Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 9, Number 3, Fall 2020, pp. 132-147
Received: 10/21/2019, Revised: 03/12/2020, Accepted: 05/07/2020

Investigating the Impact of Ensemble Machine Learning Methods on Spam Review Detection Based on Behavioral Features

Shahriar Mohammadi[1*], Mir Reza Mousavi[2]
1*- Department of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran.
2- Department of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran.
[1*]Mohammadi@kntu.ac.ir, [2]Mirreza.mousavi@email.kntu.ac.ir

Corresponding author's address: Shahriar Mohammadi, Faculty of Industrial Engineering, K.N.Toosi University of Technology, MollaSadra St, Vanak Sq, Tehran, Iran.

Abstract- One of the most influential links on the Internet is the feedback provided by consumers as an experience of using the product to the people who want to buy that product. Beneficiaries use this opportunity to transfer inaccurate experience in order to promote or demote the value of a particular service or product unjustly, and this is the cause of placing their reviews between spam reviews category. Therefore, identifying these reviews using machine learning techniques and ensemble learners has become a hot topic among researchers. The purpose of this study is to investigate the impact of using ensemble machine learning methods on identifying such reviews using behavioral features. Recent studies have shown that the ensemble methods used in this study in combination with text-based features in addition to imposing more computational expense are not able to improve the performance of the best base learners. In this study, in addition to identifying the best base and ensemble learners in using behavioral features, we seek to determine whether these features combination with ensemble learners can achieve greater accuracy or a significant change in model performance. For this purpose, seven base learners and four ensemble learners such as Bagging, Boosting, Random Forest and Extra Tree were used and the results were compared with the results of using text-based features. Our evaluations show that using the decision tree as a base learner, along with the method of boosting in unbalanced dataset and bagging in balanced dataset, yields better results and we can achieve more tangible change in the performance of the best base algorithms by ensemble learners in using behavioral features over text-based.

Keywords- Spam Reviews, Machine Learning, Ensemble Methods, Behavioral Features.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۹، شماره ۳، پائیز ۱۳۹۹ ، صفحه ۱۳۲-۱۴۷
دریافت: ۱۳۹۸/۰۷/۲۹، بازنگری: ۱۳۹۸/۱۲/۲۲، پذیرش: ۱۳۹۹/۰۲/۱۸

بررسی تأثیر استفاده از روش‌های یادگیری ماشین تجمعی در شناسایی نظرهای هرز بر اساس ویژگی‌های رفتاری

شهریار محمدی*[۱]، میررضا موسوی[۲]
*۱- دانشکده مهندسی صنایع، دانشگاه صنعتی خواجه‌نصیرالدین طوسی، تهران، ایران.
۲- دانشکده مهندسی صنایع، دانشگاه صنعتی خواجه‌نصیرالدین طوسی، تهران، ایران.
[1*]Mohammadi@kntu.ac.ir, [2]Mirreza.mousavi@email.kntu.ac.ir

* نشانی نویسنده مسئول: شهریار محمدی، تهران، میدان ونک، خیابان ملاصدرا، دانشگاه صنعتی خواجه‌نصیرالدین طوسی، دانشکده مهندسی صنایع.

چکیده- یکی از تأثیرگذارترین ارتباط‌ها در اینترنت، نظرهایی است که توسط افراد مصرف‌کننده یک محصول به‌عنوان تجربه استفاده، در اختیار افراد خواهان خرید محصول قرار می‌گیرد. استفاده سودجویان از این فرصت انتقال تجربه، به‌منظور ارتقا یا تنزل ارزش یک خدمت یا محصول خاص به‌ناحق، باعث قرارگیری نظرهای آن‌ها در دسته نظرهای هرز می‌شود. ازاین‌رو شناسایی این نظرها با استفاده از روش‌های یادگیری ماشین و یادگیرنده‌های تجمعی به مبحثی داغ در میان محققان تبدیل شده است. هدف این مطالعه بررسی تأثیر استفاده از روش‌های یادگیری ماشین تجمعی در شناسایی اینگونه نظرها با استفاده از ویژگی‌های رفتاری است. بررسی‌های اخیر نشان داده است که روش‌های تجمعی مورد استفاده در این مطالعه در ادغام با ویژگی‌های متنی علاوه بر تحمیل بار محاسباتی بیشتر قادر به ارتقای عملکرد بهترین الگوریتم‌های پایه نیستند. در این مطالعه علاوه‌بر شناسایی بهترین یادگیرنده‌های پایه و تجمعی در استفاده از ویژگی‌های رفتاری به‌دنبال آن هستیم که آیا می‌توان با استفاده از این ویژگی‌ها و یادگیرنده‌های تجمعی به دقتی بیشتر و یا تغییر محسوسی در عملکرد مدل دست یابیم. بدین منظور از هفت یادگیرنده پایه و چهار یادگیرنده تجمعی دسته‌بندی، تقویت‌سازی، جنگل تصادفی و درخت اضافی استفاده شد و نتایج حاصل با نتایج استفاده از ویژگی‌های متنی مورد مقایسه قرار گرفت. ارزیابی‌ها نشان‌دهنده عملکرد بهتر یادگیرنده پایه درخت تصمیم به‌همراه روش تجمعی تقویت‌سازی در حالت استفاده از مجموعه‌داده نامتوازن و روش تجمعی دسته‌بندی در استفاده از مجموعه‌داده متوازن و هم‌چنین تغییر محسوس‌تر عملکرد بهترین الگوریتم پایه، توسط یادگیرنده‌های تجمعی، در استفاده از ویژگی‌های رفتاری نسبت به متنی است.

واژه‌های کلیدی: نظرهای هرز، یادگیری ماشین، روش‌های تجمعی، ویژگی‌های رفتاری.

۱- مقدمه

با گسترش ارتباطات در دنیای اینترنت، تأثیر روزافزون نظرهای برخط بر تصمیم‌گیری افراد واضح و واضح‌تر می‌شود. افرادی که قصد خرید محصول و یا دریافت خدمتی را از طریق این فضا دارند قبل از هرگونه تصمیم‌گیری نهایی، اقدام به بررسی نظرهای افرادی می‌کنند که تجربه خرید و استفاده از محصول را دارا هستند. میزان اهمیت و تاثیری که این نظرها در رویکرد اتخاذی خریداران احتمالی دارند باعث شده است که زمینه فعالیت برای افراد سودجو نیز فراهم شود. هدف اصلی و اولیه نظرهای برخط، کمک به خریداران و تسهیل در تصمیم‌گیری این افراد است و از طرف دیگر، بازتاب‌های دریافتی از کاربران به تولیدکنندگان نیز در شناسایی نقاط ضعف و قوت محصول‌های خود کمک شایانی می‌کند؛ اما این حقیقت که هر کسی می‌تواند بدون هیچ

پیش‌نظارتی اقدام به نظردهی در مورد محصول یا خدمتی کند، باعث شده است که فرصت مناسبی برای برخی افراد به وجود بیاید که با استفاده از آن بتوانند ارزش محصول رقبای خود را تنزل و در مقابل درآمد خود را با ارتقای ناحق ارزش محصول خود، افزایش دهند. این نظرها که با عنوان نظرهای هرز[۱] شناخته می‌شوند، با هدف تاثیرگذاری بر روی اعتبار نظرهای برخط و از بین بردن جامعیت این نظرها ایجاد و با گسترش از طریق مجراهای ارتباطی، موجب گمراهی افراد و لطمه به اعتبار و شهرت یک تجارت یا محصول خاص می‌گردند.

وبگاه‌های آمازون[۲] و یلپ[۳] ازجمله وبگاه‌هایی هستند که با توجه به رویکردی که دارند در بررسی و شناسایی اینگونه نظرها اهتمام بیشتری ورزیده‌اند و بدین منظور پایگاه داده نظرهای خود را در اختیار محققان این حوزه قرار داده‌اند. بر اساس تحقیق‌های صورت گرفته حدود ۲۰٪ از نظرها در وبگاه یلپ هرز هستند [۱].

ایجاد تمایز مابین نظرهای هرز و سالم توسط انسان به‌شدت دشوار و تقریباً غیرممکن است. به‌همین منظور تاکنون تحقیق‌های فراوان و روش‌های متعددی به‌منظور تمییز دادن نظرهای هرز از نظرهای سالم، انجام و معرفی شده است. بیشتر روش‌های معرفی شده از الگوریتم‌های یادگیری ماشین، به‌منظور شناسایی نظرهای هرز بهره می‌برند و می‌توان این روش‌ها را در سه طبقه دسته‌بندی کرد.

تعدادی از این روش‌ها از ویژگی‌های مبتنی بر الگوهای زبانی و متنی استفاده می‌کنند که عمدتاً بر اساس بردارهای یک‌تایی[۴] و دوتایی[۵] از کلمات هستند [۴-۲]. تعدادی دیگر، از ویژگی‌های استخراج شده از رفتارهای نویسندگان این نظرها (هرزنویس‌ها) بهره می‌برند که این رفتارها را می‌توان با تمرکز بر روی فراداده‌های[۶] مرتبط با نظر از قبیل تاریخ ارسال نظر، بازخوردهای نظر و بازه‌های زمانی ارسال نظر، استخراج و در شناسایی نظرهای هرز و هرزنویس‌ها مورد استفاده قرار داد [۶-۴]. در نهایت تعدادی نیز از الگوریتم‌های مبتنی بر گراف به‌منظور تشکیل شبکه‌ای ارتباطی مابین نظرها، محصول‌ها و افراد نظردهنده استفاده می‌کنند و روش‌هایی را بر این اساس به‌منظور طبقه‌بندی نظرها ارائه می‌دهند [۹-۷].

در حالت کلی الگوریتم‌های یادگیری ماشین در سه دسته تقسیم‌بندی می‌شوند: الگوریتم‌های نظارت‌شده[۷]، نیمه نظارت‌شده[۸] و نظارت‌نشده[۹]. در استفاده از الگوریتم‌هایی که در دسته روش‌های نظارت‌شده قرار دارند باید داده‌هایی در اختیار ماشین قرار گیرند که دارای دسته‌بندی و برچسب‌گذاری مشخصی باشند که ماشین بتواند با استفاده از این داده‌ها مدلی را به‌وجود بیاورد که در شناسایی و طبقه‌بندی نظرهایی که در آینده ایجاد می‌شوند (به دو دسته هرز و سالم) بهترین عملکرد را داشته باشد؛ اما در روش نظارت‌نشده، داده‌هایی که در اختیار ماشین قرار می‌گیرند دارای دسته‌بندی و برچسب مشخصی نیستند و وظیفه ماشین ایجاد یک رابطه مابین این داده‌ها و اختصاص هر کدام به یک دسته خاص است. در روش نیمه نظارتی نیز، داده‌ها در هر دو صورت برچسب‌گذاری‌شده و نشده در اختیار ماشین قرار می‌گیرند [۱۰].

با وجود تمامی تلاش‌هایی که تاکنون در این عرصه صورت گرفته است اما هنوز جنبه‌هایی وجود دارند که مورد بررسی قرار نگرفته‌اند و عملکرد این روش‌ها در شناسایی و طبقه‌بندی نظرهای هرز در هاله‌ای از ابهام قرار دارد. یکی از زمینه‌های نو که اخیراً مورد توجه قرار گرفته، استفاده از روش‌های یادگیری ماشین تجمعی است.

روش‌های تجمعی، به خانواده‌ای از روش‌ها اطلاق می‌شود که چندین الگوریتم طبقه‌بندی پایه را به‌منظور ایجاد یادگیرنده‌های قوی‌تر و عمومی‌تر ادغام می‌کنند [۱۱]. در بررسی‌های اخیر مشخص شده است که روش‌های تجمعی همانند روش‌های دسته‌بندی و تقویتی، عملکرد بهترین الگوریتم‌های پایه را در زمینه‌های مرتبط با الگوهای متنی و زبانی افزایش می‌دهند اما میزان بهبود عملکرد آن‌ها در مقایسه با بار محاسباتی تحمیل شده قابل توجیه نیست و با توجه به دانش ما تاکنون عملکرد این روش‌ها با استفاده از ویژگی‌های رفتاری (مرتبط با محصول، کاربر و نظر) مورد ارزیابی قرار نگرفته است.

از طرف دیگر، بیشتر مطالعه‌های صورت گرفته از نظرهای ساختگی که توسط سامانه‌هایی همانند AMT[10] تولید می‌شوند در پژوهش‌های خود استفاده می‌کنند که شاخص ضعیفی برای نمایش کارایی در نظرهای دنیای واقعی است. از این رو در مطالعه پیش‌رو از نظرهای دنیای واقعی، متعلق به وبگاه یلپ، استفاده می‌شود.

