های اجتماعی امری بسیار مهم است. یکی از روش‌های مهم و تأثیرگذار برای پیش‌بینی علاقهٔ کاربر در شبکه‌های اجتماعی، تکنیک‌های طبقه‌بندی است که در اغلب کاربردها و پژوهش‌های موجود در پیشینهٔ تحقیق، هنوز در معیارهایی مانند دقت و صحت پیش‌بینی ضعف دارند. در این مقاله، به‌منظور پیش‌بینی علاقهٔ کاربر در شبکه‌های اجتماعی توییتر، از روش طبقه‌بندی تجمعی مبتنی بر رأی‌گیری که دارای دو گام اساسی است، استفاده شده است. در گام نخست، با بهره‌گیری از الگوریتم‌های طبقه‌بندی پایه‌ای شامل نزدیک‌ترین همسایه، درخت تصمیم، جنگل تصادفی و بیزین ساده، خروجی‌های هر طبقه‌بندی حاصل می‌شوند. در گام دوم، خروجی نهایی طبقه‌بندی تجمعی با استفاده از روش رأی‌گیری محاسبه می‌شود. نتایج آزمایش‌ها بر روی مجموعهٔ کلان‌داده‌های شبکهٔ اجتماعی توییتر و براساس معیارهای دقت، صحت و پوشش، استدلال بر این دارد که روش پیشنهادی طبقه‌بندی تجمعی مبتنی بر رأی‌گیری، نتایج مطلوب‌تری را نسبت‌به الگوریتم‌های دیگر داشته است.

کلیدواژه: کلان‌داده، شبکهٔ اجتماعی، پیش‌بینی علاقهٔ کاربر، طبقه‌بندی تجمعی، توییتر.

* دانشجوی دکتری مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی، واحد نیشابور، نیشابور، ایران m.chaker71@iau-neishabour.ac.ir
** استادیار گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی(نویسنده مسئول)، واحد قوچان، قوچان، ایران r.ghaemi@iauq.ac.ir

نحوهٔ استناد به این مقاله: چاکرالحسینی، محمد و قائمی، رضا (۱۴۰۳). بهبود پیش بینی علاقهٔ کاربران در کلان داده توییتر با استفاده از طبقه بند تجمعی. رسانه، ۳۵(۲)، ۱۰۷ ـ ۱۳۱ .

مقدمه

امروزه، بیشتر امور و کارهای ضروری انسان‌ها در بستر اینترنت و شبکه‌های اجتماعی[1] قرار گرفته است. شبکه‌های اجتماعی با توجه به استقبال زیاد کاربران از این محیط‌ها در جهت تبادل پیام و اشتراک‌گذاری تصویر و ویدئو، با رشد گسترده‌ای در سال‌های اخیر مواجه بوده‌اند که توییتر[2] یکی از پرکاربرترین این شبکه‌های اجتماعی است (بیهقی[3] و همکاران، ۲۰۱۸: ۸). شبکه اجتماعی توییتر، فناوری مدرن برقراری ارتباط برای انسان را شکل داده و این غیرقابل‌انکار است که به‌طور فزاینده‌ای صرف‌نظر از موقعیت جغرافیایی، خودش را به زندگی انسان‌ها القا کرده است. طبق آخرین آمار، حدود یک میلیارد نفر از شبکه‌های اجتماعی استفاده می‌کنند که سهم توییتر، حدود ۳۳۱ میلیون کاربر فعال تا سال ۲۰۱۹ بوده است. به‌علاوه، داده‌های تولیدشده از فعالیت روزمرهٔ کاربران در شبکه‌های اجتماعی باعث ایجاد حجم زیادی داده می‌شود و لذا، کلان‌داده‌ها[4] در شبکه‌های اجتماعی نقش به‌سزایی دارد و بسیار مورد استفاده قرار می‌گیرد (بیهقی و همکاران، ۲۰۱۸: ۸).

داده‌های بزرگ در کلان‌داده‌ها، از حجم عظیمی از داده‌ها به‌صورت ساخت‌یافته، نیمه‌ساخت‌یافته و یا بدون‌ساختار در حوزه‌های متفاوتی، مانند توییت‌ها، نظرها، پست‌ها و غیره جمع‌آوری می‌شوند. از این‌رو، از رایج‌ترین کاربردهای کلان‌داده‌های ایجادشده توسط شبکه‌های اجتماعی مانند توییتر، تجزیه‌وتحلیل نظرکاوی[5] و برای نمونه، به‌دست‌آوردن بازخورد مشتریان از محصول‌ها است که برای سازمان‌ها بسیار باارزش است. بنابراین، در تجزیه‌وتحلیل داده‌های حجیم شبکه‌های اجتماعی از تکنیک‌های یادگیری ماشین[6]، داده‌کاوی[7] و متن‌کاوی[8] استفاده می‌شوند. از پرکاربردترین این تکنیک‌ها به‌منظور پیش‌بینی علاقهٔ کاربران در شبکه‌های اجتماعی، می‌توان به طبقه‌بندی[9] و خوشه‌بندی[10] اشاره کرد (مورثی[11]، ۲۰۱۷: ۹).

تاکنون با استفاده از تکنیک‌های طبقه‌بندی، تحقیق‌های زیادی روی کلان‌دادهٔ توییتر صورت گرفته است؛ نظیر تشخیص واکسینه‌شدن افراد از کرونا، تجزیه‌وتحلیل تجارت الکترونیک، پیش‌بینی نتایج انتخابات ریاست‌جمهوری و میزان استقبال یا دافعهٔ انسان‌ها در دنیا نسبت‌به

1. Social Networks
2. Twitter
3. Bayhaqy
4. Big Data
5. Opinion Mining
6. Machine Learning
7. Data Mining
8. Text Mining
9. Classification
10. Clustering
11. Murthy

یـک تصمیم یـا یک رویداد مهم که کلیه این پیش‌بینی‌ها از طریق توییت‌های شـبکهٔ اجتماعی توییتر صورت گرفته اسـت (مورثی، ۲۰۱۷ : ۹). در ایـن مقاله، به پیش‌بینی و طبقه‌بندی علاقهٔ کاربران در شـبکهٔ اجتماعی توییتر پرداخته می‌شـود. در این راستا، از الگوریتم‌های طبقه‌بندی پایه و قدرتمندی شـامل K- نزدیک‌ترین همسـایه[1]، بیزین سـاده[2]، جنگل تصادفی[3] و درخت تصمیم[4] بهره‌گیری شـده اسـت. به‌علاوه، به‌منظور بهبود معیارهای دقت[5]، پوشـش[6] و صحت[7] پیش‌بینـی علاقهٔ کاربر در شـبکه اجتماعی توییتر، از طبقه‌بندی تجمعـی[8] مبتنی بر رأی‌گیری[9] استفاده شده است.

در بخش دوم به مرور ادبیات و پیشـینهٔ تحقیق شـامل کلان‌داده و طبقه‌بندها پرداخته شده است. بخش سوم، الگوریتم‌های طبقه‌بندی پایه‌ای مورد استفاده و همچنین، الگوریتم طبقه‌بندی تجمعی پیشـنهادی تشـریح شـده‌اند. در بخش چهـارم، نتایج حاصل‌شـده از الگوریتم‌های طبقه‌بنـدی پایه و همچنین، طبقه‌بند تجمعی مبتنی بر رأی‌گیری پیشـنهادی مورد ارزیابی قرار گرفته است. در نهایت بخش پنجم، به نتیجه‌گیری و کارهای آینده پرداخته است.

ادبیات و پیشینهٔ تحقیق

در ایـن بخش، در مـورد مفاهیم اصلی کلان‌داده‌ها، شـبکه‌های اجتماعی و طبقه‌بندی توضیح داده شـده است و سـپس، عملکرد الگوریتم‌های طبقه‌بندی پایه‌ای مورد استفاده در این مقاله شـامل K- نزدیک‌ترین همسایه، بیزین ساده، جنگل تصادفی و درخت تصمیم تشریح شده‌اند. در ادامه، پیشـینهٔ تحقیق در زمینهٔ پیش‌بینی و طبقه‌بندی علاقهٔ کاربران در شـبکه‌های اجتماعی به‌خصوص رسانه توییتر ارائه شده است.

