

## The Opinion Mining of Digikala Reviews by Semi-Supervised Support Vector Machine

Zohre Karimi<sup>1\*</sup>, Hadith Haghiri<sup>2</sup>

1. Assistant Professor, Department of Engineering, Damghan University, Damghan, Iran.

*Corresponding Author*, [z.karimi@du.ac.ir](mailto:z.karimi@du.ac.ir)

2. B.A, Computer Engineering, Damghan University, Damghan, Iran. [hhaghiri87@gmail.com](mailto:hhaghiri87@gmail.com)

### Abstract

**Introduction:** The widespread use of the internet and social media platforms has led to an explosion of digital data, including users' opinions about various services and products. These opinions are valuable sources of information for businesses and organizations to understand the needs and preferences of their customers. Supervised machine learning models have been proven to be effective in analyzing users' opinions. However, to achieve efficient results, a sufficient amount of labeled training data is necessary. Labeling data requires a considerable amount of time and resources, which can be a significant challenge for many organizations. This is where the concept of semi-supervised learning comes in, which utilizes both labeled and unlabeled data to improve the performance of the model.

**Method:** In this paper, a semi-supervised approach to analyze users' Persian opinions has been proposed. The method takes advantage of the abundant unlabeled data available in addition to a small number of labeled data in the training phase. The proposed method uses the support vector machine (SVM) algorithm, which has been shown to be effective in opinion mining in related research. The proposed method extracts emotional words from comments using sentiment lexicons and then extracts term frequency-inverse of document frequency vectors. The semi-supervised SVM algorithm is then applied to these vectors to estimate the polarity of sentiments.

**Results:** To evaluate the performance of the proposed method, it has been tested on the Digikala comments dataset and compared with the supervised SVM algorithm and semi-supervised self-training method for different numbers of labeled data based on accuracy, precision, recall, and F1 criteria. The results indicate that the proposed semi-supervised method outperforms the supervised SVM algorithm and the semi-supervised method of self-training. The impact of the size of unlabeled data is also investigated in the experiments.

**Discussion:** One of the advantages of the proposed method is that it can estimate the polarity of opinions that have not been trained in the training phase, which is not possible in some graph-based methods. Furthermore, it is not affected by the error of training with labeled data in self-training methods. In conclusion, the proposed semi-supervised method provides an efficient solution for analyzing users' opinions in Persian. This method can be used by businesses and organizations to gain insights into their customers' opinions and improve their products and services accordingly.

**Keywords:** Semi-supervised Learning, Semi-supervised Support vector Machine, Opinion mining, Sentiment Analysis, Digikala opinions.

## عقیده‌کاوی نظرات دیجی کالا با استفاده از روش نیمه‌نظارتی مبتنی بر ماشین بردار پشتیبان

دوره چهارم، بهار ۱۴۰۲  
شماره اول، صص: ۵۱-۶۱

تاریخ دریافت: ۱۴۰۱/۱۱/۲۵  
تاریخ پذیرش: ۱۴۰۲/۰۱/۰۸

زهره کریمی<sup>۱\*</sup>، حدیث حقیری<sup>۲</sup>

۱. استادیار، دانشکده فنی و مهندسی، دانشگاه دامغان، ایران. (نویسنده مسئول) [z.karimi@du.ac.ir](mailto:z.karimi@du.ac.ir)
۲. کارشناسی مهندس کامپیوتر، دانشکده فنی و مهندسی، دانشگاه دامغان، ایران. [hhaghiri87@gmail.com](mailto:hhaghiri87@gmail.com)

**چکیده:** رشد فراوان نظرات دیجیتال کاربران در مورد خدمات و محصولات منجر به توسعه روش‌های عقیده‌کاوی شده است. مدل‌های یادگیری ماشین نظارتی در این زمینه به نتایج خوبی دست یافته‌اند. هرچند، این روش‌ها نیاز به تعداد کافی داده‌های آموزشی برچسب‌دار دارند که آماده‌سازی آن‌ها نیازمند صرف هزینه و زمان زیاد است. در این مقاله یک رویکرد نیمه‌نظارتی جهت تحلیل نظرات فارسی کاربران پیشنهاد شده که از داده‌های بدون برچسب فراوان همراه با تعداد کمی داده برچسب‌دار در مرحله آموزش بهره‌می‌گیرد. با توجه به عملکرد مناسب روش ماشین بردار پشتیبان نظارتی جهت عقیده‌کاوی، به کارگیری روش نیمه‌نظارتی ماشین بردار پشتیبان پیشنهاد شده است. این روش در مقایسه با روش‌های موجود با چالش تقویت خطا مواجه نبوده و قادر به تخمین قطبیت نظراتی که در مرحله آموزش دیده نشده‌اند، نیز است. روش پیشنهادی روی مجموعه داده نظرات دیجی کالا مورد ارزیابی قرار گرفته و با الگوریتم ماشین بردار پشتیبان بر اساس ملاک‌های دقت، صحت، بازخوانی و F1 مقایسه شده است. نتایج به دست آمده حاکی از عملکرد بهتر روش نیمه‌نظارتی در مقایسه با روش نظارتی و نیز روش نیمه‌نظارتی خودآموزی است.

**واژه‌های کلیدی:** یادگیری نیمه‌نظارتی، ماشین بردار پشتیبان نیمه‌نظارتی، عقیده‌کاوی، تحلیل احساس، نظرات دیجی کالا.

## ۱. مقدمه

عقیده‌کاوی، نظرکاوی یا تحلیل احساس به مطالعه و کاربرد روش‌های محاسباتی مرتبط با متن از جمله پردازش زبان طبیعی و متن‌کاوی جهت مشخص کردن نظرات افراد در مورد محصولات، خدمات و رویدادهای گوناگون می‌پردازد. تصمیم‌گیری‌های گوناگون در زندگی افراد از زمان‌های گذشته تاکنون با بهره‌گیری از نظرات سایر افراد صورت می‌گرفته‌است. با پیشرفت تکنولوژی و گسترش تجارت الکترونیک و ثبت نظرات افراد به‌صورت دیجیتال، نیاز به پردازش خودکار نظرات دیجیتال، بسیار ضروری است [۱، ۲]. نتایج این پردازش‌ها برای تولیدکنندگان محصول، ارائه‌دهندگان خدمات مختلف و مدیران در کاربردهای گوناگون اهمیت زیادی دارد. عقیده‌کاوی یکی از موضوعات به‌روز پژوهشی است که در سال‌های اخیر بسیار مورد توجه قرار گرفته‌است. رویکردهای موجود در دو گروه رویکردهای مبتنی بر واژگان و رویکردهای مبتنی بر یادگیری ماشین قرار می‌گیرند [۳]. روش‌های مبتنی بر واژگان از منابع زبانی شامل واژگان احساسی که دربرگیرنده واژگان احساسی مثبت و منفی هستند استفاده می‌کنند و کیفیت آن‌ها بسیار متأثر از منبع واژگان استفاده‌شده در آن‌ها است [۴، ۵]. روش‌های مبتنی بر یادگیری ماشین، الگوریتم‌های یادگیری ماشین یا پردازش متن را روی ویژگی‌های زبانی یا گرامری استخراج شده از متن نظرات اعمال می‌کنند. اخیراً نیز روش‌هایی پیشنهاد شده‌اند که روش‌های یادگیری ماشین و روش‌های مبتنی بر واژگان را با یکدیگر ترکیب می‌کنند.