با توجه به اینکه از جمله مشکلاتی که در نظرهای دنیای واقعی وجود دارد عدم توازن در دسته‌های مختلف این داده‌ها است به‌منظور حل این مشکل نیاز است تا روشی برای ایجاد داده‌های متوازن نیز مورد استفاده قرار گیرد که در این مطالعه از روش نمونه‌گیری تصادفی و ارائه شده در [۱۲] استفاده می‌شود. لازم به ذکر است که بررسی‌های این مطالعه هم بر روی مجموعه داده متوازن و هم بر روی مجموعه داده نامتوازن صورت می‌گیرد.

اهدافی که در این کار به‌دنبال آن هستیم عبارتند از:

- شناسایی ویژگی‌های مختلف رفتاری موجود در مطالعه‌های پیشین و ایجاد مجموعه‌ای از بهترین ویژگی‌ها در شناسایی نظرهای هرز.

- اعمال چندین یادگیرنده پایه ماشین بر روی این مجموعه ویژگی‌ها و شناسایی بهترین عملکرد مابین آن‌ها.
- اعمال چندین یادگیرنده تجمعی و بررسی میزان تأثیر آن‌ها در کارایی هر کدام از یادگیرنده‌های پایه با توجه به وجود یا عدم وجود توازن در داده‌های مورد تحلیل.
- مقایسه میزان تغییر ایجاد شده توسط یادگیرنده‌های تجمعی در عملکرد بهترین الگوریتم‌های پایه، با توجه به استفاده از ویژگی‌های رفتاری یا متنی.

در نهایت مشخص خواهد شد که عملکرد یادگیرنده‌های تجمعی در استفاده از کدام ویژگی‌ها و کدام یادگیرنده‌های پایه بهتر است و آیا بار محاسباتی تحمیل‌شده توسط این روش‌ها در ادغام با ویژگی‌های رفتاری قابل توجیه است یا بهتر است به‌دنبال استفاده تنها از یادگیرنده‌های پایه و یا ایجاد سایر روش‌های تجمعی باشیم.

کار تحقیقاتی پیش رو بر مبنای استفاده از الگوریتم‌های یادگیری ماشین نظارت‌شده سامان یافته است. یکی از مشکل‌هایی که در استفاده از الگوریتم‌های نظارت‌شده وجود دارد، کمبود داده‌ها و نظرهای برچسب‌گذاری شده قابل اعتماد است. همان‌طوری که پیش از این ذکر شد، مجموعه داده‌هایی که به‌منظور بررسی نتایج مورد استفاده قرار گرفته است متعلق به وبگاه یلپ است که شامل نظرها و احساس افراد در مورد کیفیت خدمات رستوران‌ها و هتل‌ها است و توسط الگوریتم‌های فیلترسازی توصیه شده توسط وبگاه یلپ، در دو دسته هرز و سالم، برچسب‌گذاری شده است.

در ابتدای این مطالعه ویژگی‌های رفتاری که به‌منظور ایجاد مدل در الگوریتم‌های یادگیری ماشین مورد نیاز است، به‌صورت مختصر معرفی و روش استخراج آن‌ها ذکر می‌شود. سپس چندین یادگیرنده پایه از جمله K نزدیک‌ترین همسایگان، درخت تصمیم و ماشین بردار پشتیبان و ... معرفی و بر روی این مجموعه داده‌ها (با توجه به توازن و عدم توازن داده‌ها) آموزش داده می‌شوند. درنهایت نتایج حاصل از مرحله آموزش و آزمون تکنیک‌های دسته‌بندی، تقویتی، جنگل تصادفی و درخت اضافی محاسبه می‌شوند تا با نتایج حاصل از ویژگی‌های متنی مقایسه شوند.

ادامه این کار تحقیقاتی به‌صورت زیر است. در بخش دوم مفاهیم پایه عنوان می‌شوند. در بخش سوم مروری بر پژوهش‌های پیشین و در بخش چهارم توصیف دقیقی از تمامی مراحل پژوهش ارائه می‌شود. در بخش پنجم نیز تحلیل‌های آماری به‌همراه مقایسه‌های صورت گرفته مابین ویژگی‌های متنی و رفتاری در ادغام با یادگیرنده‌های مورد استفاده، مطرح می‌گردد و بخش آخر نیز شامل جمع‌بندی و پیشنهادهایی در رابطه با کارهای آتی است.

۲- مفاهیم پایه

به فعالیت‌هایی که سعی در گمراه کردن خوانندگان نظرها، با ارائه اطلاعات ناشایست در رابطه با برخی از محصول‌ها یا خدمت‌ها، به‌منظور تبلیغ یا آسیب رساندن به شهرت آن‌ها را دارند، نظرهای هرز اطلاق می‌شوند [۱۳]. در [۱۴] نظرهای هرز به سه دسته تقسیم می‌شوند: نظرهای غیرواقعی[۱۱]، نظرها بر روی نام‌های تجاری[۱۲] و نظرهای غیر[۱۳].

دسته اول نظرها، قصد منحرف کردن خریدار احتمالی با اغراق در مورد ارزش محصول به‌صورت تخریبی یا ترویجی را دارند [۱۵]. دسته دوم نظرها، تمرکز خود را بر روی نام تجاری محصول و نه ویژگی‌های مرتبط با محصول قرار می‌دهند [۱۵]. دسته سوم نیز نظرهایی هستند که عمدتاً شامل تبلیغ‌ها، سؤال و جواب یا متن‌های کاملاً غیرمرتبط با محصول هستند [۱۵]. در این کار، تمرکز ما بر روی شناسایی دسته اول و دوم است.

۳- مروری بر پیشینه تحقیق

تاکنون کارهای مطالعاتی بسیاری در زمینه استفاده از روش‌های تجمعی به‌منظور طبقه‌بندی نظرهای هرز صورت گرفته است که عمده این تحقیق‌ها، تمرکز خود را تنها بر روی استفاده از ویژگی‌های مبتنی بر متن و الگوهای زبانی و رویکردهای واژگانی مربوط به نظرهای برخط قرار داده‌اند و کمتر توجهی به اهمیت استفاده از ویژگی‌های رفتاری در این زمینه تحقیقاتی شده است.

از کارهایی که تاکنون در این زمینه صورت گرفته است می‌توان به روش ارائه شده در [۱۱] اشاره کرد. در این مطالعه تأثیر روش‌های تجمعی همانند دسته‌بندی، تقویتی و جنگل تصادفی بر روی شناسایی و طبقه‌بندی این نظرها مورد بررسی قرار گرفته است ولی به‌منظور ارائه روش اختصاصی، تنها ویژگی‌هایی که به متن نظرها مرتبط است، توسط این کار استفاده شده است. یادگیرنده‌های پایه‌ای که در این کار مورد بررسی قرار گرفته‌اند عبارتند از بیز ساده چندجمله‌ای، درخت تصمیم C4.5، رگرسیون منطقی و ماشین بردار پشتیبان که در نهایت با بررسی‌های صورت گرفته، مشخص شده است که بیز ساده چندجمله‌ای، به‌تنهایی و بدون استفاده از یادگیرنده‌های تجمعی در شناسایی این نظرها، عملکرد بهتری از خود ارائه می‌دهد.

در [۱۶] نیز روش برداری رشته به کلمه[۱۴] به‌منظور ساخت مجموعه ویژگی‌های متنی مورد استفاده قرار گرفته است. به‌منظور بررسی عملکرد این روش نیز از محاسبه سطح زیر منحنی[۱۵] (AUC) بهره گرفته شده است. بررسی‌های انجام شده، نشان داده است که روش‌های تجمعی ذکر شده، عملکرد الگوریتم‌های پایه

ماشین بردار پشتیبان، درخت تصمیم C4.5 و رگرسیون منطقی را تحت‌تأثیر قرار داده و موجب بهبود عملکرد این الگوریتم‌ها شده‌اند ولی در نهایت بیشترین سطح زیرمنحنی و کمترین انحراف معیار استاندارد متعلق به تجمیع الگوریتم بیز ساده چندجمله‌ای با روش تجمعی تقویتی بوده است که بیانگر عملکرد بهتر این الگوریتم در تجمیع با روش تقویت‌سازی در مقایسه با سایر روش‌ها است.

از دیگر مطالعه‌ها در این زمینه می‌توان به [۱۷] اشاره کرد که اقدام به ارائه روشی تجمعی با در نظر گرفتن اهمیت ویژگی‌های مورد استفاده کرده است. در این مطالعه بعد از استخراج ویژگی‌های مرتبط و موردنیاز مرحله آموزش الگوریتم‌های یادگیری ماشین، بهترین این ویژگی‌ها با استفاده از روش‌هایی نظیر مجذور خی(کای)[۱۶]، بهینه‌سازی ازدحام ذرات[۱۷]، گام‌به‌گام حریصانه[۱۸] و جستجوی فاخته[۱۹] انتخاب می‌شوند و با استفاده از چندین یادگیرنده پایه و هم‌چنین روش تجمعی ارائه شده، مورد ارزیابی قرار می‌گیرند. در نهایت نتایج حاصله و کارایی روش ارائه شده با استفاده از معیارهایی همچون دقت (Precision)، فراخوانی (Recall)، معیار اف (F-measure) و مشخصه عملکرد سیستم[۲۰] (ROC) اندازه گیری می‌شود که بر حسب نتایج حاصله به دقت ۸۵.۱ درصد و با استفاده از روش گزینش بهترین ویژگی مجذور خی دست می‌یابند. در ضمن ویژگی‌های استخراجی بر اساس روش فراوانی عبارت - معکوس فراوانی سند[۲۱] (TFIDF) از متن نظرها استخراج و مورد استفاده قرار گرفته‌اند. این روش به‌معنای فراوانی وزنی کلمه کلیدی است. هدف این روش نیز نشان دادن اهمیت کلمه کلیدی مورد نظر، از طریق مقایسه تعداد تکرار کلمه در متن با تکرار آن کلمه در مجموعه‌ای بزرگتر از مستندها (کل متن نظرها) است.

در [۱۸] نیز روشی تجمعی به‌منظور شناسایی پیام‌های الکترونیکی هرز با استفاده از رویکرد تقسیم فضای لغت[۲۲] معرفی شده است. در این چارچوب از روش‌های مختلف انتخاب ویژگی‌ها نظیر محاسبه قدرت واژه[۲۳]، بهره اطلاعاتی[۲۴] و روش‌های متفاوت ساخت ویژگی‌ها به‌همراه الگوریتم‌های پایه بیز ساده، ماشین بردار پشتیبان و درخت تصمیم به‌همراه یادگیرنده‌های تجمعی تقویت‌سازی، جنگل تصادفی و رأی‌گیری[۲۵] استفاده شده است و در نهایت روش معرفی شده توانسته است به دقت بالای ۹۷ درصد و معیار اف بالای ۹۶ درصد دست یابد.

تا بدین‌جای کار مطالعه‌هایی که از روش‌های تجمعی و ویژگی‌های عمدتاً متنی استفاده کرده بودند، معرفی شدند؛ اما به‌منظور بررسی تأثیر روش‌های تجمعی بر روی شناسایی نظرهای هرز با ویژگی‌های رفتاری، لازم است تحقیق‌هایی که از ویژگی‌های رفتاری استفاده کرده‌اند نیز معرفی و روش کار آن‌ها بررسی شود.

در [۱۹] علاوه‌بر ویژگی‌های مبتنی بر محتوا، ویژگی‌های رفتاری نیز استفاده شده‌اند. در این مطالعه مدلی مبتنی بر شبکه معرفی شده است که با استخراج هشت ویژگی متنی و رفتاری و با استفاده از مفهومی به‌نام فرامسیر[۲۶] اقدام به ایجاد شبکه‌ای ناهمگون[۲۷] از نظرها می‌کند. در ادامه اقدام به تبدیل این نظرها به گره‌هایی در شبکه و اتصال این گره‌ها به‌همدیگر با استفاده از ویژگی‌های استخراج شده می‌کند. سپس با نگاشت این مسئله به یک مسئله طبقه‌بندی در شبکه، اقدام به برچسب‌زنی نظرها و محاسبه اهمیت ویژگی‌های استخراج شده می‌کند.

با توجه به کمبود نظرهای برچسب‌گذاری شده توسط انسان یا ماشین، در [۲۰] مدلی نظارت‌نشده به‌منظور شناسایی هرزنویس‌ها بر مبنای ویژگی‌های رفتاری ارائه شده است. این مدل از چارچوبی مبتنی بر الگوریتم بیزین استفاده می‌کند که نیازی به داده برچسب‌گذاری شده ندارد. معیارهای ارزیابی کارایی ارائه شده در این مطالعه، نشان‌دهنده اهمیت مدل‌های نظارت‌نشده و دقت بالایی است که می‌توان در این مدل‌ها بدان دست یافت.