۱. کلان‌داده‌ها

کلان‌داده اصطلاحی است برای جمع‌آوری مجموعه‌های اطلاعاتی گران و پیچیده‌ای که پردازش آن با اسـتفاده از برنامه‌های کاربردی مرسـوم پردازش اطلاعات، دشوار است. چالش‌هایی از جمله پیچیدگی حافظه و زمانی بالای ذخیره‌سازی، مدیریت، پرس‌وجو، اشتراک‌گذاری، تبادل، ادراک و نقـض حفاظت در کلان‌داده‌ها وجود دارنـد. برای کاهش الگوهای تجاری، پیش‌بینی

1. K-Nearest Neighbor
2. Naïve Bayesian
3. Random Forest
4. Decision Tree
5. Accuracy
6. Recall
7. Precision
8. Classification Ensemble
9. Voting

بیماری‌ها، شناسایی تضادها و غیره، در مقایسه با مجموعه‌داده‌های کوچک‌تر، به مجموعه‌های دادهٔ بزرگ‌تری نیاز است. کار با حجم عظیمی از اطلاعات با استفاده از چارچوب‌های مدیریت پایگاه دادهٔ اجتماعی و اندازه‌گیری‌های صفحهٔ نمایش و بسته‌های ادراک مشکل است و نیاز به برنامه‌نویسی موازی بسیار زیادی دارد که روی ده‌ها، صدها یا حتی تعداد زیادی سرویس‌دهنده اجرا می‌شوند (سری و آنوشا[1]، ۲۰۱۶ : ۱).

برخلاف داده‌های سنتی، کلان‌داده به داده‌های بزرگ در حال رشد اشاره دارد و مجموعه داده‌هایی که دارای ساختارهای ناهمگن هستند، شامل داده‌های ساخت‌یافته، بدون‌ساختار و نیمه‌ساخت‌یافته. کلان‌داده ماهیت پیچیده‌ای دارد که به فناوری‌های قدرتمند و الگوریتم‌های پیشرفته‌ای نیاز دارد (اوسوس و بتجلن[2] و همکاران، ۲۰۱۸: ۲). شکل ۱ پنج اصل در کلان‌داده‌ها شامل ظرفیت، تنوع، صحت، ارزش و سرعت را نشان داده است. از آنجایی‌که داده‌ها از منابع بسیار متفاوتی جمع‌آوری می‌شوند، صحت به مفهوم کیفیت، دقت و قابلیت اطمینان داده‌های جمع‌آوری شده است. همچنین، ارزش به توانایی سازمان برای تبدیل داده‌های عظیم به کسب‌وکار واقعی اشاره دارد، چراکه داده‌های دقیق باعث می‌شوند کسب‌وکارها به رفع نیازها وانتظارات مشتریان خود نزدیک‌تر شوند (پاپاکیراکو و باربناکس [3]، ۲۰۲۲: ۵)

شکل ۱. پنج اصل در کلان‌داده

۲. شبکه‌های اجتماعی

شبکه و رسانه‌های اجتماعی، مجموعه‌ای از برنامه‌های کاربردی مبتنی بر اینترنت است که بر اساس ایدهٔ وب ۲ ایجاد شد و ابتدا در سال ۲۰۰۴ برای توصیف محتویات و برنامه‌های کاربردی مورد استفاده قرار گرفت و می‌توان آن را به‌جای این‌که به‌طور سنتی توسط افراد ایجاد، تهیه و منتشر شود توسط کاربران به‌طور مداوم و به‌طرق مختلف از طریق مشارکت و همکاری تغییر داد. استفادهٔ گسترده از نرم‌افزار و سخت‌افزار موجود برای دسترسی به

1. Sri & Anusha
2. Oussous & Benjelloun
3. Papakyriakou & Barbounakis

چارچوب‌های شـبکه و رسانه‌های اجتماعی از طریق اینترنت، منجر به ایجاد و تبادل محتوای تولیدشده توسط کاربران شده است (قانی[1] و همکاران، ۲۰۱۹: ۳).

شـبکه‌های اجتماعی به‌عنوان خدمات مبتنی بر وب تعریف می‌شـوند که ابتدا در آن، افراد مجاز به ایجاد نمایهٔ عمومی یا نیمه‌عمومی خود هسـتند. سـپس، این کاربران مجاز به ارتباط با دیگران‌اند و در نهایت برای تشکیل یک شبکه، این افراد مجاز به مشاهده و ارتباط با سایر کاربران و فعالیت‌هایی هسـتند که در شبکه آن‌ها منتشر می‌شـوند. نمونه‌هایی از شبکه‌های اجتماعی که حجم زیادی از داده‌های بدون‌ساختار را دارا هستند عبارت است از فیس‌بوک، توییتر، اینستاگرام، لینکدین، وب‌نوشت‌ها، ویکی‌ها و یوتیوب (قانی و حمید و همکاران. ۲۰۱۹: ۳).

پیش‌بینی و طبقه‌بندی

داده‌کاوی فرایند اسـتنتاج دانش از داده‌های به‌نسبت عظیم است. داده‌کاوی دارای سه تکنیک اصلی طبقه‌بندی، خوشـه‌بندی، الگوهای مکرر و قوانین وابستگی آن‌ها است. در خوشه‌بندی، مجموعه‌ای از داده‌ها بر اسـاس شباهت و تفاوت بین ویژگی‌هایشان گروه‌بندی می‌شوند. مدل خوشـه‌بندی حاصل می‌تواند برای طبقه‌بندی داده‌های آینده مورد استفاده قرار گیرد. داده‌کاوی فرایند اسـتخراج اطلاعات از یک مجموعه‌داده و تبدیل آن به یک سـاختار قابل‌درک اسـت. وظیفهٔ داده‌کاوی، تحلیل خودکار یا نیمه‌خودکار مقادیر زیادی از داده‌ها برای استخراج الگوهای جالب ناشناخته قبلی است. داده‌کاوی شامل شش وظیفهٔ اصلی است که عبارت‌اند از تشخیص ناهنجاری، یادگیری قوانین وابسـتگی، خوشـه‌بندی، طبقه‌بندی، رگرسـیون و خلاصه‌سازی. طبقه‌بندی یک تکنیک پرکاربرد در داده‌کاوی اسـت و به‌طور گسـترده‌ای در زمینه‌های مختلف مورد اسـتفاده قرار گرفته است که برای پیش‌بینی عضویت گروهی از نمونه‌های آزمون در یک کلاس مشخص، استفاده می‌شود. در این مقاله، چندین الگوریتم طبقه‌بندی پایه‌ای استفاده شده است، شامل K- نزدیک‌ترین همسایه، بیزین ساده، جنگل تصادفی و درخت تصمیم (کساورج و سوکوماران[2]، ۲۰۱۳: ۴).

طبقه‌بندی، مسئلهٔ پیش‌بینی گسستهٔ متغیر تصادفی Y از متغیر تصادفی X است که طبقه‌بندی الگـو یا تشـخیص الگو نیز نامیده می‌شـود. الگوریتم طبقه‌بندی، داده‌ها را بـه تعداد معینی از کلاس‌هـا دسـته‌بندی می‌کند و برچسـبی را به هـر کلاس اختصاص می‌دهد. ایـدهٔ اصلی در الگوریتم‌های طبقه‌بندی به‌منظور پیش‌بینی کلاس هدف توسط تجزیه‌وتحلیل مجموعه‌داده‌های آموزشی، دسته‌بندی داده‌ها به تعداد معینی از کلاس‌ها است (پاپاکیراکو و باربناکس،۲۰۲۲: ۵).

1. Ghani
2. Kesavaraj & Sukumaran

۱. K- نزدیک‌ترین همسایه

الگوریتم طبقه‌بندی K- نزدیک‌ترین همسایه، براساس یادگیری توسط مقایسه عمل می‌کند که در آن، ویژگی‌های n بعدی هر نمونه داده، نشان‌دهندهٔ یک نقطه در فضای n بعدی است. نمونه‌های آموزشی در یک فضای الگوی n بعدی ذخیره می‌شوند. هنگامی‌که یک نمونه داده آزمون توسط K- نزدیک‌ترین همسایه طبقه‌بندی می‌شود، فضای الگو را برای K نمونه آموزشی که نزدیک‌ترین به نمونهٔ آزمون هستند، جست‌وجو می‌کند و فاصلهٔ نمونهٔ آزمون را برحسب فاصله اقلیدسی مطابق رابطهٔ (۱)، با سایر نمونه‌های آموزشی محاسبه می‌کند. نمونهٔ آزمون، دارای رایج‌ترین کلاس در میان K- نزدیک‌ترین همسایه خود است. زمانی‌که K=۱ باشد، نمونهٔ آزمون به کلاس نزدیک‌ترین نمونهٔ آموزشی به آن اختصاص داده می‌شود (فیو[1]، ۲۰۰۹: ۶).

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^{2}} \rightarrow \qquad (۱)$$

در رابطهٔ (۱)، متغیرهای X و Y دو نقطه مورد نظرهستند که با محاسبهٔ مجموع اختلاف آن‌ها، فاصلهٔ بین دو نقطه محاسبه می‌شود. الگوریتم طبقه‌بندی K- نزدیک‌ترین همسایه مبتنی بر نمونه و از نوع یادگیرنده‌های تنبل است. طبقه‌بندی نمونه آزمون برای طبقه‌بندهای تنبل ممکن است با وجود تعداد زیادی همسایهٔ بالقوه، متحمل هزینه‌های زمانی زیادی شود. برای رفع این چالش، از الگوریتم طبقه‌بندی K- نزدیک‌ترین همسایهٔ وزن‌دار استفاده می‌شود که در آن، براساس اهمیت هر نمونه آموزشی یا هر ویژگی داده، وزنی به آن انتساب داده شود و زمانی‌که نمونهٔ آموزشی یا ویژگی نامربوط زیادی وجود داشته باشد، ممکن است مشکل به‌وجود آورد (فیو، ۲۰۰۹: ۶).