الگوریتم‌های یادگیری ماشین عموماً در دو دسته روش‌های نظارتی و بدون ناظر قرار می‌گیرند. روش‌های نظارتی از داده‌های برچسب‌دار به‌عنوان داده‌های آموزشی استفاده می‌کنند و نیاز به در دسترس بودن داده‌های برچسب‌دار به میزان کافی دارند. هر چند در نظرکاوی برچسب داده‌ها به‌صورت دستی توسط خبره انسانی تعیین می‌شوند و لذا نیاز به صرف هزینه و زمان زیاد است. روش‌های بدون ناظر از اطلاعات برچسب داده‌ها استفاده نمی‌کنند و لذا نتایج آن‌ها لزوماً در راستای موردنظر نبوده و دقت کافی را ندارد. الگوریتم‌های نظارتی گوناگون از جمله ماشین بردار پشتیبان (SVM)، بی‌زین ساده و سیستم استنتاج عصبی-فازی تطبیقی از جمله روش‌های نظارتی هستند که تاکنون جهت نظرکاوی به‌کاررفته‌اند [۶، ۷، ۸]. روش‌های نیمه‌نظارتی از تعداد کمی داده برچسب‌دار و تعداد زیادی داده بدون برچسب جهت آموزش مدل استفاده می‌کنند [۹، ۱۰]. این روش‌ها دانش موردنظر برای یادگیری مدل را از تعداد زیادی داده بدون برچسب در کنار تعداد اندکی داده برچسب‌دار در دسترس به‌دستی می‌آوردند و لذا در مقایسه با روش‌های نظارتی نیاز به تعداد کمتری داده‌ی برچسب‌دار دارند.

از منظری دیگر، روش‌های عقیده‌کاوی در سه سطح مبتنی بر سند، مبتنی بر جمله و مبتنی بر جنبه انجام می‌شوند. هدف روش‌های مبتنی بر سند، پیش‌بینی قطبیت یا تمایل موجود در کل یک نظر است. به‌عبارت-دیگر هر نظر به‌عنوان یک سند در نظر گرفته می‌شود. در روش‌های مبتنی بر جمله، قطبیت جملات موجود در هر نظر به صورت جداگانه

پیش‌بینی می‌شود و در روش‌های مبتنی بر جنبه، ابتدا جنبه‌های موجود در نظرات استخراج شده و سپس تمایل مربوط به هر جنبه پیش‌بینی می‌شود.

در این مقاله، عقیده‌کاوی مجموعه نظرات دیجی‌کالا با بهره‌گیری از رویکرد یادگیری نیمه‌نظارتی در سطح سند مدنظر است. انواع روش‌های نیمه‌نظارتی از جمله روش خودآموزی، روش‌های مبتنی بر گراف، روش‌های هم‌آموزی<sup>۱</sup> و ماشین بردار پشتیبان نیمه‌نظارتی (S3VM) جهت دسته‌بندی نظرات انگلیسی به‌کاررفته‌اند [۱۱، ۱۲، ۱۳، ۱۴] با این حال، پژوهش‌های انجام‌شده تاکنون در مورد زبان فارسی بسیار اندک است. «نجف زاده، راحتی قوچانی و قائمی» [۱۵] از روش خودآموز<sup>۲</sup> جهت دسته‌بندی نیمه‌نظارتی نظرات استفاده کرده‌اند. روش خودآموز ابتدا یک دسته‌بند با ناظر را از داده‌های اندک برچسب‌دار در دسترس یاد گرفته و از این دسته‌بند جهت برچسب‌گذاری داده‌های بدون برچسب استفاده می‌کند. سپس داده‌هایی را که با اطمینان بالایی درست دسته‌بندی کرده‌است به داده‌های برچسب‌دار موجود اضافه کرده و این روند را تکرار می‌کند. با توجه به این‌که دسته‌بند باناظر اولیه روی تعداد اندکی داده آموزش دیده‌است، لذا قابلیت برچسب‌گذاری داده‌ها به‌صورت درست را ندارد و از آن‌جاکه از نتایج برچسب‌گذاری خود برای یادگیری دسته‌بند تکرار بعد استفاده می‌کند، خطای خود را در هر مرحله تقویت می‌کند. «عسگریان، کاهانی و شریفی» [۱۶] نیز یک روش نیمه‌نظارتی مبتنی بر گراف را جهت استخراج یک منبع واژگان احساسی زبان فارسی به‌کار برده‌اند. روش‌های نیمه‌نظارتی مبتنی بر گراف، چون بر مبنای ایجاد گراف روی داده‌های برچسب‌دار و بدون برچسب اولیه عمل می‌کنند، فقط قابلیت برچسب‌گذاری داده‌های بدون برچسب را دارند که در مرحله آموزش در دسترس آن‌ها قرار گرفته‌است و قابلیت برچسب‌گذاری داده‌های برچسب‌دار دیگر را ندارند.

هدف اصلی این مقاله ارائه یک روش نظرکاوی عقاید فارسی در سطح سند است که شامل این ویژگی‌ها باشد: (۱) قابلیت بهره‌گیری از تعداد کم داده برچسب‌دار و تعداد زیاد داده بدون برچسب در مرحله آموزش (نیمه‌نظارتی بودن) (۲) در معرض چالش تقویت خطا در روش نیمه‌نظارتی خودآموز نباشد. (۳) قابلیت پیش‌بینی تمایل موجود در نظراتی علاوه بر نظرات مرحله آموزش (غلبه بر محدودیت روش‌های نیمه‌نظارتی متداول مبتنی بر گراف).