در [۲۱] با توجه به اینکه الگوریتم فیلترسازی وبگاه یلپ به‌صورت محرمانه است مطالعه‌ای به‌منظور شناسایی نحوه عملکرد الگوریتم این وبگاه صورت گرفته است. در این مطالعه به‌منظور بررسی تأثیر هر کدام از ویژگی‌های مبتنی بر متن و یا رفتار، بر نظرهای فیلترشده این وبگاه، ارزیابی‌هایی صورت گرفته که نتایج حاصل حاکی از آن است که تأثیر استفاده از ویژگی‌های رفتاری در این الگوریتم‌ها بیشتر از ویژگی‌های زبانی و متنی است.

از دیگر مطالعه‌های صورت گرفته در این حوزه می‌توان به استفاده از شبکه عصبی عمیق[۲۸] و ویژگی‌های متنی استخراج شده به‌وسیله بردار n تایی کلمات در [۲۲]، استفاده از مدل مخفی مارکوف و ویژگی رفتاری ازدحام نظرها در [۲۳]، استفاده از خودرمزگذار بازگشتی[۲۹] نیمه نظارت‌شده و ویژگی‌های متنی در [۲۴] و استفاده از وزن‌دهی ویژگی‌ها و ایجاد مجموعه‌ای از آن‌ها در [۲۵] اشاره کرد. در [۲۸-۲۶] نیز روش‌هایی فازی برای طبقه‌بندی ارائه شده‌اند.

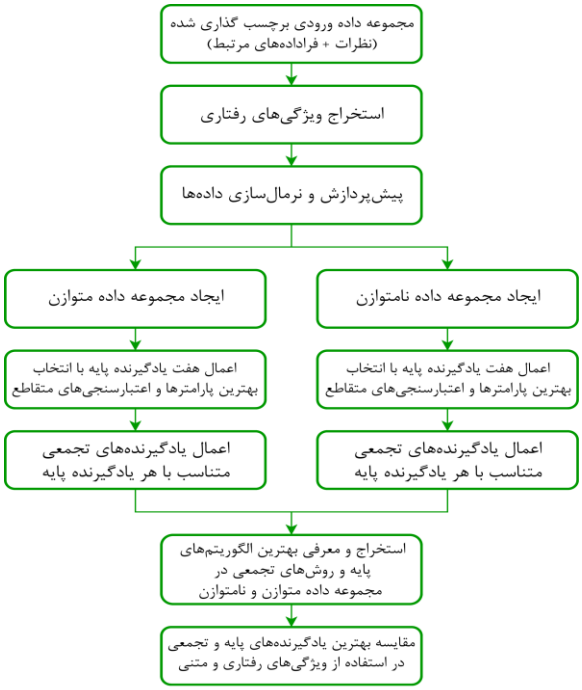
در جدول ۱ پیشینه تحقیق که شامل مقاله‌های مرتبط از سال ۲۰۰۹ تا ۲۰۲۰ است، به‌صورت مروری ارائه شده است.

۴- روش تحقیق

نمای کلی روش تحقیق و چارچوب پیشنهادی در شکل ۱ شکل ۱قابل مشاهده است.

جدول ۱: مروری بر پیشینه تحقیق.

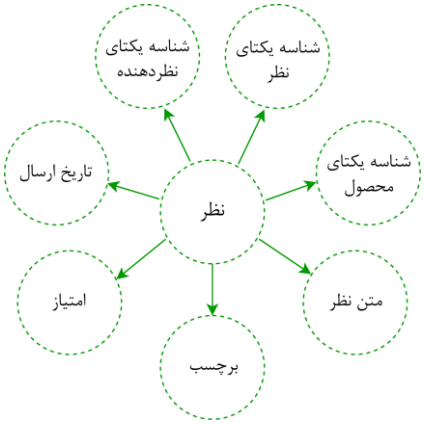
شماره مرجع	سال انتشار	ویژگی‌های هرزشناسی	نتایج
[۱۶]	۲۰۰۹	بردار سبد کلمه‌ها[۳۰]، بردار رشته به کلمه	بهترین الگوریتم: بیز ساده چندجمله‌ای در تجمیع با روش تجمعی تقویتی
[۲۰]	۲۰۱۳	مجموعه‌ای از نه ویژگی رفتاری و متنی	دست‌یابی به مقادیر مناسب هر کدام از ویژگی‌های استفاده شده به‌منظور تمییز دادن هرزنویس‌ها
[۲۱]	۲۰۱۳	استفاده از مجموعه‌ای گسترده از ویژگی‌های رفتاری و چندین ویژگی متنی	عملکرد بهتر ویژگی‌های رفتاری در مقایسه با ویژگی‌های متنی در مجموعه نظرهای دنیای واقعی
[۵]	۲۰۱۳	مجموعه‌ای از ویژگی‌های رفتاری قابل اعمال به‌منظور شناسایی ازدحام نظرها	ارائه روشی به‌منظور شناسایی هرزنویس‌ها به‌صورت مؤثر و ارائه یک روش جدید به‌منظور ارزیابی نتایج حاصله
[۱۱]	۲۰۱۶	بردار سبد کلمه‌ها، TFIDF	بهترین الگوریتم: بیز ساده چندجمله‌ای با AUC = ۰.۹ بدون استفاده از روش تجمعی تقویتی به دلیل بار زمانی تحمیلی
[۲۴]	۲۰۱۶	استفاده از ساختار سلسله‌مراتبی و معانی ترکیبی به‌منظور درک مفاهیم متون	عملکرد خوب مدل در استفاده از مجموعه داده‌های نامتوازن و دست‌یابی به فراخوانی بالای ۰.۹۱
[۱۹]	۲۰۱۷	مجموعه‌ای از هشت ویژگی رفتاری و متنی	۲ درصد بهبود در روش نیمه‌نظارت‌شده و در مقایسه با روش SPeagle و معرفی معیاری به‌منظور محاسبه اهمیت ویژگی‌های مورد استفاده
[۲۳]	۲۰۱۷	مجموعه‌ای از ویژگی‌های مبتنی بر ازدحام نظرها	دست‌یابی به فراخوانی بالای ۰.۸۷ در استفاده از مدل مخفی مارکوف و روش پیشنهادی
[۱۷]	۲۰۱۸	TFIDF	دقت ۸۵.۱ درصد در استفاده از روش تجمعی پیشنهادی در ادغام با روش گزینش بهترین ویژگی مجذور خی
[۱۸]	۲۰۱۸	بردار سبد کلمه‌ها، روش CFC، روش LC	دقت بالای ۹۷ درصد در مدل پیشنهادی با استفاده از روش تجمعی رأی‌گیری
[۲۲]	۲۰۱۹	استفاده از ویژگی‌های برداری n تایی و n تایی پرشی[۳۱] از کلمات	دست‌یابی به فراخوانی بالای ۰.۸۸ و سطح زیر منحنی بالای ۰.۹۵
[۲۵]	۲۰۲۰	ادغام مجموعه‌ای از ویژگی‌های رفتاری و متنی	دست‌یابی به صحت بالای ۰.۹۸ در استفاده از مجموعه‌ای وزن‌دهی‌شده ازویژگی‌های مبتنی بر نظر، نظردهنده و محصول



شکل ۱: نمای کلی روش تحقیق.

۴-۱- مجموعه داده

ساختاری که مجموعه داده‌های مورد استفاده ما از آن پیروی می‌کنند مطابق شکل ۲ است. هر نظری که در این مجموعه داده ثبت شده است توسط شناسه یکتای آن، از سایر نظرها متمایز می‌شود. در ضمن هر کدام از نظردهندگان نیز دارای شناسه یکتای خود هستند که به‌همراه شناسه انحصاری محصولی که در مورد آن ثبت نظر می‌کنند در کنار سایر موارد، ذخیره شده است. این شناسه‌ها ترکیبی از حروف بزرگ و کوچک هستند.



شکل ۲: ساختار مجموعه داده‌های مورد استفاده.

سایر پارامترها نیز شامل تاریخ ثبت نظر (به روز، ماه و سال)، متن نظر و امتیازی است که توسط کاربر در بازه ۱ تا ۵ برای محصول ثبت شده است. برچسب نظر نیز مشخص‌کننده کلاس نظر است که توسط الگوریتم‌های فیلترسازی وبگاه یلپ تعیین شده است. در

صورتی که نظر متعلق به دسته نظرهای هرز باشد با عدد ۱ و در غیراینصورت با عدد ۰ نشانه‌گذاری شده است. ویژگی‌های رفتاری که در ادامه توضیح داده می‌شوند نیز با استفاده از پارامترهای موجود محاسبه می‌شوند.

۴-۲- ویژگی‌های رفتاری استخراج شده

بر اساس [۱۹] تأثیر استفاده از ویژگی‌های رفتاری مبتنی بر نظر و نظردهنده، بیشتر از سایر ویژگی‌ها، نظیر ویژگی‌های مبتنی بر الگوهای زبانی و متنی نظرها است. بدین منظور عمده تمرکز ما در این کار بر روی استخراج این قبیل ویژگی‌ها و بررسی تأثیر آن‌ها معطوف شده است. در فرایند استفاده از آموزش ماشین به‌منظور شناسایی و پیش‌بینی برچسب نظرها، نیاز به استفاده از ویژگی‌هایی است که الگوریتم ماشین مدنظر ما بتواند با استفاده از مقادیر اختصاص داده شده به این ویژگی‌ها، مدلی را ایجاد کند که دارای کمترین خطا و بیشترین کارایی در شناسایی نظرهایی باشد که در آینده ایجاد می‌شوند. البته باید این موضوع را نیز مدنظر داشت که بیش‌برازش[۳۲] به‌معنای یادگیری صرف داده‌های آموزشی و عدم پیش‌بینی درست داده‌هایی که با داده‌های مرحله آموزش ماشین اندکی فاصله دارند و زیربرازش[۳۳] به‌معنای یادگیری کلی داده‌های مرحله آموزش و بروز خطای قابل توجه، حتی در پیش‌بینی داده‌هایی که آموزش به‌وسیله آن‌ها صورت گرفته، اتفاق نیافتد. یکی از مزایای روش‌های یادگیری ماشین تجمعی جلوگیری از بروز این قبیل مسائل است. ویژگی‌های رفتاری محاسبه شده عبارتند از:

۴-۲-۱- تعداد نظرهای ارسال شده توسط هر کاربر[۳۴] (NR)

کاربران عادی در حالت طبیعی تعداد نظرهای کمی در بازه‌های زمانی عضویت و فعالیت در وبگاه ثبت می‌کنند. تعداد نظرهای بالا می‌تواند نشان دهنده رفتار غیرطبیعی نظردهنده باشد.

۴-۲-۲- بیشترین تعداد نظرهای ارسال شده در یک روز (MNR)[۳۵]

ارسال تعداد نظرهای بالا در یک روز می‌تواند بیانگر یک رفتار غیرعادی از کاربر باشد [۲۰, ۲۱]. به‌منظور محاسبه این ویژگی بعد از گروه‌بندی نظرها بر اساس کاربران ارسال‌کننده، گروه‌بندی دیگری نیز بر اساس تاریخ ارسال نظر، برای هر کاربر، صورت می‌گیرد و بیشترین تعداد نظرهای ارسالی در یک روز توسط هر کاربر مشخص می‌شود.

۴-۲-۳- درصد نظرهای مثبت هر کاربر[۳۶] (PPR)

نظرهای مثبت، مشخص‌کننده نظرهایی هستند که امتیاز ۴ و ۵ ستاره در آنها ثبت شده است. مطالعه‌ها نشان داده است که هرزنویس‌ها تمایل زیادی به امتیازدهی خیلی بالا (۴ و ۵ ستاره) یا خیلی پایین (۱ و ۲ ستاره) به‌منظور ارتقا یا تنزل کاذب ارزش یک محصول را دارند [۲۱]. این ویژگی با محاسبه تعداد نظرهایی از هر کاربر که دارای امتیازدهی بالا است، به نسبت تعداد کل نظرهای ارسالی توسط آن کاربر به‌دست می‌آید.

۴-۲-۴- درصد نظرهای منفی هر کاربر[۳۷] (PNR)

این ویژگی نیز تعداد نظرهای منفی (۱ و ۲ ستاره) هر کاربر را به تعداد کل نظرهای کاربر تقسیم و درصد نظرهای منفی را همانند ویژگی بالا محاسبه می‌کند [۲۱].