۲. بیزین ساده

از تکنیک دسته‌بند بیزین اغلب به‌عنوان یک راه‌کار ساده برای دسته‌بندی و تعیین روشی برای تشخیص برچسب اشیا یا نقاط استفاده می‌شود. به‌منظور به‌کارگیری دسته‌بند بیزین ساده، الگوریتم یکتایی وجود ندارد، بلکه خانواده‌ای از الگوریتم‌ها موجود است که با فرض استقلال ویژگی‌ها یا متغیرها نسبت‌به یکدیگر عمل می‌کنند. برای نمونه، اندازه یک میوه و رنگ آن که متغیرهای مستقل در نظر گرفته می‌شوند، در تعیین نوع آن میوه مؤثرهستند. بنابراین، اگر میوه‌ای با رنگ قرمز و دارای اندازه حدود ۱۰ سانتی‌متر باشد، با احتمال زیاد سیب است. یکی از مزایای قابل توجه در دسته‌بند بیزین ساده، امکان برآورد پارامترهای مدل با اندازهٔ نمونهٔ کوچک به‌عنوان مجموعه‌دادهٔ آموزشی است (فیو، ۲۰۰۹: ۶). بیزین ساده را می‌توان یک مدل برمبنای احتمال شرطی در نظر گرفت. فرض کنید $X=(x_1,\ldots,x_n)$ برداری از n ویژگی

1. Phyu

را بیان کند که به‌صورت متغیرهای مستقل هستند. به‌این ترتیب می‌توان احتمال رخداد CK، یعنی $P(CK|x_1,\ldots,x_n)$ را به‌عنوان یکی‌از حالت‌های کلاس رخدادهای مختلف به‌ازای Kهای متفاوت، در رابطه (۲) نمایش داد (فیو، ۲۰۰۹ : ۶؛ متسیس و آندرو[1] و همکاران، ۲۰۰۶ : ۷).

$$P(Ck \mid X) = (P(Ck)P(X \mid Ck))/P(X) \qquad (۲)$$

رابطهٔ (۲)، همان قضیهٔ بیز است که X بردار و CK رخداد است و بر اساس احتمالات پیشامدهای پیشین[2]، پسین[3]، درست‌نمایی[4] و شواهد[5]، در رابطهٔ (۳) بازنویسی شده است (فیو، ۲۰۰۹: ۶):

$$Posterior = \frac{Prior*Likelihood}{Evidence} \qquad (۳)$$

به‌این ترتیب، برای محاسبه احتمال $P(CK|x_1,\ldots,x_n)$ کافی است از احتمال توأم کمک گرفته شود و با احتمال شرطی با توجه به استقلال متغیرها، مطابق رابطهٔ (۴) ساده‌تر نمایش داده شود (فیو، ۲۰۰۹: ۶):

$$P(Ck, x1, x2, \ldots, xn) = \qquad (۴)$$
$$p(x1, x2, \ldots, xn, Ck)p(x2|x3, \ldots, xn, Ck) \ldots p(xn-1|xn, Ck)p(xn|Ck)\mathrm{p(Ck)}$$

به این ترتیب احتمال توأم را به‌صورت حاصل‌ضرب احتمال شرطی می‌توان نوشت:

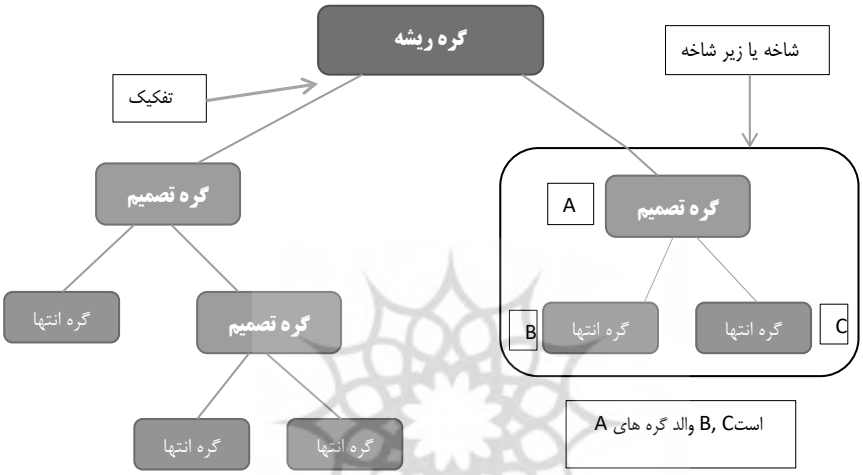
$$P(Ck, x1, x2, \ldots, xn) = p(Ck)\prod_{i=1}^{n} p(xi|Ck) \qquad (۵)$$

۳. درخت تصمیم

درخت تصمیم یکی‌از روش‌های قدرتمندی است که اغلب در زمینه‌های مختلفی نظیر یادگیری ماشین، پردازش تصویر و شناسایی الگوها مورد استفاده قرار می‌گیرد. درخت تصمیم یک مدل متوالی است که یک سری آزمون‌های پایه را به‌طور کارآمد و منسجم متحد می‌کند که در آن، یک ویژگی عددی با عددی دیگر که مقدار آستانه در هر آزمون است، مقایسه می‌شوند. درخت تصمیم یک مدل طبقه‌بندی است که از گره و شاخه‌ها تشکیل شده است و هر گره، ویژگی‌هایی را در یک دسته برای طبقه‌بندی نشان می‌دهد و هر زیرمجموعه مقداری را تعریف می‌کند که

1. Metsis& Androutsopoulos
2. Prior
3. Posterior
4. likelihood
5. Evidence

می‌تواند توسط گره گرفته شود. درخت‌های تصمیم به‌دلیل تجزیه‌وتحلیل ساده و دقت آن‌ها بر روی فرم‌های داده‌های متعدد، زمینه‌های پیاده‌سازی بسیاری را پیدا کرده‌اند. انواع متفاوتی از الگوریتم‌های درخت تصمیم وجود دارند، از جمله دوگانگی تکراری[1]، جانشین درخت طبقه‌بندی و رگرسیون[2]، رگرسیون تطبیقی چندمتغیره[3]، تعمیم‌یافته[4]، بی‌طرفانه[5]، تشخیص و تخمین تعامل[6]، درختان استنتاج شرطی[7] و غیره. شکل ۲، نمونه‌ای از یک درخت تصمیم را نشان می‌دهد (چاربوتی و عبدالله زی[8]، ۲۰۲۱ : ۱۰).



شکل ۲. درخت تصمیم

۴. جنگل تصادفی

همان‌طور که از نام این الگوریتم پیداست، جنگل تصادفی یک الگوریتم طبقه‌بندی نظارت‌شده است که داده‌ها را با ساخت تعدادی طبقه‌بندی‌کننده با هدف دستیابی به دقت بالاتر پیش‌بینی، طبقه‌بندی می‌کند (شایک و سرین واسان[9]، ۲۰۱۹: ۱۵). جنگل تصادفی برای رتبه‌بندی اهمیت متغیرها در یک مسئله رگرسیونی یا طبقه‌بندی می‌تواند استفاده شود. نخستین گام، اندازه‌گیری

1. Repetitive Duality
2. Regression
3. Multivariate Adaptive Regression
4. Generalized
5. Impartially
6. Estimation of Interaction
7. Conditional Inference Trees
8. Charbuty & Abdulazeez
9. Shaik & Srinivasan

اهمیت یک متغیر در مجموعه داده $D_n(X_i,Y_i)=1$ اســت که برازش جنگل تصادفی اســت. در طــول فراینــد، برازش خطای خارج از محدوده هر نقطه از داده ثبت می‌شــود و میانگین آن در جنگل محاســبه می‌شود (زکریا[1]، ۲۰۱۴ : ۱۲). از مزایای مهم الگوریتم جنگل تصادفی می‌توان بــه دقت بالا، کارآمدی در مدیریت پایگاه‌های داده‌های بزرگ، و کنترل داده‌های ازدســت‌رفته بدون به‌خطر انداختن دقت، اشاره کرد. نمونه‌های اولیه برای دادن اطلاعات یا داده‌های متا در رابطه بین متغیرهای مختلف استفاده می‌شود (شایک و سرین واسان، ۲۰۱۹: ۱۵).