بدین منظور، پیشنهاد می‌شود که روش S3VM جهت نظرکاوی عقاید فارسی به‌کار رود. این روش که تعمیم‌یافته ماشین بردار پشتیبان است از همان ابتدا مرز تصمیم را با استفاده از داده‌های برچسب‌دار و بدون برچسب با یکدیگر یاد می‌گیرد، لذا در معرض چالش تقویت خطا نیست و محدودیتی برای برچسب‌گذاری داده‌های بدون برچسب ندارد.

در ادامه، در ابتدا پیشینه پژوهش بررسی شده، سپس جزئیات روش به‌کاررفته در این مقاله آمده‌است. در بخش بعد، آزمایش‌های انجام‌شده و نتایج به‌دست‌آمده، ارائه‌شده و نهایتاً نتیجه پژوهش آمده‌است.

## ۲. پیشینه پژوهش

تاکنون پژوهش‌های زیادی در زمینه عقیده‌کاوی انجام شده است [۱۷]، ۱۸، ۱۹، ۲۰. در این بخش، پژوهش‌هایی که در دو منظر زبان و رویکرد با این مقاله مشترک هستند، مورد بررسی قرار می‌گیرند. ابتدا پژوهش‌هایی که به صورت خاص روی زبان فارسی متمرکز شده‌اند مورد بررسی قرار گرفته و سپس مروری بر رویکردهای نظرکاوی نیمه‌نظارتی صورت می‌گیرد. در نهایت هم پژوهش‌های اندکی که مبتنی بر رویکرد نیمه‌نظارتی در عقیده‌کاوی زبان فارسی هستند به صورت خاص مورد توجه قرار می‌گیرند.

پردازش متون در زبان فارسی با چالش تعامل با ویژگی‌های خاص این زبان مواجه است. به کارگیری کلمات محاوره‌ای، وجود نیم‌فاصله در برخی کلمات و کلمات مرکب از ویژگی‌های خاص زبان فارسی هستند. وجود منابع زبانی محدود در مقایسه با زبان انگلیسی و نیز قطبیت وابسته به فرهنگ نیز از چالش‌های عقیده‌کاوی در این زبان است [۲۱]. برخی از پژوهش‌های موجود بر ایجاد منابع زبانی عقیده‌کاوی فارسی متمرکز شده‌اند [۲۲]. LexiPers، PerSent و SentiPers از جمله پیکره‌های فارسی ایجاد شده هستند [۲۳، ۲۴، ۲۵، ۲۶، ۲۷، ۲۸، ۲۹]. «باقری و سرائی» روی مسئله انتخاب ویژگی در عقیده‌کاوی متون فارسی متمرکز شده و یک معیار بهبود یافته اطلاعات متقابل را برای انتخاب ویژگی پیشنهاد کرده‌اند [۲۱]. در نهایت، دسته‌بند بیزین ساده برای دسته‌بندی نظرات استفاده شده است. «وزیری پور، گیرود کریر و زاپالا» روند تغییر نظرات توثیق‌های فارسی در زمینه عناوین سیاسی را بررسی کرده‌اند. آن‌ها ابتدا از الگوریتم خوشه‌بندی براون<sup>۲</sup> [۱۸] برای انتخاب ویژگی استفاده کرده و سپس دسته‌بند SVM را اعمال کرده‌اند [۳۰]. پژوهش‌هایی نیز در عقیده‌کاوی نظرات فارسی با استفاده از روش‌های یادگیری عمیق انجام شده است. «گنبدی و رنجبر» از مدل زبانی «برت» جهت دسته‌بندی تمایلات موجود در نظرات فارسی در مورد برخی خوردوهای ایرانی استفاده کرده‌اند [۳۱]. «دستغیب، کلینی و راستی» [۳۲] ترکیبی از روش‌های یادگیری متناظر ساختاری<sup>۴</sup> و شبکه‌های عصبی کانولوشنال را جهت دسته‌بندی نظرات فارسی اعمال کرده‌اند. دستیابی به دقت مناسب در روش‌های یادگیری عمیق نیاز به در دسترس بودن داده‌های با برچسب فراوان دارد که در زبان فارسی به اندازه کافی در دسترس نیست. «قائمی، اشرفی و ممتازی» بهره‌گیری از یادگیری انتقالی جهت انتقال مدل از زبان انگلیسی به زبان فارسی را جهت مرتفع‌نمودن این مسئله پیشنهاد کرده‌اند [۳۳].

در ادامه، روش‌های نیمه‌نظارتی عقیده‌کاوی بررسی می‌شوند. «انند و نارم» [۱۲] یک روش نیمه‌نظارتی جهت نظرکاوی مبتنی بر جنبه، روی نقد فیلم‌ها انجام داده‌اند. در این پژوهش، استخراج جنبه‌ها به روش نیمه‌نظارتی مبتنی بر گراف صورت می‌گیرد. «حسن خان، قمر، بشیر» [۱۱] از یک روش مبتنی بر واژگان، بهره‌گرفته و با استفاده از ملاک‌های بهره اطلاعات<sup>۵</sup> و شباهت کسینوسی امتیازهای کلمات احساسی موجود در واژگان را بهبود داده‌اند. در نهایت، الگوریتم SVM جهت تعیین تمایل

موجود در نظر در این پژوهش اعمال شده است. برچسب لازم برای مجموعه آموزشی SVM از طریق اعمال بهبودهایی روی منبع واژگان احساسی تأمین شده است. در پژوهشی دیگر، از دسته‌بندی شبکه بیزین نیمه‌نظارتی جهت تعیین عینی یا ذهنی بودن<sup>۶</sup> نظرات استفاده شده است [۱۴]. «انصاری، احمد، دجا و ساکسنا» [۳۴] نیز یک رویکرد نیمه‌نظارتی مبتنی بر گراف جهت استخراج جنبه‌ها پیشنهاد کردند. «رن، کاجی، یوشیناگا و کیتسورگاوا» [۳۵] نیز یک روش نیمه‌نظارتی مبتنی بر گراف را جهت عقیده‌کاوی به کار گرفته‌اند. آن‌ها روی انتخاب مناسب نمونه‌های برچسب‌دار متمرکز شده و تأثیر ملاک‌های گوناگون از جمله درجه رأس‌ها و مقدار رتبه صفحه<sup>۷</sup> را با یکدیگر مقایسه کرده‌اند. یک روش مبتنی بر شبکه عصبی گراف ناهمگن نیمه‌نظارتی نیز برای دسته‌بندی نظرات پیشنهاد شده است [۳۶]. این روش، در مقایسه با روش‌های نیمه‌نظارتی ترارسان<sup>۸</sup> که فقط می‌توانند برچسب داده‌های بدون برچسب به کاررفته در مرحله آموزش را تخمین بزنند، قابلیت دسته‌بندی نظرات بدون برچسبی که در مرحله آموزش دیده نشده‌اند را نیز دارد. مطالعه جامعی در مورد اعمال انواع روش‌های نیمه‌نظارتی روی پیام‌های موجود در توییتر نیز انجام شده است [۳۷]. پژوهش‌های مورد اشاره در این پاراگراف نظرات انگلیسی را تحلیل کرده‌اند.