۴-۲-۵- ازدحام نظرها[۳۸] (BST)

در [۵, ۲۱, ۲۹] نشان داده شده است که هرزنویس‌ها معمولاً اعضای طولانی مدت یک وبگاه نیستند. اعضای واقعی، معمولاً نظرهای خود را در بازه‌های زمانی معقولی ثبت و ارسال می‌کنند [۲۰] ولی اعضای هرزنویس به‌دلیل مسائلی نظیر کسب درآمد بیشتر و تاثیرگذاری بالاتر بر روی خوانندگان، سعی در ارسال نظرهای زیاد در بازه‌های زمانی کوتاه و با ازدحام می‌کنند؛ بنابراین با رهگیری فعالیت اعضا می‌توان رفتار غیرعادی این افراد را شناسایی و به‌عنوان معیاری برای طبقه‌بندی نظرهای آن‌ها استفاده کرد. به‌منظور محاسبه این ویژگی از فرمول (۱) استفاده می‌شود.

$$\mathrm{f}_{BST}(\mathrm{i}) = \begin{cases} ۰, & L(i)-F(i) > ۲۸ \\ ۱ - \dfrac{L(i)-F(i)}{\tau}, & \text{در غیر اینصورت} \end{cases} \qquad (۱)$$

در این فرمول i بیانگر کاربر نظردهنده و $L(i)-F(i)$ نشانگر پنجره زمانی مابین آخرین نظر و اولین نظر ثبت شده (به‌صورت روزشمار) توسط کاربر است. مقدار (روز $\tau = ۲۸$) نیز با توجه به [۲۹] تخمین زده شده است.

۴-۲-۶- انحراف مطلق امتیاز نظرها[۳۹] (ARD)

انحراف مطلق امتیاز نظرها، بیانگر میزان انحرافی است که امتیاز هر نظر با میانگین امتیازهای سایر نظرهای ثبت‌شده برای آن محصول (اجماع نظر عموم) دارد [۲۰]. مطالعه‌ها نشان داده است که هرزنویس‌ها معمولاً به‌هدف تنزل یا ترفیع ارزش یک محصول، سعی در تغییر میانگین امتیازهای ثبت شده توسط سایر

نظردهندگان به‌صورت مثبت یا منفی را دارند. به‌منظور محاسبه انحراف مطلق امتیاز نظرها از فرمول (۲) استفاده می‌شود [۹].

$$f_{ARD}(i) = |d_{ij}| = \frac{\left|r_{ij} - avg_{e \in E_{*j}} r(e)\right|}{۴} \qquad (۲)$$

در این فرمول r_{ij} بیانگر امتیازی است که کاربر i برای محصول j ثبت کرده است و $avg_{e \in E_{*j}} r(e)$ بیانگر میانگین امتیازهایی است که توسط بقیه نظردهندگان برای محصول j ثبت شده است. نسبت‌گیری حاصل نتیجه به ۴ به‌منظور نرمال‌سازی نتایج در یک سیستم ۵ ستاره است [۲۰] که تمامی نتایج را به‌منظور راحتی محاسبه‌ها در یک بازه عددی کوچک قرار می‌دهد.

۴-۲-۷- انحراف امتیاز میانگین نظرها (avgRD)[۴۰]

میانگین کل انحراف مطلق امتیازی نظرهای کاربر را بر حسب تمامی محصول‌هایی که کاربر برای آن‌ها ثبت نظر کرده است، محاسبه می‌کند [۹]. در فرمول (۳) بعد از محاسبه میانگین انحراف امتیاز هر نظر، میانگین انحراف مطلق امتیازی نظرهای یک کاربر برای محصول‌هایی که به آنها امتیازدهی و ثبت نظر کرده است محاسبه می‌شود.

$$f_{avgRD}(i) = avg_{e_{ij} \in E_{i*}} |d_{ij}| \qquad (۳)$$

۴-۲-۸- امتیازدهی شدید (EXT)[۴۱]

بر اساس تحقیق‌ها، هرزنویس‌ها تمایل دارند به‌منظور تحمیل کردن نظرهای خود از امتیازدهی شدید استفاده کنند [۲۰]. به‌عنوان نمونه در یک سیستم ۵ ستاره تمایل به امتیازدهی ۱ یا ۵ دارند. به‌منظور محاسبه این ویژگی از فرمول (۴) استفاده می‌شود.

$$f_{EXT}(r_i) = \begin{cases} ۱, & \text{امتیاز نظر} \in \{۱،۵\} \\ ۰, & \text{امتیاز نظر} \in \{۲،۳،۴\} \end{cases} \qquad (۴)$$

عبارت r_i نمایش دهنده نظر کاربر i است.

۴-۲-۹- انحراف امتیاز آستانه (TRD)[۴۲]

انحراف امتیاز آستانه در صورتی مقدار یک می‌گیرد که از مقدار مشخص β بیشتر باشد [۹, ۲۰]. این مقدار مشخص از طریق روش آنتروپی کمینه بازگشتی[۴۳] در [۳۰] محاسبه شده است. نحوه اختصاص مقدار به این ویژگی در فرمول (۵) قابل مشاهده است.

$$f_{TRD}(i) = \begin{cases} ۱, & \dfrac{\left|r_{ij} - avg_{e \in E_{*j}} r(e)\right|}{۴} > \beta \\ ۰, & \text{در غیر اینصورت} \end{cases} \qquad (۵)$$

۴-۳- پیش‌پردازش و نرمال‌سازی داده

به‌منظور محدود کردن تعداد نظرها و دست‌یابی به نتایج بهتر، بررسی‌ها بر روی نظرهایی انجام می‌شوند که کاربران آن‌ها حداقل سه نظر ثبت شده در مجموعه داده داشته باشند. با این عمل که به پیش‌پردازش داده معروف است تعداد نظرها محدود و نتایج بهتری حاصل می‌شود. در ضمن مقادیری که برای هر کدام از ویژگی‌های ذکر شده، محاسبه می‌شوند ممکن است دارای مقادیر پرت و مقادیری باشند که در فرایند آموزش الگوریتم‌هایی که در ادامه اعمال می‌شوند، مشکل ایجاد کنند. از طرفی الگوریتم‌های یادگیری ماشین با داده‌هایی که در بازه مشخصی قرار دارند نتایج بهتری تولید می‌کنند. بدین منظور نیاز است که مقادیر نرمال‌سازی شوند تا مقیاس ویژگی‌ها یکسان شود. در این مطالعه مقادیر محاسبه‌شده در بازه $[۰،۱]$ نگاشت و نرمال‌سازی می‌شوند. مقادیر نرمال توسط فرمول (۶) به‌دست می‌آیند [۱۷].

$$a' = \frac{a - a_{min}}{a_{max} - a_{min}} \qquad (۶)$$

۴-۴- الگوریتم‌ها و یادگیرنده‌های پایه

در این مطالعه، یادگیرنده‌های پایه‌ای که به‌منظور ادغام با یادگیرنده‌های تجمعی مورد بررسی قرار می‌گیرند عبارتند از: K نزدیک‌ترین همسایگان (KNN)[۴۴]، درخت تصمیم (DT)[۴۵]، رگرسیون منطقی خطی (LLR)[۴۶]، رگرسیون منطقی چندجمله‌ای (PLR)[۴۷]، بیز ساده چندجمله‌ای (MNB)[۴۸]، ماشین بردار پشتیبان با هسته خطی (LSVM)[۴۹] و ماشین بردار پشتیبان با هسته تابع پایه شعاعی (RBFSVM)[۵۰] که با استفاده از زبان برنامه‌نویسی پایتون پیاده‌سازی شده‌اند.

الگوریتم KNN به‌منظور پیش‌بینی کلاس و یا برچسب داده‌های مرحله آزمون، از برچسب داده‌های نزدیک در مرحله آموزش استفاده می‌کند. در حقیقت این الگوریتم به‌منظور پیش‌بینی برچسب داده‌های ورودی طبقه‌بندی نشده، از شباهت ویژگی داده‌های برچسب‌گذاری شده استفاده می‌کند و با انجام یک رأی‌گیری مابین K داده‌ای که کمترین فاصله و بیشترین شباهت را با داده مذکور دارند عملیات طبقه‌بندی را انجام می‌دهد. یکی از پارامترهایی که تأثیر زیادی در بهبود عملکرد این الگوریتم دارد، تعداد همسایه‌ها یا مقدار K است. در ادامه عملیاتی که به‌منظور تعیین بهترین تعداد K صورت می‌گیرد شرح داده خواهد شد.

الگوریتم DT با استفاده از ویژگی‌های محاسبه شده، درخت تصمیمی را به‌منظور تمایز مابین برچسب نظرها ایجاد می‌کند [۱۱]. شاخه‌های درخت ایجاد شده نشانگر ویژگی‌ها و برگ‌ها نمایانگر برچسب نظرها است. از پارامترهای درخت‌تصمیم که با تغییر آن‌ها، عملکرد و دقت مدل افزایش می‌یابد می‌توان به حداکثر عمق و حداکثر ویژگی‌های مورد استفاده اشاره کرد.

الگوریتم LLR به‌دنبال ایجاد رابطه احتمالی مابین ویژگی‌ها و برچسب نظرها است [۱۱]. در حقیقت رگرسیون‌منطقی مشابه رگرسیون خطی عمل می‌کند و به‌دنبال پیدا کردن خطی است که کمترین فاصله و به‌اصطلاح بیشترین تناسب و برازندگی را با داده‌های برچسب گذاری شده داشته باشد ولی با این تفاوت که رگرسیون خطی برای پیش‌بینی داده‌های پیوسته کاربرد دارد ولی رگرسیون منطقی به‌منظور طبقه‌بندی داده‌های گسسته استفاده می‌شود. این الگوریتم مقدار احتمال تعلق یک داده به یک گروه خاص را محاسبه و به‌عنوان خروجی بازمی‌گرداند. در PLR ممکن است یک یا چند متغیر مستقل تأثیر بیشتری بر روی متغیر وابسته داشته باشند بدین منظور مدلی از رگرسیون منطقی استفاده می‌شود که خطی نیست [۳۱].

بیز ساده چندجمله‌ای نیز بر اساس احتمال وقوع یا عدم وقوع یک رویداد، طبقه‌بندی را انجام می‌دهد و زیرمجموعه الگوریتم بیز ساده است. در حقیقت این روش با فرض مستقل بودن هر کدام از ویژگی‌ها، عملیات برچسب‌گذاری را انجام می‌دهد و بر اساس [۱۱] بهترین الگوریتم به‌منظور طبقه‌بندی نظرها با استفاده از ویژگی‌های متنی و حتی بدون استفاده از یادگیرنده‌های تجمعی است.

ماشین بردار پشتیبان نیز به‌وسیله نگاشت هر کدام از داده‌ها به فضای n بعدی و ایجاد صفحه‌ای در فضا اقدام به گروه‌بندی داده‌ها می‌کند. هسته این روش می‌تواند به‌صورت خطی یا غیر خطی باشد. در LSVM تأثیر استفاده از هسته خطی و در RBFSVM تأثیر استفاده از هسته غیرخطی مورد بررسی قرار گرفته است.

۴-۵- روش‌های یادگیری ماشین تجمعی

در الگوریتم‌های یادگیری ماشین، داده‌ها به دو بخش داده‌های مرحله آموزش و مرحله آزمون تقسیم‌بندی می‌شوند. در مرحله آموزش، بخش اعظمی از مجموعه داده، به‌همراه برچسب اختصاص داده شده، به‌منظور یادگیری و مدل‌سازی در اختیار ماشین قرار داده می‌شود و بخش باقیمانده نیز به‌منظور ارزیابی مدل، بدون برچسب، به‌عنوان داده مرحله آزمون در اختیار ماشین قرار می‌گیرد. در نهایت برچسب پیش‌بینی شده، به‌وسیله هر کدام از یادگیرنده‌ها برای داده‌های مرحله آزمون، با برچسب‌های واقعی مقایسه و عملکرد هر کدام از روش‌ها ارزیابی می‌شود. در روش‌های تجمعی بخشی از کار نیز بر روی نحوه انتخاب داده‌های این دو مرحله تمرکز دارد.

۴-۵-۱- روش دسته‌بندی[۵۱]

معرفی این بخش نیازمند معرفی روشی به‌نام خودراه‌اندازی[۵۲] است. این روش، یک روش نمونه‌برداری به‌منظور آموزش ماشین است. در این روش، زیرمجموعه‌ای از نمونه‌ها با جایگزینی انتخاب می‌شود به‌این‌صورت که بعد از انتخاب هر نمونه، دوباره به مجموعه کلی داده‌ها بازگردانی می‌شود و مجدداً شانس انتخاب و قرارگیری در بین نمونه‌های زیرمجموعه را دارد. در ضمن اندازه زیرمجموعه انتخاب شده، معادل اندازه کلی مجموعه داده است. در روش دسته‌بندی با استفاده از روش خودراه‌اندازی چندین زیرمجموعه از داده‌ها، هم اندازه مجموعه داده کلی، انتخاب و عملیات یادگیری با استفاده از الگوریتم‌های پایه معرفی شده به‌صورت موازی بر روی هر کدام از این زیرمجموعه‌ها انجام می‌شود و در نهایت با ادغام نتایج هر کدام از این مدل‌ها، نتیجه نهایی و عملکرد کلی مدل بررسی می‌شود [۱۱]. در کارهای پیشین انجام شده مشخص شده است که این روش عملکرد مدل را در استفاده از ویژگی‌های متنی و زبانی افزایش می‌دهد [۳۲].