پیشینهٔ تحقیق

در این بخش، به مرور تحقیق‌های موجود از چند سال گذشته در خصوص طبقه‌بندی در شبکهٔ اجتماعی توییتر، پرداخته می‌شود. مقالهٔ هارجلو و گورجار [2] و همکاران (۲۰۲۰ : ۱۷) با استفاده از روش‌های طبقه‌بندی رگرســیون، بیزین ساده و ماشــین بردار پشتیبان[3] بر روی مجموعه‌داده Sentiment140 مربوط به دانشگاه استفورد توانستند احساساتی که در توییت‌ها تبادل می‌شوند را بر مبنای دو دســته مثبت و منفی طبقه‌بندی کنند. به‌دلیــل محبوبیت زیاد توییتر، افراد زیادی هرزنامه‌هایی را به‌منظور ایجاد حســاب‌های جعلی و انتشــار پیام‌هــای مخرب ایجاد می‌کنند. در تحقیــق آدوله و هان[4] و همکاران (۲۰۲۰: ۱۸) با تلفیق الگوریتم خوشــه‌بندی K–Means و روش‌های طبقه‌بندی ماشین بردار پشتیبان و جنگل تصادفی، برای تشخیص حساب‌های جعلی به‌میزان ۹۶/۳ درصد دقت، حساسیت و پوشش بهبود داشته است.

یکــی از موارد مهم رفع ابهام برای کاربران، حمایت از تصمیم آن‌ها و ســطح رضایتشــان اســت. در پژوهش فان و نگوین[5] و همکاران (۲۰۲۱ :۱۹) سطح رضایت کاربران در سه مرحله محاســبه شده است. مرحلهٔ اول، موضوعی را مشــخص می‌کند که کاربر به آن علاقه‌مند است. مرحلهٔ دوم، جنبه‌هایی از موضوع و احساســات آن‌ها در توییت‌ها اســتخراج می‌شــود. مرحلهٔ سوم، سطح رضایت کاربر با توجه به هر نوع احساسات شناسایی و محاسبه می‌شود. در نهایت، با اســتفاده از درخت تصمیم، بهبود لازم حاصل شده است. یکی از عوامل مهم تجزیه‌وتحلیل احساســات کاربران در مواجه با تجارت الکترونیک اســت. در تحقیق بایاق و ســفنیرانتو[6] و همکاران (۲۰۱۸ : ۸) با استفاده از روش‌های طبقه‌بندی نزدیک‌ترین همسایه، درخت تصمیم و بیزین ساده، احساسات کاربران در تجارت الکترونیک تجزیه‌وتحلیل شد. بیزین ساده عملکرد بهتری در طبقه‌بندی احساســات بر اســاس معیارهای دقت، حساسیت و پوشش داشته است.

1. Zakariah
2. Harjule & Gurjar
3. Support Vector Machine
4. Adewole & Han
5. Phan & Nguyen
6. Bayhaqy & Sfenrianto

طبـق پژوهـش نئوجی و گرج[1] و همـکاران (۲۰۲۱: ۲۱) آمـوزش در اندونزی امری واجب و پرداخت هزینهٔ آن الزامی اسـت. به‌علت این‌که کلیه هزینه‌های مدارس از طریق شـهریه انجام می‌پذیرد، اما بسـیاری از دانش‌آموزان تأخیر در پرداخت دارند. به‌همین منظور، با اسـتفاده از روش طبقه‌بندی نزدیک‌ترین همسـایه و بیزین ساده نشان داده شد که بیزین ساده در پیش‌بینی دقت بالاتری داشته است.

در تحقیق (عباس و سـلیه[2] و همکاران، ۲۰۲۰: ۲۲) در سـال ۲۰۲۱، روی موضوع اعتراض کشـاورزان بیش از ۲۰ هزار توییت برای شناسـایی احسـاس اعتراضـی از طریق روش‌های طبقه‌بندی بیزین سـاده، درخت تصمیم، ماشـین بردار پشـتیبان و جنگل تصادفی انجام شد که روش جنگل تصادفی دقت بالاتری نسـبت به سـایر روش‌ها داشـت. در داده‌های رسانه‌های اجتماعی، توییتر اغلب از متن مبهم اسـتفاده می‌کند که می‌تواند در شناسایی احساسات مثبت یا منفی مشـکل ایجاد کند. بیش از یک میلیارد پیام رسـانه‌های اجتماعی وجود دارد که باید در یک پایگاه دادهٔ مناسـب ذخیره‌سازی شوند و برای تجزیه‌وتحلیل، به‌درستی پردازش شوند. در پژوهش شمرت و چاکرابروتی[3] (۲۰۲۱: ۲۳) با ترکیب روش طبقه‌بندی تجمعی با بیزین ساده، درخت تصمیم، رگرسیون و شبکه عصبی پرسپترون[4] نشان داده شد که عملکرد بهتر شده است. در دو سـال اخیر، به‌دلیل همه‌گیری ویروس کرونـا در جهان، به‌منظور رفع این ویروس تلاش شـد و سه واکسـن مدرنا، فایزر و آسترازنکا از مشهورترین واکسـن‌ها تولید شدند. در تحقیق شـمرت و چاکرابروتی در سال ۲۰۲۱، با استفاده از روش طبقه‌بندی K- نزدیک‌ترین همسایه نظرات توییتری راجع به واکسن‌های مذکور، به سه دستهٔ مثبت، منفی و خنثی طبقه‌بندی شدند. مهم‌بودن ویروس کرونا به‌عنوان یک مسـئلهٔ حیاتی و همچنین، تجزیه و تحلیل احساسـات و نظرسـنجی کاربران ازطریـق توییت‌ها اهمیت زیادی پیدا کرده اسـت، اما تجزیه‌وتحلیل‌ها بر روی توییت‌های انگلیسی بوده است.

در پژوهش الحاشدی و الفوحیدی[5] و همکاران (۲۰۲۰: ۲۴) با استفاده از طبقه‌بندی بیزین سـاده برای تحلیل بر روی زبان عربی صورت گرفت. یکی از مسـائل مهمی که جنبهٔ تحقیقات جالبی به‌خصوص برای پزشـکی قانونی ارائه می‌دهد، تعیین جنسیت بر مبنای توییت‌ها است، که در تحقیق وشیس و میحان[6] (۲۰۲۰: ۲۵) با استفاده از روش پردازش زبان طبیعی[7] و همچنین، روش‌های طبقه‌بندی بیزین ساده، ماشین بردار پشتیبان و رگرسیون خطی در طبقه‌بندی جنسیت

1. Neogi, Garg
2. Abbas, Salih
3. Shamrat, Chakraborty
4. Perceptron Neural Network
5. Al-Hashedi, Al-Fuhaidi
6. Vashisth & Meehan
7. Natural Language Processing

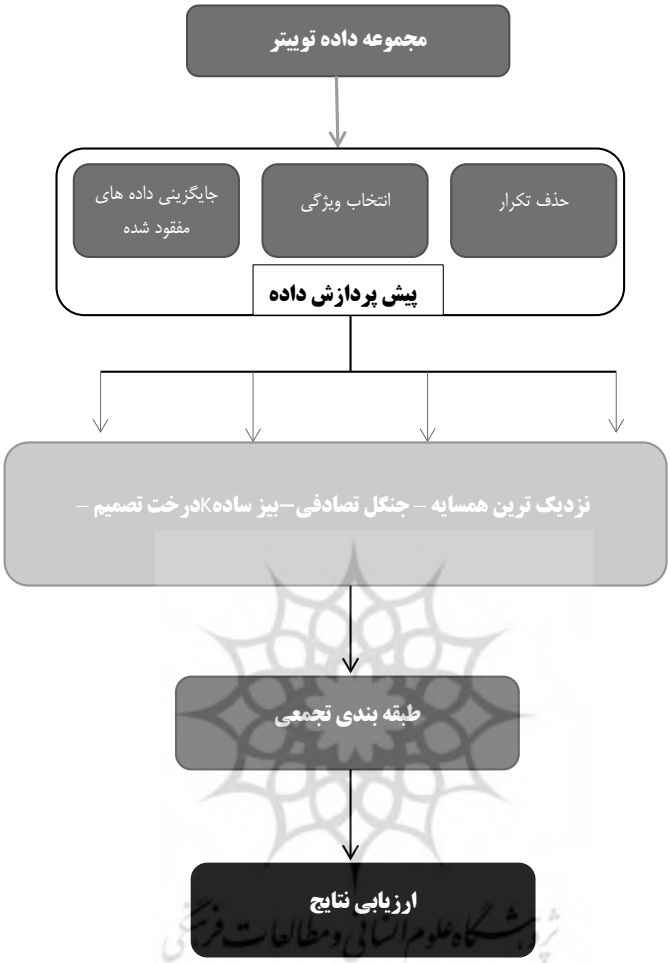
کمک‌کننده بوده اســت. در رویکرد مشتق‌شــده پژوهش پاتل و پســی[1] (۲۰۲۰: ۲۶)، تحلیلی بر روی داده‌های توییتر برای جام‌جهانی فوتبال ۲۰۱۴ برزیل انجام شــد تا احساســات مردم در سراســر جهان تجزیه‌وتحلیل شود. با اســتفاده روش‌های طبقه‌بندی بیزین ساده، ماشین بردار پشتیبان، جنگل تصادفی و K ـ نزدیک‌ترین همسایه نشان داده شد که بیزین ساده دقت بالاتری داشته است.