«عسگریان، کاهانی و شریفی» یک منبع واژگان حسی فارسی را با استفاده از الگوریتم نیمه‌نظارتی مبتنی بر مدل مخفی مارکوف و بهره‌گیری از منبع زبانی نگاهت قطبیت SentiWordNet موجود به زبان انگلیسی ایجاد کرده‌اند [۱۶]. روش یادگیری عمیق بدون ناظر بازنمایی رمزگذار دوطرفه از مبدل‌ها<sup>۹</sup> که با کمک مجموعه داده کوچکی از شایعات فارسی تنظیم شده و با یک مدل یادگیری ناظر ترکیب شده است جهت اعتبارسنجی شایعات در زبان فارسی به کاررفته است [۳۸]. هرچند پژوهش‌های ذکر شده دقیقاً در راستای کاربرد مورد نظر این مقاله نیستند اما مؤید اهمیت یادگیری نیمه‌نظارتی در عقیده‌کاوی متون فارسی هستند. دسته‌بند مبتنی بر قانون مدل مخفی مارکوف در قالب یادگیری نیمه‌نظارتی خودآموز جهت تعیین قطبیت نظرات موجود در دیجی کالا به کاررفته است [۱۵]. روش خودآموز ابتدا با استفاده از تعداد کم داده‌های برچسب‌دار موجود، مدل را یاد می‌گیرد و سپس احتمال تعلق داده‌های بدون برچسب آموزشی به هر دسته را با استفاده از مدل یاد گرفته شده، تخمین می‌زند. در تکرار بعد، داده‌هایی که با احتمال بالایی متعلق به هر دسته هستند به داده‌های آموزشی اضافه می‌شوند و این گام تا برچسب‌گذاری تمام داده‌ها ادامه می‌یابد. این الگوریتم در معرض دو چالش قرار دارد [۱۰]: چالش تقویت خطا بدین معنا که اگر برچسب‌های تخمینی اشتباه مشخص شده باشد در تکرارهای بعد، آموزش با استفاده از داده‌های خطا دار انجام می‌شود و (۲) ناکافی بودن داده بدین معنا که مدل اولیه با تعداد کمی از داده‌ها که برای تعیین مرز تصمیم کافی نیستند، آموزش داده می‌شود. «دهداربهبهانی، شاکری و فایلی» نیز از یک رویکرد نیمه‌نظارتی جهت عقیده‌کاوی استفاده کرده‌اند. آن‌ها از منابع زبانی غنی زبان انگلیسی بهره‌برده و یک شبکه معنایی وزن‌دار چندزبانه

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F1 = \frac{2 \times Precision \times recall}{Precision + recall} \quad (4)$$

TP، FN و FP بر اساس مقادیر ماتریس اغتشاش<sup>۱۴</sup> به صورت زیر مشخص می‌شود:

Actual	Predicted	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

از آنجا که داده‌های برچسب‌دار انتخاب‌شده در نتایج به دست آمده تأثیرگذار هستند، روش پیشنهادی (و نیز سایر روش‌هایی که برای مقایسه انتخاب‌شده‌اند) ده مرتبه روی داده‌های برچسب‌دار متفاوت که به صورت تصادفی انتخاب شده‌اند، اجرا می‌شود و میانگین نتایج این اجراها گزارش می‌شود تا نتایج پایدار و قابل استنادی داشته باشیم. روش پیشنهادی با الگوریتم باناظری که از همان تعداد داده‌های برچسب‌دار برای آموزش استفاده می‌کند، مقایسه می‌شود. نتایج این مقایسه، تأثیر به کارگیری داده‌های بدون برچسب در مرحله آموزش را نشان می‌دهد. علاوه بر آن، روش پیشنهادی با روش نیمه‌نظارتی خودآموز مقایسه می‌شود.

#### ۴. روش پیشنهادی

در این مقاله استفاده از داده‌های بدون برچسب در قالب یادگیری نیمه‌نظارتی جهت آموزش مدل مناسب برای نظرکاوی نظرات فارسی مدنظر است. روش پیشنهادی قابلیت پیش‌بینی قطبیت موجود در نظراتی را که در مرحله آموزش در دسترس نداشته نیز دارد. با در دسترس بودن تعداد  $n$  نظر  $\{r_1, \dots, r_n\}$ ، هدف وظیفه<sup>۱۵</sup> تعیین قطبیت در عقیده‌کاوی، یادگیری تابعی است که بتواند مثبت یا منفی بودن تمایل موجود در نظر را پیش‌بینی کند. در یادگیری نیمه‌نظارتی، فرض بر آن است که برچسب تعداد محدودی از  $n$  نظر در دسترس بوده و برچسب سایر نظرات مشخص نیست، هرچند از تمام این نظرات جهت

جدول ۱: نمونه‌ای از نظرات موجود در مجموعه داده حاوی

نظرات دیجی کالا

تمایل	متن نظر
مثبت	خوب و با دوام
مثبت	در یک کلام برای استفاده در طبیعت و به علت سبکی و
منفی	مقاوم بودن عالییه
	خریدمش ولی یک بار هم ازش استفاده نکردم.. خراب بود
	راضی نیستم
منفی	کاملاً بلا استفاده و اسباب بازی، فاقد ارزش

یادگیری مدل استفاده می‌شود. مراحل روش پیشنهادی در شکل (۱) نشان داده شده است. ابتدا گام‌های پیش‌پردازشی روی داده اعمال شده تا بردار مربوط به هر نظر استخراج شود. در این مرحله از دو منبع واژگان

ناهمگن<sup>۱۶</sup> ایجاد کرده و با استفاده از الگوریتم نیمه‌نظارتی قدم‌زدن تصادفی دسته‌بندی را انجام داده‌اند. الگوریتم قدم‌زدن تصادفی<sup>۱۱</sup> الگوریتمی ترارسان بوده و قابلیت برچسب‌گذاری داده‌های بدون برچسبی که در مرحله آموزش وجود نداشته‌اند، ندارد [۲۳]. در این پژوهش اعمال الگوریتم نیمه‌نظارتی که در معرض چالش‌های بیان‌شده برای خودآموز نبوده و قابلیت اعمال به نظرات دیده‌نشده در مرحله آموزش را داشته باشد مدنظر است.