۴-۵-۲- روش تقویتی[۵۳]

این روش نیز با استفاده از روش خودراه‌اندازی اقدام به نمونه‌گیری از مجموعه داده اصلی می‌کند. روش تقویتی یک رویکرد مبتنی بر تکرار است [۳۳]. به‌این‌صورت که در مرحله اول به‌منظور آموزش ماشین به هر کدام از نمونه‌ها وزنی اختصاص می‌دهد و اقدام به آموزش با استفاده از الگوریتم‌های پایه می‌کند. سپس در مراحل بعدی نیز همین عمل را تکرار می‌کند با این تفاوت که نمونه‌هایی که در مراحل قبلی درست پیش‌بینی نشده‌اند وزن بیشتری را دریافت و احتمال بیشتری در انتخاب به‌عنوان اعضای زیرمجموعه‌ها را دارند [۱۱].

۴-۵-۳- روش جنگل تصادفی[۵۴] و درخت اضافی[۵۵]

همانند روش‌های پیشین که از الگوریتم‌های پایه به‌منظور ایجاد مدل قوی‌تر استفاده می‌کردند، روش‌های جنگل تصادفی و درخت اضافی نیز از الگوریتم درخت تصمیم به‌منظور ایجاد مدل‌های با کارایی بالاتر استفاده می‌کنند. تفاوت جنگل تصادفی با درخت اضافی در دو مورد مهم است. اولاً روش درخت اضافی از نمونه‌برداری خودراه‌اندازی استفاده نمی‌کند و داده‌ها با جایگزینی انتخاب نمی‌شوند و دوما در جنگل تصادفی گره‌های ایجاد شده بر اساس بهترین ویژگی‌ها ایجاد می‌شود در حالی که در درخت اضافی این گره‌ها به‌صورت تصادفی انتخاب می‌شوند [۳۴]. هم‌چنین در جنگل تصادفی باید تعداد تخمین‌زننده‌ها یا به‌عبارتی تعداد درخت‌های تصمیمی که با ادغام نتایج آن‌ها، نتیجه نهایی حاصل می‌شود نیز مشخص شود که در این کار برابر با ۱۰۰۰ در نظر گرفته شده است.

۴-۶- معرفی اعتبارسنجی متقاطع k بخشی[۵۶]، جستجوی شبکه‌ای[۵۷] و معیارهای اندازه‌گیری دقت و عملکرد

یکی از مواردی که تأثیر بسیاری در مرحله آموزش و آزمون الگوریتم‌های یادگیری ماشین دارد نحوه تقسیم‌بندی نمونه‌ها است. با استفاده از اعتبارسنجی متقاطع k بخشی، نمونه‌های در اختیار ماشین در هر مرحله به k بخش تقسیم می‌شوند که یک بخش به‌منظور آزمون و k-1 بخش باقی‌مانده نیز به‌منظور آموزش ماشین استفاده می‌شوند. در حقیقت با استفاده از این روش k تجربه متفاوت از آموزش را برای ماشین ایجاد می‌کنیم و می‌توانیم مابین نتایج حاصل در هر مرحله میانگین، بیشترین و کمترین میزان صحت را محاسبه کنیم و بهینه‌ترین حالت را به‌دست آوریم. در این مطالعه اعتبارسنجی ۵ بخشی و ۱۰ بخشی مورد بررسی قرار می‌گیرند.

هر کدام از الگوریتم‌های پایه و روش‌های تجمعی که در این مطالعه مورد بررسی قرار می‌گیرند دارای ضرایبی هستند که به آنها ابرپارامتر[۵۸] گفته می‌شود و تعیین‌کننده پیچیدگی مدل ایجاد شده هستند. روشی که در این مطالعه به‌منظور بهینه‌سازی مقادیر این پارامترها استفاده می‌شود، جستجوی شبکه‌ای است. در این روش با استفاده از مقادیر مختلف این پارامترها در یک بازه خاص، نتایج حاصله از هر کدام، محاسبه و در یک جدول شبکه‌ای ثبت می‌شود و در نهایت بهترین نتیجه و بهینه‌ترین پارامتر مشخص می‌گردد. در استفاده از این روش نیاز است که مقادیر اعتبارسنجی متقاطع نیز مشخص شود.

پارامترهایی که در هر کدام از یادگیرنده‌های پایه و تجمعی استفاده می‌شوند و با استفاده از این روش بهینه‌ترین مقادیر استخراج می‌شوند عبارتند از: تعداد همسایگان یا K در KNN، حداکثر عمق، حداکثر ویژگی‌های مورد استفاده و حداقل تعداد نمونه‌های مورد نیاز برای ساخت برگ در DT، پارامترهای منظم‌سازی[۵۹] و نوع جریمه[۶۰] در LLR، PLR، SVM و LSVM، پارامتر آلفا به‌منظور تناسب و برازندگی نرم‌تر در MNB و تعداد تخمین‌زننده‌ها در هر کدام از یادگیرنده‌های تجمعی.

در مدل‌های دودویی که یادگیرنده موظف است مابین دو حالت از کلاس‌ها، یکی را انتخاب کند (به‌عنوان نمونه: هرز و سالم) معیارهای مناسبی که به‌منظور بررسی کیفیت مدل وجود دارند عبارتند از: ماتریس درهم‌ریختگی[۶۱]، صحت[۶۲]، دقت[۶۳]، فراخوانی[۶۴] یا حساسیت[۶۵] و معیار اف[۶۶]. ساختار ماتریس درهم‌ریختگی در جدول ۲ قابل مشاهده است. مقادیر ردیف‌های این جدول برچسب واقعی نظرها و مقادیر ستون‌ها نشان‌دهنده برچسب پیش‌بینی شده است.

جدول ۲: ساختار ماتریس درهم‌ریختگی.

ماتریس درهم‌ریختگی		برچسب پیش‌بینی شده	
		هرز	سالم
برچسب واقعی	هرز	TP[67]	FN[68]
	سالم	FP[69]	TN[70]

در جدول ۲جدول ۲، TP نشان‌دهنده تعداد نظرهایی است که به درستی هرز تشخیص داده شده‌اند، FN نشان‌دهنده تعداد نظرهای هرزی است که به اشتباه سالم تشخیص داده شده‌اند، FP نشان‌دهنده تعداد نظرهای سالمی است که به اشتباه هرز تشخیص داده شده و TN نشان‌دهنده تعداد نظرهایی است که به درستی سالم شناسایی شده‌اند. لازم به‌ذکر است این مقادیر می‌توانند به‌صورت درصد نیز بیان شوند.

مقادیر هر کدام از معیارهای دیگر نیز بر اساس ماتریس درهم‌ریختگی محاسبه می‌شوند که به‌ترتیب، صحت توسط فرمول(۷) ، دقت توسط فرمول (۸)، فراخوانی یا حساسیت توسط فرمول (۹) و معیار اف توسط فرمول (۱۰) محاسبه می‌شوند.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۷)$$

این معیار نسبت تعداد نظرهایی را که به درستی، هرز و سالم تشخیص داده شده‌اند، به همه حالت‌های موجود محاسبه می‌کند.

$$Precision = \frac{TP}{TP + FP} \qquad (۸)$$

این معیار نیز نسبت تعداد نظرهای هرزی را که به درستی شناسایی شده‌اند، به تعداد کل نظرهایی که به درست یا اشتباه، هرز تشخیص داده شده‌اند را محاسبه می‌کند. مقدار بالای این معیار بدین معنی است که نظرهای سالم زیادی نبوده‌اند که به اشتباه هرز تشخیص داده شده باشند.

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \qquad (۹)$$

این معیار نیز نسبت تعداد نظرهایی است که به درستی هرز تشخیص داده شده‌اند، به مجموع تعداد نظرهایی که به درستی هرز تشخیص داده شده و نظرهای هرزی که به اشتباه سالم تشخیص داده شده‌اند. مقدار بالای این معیار نیز نشان دهنده این است که نظرهای هرز بسیاری، به درستی تشخیص داده شده‌اند.

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (۱۰)$$

این معیار نیز در صورتی مورد استفاده قرار می‌گیرد که نیاز به انتخاب مدلی با حالت تعادلی مابین مقدار محاسبه شده برای فراخوانی و دقت باشد.

۵- نتایج

مجموعه نظرها، به‌منظور ارزیابی نتایج این کار تحقیقاتی، از وبگاه

یلپ جمع‌آوری شده است که توسط [۹] معرفی شده است. از این مجموعه نظرها به‌منظور ارائه روشی به‌نام اسپیگل[۷۱] استفاده شده است که از روش‌های مطرح در زمینه انجام کارهای تحقیقاتی مبتنی بر گراف، به‌منظور طبقه‌بندی نظرها است. اطلاعات آماری در مورد این مجموعه داده‌ها در جدول ۳ ذکر شده است.

مجموعه داده‌های مدنظر ما، با نام یلپچی، شامل نظرهایی در مورد مجموعه‌ای از رستوران‌ها و هتل‌ها در اطراف شیکاگو است. این وبگاه، هرساله چالشی را به‌منظور ارائه الگوریتم‌هایی توسط محققان، در زمینه تمییز نظرهای هرز از سالم برگزار می‌کند که از نتایج این چالش به‌منظور ارتقای الگوریتم‌های فیلترسازی نظرهای خود استفاده می‌کند. داده‌های ما نیز توسط الگوریتم‌های فیلترسازی خود وبگاه، طبقه‌بندی و برچسب گذاری شده‌اند. لازم به‌ذکر است که نظرهای فیلترشده نیز، به‌منظور انجام تحقیقات، در نهایت در اختیار عموم قرار می‌گیرند.

در این کار از دو مجموعه داده استفاده شده است ولی با توجه به شباهت کلی نتایج در هر سه حالت استفاده از یادگیرنده‌های پایه و تجمعی و مقایسه با ویژگی‌های متنی، تنها نتایج مربوط به مجموعه داده بزرگتر ارائه می‌شود.

جدول ۳: جزئیات آماری مجموعه داده‌ها.

تجارت هدف (هتل یا رستوران)	تعداد نظرهای هرز	تعداد نظرهای سالم	تعداد کل نظرها	تعداد کاربران نظردهنده	تعداد هتل یا رستوران هدف
هتل	۷۷۸	۵۰۷۶	۵۸۵۴	۵۰۲۶	۷۲
رستوران	۸۱۴۱	۵۳۴۰۰	۶۱۵۴۱	۳۳۹۶۴	۱۲۹

در این مرحله مقدار صحت حاصل از هر کدام از الگوریتم‌های پایه و مدل‌های تجمعی آن‌ها با استفاده از روش‌های دسته‌بندی و تقویتی محاسبه و با توجه به اینکه از اعتبارسنجی متقاطع استفاده می‌شود، به‌منظور بررسی تأثیر این عامل نیز، اعتبارسنجی‌های متقاطع ۵ بخشی و ۱۰ بخشی مورد استفاده قرار می‌گیرند و نتایج حاصل به‌صورت جداگانه ارائه می‌شوند. در جدول ۴ که مقادیر صحت ذکر می‌شوند میزان بالاترین، پایین‌ترین و میانگین مقادیر، در هر کدام از بخش‌های اعتبارسنجی محاسبه و ذکر شده است.

با توجه به اینکه تعداد نظرهای هرز در مجموعه داده مورد بررسی، به‌نسبت تعداد نظرهای سالم بسیار کمتر است و این عامل باعث تاثیرگذاری بر نتایج حاصل، به‌دلیل کمبود نظرهای هرز در مرحله آموزش ماشین می‌شود بدین منظور نتایج را در دو حالت مورد بررسی قرار می‌دهیم. در مرحله اول، داده‌ها با استفاده از همین نسبت موجود در مجموعه داده، به الگوریتم‌ها ارائه می‌شوند ولی در مرحله دوم مجموعه داده‌های جدیدی از روی مجموعه داده اولیه ایجاد می‌شوند که دارای تعداد مساوی نظرهای هرز و سالم هستند و نحوه انتخاب نمونه‌ها به‌صورت تصادفی و بر اساس [۱۲] است. در نهایت میانگین صحت حاصل با توجه به مجموعه داده‌های ایجاد شده، محاسبه و بهترین الگوریتم‌های پایه و روش‌های تجمعی معرفی خواهند شد. در این حالت تأثیر توزیع مساوی و نامساوی نمونه‌ها در مجموعه داده نیز مورد بررسی قرار می‌گیرد. لازم به‌ذکر است که بررسی‌های انجام شده، تحت سیستمی با حافظه موقت ۸ گیگابایت و پردازنده ۵ هسته‌ای صورت گرفته است.