روش پیشنهادی

از چالش‌هــای موجود در شــبکه‌های اجتماعــی به‌خصوص توییتر، عدم اســتفاده مطلوب در استفاده از طبقه‌بندها است. انتخاب غیرصحیح و عدم بررسی صحیح در این مورد، در معیارهایی چون دقت، حساســیت، پوشــش و غیره، باعث خروجی ناکارآمدی می‌شــود، به‌خصوص در کلان‌داده‌ای چون توییتر که انتخاب بهترین طبقه‌بند در فرایند بهینه‌ســازی می‌تواند بسیار مؤثر باشــد. شکل ۳ عملکرد روش پیشــنهادی پیش‌بینی و طبقه‌بندی علاقهٔ کاربران را در کلان‌داده رسانهٔ اجتماعی توییتر نشان داده است.

گام اول، فراینــد جمــع‌آوری داده‌ها از توییتر و برچســب‌گذاری توییت‌ها اســت. پس‌از برچسب‌گذاری داده‌ها، گام دوم پیش‌پردازش توییت‌ها و آماده‌سازی آن‌ها برای تجزیه‌وتحلیل و تبدیل‌شــدن است که دارای چندین مرحله اســت. پیش‌پردازش توییت‌ها شامل پاک‌سازی، تبدیل نفی، تبدیل شکلک‌ها (ایموجی)، فیلترکردن و غیره است. پس‌از پیش‌پردازش توییت‌ها، در گام ســوم، توییت‌ها با اســتفاده از الگوریتم‌های پایه‌ای، اما قدرتمندی همچون بیزین ساده، K ـ نزدیک‌ترین همســایه، جنگل تصادفی و درخت تصمیم، طبقه‌بندی می‌شوند. در این مقاله و در گام چهارم، برای پیش‌بینی علاقهٔ کاربران، روش طبقه‌بندی تجمعی اســتفاده شــده اســت که خروجی الگوریتم‌های طبقه‌بندی پایه‌ای را دریافت کرده است و مبتنی بر روش رأی‌گیری، خروجی نهایی را تعیین می‌کند. در نهایت در گام پنجم، روش‌های طبقه‌بندی پایه‌ای و طبقه‌بند تجمعی پیشنهادشده، بر اساس معیارهای دقت، حساسیت و پوشش ارزیابی شده‌اند.

1. Patel & Passi



شکل ۳. روش پیشنهادی پیش‌بینی و طبقه‌بندی علاقهٔ کاربران در کلان‌دادهٔ توییتر

۱. پیش‌پردازش داده‌ها

در پیش‌پردازش، ابتدا دادهٔ خروجی از مجموعه دادهٔ توییتر، توسط روش جایگزینی داده‌های مفقودشده[1]، مقداردهی می‌شوند، به این‌صورت که داده‌های مفقودشده با مقادیر مشخص‌شده، جایگزین و مقداردهی می‌شوند. در اصل مقدارهای مفقودشده در توییتر یا بر مبنای سیستم یا بر مبنای کاربر ثبت صورت نمی‌گیرد، و یا در داده‌های آماری باید لحاظ شوند، در حالی‌که داده‌ای ثبت نشده است که این ثبت‌نشدن یا حاصل نویز و یا خطا در روال سیستم بوده است و

1. Replace Missing Value

یا حاصل عدم درج کاربر و از این‌رو، در مجموعه‌داده باعث ایجاد داده مفقودشده می‌شود. بعد از مرحلـهٔ جایگزینی، انتخاب ویژگی[1] صورت می‌پذیرد و خصوصیت‌های مهم و مؤثر توییت که به‌منظور تجزیه‌وتحلیل مورد نظر است، انتخاب می‌شود. در نهایت در مرحلهٔ پیش‌پردازش، حـذف تکـرار[2] انجام می‌گیرد، به‌طوری‌کـه هر رکورد که دارای خصوصیت تکراری باشـد را حذف می‌کند و از مقادیر زائد داده و حجم پردازش بیهوده آن‌ها جلوگیری می‌شود.

طبقه‌بندهای منفرد پایه

در این بخش، به تشـریح و توسعهٔ هر یک از الگوریتم‌های طبقه‌بندی پایه‌ای و قدرتمند شامل بیزین سـاده، K- نزدیک‌ترین همسـایه، جنگل تصادفی و درخت تصمیـم، روی توییت‌های کلان‌داده رسـانه اجتماعی توییتر پرداخته شده اسـت. به‌عنوان اولین الگوریتم طبقه‌بندی پایه به‌منظور پیش‌بینی علاقهٔ کاربران، از طبقه‌بند بیزین ساده استفاده شده است که در شکل ۴ نشان داده شده است (فیو، ۲۰۰۹: ۶).

ALGORITHM NB TRAINING

1. Let V be the vocabulary of ALL words in the documents in D
2. For each category $c_i \in C$
   - Let Di be the subset of documents in D in category c_i
   - $P(c_i) = |D_i| / |D|$
   - Let T_i be the concatenation of all the documents in D_i
   - Let n_i be the total number of word occurrences in T_i
   - For each word $w_j \in V$
     - Let n_{ij} be the number of occurrences of w_j in T_i
     - Let $P(w_j | c_i) = (n_{ij} + 1) / (n_i + |V|)$

ALGORITHM NB TESTING

1. Given a test document X
2. Let n be the number of word occurrences in X
3. Return the category:

$$\arg\max_{c_i \in C} P(c_i) \prod_{i=1}^{n} P(a_i | c_i)$$

where a_i is the word occurring at the i^{th} position in X

شکل ۴. الگوریتم طبقه‌بندی بیزین ساده (پاتیل و پاوار[3]، ۲۰۱۲: ۲۸)

1. Feature Selection
2. Remove Duplicate
3. Patil & Pawar

دومین الگوریتم طبقه‌بندی پایه برای پیش‌بینی علاقهٔ کاربران، طبقه‌بند K- نزدیک‌ترین همسایه است که مبتنی بر نمونه است و از نوع یادگیرنده‌های تنبل است که در شکل ۵ نشان داده شده است (فیو، ۲۰۰۹ :۶).

```
K ← number of nearest neighbors
  For each object X in the test set do
      calculate the distance D(X,Y) between X and every object Y in the training set
      neighborhood ← the k neighbors in the training set closest to X
      X.class ← SelectClass (neighborhood)
  End for
```

شکل ۵. الگوریتم طبقه‌بندی K ـ نزدیک‌ترین همسایه (آرچانا و الانگوان[1]، ۲۰۱۴ : ۱۳)

به‌عنوان سومین الگوریتم طبقه‌بندی پایه به‌منظور پیش‌بینی علاقهٔ کاربران، از طبقه‌بند درخت تصمیم که در مقابل داده‌های نویزدار حساس نیست، استفاده شده است و در شکل ۶ نشان داده شده است (هان و کامبر[2]، ۲۰۰۶ : ۱۶).

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

شکل ۶. الگوریتم طبقه‌بندی درخت تصمیم

چهارمین الگوریتم طبقه‌بندی پایه برای پیش‌بینی علاقهٔ کاربران، طبقه‌بند جنگل تصادفی است و در رتبه‌بندی اهمیت متغیرها در یک مسئلهٔ رگرسیونی یا طبقه‌بندی کاربرد فراوانی داشته است. این الگوریتم در شکل ۷ نشان داده شده است (زکریا، ۲۰۱۴ : ۱۲).