#### ۳. روش پژوهش

در این پژوهش با مطالعه منابع علمی مرتبط و معتبر، روشی جهت تخمین تمایل موجود در نظرات متنی با استفاده از روش‌های یادگیری ماشین با هدف استفاده از داده‌های برچسب‌دار توسعه داده می‌شود. هدف مقاله پرداختن به این مسئله است که آیا می‌توان با تعداد کمتری نظر برچسب‌گذاری شده به زبان فارسی در مرحله آموزش مدل به کارایی حداقل برابر یا حتی بهتر از روش‌های باناظری که از همان تعداد داده برچسب‌دار استفاده می‌کنند، رسید؟ مدلی مدنظر است که در نهایت، قابلیت تخمین نظرات جدید را داشته باشد و در معرض خطای ناشی از تعداد کم داده‌های برچسب‌دار نباشد.

نتایج اعمال روش پیشنهادی روی مجموعه داده حاوی نظرات دیجی کالا ارزیابی شده است. این مجموعه داده حاوی ۱۰۰۰۰ نظر بدون برچسب را دیجی نکست، مرکز نوآوری و سرمایه‌گذاری دیجی کالا منتشر کرده است. زیرمجموعه‌ای از این نظرات به صورت دستی برچسب‌گذاری شده و ارزیابی روش پیشنهادی روی این زیرمجموعه داده انجام شده است. نمونه‌ای از این نظرات در جدول ۱ آمده است.

ابتدا نظرات تبدیل به بردارهای عددی شده و سپس الگوریتم یادگیری ماشین نیمه‌نظارتی مدنظر روی آن‌ها آموزش داده می‌شود. تعدادی داده برچسب‌دار به صورت تصادفی از بین داده‌ها انتخاب شده و همراه با تعدادی داده بدون برچسب در مرحله آموزش استفاده می‌شوند. پیاده‌سازی با استفاده از زبان پایتون انجام شده است. کتابخانه‌های سکلرن<sup>۱۲</sup> و سمی سوپروایزید<sup>۱۳</sup> از کتابخانه‌های اصلی هستند که مورد استفاده قرار گرفته‌اند.

روش به کاررفته برای یادگیری نیمه‌نظارتی، SVM<sup>۱۷</sup>، الگوریتم توسعه‌یافته ماشین بردار پشتیبان است که قابلیت تعامل با تمام چالش‌های بیان‌شده در این مقاله را دارد. به صورت همزمان از داده‌های برچسب‌دار و بدون برچسب یادمی‌گیرد و مدل به دست آمده قابلیت پیش‌بینی داده‌های بدون برچسب جدید را نیز دارد. علاوه بر آن، عملکرد مناسب ماشین بردار پشتیبان در بین سایر روش‌های باناظر در کاربرد نظرکاوی از دلایل ما برای انتخاب این روش بود.

ارزیابی بر اساس ملاک‌های استاندارد روش‌های دسته‌بندی شامل دقت، صحت، بازخوانی و F1 انجام شده است.

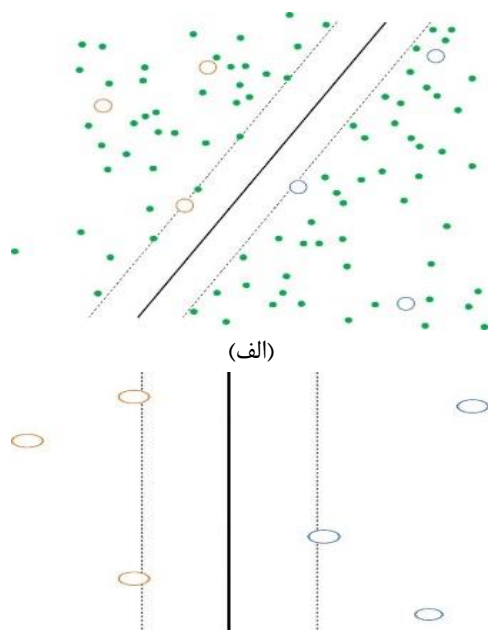
$$Precision = \frac{TP}{TP + FP} \quad (1)$$

بدون برچسب یادگرفته‌است بهتر از مرز تصمیمی است که در یادگیری آن فقط از داده‌های برچسب‌دار استفاده شده‌است. داده‌های بدون برچسبی که خارج از حاشیه<sup>۱۱</sup> و به‌دور از مرز تصمیم هستند می‌توانند در تخمین بهتر مرز تصمیم کمک‌کننده باشند.

در بیان رسمی مسئله، داده‌های بدون برچسبی که در بین دو حاشیه قرار می‌گیرند جریمه می‌شوند. به عبارت دیگر، مرز تصمیم باید در ناحیه با چگالی پایین مجموعه داده باشد به طوری که تعداد کمی از نمونه‌های بدون برچسب به آن نزدیک باشد. با در نظر گرفتن مجموعه نظرات به صورت  $\{x_i \in R^D\}_{i=1}^n$  و با فرض اینکه برچسب  $l$  نظر اول،  $\{y_i \in \{+1, -1\}\}_{i=1}^l$ ، در دسترس بوده و  $u$  نظر بعدی بدون برچسب باشند<sup>۱۲</sup> ( $n = l + u$ )، مسئله بهینه‌سازی SVM به صورت رابطه (۶) است [۹]:

$$\min_{w,b} \sum_{i=1}^l \max(1 - y_i(w^T x_i + b), 0) + \lambda_1 \|w\|^2 + \lambda_2 \sum_{j=l+1}^{l+u} \max(1 - |w^T x_j + b|, 0) \quad (6)$$

جمله اول، تابع هزینه هینگ<sup>۱۱</sup> است که نمونه‌های برچسب‌داری را که برچسب تخمینی آن‌ها برابر برچسب واقعی آن‌ها نیست جریمه می‌کند. جمله دوم، جمله منظم‌ساز<sup>۱۲</sup> برای جلوگیری از بیش‌برازش است. اولین دو جمله، جملاتی هستند که مسئله بهینه‌سازی SVM را شکل می‌دهند و جمله سوم مخصوص یادگیری نیمه‌نظارتی است. این جمله تابع هزینه کلاه<sup>۱۳</sup> را محاسبه می‌کند که داده‌های بدون برچسب را در صورتی که در بین دو حاشیه قرار گیرد، جریمه می‌کند. تابع هزینه هینگ و کلاه در شکل (۳) نشان داده شده‌اند.  $\lambda_1$  و  $\lambda_2$  پارامترهایی هستند که اهمیت جملات مربوطه را نسبت به یکدیگر نشان می‌دهند. همانند SVM، SVM نیز می‌تواند با بهره‌گیری از توابع هسته<sup>۱۴</sup>، دسته‌هایی را که به صورت غیرخطی از یکدیگر جدا شده‌اند از همدیگر تمیز دهد.