جدول ۴: صحت روش‌های پایه و تجمعی در مجموعه داده نامتوازن.

الگوریتم	تعداد بخش‌های اعتبارسنجی متقاطع	کمترین صحت	بیشترین صحت	میانگین صحت‌ها
KNN	cv = ۵	۰.۹۸۲۵	۰.۹۸۳۴	۰.۹۸۳۰
دسته‌بندی با KNN		۰.۹۸۰۶	۰.۹۸۲۵	۰.۹۸۱۳
تقویت‌سازی با KNN		۰.۹۸۲۰	۰.۹۸۴۵	۰.۹۸۳۳
KNN	cv =۱۰	۰.۹۸۱۹	۰.۹۸۵۰	۰.۹۸۳۶
دسته‌بندی باKNN		۰.۹۸	۰.۹۸۳۵	۰.۹۸۱۵
تقویت‌سازی باKNN		۰.۹۸۰۱	۰.۹۹۱۲	۰.۹۸۴۸
DT	cv =۵	۰.۹۸۷۸	۰.۹۹۴۰	۰.۹۸۹۷
دسته‌بندی با DT		۰.۹۹۰۳	۰.۹۹۳۴	۰.۹۹۲۱
تقویت‌سازی با DT		۰.۹۸۷۲	۰.۹۹۴۶	۰.۹۹۰۴
DT	cv =۱۰	۰.۹۹	۰.۹۹۵۶	۰.۹۹۲۸
دسته‌بندی باDT		۰.۹۸۶۲	۰.۹۹۴۳	۰.۹۹۱۴
تقویت‌سازی باDT		۰.۹۹۱۸	۰.۹۹۷۵	۰.۹۹۵۰
LLR	cv =۵	۰.۹۸۱۵	۰.۹۸۲۲	۰.۹۸۱۷
دسته‌بندی باLLR		۰.۹۸۱۵	۰.۹۸۲۲	۰.۹۸۱۷
تقویت‌سازی با LLR		۰.۹۸۱۵	۰.۹۸۱۵	۰.۹۸۱۵
LLR	cv =۱۰	۰.۹۸۱۲	۰.۹۸۲۵	۰.۹۸۱۷
دسته‌بندی باLLR		۰.۹۸۱۲	۰.۹۸۲۵	۰.۹۸۱۷
تقویت‌سازی باLLR		۰.۹۸۱۲	۰.۹۸۱۸	۰.۹۸۱۵
MNB	cv =۵	۰.۹۸۱۵	۰.۹۸۱۵	۰.۹۸۱۵
دسته‌بندی باMNB		۰.۹۸۱۵	۰.۹۸۱۵	۰.۹۸۱۵
تقویت‌سازی با MNB		۰.۹۸۱۵	۰.۹۸۱۵	۰.۹۸۱۵
MNB	cv =۱۰	۰.۹۸۱۲	۰.۹۸۱۸	۰.۹۸۱۵
دسته‌بندی باMNB		۰.۹۸۱۲	۰.۹۸۱۸	۰.۹۸۱۵
تقویت‌سازی باMNB		۰.۹۸۱۲	۰.۹۸۱۸	۰.۹۸۱۵

۵-۱- مجموعه داده نامتوازن

درجدول ۵ نیز صحت حاصله از هرکدام از روش‌های تجمعی جنگل‌تصادفی با ۱۰۰۰ تخمین زننده و درخت‌اضافی ذکر شده است.

جدول ۵: صحت روش‌های تجمعی جنگل تصادفی و درخت اضافی در مجموعه داده نامتوازن.

الگوریتم	cv	کمترین صحت	بیشترین صحت	میانگین صحت‌ها
جنگل تصادفی	۵	۰.۹۸۷۵	۰.۹۹	۰.۹۸۹۰
جنگل تصادفی	۱۰	۰.۹۸۵۶	۰.۹۹۳۱	۰.۹۹۰۱
درخت اضافی	۵	۰.۹۷۵۶	۰.۹۸۳۴	۰.۹۷۹۱
درخت اضافی	۱۰	۰.۹۷۸۷	۰.۹۸۶۸	۰.۹۸۲۲

مقدار صحت حاصل از الگوریتم‌های MNB، PLR، LSVM و RBFSVM و روش‌های تجمعی آن‌ها تقریباً مشابه یکدیگر است و با توجه به بررسی‌های بیشتر صورت گرفته نظیر محاسبه ماتریس درهم‌ریختگی، دقت، فراخوانی و معیار اف برای این روش‌ها، مشخص شد که این الگوریتم‌ها در حالت استفاده از مجموعه داده نامتوازن به‌همراه الگوریتم LLR و روش‌های تجمعی آن، با اینکه در شناسایی نظرهای سالم عملکرد خوبی دارند ولی در شناسایی نظرهای هرز که موضوع اصلی این کار تحقیقاتی است کارایی خوبی از خود نشان نمی‌دهند و مقدار صحت بالای به‌دست آمده در جدول ۴ نیز، به‌دلیل شناسایی بهتر نظرهای سالم است که نسبت بالایی از مجموعه داده ما را شامل شده است. در نتیجه ادامه کار و بررسی‌های بیشتر بر روی الگوریتم‌های KNN و DT و روش‌های تجمعی آنها صورت می‌گیرد و نتایج به‌دست آمده با یکدیگر مورد قیاس قرار می‌گیرند تا بهترین الگوریتم و تاثیرگذارترین روش تجمعی مرتبط مشخص گردد.

لازم به‌ذکر است که در ادامه مواردی نظیر ماتریس درهم‌ریختگی هر کدام از روش‌های باقیمانده به‌همراه مقدار دقت و فراخوانی محاسبه و ذکر می‌شوند.

در جدول ۶ ماتریس درهم‌ریختگی هر کدام از روش‌های باقیمانده به‌منظور مقایسه ارائه شده است. در این جدول PR بیانگر پیش‌بینی نظرها به‌صورت واقعی[۷۲]، PS بیانگر پیش‌بینی نظرها به‌صورت هرز[۷۳]، AR بیانگر برچسب واقعی نظرها به‌صورت واقعی[۷۴] و AS نشان‌دهنده برچسب واقعی نظرها به‌صورت هرز[۷۵] است. به‌عنوان نمونه تقاطع AS با PS نمایش دهنده درصد نظرهای هرزی است که توسط مدل به‌درستی پیش‌بینی شده یا همان مقدار TP است.

مقادیر جدول ۷ نیز نشانگر دقت، فراخوانی و معیار اف در هر کدام از الگوریتم‌های KNN و DT و روش‌های تجمعی آن‌ها است. لازم به‌ذکر است که این مقادیر هم برای شناسایی نظرهای هرز و هم برای شناسایی نظرهای سالم قابل محاسبه است و نظر به اینکه هدف این مطالعه شناسایی نظرهای هرز است فقط برای این دسته از نظرها ذکر و مقایسه می‌شوند.

جدول ۶: ماتریس درهم‌ریختگی الگوریتم‌های KNN، DT و روش‌های تجمعی آن‌ها.

الگوریتم	cv		PR	PS
KNN	۵	AR	۰.۹۹۹۰	۰.۰۰۰۹
		AS	۰.۶۸۳۶	۰.۳۱۶۳
دسته‌بندی با KNN		AR	۰.۹۹۵۴	۰.۰۰۴۵
		AS	۰.۶۵۳۰	۰.۳۴۶۹
تقویت‌سازی با KNN		AR	۰.۹۹۹۲	۰.۰۰۰۷
		AS	۰.۶۲۲۴	۰.۳۷۷۵
KNN	۱۰	AR	۰.۹۹۹۰	۰.۰۰۰۹
		AS	۰.۶۸۳۶	۰.۳۱۶۳
دسته‌بندی با KNN		AR	۰.۹۹۵۴	۰.۰۰۴۵
		AS	۰.۶۵۳۰	۰.۳۴۶۹
تقویت‌سازی با KNN		AR	۰.۹۹۹۲	۰.۰۰۰۷
		AS	۰.۶۲۲۴	۰.۳۷۷۵
DT	۵	AR	۰.۹۹۶۹	۰.۰۰۳۰
		AS	۰.۲۴۴۸	۰.۷۵۵۱
دسته‌بندی با DT		AR	۰.۹۹۹۶	۰.۰۰۰۳
		AS	۰.۲۵۵۱	۰.۷۴۴۸
تقویت‌سازی با DT		AR	۰.۹۹۶۹	۰.۰۰۳۰
		AS	۰.۲۱۴۲	۰.۷۸۵۷
DT	۱۰	AR	۰.۹۹۸۲	۰.۰۰۱۷
		AS	۰.۲۷۵۵	۰.۷۲۴۴
دسته‌بندی با DT		AR	۰.۹۹۹۸	۰.۰۰۰۱
		AS	۰.۲۸۵۷	۰.۷۱۴۲
تقویت‌سازی با DT		AR	۰.۹۹۹۶	۰.۰۰۰۳
		AS	۰.۲۰۴۰	۰.۷۹۵۹
درخت اضافی	۵	AR	۰.۹۹۱۷	۰.۰۰۸۲
		AS	۰.۴۸۹۷	۰.۵۱۰۲
درخت اضافی	۱۰	AR	۰.۹۹۵۸	۰.۰۰۴۱
		AS	۰.۳۳۶۷	۰.۶۶۳۲
جنگل تصادفی	۵	AR	۰.۹۹۹۸	۰.۰۰۰۱
		AS	۰.۳۸۷۷	۰.۶۱۲۲
جنگل تصادفی	۱۰	AR	۰.۹۹۹۸	۰.۰۰۰۱
		AS	۰.۳۸۷۷	۰.۶۱۲۲

در جدول ۶ مقادیری که از اهمیت بالایی به‌منظور مقایسه الگوریتم‌ها برخوردارند تقاطع ستون PS با ردیف‌های AS هر کدام از روش‌ها است که همانطوری که مطرح شد، بیانگر درصد نظرهای هرزیست که به‌درستی پیش‌بینی شده‌اند. به‌عبارت دیگر این

مقدار، همان مقدار فراخوانی در جدول ۷ است. بر اساس مقادیر به‌دست‌آمده، روش تجمعی تقویت‌سازی به‌همراه الگوریتم پایه DT با تعداد ۱۰ بخش اعتبارسنجی متقاطع و درصد پیش‌بینی ۷۹.۵۹ درصد بالاترین میزان فراخوانی را درمیان سایر روش‌ها دارا است. در ضمن بررسی‌های بیشتر نیز نشان‌دهنده این است که روش تجمعی تقویت‌سازی نسب به دسته‌بندی تأثیر بیشتری بر افزایش کارایی الگوریتم‌ها در این حالت دارد. با افزایش تعداد بخش‌های اعتبارسنجی متقاطع نیز، بسته به الگوریتم مورد استفاده، می‌توان فراخوانی را همانطوری که در روش درخت اضافی نیز قابل مشاهده است، به‌میزان قابل توجهی افزایش داد.

جدول ۷: مقادیر دقت، فراخوانی و معیار اف برای الگوریتم‌های KNN، DT و روش‌های تجمعی آن‌ها.