1. Archana & Elangovan
2. Han & Kamber

رسانه / سال سی‌وپنجم / شمارهٔ ۲ / پیاپی ۱۳۵ /

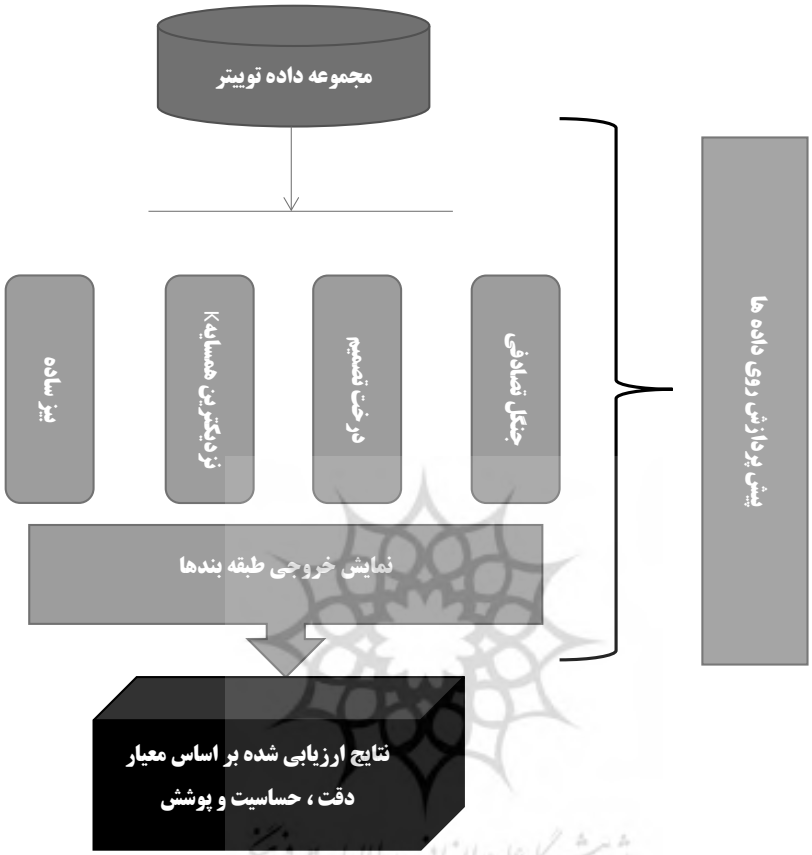
```
To generate c classifiers:
for i = 1 to c do
   Randomly sample the training data D with replacement to produce D_i
   Create a root node, N_i containing D_i
   Call BuildTree( N_i )
end for

BuildTree(N):
if N contains instances of only one class then
   return
else
   Randomly select x% of the possible splitting features in N
   Select the feature F with the highest information gain to split on
   Create f child nodes of N, N_1, ..., N_f, where F has f possible values (F_1, ..., F_f)
   for i = 1 to f do
      Set the contents of N_i to D_i, where D_i is all instances in N that match
      F_i
      Call BuildTree( N_i )
   end for
end if
```

شکل ۷. الگوریتم طبقه‌بندی جنگل تصادفی (پینتوو کریا[1]، ۲۰۱۸ : ۲۳)

1. Pinto & Correia

مدل پیشنهادی طبقه‌بند تجمعی



شکل ۸. مدل پیشنهادی

طبقه‌بندهای ترکیبی از ترکیبِ چندین طبقه‌بند استفاده می‌کنند. در واقع این طبقه‌بندها، هرکدام مدل خود را بر روی داده‌ها می‌سازند و این مدل را ذخیره‌سازی می‌کنند. در نهایت برای طبقه‌بندیِ نهایی، یک رأی‌گیری در میان این طبقه‌بندها انجام می‌شود و آن طبقه‌ای که بیشترین میزان رأی را بیاورد، طبقهٔ نهایی محسوب می‌شود. به‌طورکلی، این روش تلاش می‌کند تا از خطاهای طبقه‌بندی‌کننده‌های پایه‌ای، با هدف دستیابی به خروجی دقیق‌تر انجام پذیرد (مارتینز، واسکز[1]: ۳۰)، (تروساس، کروسا و همکاران، ۲۰۱۶). روش تجمع خودراه‌انداز[2] نمونه‌های تصادفی از مجموعه‌داده‌های آموزشی

1. Martınez-Cámara, Gutiérrez-Vázquez
2. Bootstrap Aggregation

را تولید می‌کند و الگوریتم‌های یادگیری پایه‌ای را روی هر نمونه اعمال می‌کند. سپس نتایج این طبقه‌بندی‌کننده‌های چندگانه با استفاده از رأی‌گیری میانگین یا اکثریت ترکیب می‌شوند. این روش واریانس رویه‌های ناپایدار نظیر درخت‌های تصمیم را کاهش می‌دهد و در نتیجه، معیار دقت را بهبود می‌بخشد (تروساس، کروسا و همکاران، ۲۰۱۶: ۳۱).

روش تقویت[1]، یک روش مجموعه‌ای است که مدل‌های مکمل با آموزش هر مدل جدید بر روی مدل‌های قبلی که به اشتباه طبقه‌بندی شده‌اند، تولید می‌شوند. این روش تا زمانی که به محدودیتی در تعداد مدل‌ها یا معیار دقت برسد، تکرار می‌شود. اگرچه روش تقویت در برخی موارد بهتر از کیسه‌کردن عمل کرده است، اما به احتمال زیاد بیش‌ازحد داده‌های آموزشی را تطبیق می‌دهد. به‌عنوان طبقه‌بندی‌کننده ضعیف، به‌طور معمول از طبقه‌بندی‌کننده‌های مبتنی بر قانون، درخت‌های یک تا دو سطحی، شبکه‌های عصبی بدون لایه‌های پنهان و غیره استفاده می‌شود (تروساس، کروسا و همکاران، ۲۰۱۶: ۳۱).

روش پشته[2]، الگوریتم‌های یادگیری متفاوتی را برروی داده‌های آموزشی اجرا می‌کند و سپس، از یک متاطبقه‌بند استفاده می‌کند که پیش‌بینی‌های هر طبقه‌بندی‌کننده را به‌عنوان ورودی اضافی می‌گیرد. این می‌تواند منجر به کاهش خطای سوءگیری یا واریانس بسته به یادگیرنده ترکیبی مورد استفاده قرار گیرد. انباشتن اغلب عملکرد بهتری نسبت به هریک از مدل‌های آموزش‌دیده دارد. یک مدل رگرسیون لجستیک تک‌لایه اغلب به‌عنوان ترکیب‌کننده استفاده می‌شود (تروساس، کروسا و همکاران، ۲۰۱۶: ۳۱).

روش رأی‌گیری[3] تلاش می‌کند تا مشکل تطابق را حل کند. سپس می‌توان از یک تولید رأی‌گیری ساده برای تخصیص داده‌ها برای اجماع نهایی استفاده کرد. رأی‌گیری می‌تواند برای تعیین عضویت هر دسته برای هر شیء اعمال شود. با این‌حال، در روش رأی‌گیری داده‌های مختلفی که از مدل‌های متفاوتی تولید شده است را بر اساس رأی‌گیری اکثریت بر مبنای معیارهای مورد نظر ازجمله دقت انتخاب می‌کند و به‌عنوان خروجی نهایی ارائه می‌دهد (قائمی، سلیمان[4] و همکاران، ۲۰۰۹: ۲۷)، (تروساس، کروسا و همکاران، ۲۰۱۶: ۳۱).

در این مقاله، با استفاده از روش طبقه‌بندی تجمعی و تکنیک رأی‌گیری بر مبنای اکثریت آرا ازمیان خروجی هر یک از طبقه‌بندهای به‌کارگرفته‌شده شامل K- نزدیک‌ترین همسایه، جنگل تصادفی، درخت تصمیم، و بیز ساده، و برمبنای معیارهای دقت، صحت و پوشش آن طبقه‌بندی که اکثریت آرا را داشته است، یعنی بهترین خروجی ممکن به‌منظور پیش‌بینی رفتار کاربران در مجموعه داده توییتر را ارائه می‌دهد.

1. Boosting
2. Staking
3. Voting
4. Ghaemi &Soliman

ارزیابی نتایج آزمایش

در ایـن بخش، تنظیمـات محیط سـخت‌افزاری و شبیه‌سـازی آزمایش، مجموعـه داده‌های آزمایش‌شـده و معیارهایی که مورد ارزیابی قرار گرفته شده است، توصیف شده‌اند. در نهایت، نتایج آزمایش مورد ارزیابی قرار گرفته شده است و بین نتایج حاصل از الگوریتم‌های طبقه‌بندی، مقایسه صورت پذیرفته است.

۱. تنظیمات محیط سخت‌افزاری و شبیه‌سازی آزمایش

کلیهٔ شبیه‌سـازی‌های آزمایش‌ها در این مقاله، روی سیسـتمی با پردازندهٔ ۵ هسته‌ای و سرعت ۴/۴ گیگاهرتز، حافظهٔ اصلی (RAM) ۸ گیگابایت و حافظه فیزیکی ۱ ترابایت، و سیستم عامل ویندوز ۷ نسخه ۶۴ بیتی، اجرا و آزمایش شده است. به‌علاوه، کلیه شبیه‌سازی‌های الگوریتم‌های طبقه‌بندی پایه و مدل پیشـنهادی طبقه‌بندی تجمعی، با اسـتفاده از شبیه‌ساز داده‌کاوی Rapid Miner نسخه ۰/۹ توسعه و پیاده‌سازی شده است.

۲. مجموعه داده آزمایش‌شده

کلیه الگوریتم‌های طبقه‌بندی پایه‌ای و مدل پیشنهادی طبقه‌بندی تجمعی شبیه‌سازی‌شده، بر روی مجموعه‌دادهٔ Tweets.csv که مربوط به کلان‌داده رسـانهٔ اجتماعی توییتر است، اجرا و آزمایش‌شده است. مجموعه دادهٔ Tweets دارای تعداد ۵۲۵۴۳ نمونه و همچنین ۱۰ ویژگی محتوای موجود شامل محتوا، کشـور، تاریخ، زمان، شناسـه، زبان، تعداد مورد علاقه‌ها، تعداد اشتراک گذاشته‌شده و نوع محتوا (رشته‌ای یا عددی) است. به‌علاوه، مجموعه دادهٔ Tweets دارای ۱۰ کلاس مختلف است.