احساسی موجود برای زبان فارسی استفاده شده‌است. سپس الگوریتم دسته‌بندی SVM با استفاده از ترکیبی از داده‌های برچسب‌دار و بدون برچسب یادگرفته می‌شود. در ادامه جزئیات این مراحل بیان شده‌است.

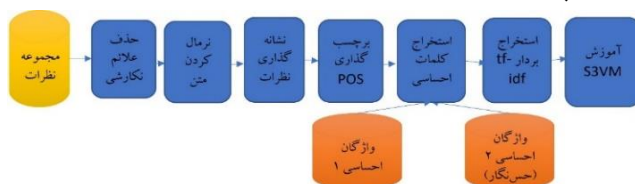
#### ۱.۴. گام پیش پردازش

ابتدا بعد از حذف علائم نگارشی و نرمال‌سازی متن، جداسازی واژگان<sup>۱۶</sup> متن انجام شده و کلمات موجود در متن استخراج می‌شود. همچنین برچسب‌گذاری اجزا کلام<sup>۱۷</sup> نیز انجام می‌شود. سپس کلماتی که بار احساسی مثبت یا منفی دارند با بهره‌گیری از دو منبع واژگان احساسی استخراج شده‌است. یکی از این منابع شامل ۲۹۴۸ واژه مثبت و منفی است<sup>۱۸</sup>. منبع واژگان دیگر، حس‌نگار است که شامل تعداد زیادی از لغات با بار احساسی مثبت و منفی با در نظر گرفتن نقش آن‌ها در جمله است [۱۶].

بعد از استخراج کلمات احساسی، بردار فراوانی عبارت-معکوس فراوانی سند برای هر کلمه احساسی موجود در هر نظر، بر اساس رابطه (۵) محاسبه می‌شود:

$$tf - idf_{i,j} = tf_{i,j} \times \log\left(\frac{n}{df_i}\right) \quad (5)$$

$n$  تعداد کل نظرات است و  $tf_{i,j}$  فراوانی عبارت  $i$  در سند  $j$  را نشان می‌دهد و  $df_i$  نشان‌دهنده تعداد اسناد شامل عبارت  $i$  است. منظور از عبارت در اینجا کلمات احساسی استخراج شده‌است که همان ویژگی‌های بردارها را تشکیل می‌دهند. بدین ترتیب، هر نظر  $r_i$  به صورت بردار  $x_i = (x_{i1}, \dots, x_{iD})$  بازنمایی می‌شود که در آن  $x_{ij} = tf - idf_{i,j}$  و  $D$  برابر تعداد کل عبارات احساسی موجود در نظرات است. بردارهای به دست آمده نهایتاً نرمال می‌شوند تا مقادیری بین ۰ و ۱ داشته باشند.



شکل ۱: مراحل الگوریتم پیشنهادی

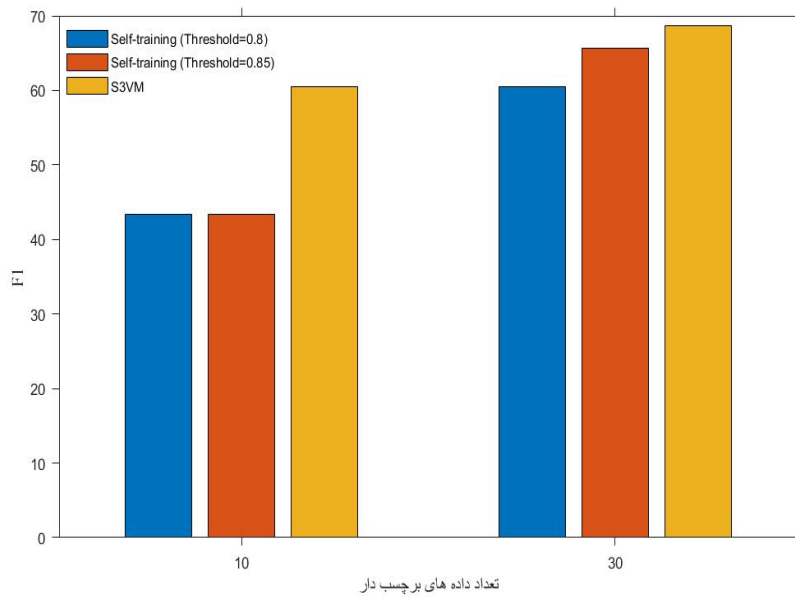
#### ۲.۴. مدل ماشین بردار پشتیبان نظارتی

بعد از استخراج بردارهای عددی نظرات، الگوریتم SVM به داده‌ها اعمال می‌شود تا مرز تصمیم مناسب برای تمایز دو قطبیت مثبت و منفی یادگرفته شود. الگوریتم SVM توسعه یافته SVM جهت بهره‌گیری از داده‌های بدون برچسب در مرحله آموزش است [۹]. الگوریتم SVM یک مدل یادگیری نظارتی شناخته شده مبتنی بر یادگیری ماشین آماری است که بر مبنای این فرض عمل می‌کند که مرز تصمیم دسته‌ها در نواحی با چگالی پایین قرار دارد. شهود استفاده از داده‌های بدون برچسب در مرحله آموزش در شکل (۲) نشان داده شده‌است. همان‌طور که در شکل مشخص است، داده‌های برچسب‌دار دو شکل الف و ب یکی هستند اما مرز تصمیمی که با کمک داده‌های