الگوریتم	cv	دقت	فراخوانی	معیار اف
KNN	۵	۰.۸۶	۰.۳۲	۰.۴۶
دسته‌بندی با KNN		۰.۵۹	۰.۳۵	۰.۴۴
تقویت‌سازی با KNN		۰.۹	۰.۳۸	۰.۵۳
KNN	۱۰	۰.۸۶	۰.۳۲	۰.۴۶
دسته‌بندی با KNN		۰.۵۹	۰.۳۵	۰.۴۴
تقویت‌سازی با KNN		۰.۹	۰.۳۸	۰.۵۳
DT	۵	۰.۸۲	۰.۷۶	۰.۷۹
دسته‌بندی با DT		۰.۹۷	۰.۷۴	۰.۸۴
تقویت‌سازی با DT		۰.۸۳	۰.۷۹	۰.۸۱
DT	۱۰	۰.۸۹	۰.۷۲	۰.۸۰
دسته‌بندی با DT		۰.۹۹	۰.۷۱	۰.۸۳
تقویت‌سازی با DT		۰.۹۷	۰.۸۰	۰.۸۸
درخت اضافی	۵	۰.۵۴	۰.۵۱	۰.۵۲
درخت اضافی	۱۰	۰.۷۵	۰.۶۶	۰.۷۰
جنگل تصادفی	۵	۰.۹۸	۰.۶۱	۰.۷۵
جنگل تصادفی	۱۰	۰.۹۸	۰.۶۱	۰.۷۵

۵-۲- مجموعه داده متوازن

در این حالت، کل نظرهای هرز و به‌همان تعداد نظرهای سالم را به‌صورت تصادفی از مجموعه داده اولیه و بر اساس روش [۱۲] انتخاب و ایجاد چند مجموعه داده می‌دهیم به‌صورتی که مجموع تعداد نظرهای سالم در این چند مجموعه داده، برابر با تعداد نظرهای سالم در مجموعه داده اولیه باشد و سپس هر کدام از الگوریتم‌های پایه و روش‌های تجمعی را بر روی هر کدام از این مجموعه داده‌ها اعمال و میانگین نتایج را محاسبه می‌کنیم تا موثرترین الگوریتم پایه و روش تجمعی را در حالت استفاده از مجموعه داده‌ای با تعداد نمونه‌های مساوی، به‌منظور پیش‌بینی نظرهای هرز، شناسایی و معرفی نماییم.

در جدول ۸ میانگین صحت حاصل هر کدام از روش‌های پایه و تجمعی، با تعداد بررسی بر روی ۳۳ مجموعه داده‌ای که از مجموعه داده اصلی استخراج شده‌اند، ذکر می‌شود.

نتایج حاصل در روش RBFSVM تقریباً مشابه LSVM است و به‌همین دلیل ذکر نشده است. در جدول ۹ نیز نتایج درخت اضافی و جنگل تصادفی با ۱۰۰۰ تخمین‌زننده ذکر می‌شوند.

با توجه به مقادیر صحت‌های به‌دست آمده، مشخص است که به‌غیر از الگوریتم درخت تصمیم و روش‌های تجمعی آن به‌غیر از درخت اضافی، سایر روش‌ها اختلاف بسیار زیادی با این روش‌ها دارند و دارای عملکرد خوبی نیستند. بدین لحاظ ادامه بررسی‌ها بر روی الگوریتم درخت تصمیم و روش‌های تجمعی دسته‌بندی، تقویت‌سازی و جنگل تصادفی صورت می‌گیرد. در جدول ۱۰ فراخوانی تشخیص نظرهای هرز در این الگوریتم‌ها محاسبه و ذکر شده‌اند.

جدول ۸: میانگین صحت روش‌های پایه و تجمعی در مجموعه داده‌های متوازن.

الگوریتم	cv	کمترین صحت	بیشترین صحت	میانگین صحت‌ها
KNN	۵	۰.۶۹۵۰	۰.۷۵۶۳	۰.۷۲۶۶
دسته‌بندی باKNN		۰.۶۹۴۴	۰.۷۵۸۴	۰.۷۲۸۱
تقویت‌سازی باKNN		۰.۶۹۱۵	۰.۷۵۹۸	۰.۷۲۹۰
KNN	۱۰	۰.۶۷۶۸	۰.۷۸۵۹	۰.۷۳۱۶
دسته‌بندی باKNN		۰.۶۷۸۶	۰.۷۸۹۷	۰.۷۳۴۶
تقویت‌سازی باKNN		۰.۶۷۹۴	۰.۷۹۰۴	۰.۷۳۵۲
DT	۵	۰.۷۸۵۲	۰.۸۵۱۶	۰.۸۲۰۱
دسته‌بندی باDT		۰.۸۳۹۷	۰.۸۸۱۰	۰.۸۶۸۹
تقویت‌سازی با DT		۰.۸۱۹۴	۰.۸۷۲۷	۰.۸۵۷۲
DT	۱۰	۰.۷۵۹۰	۰.۸۷۸۲	۰.۸۲۸۵
دسته‌بندی با DT		۰.۸۲۲۷	۰.۹۱۶۷	۰.۸۷۰۲
تقویت‌سازی با DT		۰.۷۹۶۳	۰.۸۹۰۳	۰.۸۵۸۹
LLR	۵	۰.۶۶۶۸	۰.۷۶۲۰	۰.۷۲۲۱
دسته‌بندی باLLR		۰.۶۷۰۲	۰.۷۶۱۰	۰.۷۲۵۰
تقویت‌سازی باLLR		۰.۶۵۱۲	۰.۷۳۴۸	۰.۷۰۰۶
LLR	۱۰	۰.۶۴۶۸	۰.۷۷۴۵	۰.۷۲۰۹
دسته‌بندی باLLR		۰.۶۵۳۹	۰.۷۷۹۴	۰.۷۲۳۹
تقویت‌سازی باLLR		۰.۶۲۹۱	۰.۷۵۹۱	۰.۷۰۱۱
PLR	۵	۰.۶۹۴۲	۰.۷۷۰۴	۰.۷۳۴۹
دسته‌بندی باPLR		۰.۶۹۷۲	۰.۷۷۱۱	۰.۷۳۶۷
تقویت‌سازی باPLR		۰.۶۸۱۱	۰.۷۳۲۵	۰.۷۰۸۰

PLR		۰.۶۸۹۴	۰.۷۹۰۸	۰.۷۳۴۰
دسته‌بندی باPLR	۱۰	۰.۶۷۱۳	۰.۷۹۲۲	۰.۷۳۶۵
تقویت‌سازی باPLR		۰.۶۵۶۸	۰.۷۷۸۱	۰.۷۱۳۰
MNB		۰.۶۵۰۲	۰.۷۱۷۷	۰.۶۸۵۷
دسته‌بندی باMNB	۵	۰.۶۵۰۴	۰.۷۲۰۹	۰.۶۸۷۴
تقویت‌سازی باMNB		۰.۶۷۳۶	۰.۷۵۶۷	۰.۷۲۰۲
MNB		۰.۶۳۷۳	۰.۷۳۰۱	۰.۶۸۵۲
دسته‌بندی باMNB	۱۰	۰.۶۳۶۹	۰.۷۳۳۳	۰.۶۸۷۳
تقویت‌سازی باMNB		۰.۶۴۵۰	۰.۷۷۵۷	۰.۷۱۷۸
LSVM		۰.۶۱۸۶	۰.۶۷۴۹	۰.۶۴۸۵
دسته‌بندی با LSVM	۵	۰.۶۱۹۵	۰.۶۷۷۰	۰.۶۵
تقویت‌سازی باLSVM		۰.۶۰۲۰	۰.۶۵۱۲	۰.۶۲۸۸
LSVM		۰.۶۰۵۲	۰.۷۰۳۱	۰.۶۴۹۳
دسته‌بندی با LSVM	۱۰	۰.۶۰۸۱	۰.۷۰۳۳	۰.۶۵۰۷
تقویت‌سازی با LSVM		۰.۵۸۸۵	۰.۶۸۵۶	۰.۶۳۱۳

جدول ۹: صحت روش‌های تجمعی جنگل‌تصادفی و درخت‌اضافی در حالت استفاده از مجموعه داده‌های متوازن.

الگوریتم	cv	کمترین صحت	بیشترین صحت	میانگین صحت‌ها
جنگل تصادفی	۵	۰.۸۱۹۱	۰.۸۷۵۲	۰.۸۴۶۸
جنگل تصادفی	۱۰	۰.۸۰۷۶	۰.۹۰۷۹	۰.۸۵۶۵
درخت اضافی	۵	۰.۷۲۳۵	۰.۷۸۲۵	۰.۷۵۹۵
درخت اضافی	۱۰	۰.۷۰۳۲	۰.۸۲۶۴	۰.۷۶۱۴

جدول ۱۰: فراخوانی تشخیص نظرهای هرز در یادگیرنده پایه درخت تصمیم و روش‌های تجمعی دسته‌بندی، تقویت‌سازی و جنگل تصادفی.

الگوریتم	cv	فراخوانی نظرهای هرز
DT	۵	۰.۸۳
دسته‌بندی با DT		۰.۸۷
تقویت‌سازی با DT		۰.۸۴
DT	۱۰	۰.۸۳
دسته‌بندی با DT		۰.۸۷
تقویت‌سازی با DT		۰.۸۴
جنگل تصادفی	۵	۰.۸۷
جنگل تصادفی	۱۰	۰.۸۷

با توجه به میزان فراخوانی به‌دست آمده در شناسایی نظرهای هرز می‌توان گفت که تجمیع دسته‌بندی با درخت تصمیم، عملکردی نزدیک به جنگل تصادفی در شناسایی نظرهای هرز ارائه می‌کند. از این رو به‌منظور اینکه مشخص کنیم کدام الگوریتم عملکرد بهتری دارد به‌سراغ مقایسه فراخوانی در شناسایی نظرهای سالم می‌رویم.

جدول ۱۱ نشانگر فراخوانی به‌دست آمده توسط هر کدام از الگوریتم‌ها در شناسایی نظرهای سالم است.

با توجه به نتایج به‌دست آمده می‌توان عنوان کرد که حالت تجمعی دسته‌بندی با درخت تصمیم، نتایج بهتری در مقایسه با جنگل تصادفی چه در شناسایی نظرهای هرز و چه سالم ارائه می‌کند.

جدول ۱۱: فراخوانی تشخیص نظرهای سالم در یادگیرنده پایه درخت تصمیم و روش‌های تجمعی دسته‌بندی، تقویت‌سازی و جنگل تصادفی.

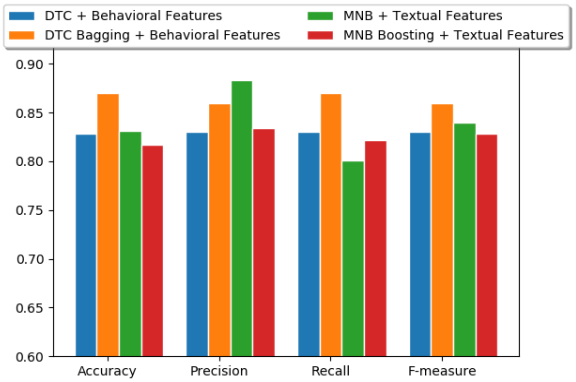
الگوریتم	cv	فراخوانی نظرهای سالم
DT	۵	۰.۸۳
دسته‌بندی با DT	۵	۰.۸۶
تقویت‌سازی با DT	۵	۰.۸۴
DT	۱۰	۰.۸۳
دسته‌بندی با DT	۱۰	۰.۸۶
تقویت‌سازی با DT	۱۰	۰.۸۴
جنگل تصادفی	۵	۰.۸۳
جنگل تصادفی	۱۰	۰.۸۳

در نهایت می‌توان عنوان کرد که با استفاده از الگوریتم درخت تصمیم، در صورتی که مجموعه داده‌ای متوازن در اختیار داشته باشیم می‌توان با استفاده از روش تجمعی دسته‌بندی به فراخوانی ۸۷٪ و در صورتی که مجموعه داده به‌صورت نامتوازن باشد با روش تجمعی تقویت‌سازی به فراخوانی ۷۹٪ در شناسایی نظرهای هرز دست یافت.

به‌منظور بررسی بیشتر نیز، ویژگی‌های متنی با استفاده از روش TFIDF استخراج و الگوریتم MNB به‌عنوان بهترین الگوریتم و روش تجمعی تقویت‌سازی به‌عنوان بهترین روش تجمعی در ادغام با این الگوریتم، با توجه به [۱۱] پیاده‌سازی شدند که نتایج حاصله در شکل ۳ و در مقایسه با بهترین الگوریتم‌های پایه و تجمعی در استفاده از ویژگی‌های رفتاری، قابل ملاحظه است. لازم به‌ذکر است که مقایسه‌ها فقط در مجموعه داده متوازن صورت می‌گیرد زیرا در استفاده از مجموعه داده نامتوازن نیاز به بررسی حالت‌های مختلفی است که تعداد نظرهای هرز به چه نسبتی از نظرهای سالم باشند که این موضوع، روند مطالعه را طولانی‌تر می‌کند.

همانطوری که در شکل ۳ قابل ملاحظه است تأثیر استفاده از روش تجمعی دسته‌بندی بر روی DT در استفاده از ویژگی‌های رفتاری، در تمامی معیارهای ارزیابی، به‌مراتب بیشتر از تأثیر روش تقویتی بر روی MNB در استفاده از ویژگی‌های متنی است. روش تقویتی در ادغام با MNB، تنها موجب نزدیکی مقدار دقت و فراخوانی به‌همدیگر می‌شود که به‌معنای ایجاد مساوات بیشتر در

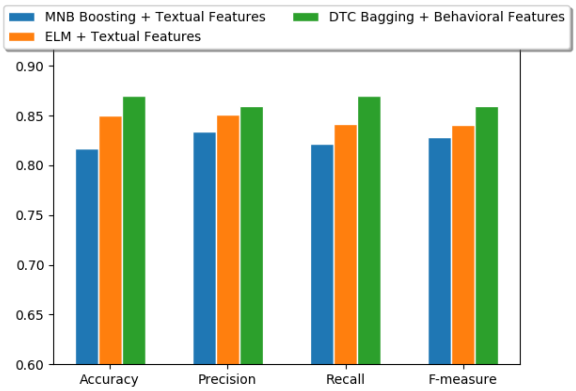
طبقه‌بندی نظرهای هرز و سالم است.