۳. معیارهای ارزیابی‌شده

در این مقاله، صحت، دقت و پوشـش معیارهایی هسـتند که الگوریتم‌هـای طبقه‌بندی پایه‌ای و مدل پیشـنهادی طبقه‌بندی تجمعی، بر اسـاس این سـه معیار مورد ارزیابی قـرار گرفته‌اند. مطابق رابطهٔ (۶)، معیار صحت برابر اسـت با حاصل‌تقسیم تعداد نمونه‌هایی که به‌طور صحیح پیش‌بینی شده‌اند، بر تعداد کل پیش‌بینی‌های انجام‌شده (هان، کامبر، ۲۰۰۶: ۱۶).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (۶)$$

که در آن، TP تعداد نمونه‌های مثبت صحیح پیش‌بینی‌شـده، TN تعداد نمونه‌های منفی صحیح پیش‌بینی‌شده، FP تعداد نمونه‌های مثبت غلط پیش‌بینی‌شده و FN تعداد نمونه‌های منفی غلط پیش‌بینی‌شده هستند.

در معیار دقت، هرچه میزان پیش‌بینی غلط از میزان پیش‌بینی صحیح بیشتر باشد، مقدار دقت کمتر است. دقت برابر است با حاصل تقسیم تعداد نمونه‌های مثبت صحیح پیش‌بینی‌شده بر تعداد نمونه‌های مثبت صحیح پیش‌بینی‌شده و تعداد نمونه‌های منفی غلط پیش‌بینی‌شده که در رابطه (۷) نشان داده شده است (هان، کامبر، ۲۰۰۶ : ۱۶).

$$Precision = \frac{TP}{TP+FP} \qquad (۷)$$

مطابق با رابطه (۸) (هان، کامبر، ۲۰۰۶ : ۱۶)، معیار پوشش برابر است با تعداد نمونه‌های مثبت صحیح پیش‌بینی‌شده، تقسیم بر مجموع تعداد نمونه‌های مثبت صحیح پیش‌بینی‌شده و تعداد نمونه‌های منفی غلط پیش‌بینی‌شده. هرچه میزان تعداد نمونه‌هایی که غلط پیش‌بینی شده است، بیشتر از تعداد نمونه‌های صحیح پیش‌بینی شده باشد، میزان پوشش کمتر است.

$$Recall = \frac{TP}{TP+FN} \qquad (۸)$$

۴. ارزیابی نتایج آزمایش

در این بخش، مطابق با آزمایش‌های انجام‌شده و اجرای چهار الگوریتم طبقه‌بندی پایه‌ای و قدرتمند شامل K- نزدیک‌ترین همسایه، بیزین ساده، جنگل تصادفی و درخت تصمیم و همچنین روش پیشنهادی طبقه‌بندی تجمعی مبتنی بر رأی‌گیری، روی مجموعه دادهٔ توییتر به‌نام Tweets، به ارزیابی این الگوریتم‌ها براساس معیارهای دقت، صحت، و پوشش پرداخته شده است. در روش پیشنهادی طبقه‌بندی تجمعی، از روش رأی‌گیری برای به‌دست آوردن خروجی نهایی طبقه‌بندی با استفاده از خروجی هر یک از طبقه‌بندهای پایه‌ای، استفاده شده است و انتظار می‌رود که معیارهای صحت، دقت، و پوشش بهبود حاصل شود.

در جدول ۱ نتایج آزمایش اجرای چهار الگوریتم طبقه‌بندی پایه‌ای شامل K- نزدیک‌ترین همسایه، بیزین ساده، جنگل تصادفی و درخت تصمیم و همچنین، روش پیشنهادی طبقه‌بندی تجمعی مبتنی بر رأی‌گیری و بر مبنای حداکثر آرا، روی مجموعه دادهٔ توییتر Tweets و بر اساس معیار دقت نشان داده شده است. طبق نتایج به‌دست آمده از خروجی هر یک از طبقه‌بندهای پایه‌ای بر اساس معیار دقت روی مجموعه دادهٔ Tweets، بدترین عملکرد متعلق به طبقه‌بند بیزین ساده با مقدار ۳۷/۹۲ درصد و بهترین عملکرد متعلق به طبقه‌بند جنگل تصادفی با مقدار ۹۲/۹۷ درصد بوده است. با استفاده از روش طبقه‌بند تجمعی پیشنهادی مبتنی بر روش رأی‌گیری، بهترین خروجی از میان خروجی‌های کلیه طبقه‌بندها انتخاب می‌شود. همان‌طور که مشاهده می‌شود، بهترین عملکرد براساس معیار دقت با مقدار ۹۲/۹۷، متعلق به خروجی

طبقه‌بنـد پایه جنـگل تصادفی و همچنیـن، روش طبقه‌بند تجمعی پیشـنهادی مبتنی بر روش رأی‌گیری بوده است.

جدول ۱. نتایج الگوریتم‌های طبقه‌بندی پایه‌ای و تجمعی براساس معیار دقت

K- نزدیک‌ترین همسایه	بیزین ساده	جنگل تصادفی	درخت تصمیم	طبقه‌بند تجمعی
۹۲/۵۳درصد	۳۷/۹۲درصد	۹۲/۹۷درصد	۹۲/۹۶درصد	۹۲/۹۷درصد

در جـدول ۲ نتایـج آزمایش اجرای چهار الگوریتم طبقه‌بندی پایه‌ای شـامل K ـ نزدیک‌ترین همسـایه، بیزین ساده، جنگل تصادفی و درخت تصمیم و همچنین، روش پیشنهادی طبقه‌بندی تجمعـی مبتنی بـر رأی‌گیری و برمبنای حداکثـر آرا، روی مجموعـه داده توییتر Tweets و بر اسـاس معیار صحت نشـان داده شده اسـت. طبق نتایج به‌دسـت آمده از خروجی هریک از طبقه‌بندهای پایه‌ای بر اساس معیار صحت روی مجموعه داده Tweets، بدترین عملکرد متعلق به طبقه‌بند بیزین سـاده با مقدار ۱۵/۱۴ درصد، و بهترین عملکرد متعلق به طبقه‌بندهای پایه‌ای جنگل تصادفی و درخت تصمیم با مقدار ۶/۰۶ درصد بوده اسـت. با استفاده از روش طبقه‌بند تجمعـی پیشـنهادی مبتنی بر روش رأی‌گیری و بر مبنای حداکثـر آرا، بهترین خروجی از میان خروجی‌های کلیه طبقه‌بندها انتخاب می‌شـود. همان‌طور که مشـاهده می‌شـود، از آنجایی‌که خروجـی طبقه‌بندهای جنـگل تصادفی و درخت تصمیـم با یکدیگر برابر بودنـد، لذا بعد از رأی‌گیری ازمیان طبقه‌بندها، خروجی‌های طبقه‌بندهای پایه‌ای جنگل تصادفی و درخت تصمیم را به‌عنوان خروجی بهینه و نهایی در نظر گرفته است. بنابراین، عملکرد روش طبقه‌بند تجمعی پیشنهادی مبتنی بر روش رأی‌گیری بر اساس معیار صحت، مقدار ۶/۰۶ درصد بوده است.

جدول ۲. نتایج الگوریتم‌های طبقه‌بندی پایه‌ای و تجمعی براساس معیار صحت

K- نزدیک‌ترین همسایه	بیزین ساده	جنگل تصادفی	درخت تصمیم	طبقه‌بند تجمعی
۶/۷۳ درصد	۱۵/۱۴ درصد	۶/۰۶ درصد	۶/۰۶ درصد	۶/۰۶ درصد

در جـدول ۳ نتایـج آزمایش اجرای چهار الگوریتم طبقه‌بندی پایه‌ای شـامل K ـ نزدیک‌ترین همسـایه، بیزین ساده، جنگل تصادفی و درخت تصمیم و همچنین روش پیشنهادی طبقه‌بندی تجمعی مبتنی بر رأی‌گیری و برمبنای حداکثر آرا، روی مجموعه دادهٔ توییتر Tweets و براساس معیار پوشش نشان داده شده است. طبق نتایج به‌دست آمده از خروجی هریک از طبقه‌بندهای پایه‌ای براسـاس معیار پوشـش روی مجموعه داده Tweets، بدترین عملکرد متعلق به طبقه‌بند پایه‌ای بیزین ساده با مقدار ۷/۴۲ درصد و بهترین عملکرد متعلق به طبقه‌بندهای پایه‌ای جنگل

تصادفی و درخت تصمیم با مقدار ۵/۷۴ درصد بوده است. با استفاده از روش طبقه‌بند تجمعی پیشنهادی مبتنی بر روش رأی‌گیری، بهترین خروجی از میان خروجی‌های کلیه طبقه‌بندها انتخاب می‌شود. همان‌طور که مشاهده می‌شود، ازآنجایی که خروجی طبقه‌بندهای جنگل تصادفی و درخت تصمیم با یکدیگر برابر بودند، لذا بعد از رأی‌گیری از میان طبقه‌بندها، خروجی‌های طبقه‌بندهای پایه‌ای جنگل تصادفی و درخت تصمیم را به‌عنوان خروجی بهینه و نهایی در نظر گرفته است. بنابراین، عملکرد روش طبقه‌بند تجمعی پیشنهادی مبتنی بر روش رأی‌گیری براساس معیار صحت، مقدار ۵/۷۴ درصد بوده است.