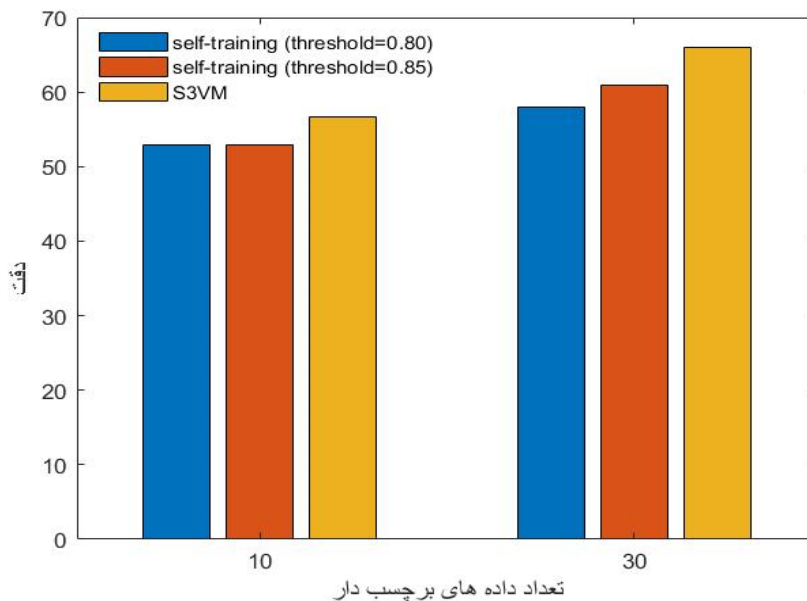
جدول ۳. مقایسه نتایج الگوریتم SVM با SVM به ازای مقادیر مختلف داده‌های برچسب‌دار

SVM				S3VM				تعداد داده‌های برچسب‌دار
F1 (%)	صحت (%)	بازخوانی (%)	دقت (%)	F1 (%)	صحت (%)	بازخوانی (%)	دقت (%)	
۴۲.۸۹ (۰.۱۰۴۰)	۳۵.۹۸ (۰.۱۵۱۰)	۵۳.۱۰ (۰.۰۳۸۵)	۵۳.۱۰ (۰.۰۳۸۵)	۶۰.۴۹ (۰.۰۸۹۸)	۶۴.۸۳ (۰.۱۰۳۰)	۵۶.۷۰ (۰.۰۷۱۵)	۵۶.۷۰ (۰.۰۷۱۵)	۱۰
۵۰.۸۶ (۰.۱۱۴۴)	۴۸.۳۷ (۰.۲۱۴۰)	۵۳.۷۵ (۰.۰۵۸۱)	۵۳.۷۵ (۰.۰۵۸۱)	۶۵.۸۰ (۰.۱۰۷۸)	۶۸.۵۱ (۰.۱۰۳۶)	۶۳.۳۰ (۰.۰۷۵۸)	۶۳.۳۰ (۰.۰۷۵۸)	۲۰
۶۰.۱۵ (۰.۱۱۰۱)	۶۰.۵۵ (۰.۱۸۱۹)	۵۹.۷۵ (۰.۰۵۷۱)	۵۹.۷۵ (۰.۰۵۷۱)	۶۸.۶۳ (۰.۰۵۶۶)	۷۱.۴۴ (۰.۰۵۴۴)	۶۶.۰۵ (۰.۰۴۲۰)	۶۶.۰۵ (۰.۰۴۲۰)	۳۰
۶۶.۶۱ (۰.۰۷۳۸)	۷۰.۷۳ (۰.۰۳۱۴)	۶۲.۹۵ (۰.۰۳۹۷)	۶۲.۹۵ (۰.۰۳۹۷)	۷۱.۱۲ (۰.۰۷۷۱)	۰۷.۷۴ (۰.۰۴۱۵)	۶۸.۳۹ (۰.۰۵۹۴)	۶۸.۳۹ (۰.۰۵۹۴)	۴۰
۷۰.۲۶ (۰.۰۸۸۱)	۷۳.۱۵ (۰.۰۳۷۰)	۶۷.۵۹ (۰.۰۶۰۲)	۶۷.۵۹ (۰.۰۶۰۲)	۷۲.۰۳ (۰.۰۷۹۷)	۷۵.۲۸ (۰.۰۲۳۵)	۶۹.۰۵ (۰.۰۵۸۲)	۶۹.۰۵ (۰.۰۵۸۲)	۵۰
۷۱.۳۷ (۰.۰۵۸۵)	۷۳.۶۴ (۰.۰۲۳۶)	۶۹.۰۵ (۰.۰۴۳۸)	۶۹.۰۵ (۰.۰۴۳۸)	۷۱.۴۳ (۰.۰۷۵۵)	۷۵.۴۵ (۰.۰۵۰۴)	۶۷.۶۵ (۰.۰۵۵۹)	۶۷.۶۵ (۰.۰۵۵۹)	۶۰
۷۰.۸۸ (۰.۰۹۶۵)	۷۳.۵۵ (۰.۰۳۳۶)	۶۸.۳۹ (۰.۰۶۹۱)	۶۸.۳۹ (۰.۰۶۹۱)	۷۴.۸۹ (۰.۰۵۰۷)	۷۷.۲۱ (۰.۰۲۶۵)	۷۲.۷۰ (۰.۰۴۲۶)	۷۲.۷۰ (۰.۰۴۲۶)	۷۰
۷۳.۶۴ (۰.۰۵۲۶)	۷۶.۰۲ (۰.۰۲۷۴)	۷۱.۴۰ (۰.۰۴۰۳)	۷۱.۴۰ (۰.۰۴۰۳)	۷۴.۶۸ (۰.۰۶۱۸)	۷۶.۲۸ (۰.۰۳۱۳)	۷۳.۱۵ (۰.۰۴۷۶)	۷۳.۱۵ (۰.۰۴۷۶)	۸۰
۷۳.۰۷ (۰.۰۴۵۹)	۷۴.۳۸ (۰.۰۳۹۲)	۷۱.۹۰ (۰.۰۴۰۴)	۷۱.۹۰ (۰.۰۴۰۴)	۷۶.۹۶ (۰.۰۶۶۳)	۷۸.۸۵ (۰.۰۳۱۲)	۷۵.۱۵ (۰.۰۵۵۳)	۷۵.۱۵ (۰.۰۵۵۳)	۹۰
۷۵.۳۰ (۰.۰۵۲۷)	۷۸.۱۴ (۰.۰۳۵۹)	۷۲.۶۵ (۰.۰۴۶۵)	۷۲.۶۵ (۰.۰۴۶۵)	۷۷.۳۱ (۰.۰۷۴۴)	۷۹.۰۵ (۰.۰۲۴۰)	۷۵.۶۵ (۰.۰۵۹۴)	۷۵.۶۵ (۰.۰۵۹۴)	۱۰۰



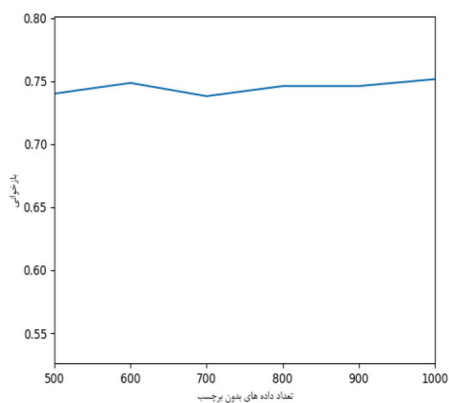
(الف)