شکل ۳: مقایسه بهترین یادگیرنده‌های پایه و تجمعی از نظر صحت، دقت، فراخوانی و معیار اف در استفاده از ویژگی‌های متنی و رفتاری.

در نتیجه می‌توان عنوان کرد که استفاده از روش تجمعی تقویت‌سازی با ویژگی‌های متنی تنها موجب تحمیل بار محاسباتی بیشتر می‌شود در صورتی که روش تجمعی دسته‌بندی با ویژگی‌های رفتاری موجب بهبود عملکرد نیز می‌شود در حالی که یادگیرنده پایه DT با ویژگی‌های رفتاری و MNB با ویژگی‌های متنی تقریباً به نتایج مشابهی دست یافته‌اند.

در شکل ۴ نیز مقایسه‌ای مابین روش تجمعی دسته‌بندی در ادغام با DT در استفاده از ویژگی‌های رفتاری، روش تقویتی در ادغام با MNB در استفاده از ویژگی‌های متنی و روش ELM معرفی‌شده توسط [۱۷] در استفاده از ویژگی‌های متنی ارائه شده است که نشان‌دهنده عملکرد بهتر روش ارائه شده است.



شکل ۴: مقایسه بهترین یادگیرنده‌های تجمعی از نظر صحت، دقت، فراخوانی و معیار اف در استفاده از ویژگی‌های متنی و رفتاری.

لازم به‌ذکر است که بررسی‌ها بیانگر این است که یادگیرنده‌های تجمعی، بیشترین تاثیرگذاری را بر روی یادگیرنده‌های پایه با نتایج و دقت بالاتر دارند. به‌عنوان نمونه در این کار مطالعاتی تاثیرگذاری یادگیرنده‌های تجمعی بر روی روش پایه درخت تصمیم که به‌تنهایی عملکرد قابل قبولی دارد نسبت به سایر روش‌هایی که عملکرد ضعیفتری دارند بیشتر است.

۶- جمع‌بندی و کارهای آتی

بر اساس بررسی‌های صورت گرفته مشخص شد که تاکنون بیشتر مطالعه‌ها بر روی تأثیر استفاده از یادگیرنده‌های تجمعی با استفاده از ویژگی‌های متنی و الگوهای زبانی نظرها تمرکز کرده‌اند و نیاز است که تأثیر این روش‌ها با استفاده از ویژگی‌های رفتاری نیز در نظر گرفته شود. نتایج به‌دست‌آمده در این مطالعه، حاکی از آن است که بهترین الگوریتم پایه به‌منظور طبقه‌بندی نظرهای هرز با استفاده از ویژگی‌های رفتاری، درخت تصمیم است و یادگیرنده تجمعی تقویت‌سازی در حالت استفاده از مجموعه داده نامتوازن و دسته‌بندی در حالت استفاده از مجموعه داده متوازن بیشترین تأثیر را در ارتقای عملکرد این الگوریتم پایه دارد. در ضمن استفاده از ویژگی‌های رفتاری در شناسایی نظرهای هرز عملکرد بهتری را نسبت به ویژگی‌های متنی ارائه می‌دهد و میزان بهبودی که یادگیرنده‌های تجمعی در عملکرد بهترین الگوریتم‌های یادگیری ماشین پایه، در استفاده از ویژگی‌های رفتاری، ایجاد می‌کنند به‌مراتب بیشتر از ویژگی‌های متنی است.

اما با تمامی موارد مطرح شده و نتایج به‌دست‌آمده باید این موضوع در نظر گرفته شود که استفاده از یادگیرنده‌های تجمعی چه با ویژگی‌های رفتاری و چه با ویژگی‌های متنی، دارای بار محاسباتی، زمانی و هزینه‌هایی است که به‌اصطلاح، هزینه طبقه‌بندی نادرست[۷۶] نامیده می‌شود و باید روش‌های تجمعی ارائه شوند تا ضمن کاهش این هزینه‌ها به عملکردی مشابه یا حتی بهتر از روش‌های تجمعی موجود دست پیدا کرد. اخیرا استفاده از تصمیم‌گیری‌های فازی و اعمال هزینه‌های مختلف زمانی و محاسباتی برای طبقه‌بندی‌های نادرست، که در نهایت منجر به کاهش هزینه‌های کلی می‌شود به مبحثی داغ در میان محققان این حوزه تبدیل شده است که لازم است بر روی شناسایی نظرهای هرز نیز اعمال شود.

در ضمن کمبود مجموعه نظرهای برچسب‌گذاری شده که لازمه الگوریتم‌های یادگیری ماشین نظارت‌شده است یکی دیگر از مشکل‌های استفاده از این روش‌هاست که با تمرکز بیشتر بر روی اعمال الگوریتم‌های نظارت‌نشده و ایجاد یادگیرنده‌های تجمعی بر اساس آن‌ها قابل بررسی است.

در نهایت با توجه به اینکه نظرهای دنیای واقعی دارای عدم توازن در پراکندگی هستند، نیاز است که یادگیرنده‌های تجمعی ارائه‌شده در آینده این موضوع را نیز در نظر داشته و راه‌حل‌های پیشنهادی خود را ارائه دهند.

مراجع

[1] J. D'Onfro, "A whopping 20% of yelp reviews are fake", Business Insider, Sep. 25, 2013. Accessed on: Aug. 23, 2019. [Online]. Available: https://www.businessinsider.com/20-percent-of-yelp-reviews-fake-2013-9.

[2] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 201-210 .

[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language Technologies-volume 1*, 2011: Association for Computational Linguistics, pp. 309-319 .

[4] C. Xu and J. Zhang, "Combating product review spam campaigns via multiple heterogeneous pairwise features," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015: SIAM, pp. 172-180 .

[5] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013 .

[6] B. Viswanath *et al.*, "Towards detecting anomalous user behavior in online social networks," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 2014, pp. 223-238 .

[7] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013 .

[8] H. Li ,Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *2014 IEEE International Conference on Data Mining*, 2014: IEEE, pp. 899-904 .

[9] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 985-994 .

[10] I. Salian, "Supervize me: What's the difference between supervised, unsupervised, semi-supervised and reinforcement learning?", Nvidia, Aug. 2, 2018. Accessed on: 10. Aug, 2019. [Online]. Available: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning./

[11] B. Heredia, T. M. Khoshgoftaar, J. Prusa, and M. Crawford, "An investigation of ensemble techniques for detection of spam reviews," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016: IEEE, pp. 127-133 .

[12] J. D. Prusa, T. M. Khoshgoftaar, and N. Seliya, "Enhancing ensemble learners with data sampling on high-dimensional imbalanced tweet sentiment data," in *The Twenty-Ninth International Flairs Conference*, 2016 .

[13] J. Jin, P. Ji, and Y. Liu, "Recommending rating values on reviews for designers," in *Encyclopedia of Business Analytics and Optimization*: IGI Global, 2014, pp. 1998-2009.

[14] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 219-230 .

[15] S. Mani, S. Kumari, A. Jain, and P. Kumar, "Spam review detection using ensemble machine learning," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2018: Springer, pp. 198-209 .

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[17] F. Khurshid, Y. Zhu, Z. Xu, M. Ahmad, and M. Ahmad, "Enactment of ensemble learning for review spam detection on selected features," *International Journal of Computational Intelligence Systems*, vol. 12, no. 1, pp. 387-394, 2018.

[18] Y. Tan, Q. Wang, and G. Mi, "Ensemble decision for spam detection using term space partition approach," *IEEE Transactions on Cybernetics* ,vol. 50, no. 1, pp. 297-309, 2018.

[19] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no ,۷ .pp. 1585-1595, 2017.

[20] A. Mukherjee *et al.*, "Spotting opinion spammers using behavioral footprints," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 632-640 .

[21] A. Mukherjee, V .Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013 .

[22] A. Barushka and P. Hajek, "Review spam detection using word embeddings and deep neural networks," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2019: Springer, pp. 340-350 .

[23] H. Li *et al.*, "Bimodal distribution and co-bursting in review spam detection," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1063-1072 .

[24] B. Wang, J. Huang, H. Zheng, and H. Wu, "Semi-supervised recursive autoencoders for social review spam detection," in *2016 12th International Conference on Computational Intelligence and Security*) *CIS)*, 2016: IEEE, pp. 116-119 .

[25] M. Z. Asghar, A. Ullah, S. Ahmad, and A. Khan, "Opinion spam detection framework using hybrid classification scheme," *Soft computing*, vol. 24, no. 5, pp. 3475-3498, 2020.

[26] X. Jia, Z. Deng, F. Min, and D. Liu, "Three-way decisions based feature fusion for Chinese irony detection," *International Journal of Approximate Reasoning*, vol. 113, pp. 324-335, 2019.

[27] H.-R. Zhang and F. Min, "Three-way recommender systems based on random forests," *Knowledge-Based Systems* ,vol. 91, pp. 275-286, 2016.

[28] Y. Zhang, D. Miao, J. Wang, and Z. Zhang, "A cost-sensitive three-way combination technique for ensemble learning in sentiment classification," *International Journal of Approximate Reasoning*, vol. 105, pp. 85-97, 2019.

[29] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 191-200 .

[30] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," 1993.

[31] "Logistic polynomial regression in R", Educational Research Techniques, Dec. 29, 2017. Accessed on: Aug. 1, 2019. [Online]. Available: https://educationalresearchtechniques.com/2017/12/29/logistic-polynomial-regression-in-r./

[32] J. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Using ensemble learners to improve classifier performance on tweet sentiment data," in *2015 IEEE International Conference on Information Reuse and Integration*, 2015 :IEEE, pp. 252-257 .

[33] Y. Freund and R. E. Schapire, "Schapire R: Experiments with a new boosting algorithm," in *in: Thirteenth International Conference on ML*, 1996: Citeseer .

[34] N. Bhandari, "How does extratreesclassifier reduce the risk of overfitting?", Medium, Oct. 22, 2018. Accessed on: Sep. 10, 2019. [Online]. Available: https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7.

پاورقی‌ها:

[1] Spam Reviews
[2] Amazon
[3] Yelp
[4] Unigram
[5] Bigram
[6] Metadata
[7] Supervised
[8] Semi-Supervised
[9] Unsupervised
[10] Amazon Mechanical Turk
[11] Untruthful Reviews
[12] Reviews on Brands
[13] Non-Reviews
[14] String to Word Vector
[15] Area Under Curve
[16] Chi-Square
[17] Particle Swarm Optimization
[18] Greedy Stepwise
[19] Cuckoo Search
[20] Receiver Operating Characteristic
[21] Term Frequency-Inverse Document Frequency
[22] Term Space Partition
[23] Term Strength
[24] Information Gain
[25] Voting
[26] Meta-path
[27] Heterogeneous Network
[28] Deep Neural Network
[29] Recursive Autoencoders
[30] Bag of Words Vector
[31] Skip-gram
[32] Overfitting
[33] Underfitting
[34] Number of Reviews
[35] Maximum Number of Reviews
[36] Percentage of Positive Reviews
[37] Percentage of Negative Reviews
[38] Burstiness
[39] Absolute Rating Deviation
[40] Average Rating Deviation
[41] Extreme Rating
[42] Threshold Rating Deviation
[43] Recursive Minimal Entropy
[44] K Nearest Neighbors
[45] Decision Tree
[46] Linear Logistic Regression
[47] Polynomial Logistic Regression
[48] Multinomial Naïve Bayes
[49] Linear Support Vector Machine
[50] Radial Based Function SVM
[51] Bagging
[52] Bootstrapping
[53] Boosting
[54] Random Forest
[55] Extra Tree
[56] K Fold Cross Validation
[57] Grid Search
[58] Hyper Parameter
[59] Regularization
[60] Penalty
[61] Confusion Matrix
[62] Accuracy
[63] Precision
[64] Recall
[65] Sensitivity
[66] F-measure
[67] True Positive
[68] False Negative
[69] False Positive
[70] True Negative
[71] SpEagle
[72] Predicted Real
[73] Predicted Spam
[74] Actual Real
[75] Actual Spam
[76] Misclassification Cost