جدول ۳. نتایج الگوریتم‌های طبقه‌بندی پایه‌ای و تجمعی براساس معیار پوشش

K- نزدیک‌ترین همسایه	بیزین ساده	جنگل تصادفی	درخت تصمیم	طبقه‌بند تجمعی
۶/۷۳ درصد	۷/۴۲ درصد	۵/۷۴ درصد	۵/۷۴ درصد	۵/۷۴ درصد

نتیجه‌گیری و کارهای آینده

امروزه استقبال شدید از رسانه‌ها و شبکه‌های اجتماعی و همچنین افزایش حجم ذخیره‌سازی و مبادلات پیام‌ها، باعث ایجاد کلان‌داده‌ها شده است. از کلان‌داده‌های شبکه‌های اجتماعی برای پیش‌بینی علاقهٔ کاربران، شناسایی رفتار آن‌ها و پیش‌بینی موارد کاربردی دیگر استفاده می‌شود. اما سامان‌دهی این کلان‌داده‌ها به‌منظور استفادهٔ بهینه، خود یک چالش بزرگ است. یکی از تکنیک‌های سامان‌دهی این داده‌ها برای آماده‌سازی داده‌ها به‌منظور پیش‌بینی علاقه یا رفتار کاربران و غیره، تکنیک طبقه‌بندی است. در این مقاله، با استفاده از الگوریتم‌های پایه‌ای طبقه‌بندی شامل K- نزدیک‌ترین همسایه، بیزین ساده، جنگل تصادفی و درخت تصمیم و همچنین، روش پیشنهادی طبقه‌بندی تجمعی مبتنی بر روش رأی‌گیری، روی مجموعه دادهٔ توییتر Tweets و بر اساس معیارهای دقت، صحت و پوشش نشان داده شد که روش پیشنهادی طبقه‌بندی تجمعی مبتنی بر روش رأی‌گیری نسبت‌به چهار الگوریتم دیگر طبقه‌بندی پایه‌ای، بر اساس هرسه معیار دقت، صحت و پوشش، دارای نتایج مطلوب‌تری بوده است که نشان از کارآمدی عملکرد روش پیشنهادی طبقه‌بندی تجمعی مبتنی‌بر روش رأی‌گیری است.

برای کارهای آتی، می‌توان از الگوریتم‌های دیگر طبقه‌بندی نظیر ماشین بردار پشتیبان استفاده کرد. علاوه بر این، روش‌های خوشه‌بندی و تلفیق کاربرد روش‌های خوشه‌بندی با طبقه‌بندی به‌منظور تجزیه‌وتحلیل داده‌ها و پیش‌بینی علاقه یا رفتار کاربران در شبکه‌های اجتماعی می‌تواند از کارهای آینده در این مقاله به‌شمار رود.

منابع

Abbas, A. Kh., Ali Khalil, S., Harith, A. H., Qasim, M. H., & Saba Alaa, A. (2020). Twitter sentiment analysis using an ensemble majority vote classifier. *Journal of Southwest Jiaotong University*, 55(1).

Adewole, K. S., Tao, H., Wanqing, W., Houbing, S., & Arun Kumar, S. (2020). Twitter spam account detection based on clustering and classification methods." *The Journal of Supercomputing*, 76(7), 4802-4837.

Al-Hashedi, A., Belal A. F., Abdulqader, M., Mohsen, Y. A., Hasan Ali, G., A.K, Wedad, A., & Naseebah, M. (2022). Ensemble classifiers for Arabic sentiment analysis of social network (Twitter data) towards COVID-19-related conspiracy theories. *Applied Computational Intelligence and Soft Computing* 2022.

Archana, S., & Elangovan, K. (2014). Survey of classification techniques in data mining." *International Journal of Computer Science and Mobile Applications*, 2(2) , 65-71.

Bayhaqy, A., Sfenrianto, S., Kaman, N., & Emil R. K. (2018). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. In 2018 *international conference on orange technologies* (ICOT), 1-6.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends* 2, 1 , 20-28.

Ghaemi, R., Md Nasir, S., Hamidah, I., & Norwati M. (2009). "A survey: clustering ensembles techniques. *International Journal of Computer and Information Engineering* 3, 2, 365-374.

Ghani, Norjihan Abdul, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. "Social media big data analytics: A survey." *Computers in Human Behavior* 101 (2019): 417-428.

Han, J., & Micheline K. (2006). Data mining: concepts and techniques, 2nd. *University of Illinois at Urbana Champaign: Morgan Kaufmann* .

Harjule, P., Astha, G., Harshita, S., & Priya T. (2020). Text classification on Twitter data. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, 160-164.

Kesavaraj, G., & Sreekumar S. (2013). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, 1-7.

Kesavaraj, G., & Sreekumar S.(2013). A study on classification techniques in data mining. In 2013 *fourth international conference on computing, communications and networking technologies* (ICCCNT), 1-7.

Martınez-Cámara, E., Gutiérrez-Vázquez, Y. Javi, F, Montejo-Ráez, A., & Munoz-Guillena, R. (2015). Ensemble classifier for twitter sentiment analysis. NLP *Applications: completing the puzzle*, 1-12.

Metsis, V., Ion, A., & Georgios, P. (2006). Spam filtering with naive bayes-which naive bayes? In CEAS, 17, 28-69.

Muqorobin, M., Siti, R., Isnawati, M., & Nendy, A. R. R. (2020). Classification of Community Complaints Against Public Services on Twitter. *International Journal of Computer and Information System (IJCIS)*, 1(1), 7-10.

Murthy, D. (2017). The ontology of tweets: Mixed methods approaches to the study of Twitter. *The SAGE handbook of social media research methods*, 559-572.

Neogi, A. S., Kirti Anilkumar, G., Ram Krishn, M., &) Yogesh, K. D. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, 1(2).

Oussous, A., Benjelloun, F., Ait Lahcen, R., & Belfkih. S. (2018). Big Data technologies: A survey. *King Saud University-Computer and Information Sciences* 30, 4, 431-448.

Papakyriakou, D., & Ioannis S. (2022). Data Mining Methods: A Review. *International Journal of Computer Applications* 975: 8887.

Patel, R., & Kalpdrum P. (2020). Sentiment analysis on twitter data of world cup soccer tournament using machine learning. *IoT* 1, 2 (14).

Patil, A. S., & Pawar, B. V. (2012) Automated classification of web sites using Naive Bayesian algorithm.

In *Proceedings of the international multiconference of engineers and computer scientists*, 1, 519-523.
Phan, H. T., Ngoc Th. N., Van Cuong T., & Dosam H. (2021). An approach for a decision-making support system based on measuring the user satisfaction level on twitter. *Information Sciences 561* , 243-273.
Phyu, Th.N. (2009). Survey of classification techniques in data mining. In *Proceedings of the international multiconference of engineers and computer scientists*, 1(5), Citeseer.
Pinto, L. R. M., & Henrique Andrade Correia,L. (2018). Analysis of machine learning algorithms for spectrum decision in cognitive radios. In *2018 15th International symposium on wireless communication systems (ISWCS)*, 1-6.
Sen, P. Ch., Mahimarnab, H., & Mitadru, Gh. (2020). "Supervised classification algorithms in machine learning: A survey and review." In *Emerging technology in modelling and graphics*, 99-111.
Shaik, A. B., & Sujatha S. (2019). A brief survey on random forest ensembles in classification model. In *International Conference on Innovative Computing and Communications*, 253-260.
Shamrat, F. M. J. M., Sovon Chakraborty, M. M., Imran, J., Naeem, M., Md Masum, B, Protiva, D., & Rahman, O. M. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 23(1), 463-470.
Sri, P.A., & Anusha, M. (2016). Big data-survey. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* 4 (1), 74-80.
Troussas, Ch., Akrivi, K., & Virvou, M .(2016). Evaluation of ensemble-based sentiment classifiers for Twitter data. In *2016 7th international conference on information, intelligence, systems & applications (IISA)*, 1-6.
Vashisth, P., & Kevin, M. (2020). Gender classification using twitter text data. In *2020 31st Irish Signals and Systems Conference (ISSC)*, 1-6.
Zakariah, M. (2014). Classification of large datasets using Random Forest Algorithm in various applications: Survey. *International Journal of Engineering and Innovative Technology,(IJJEIT)* 4(3).