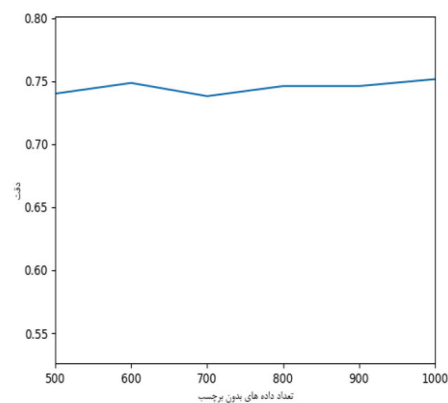


(ب)

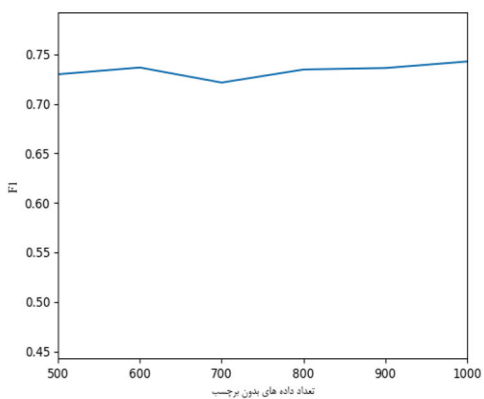
شکل ۴: مقایسه نتایج S3VM با خودآموز بر حسب (الف) F1 و (ب) دقت



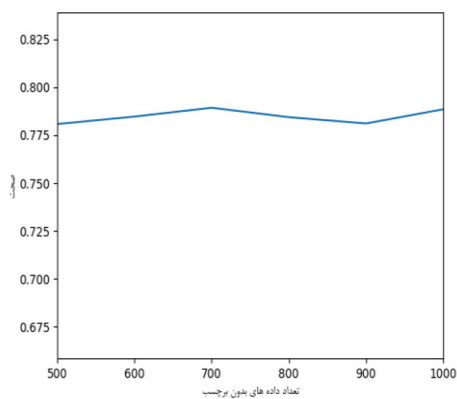
(ب)



(الف)



(د)



(ج)

شکل ۵: نمودار ملاک‌های ارزیابی روش S3VM به‌ازای تعداد متفاوت داده‌های بدون برچسب (الف) دقت، (ب) F1، (ج) صحت و (د) بازخوانی

- of Multi-dimensional Classifiers," *Neurocomputing* 92, pp. 98-115, 2012.
- [15] M. Najafzadeh, S. Rahati Quchan and R. Ghaemi, "A Semi-Supervised Framework based on Self-constructed Adaptive Lexicon for Persian Sentiment Analysis," *Signal and Data Processing*, pp. 89-102, 2018.
- [16] E. Asgarian, M. Kahani and S. Sharifi, "Hesnegar: Persian sentiment wordnet," *Signal and Data Processing*, pp. 71-86, 2018.
- [17] Z. Rajabi and M. Hourali, "Sentiment Analysis Methods in Persian Text: A survey," *Signal and Data Processing*, pp. 107-132, 2022.
- [18] E. Vaziripour, C. Giraud-Carrier and D. Zappala, "Analyzing the Political Sentiment of Tweets in Farsi," *Tenth International AAAI Conference on Web and Social Media*, 2016.
- [19] Z. Li, Y. Fan, B. Jiang, T. Lei and W. Liu, "A Survey on Sentiment Analysis and Opinion Mining for Social Multimedia," *Multimedia Tools and Applications*, pp. 6939-6967, 2019.
- [20] Z. Karimi, "Opinion Mining of Drug Reviews using Support Vector Machine for Multiple Instance Learning," *1st International and 3rd National Conference on Biomathematics*, pp. 218-224, 2022.
- [21] A. Bagheri and M. Saraee, "Persian Sentiment Analyzer: A Framework based on a Novel Feature Selection Method," *International Journal of Artificial Intelligence*, pp. 115-129, 2014.
- [22] M. Shams, A. Shakery and H. Faili, "A Nonparametric LDA-based Induction Method for Sentiment Analysis," *Artificial Intelligence and Signal Processing*, 2012.
- [23] I. Dehdarbehbahani, A. Shakery and H. Faili, "Semi-Supervised Word Polarity Identification in Resource-lean Languages," *Neural networks* 58, pp. 50-59, 2014.
- [24] K. Dashtipour, A. Hussain, Q. Zhou, A. Gelbukh, A. Y. A. Hawalah and E. Cambria, "PerSent: A Freely Available Persian Sentiment Lexicon," *International Conference on Brain Inspired Cognitive Systems*, pp. 310-320, 2016.
- [25] E. Cambria, P. Soujanya, H. Amir and L. Bing, "Computational Intelligence for Affective Computing and Sentiment Analysis [Guest Editorial]," *IEEE Computational Intelligence Magazine*, pp. 16-17, 2019.
- [26] P. Hosseini, A. Ahmadian Ramaki, M. Anvari, H. Maleki and S. A. Mirroshandel, "SentiPers: A Sentiment Analysis Corpus for Persian," *Conference on Computational Linguistics*, 2013.
- [27] B. Sabeti, P. Hosseini, G. Ghassem-Sani and S. A. Mirroshandel, "An ontology based sentiment lexicon for Persian," *Global Conference on Artificial Intelligence (GCAI)*, pp. 329-339, 2016.
- [28] M. Moradi, P. Khosravizade and V. Bahram, "Constructing tagged corpora with a web approach as a corpus," *the 2th symposium on computational Linguistics*, 2012.
- [29] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani and A. Hussain, "Sentiment Analysis of Persian Movie Reviews Using Deep Learning," *Entropy*, 2021.
- [30] P. F. Brown, P. V. de Souza, R. L. Mercer, V. J. D. Pietra and C. L. Jennifer, "Class-based n-gram Models of Natural Language," *Computational linguistics*, p. 467-479, 1992.
- [1] M. Kang, J. Ahn and K. Lee, "Opinion Mining using Ensemble Text Hidden Markov Models for Text Classification," *Expert Systems with Applications*, pp. 218-227, 2018.
- [2] S. Mokarrami Sefidab, S. A. Mirroshandel, H. Ahmadifar and M. Mokarrami, "Adversarial Attacks on a Text Sentiment Analysis Model," *Intelligent Multimedia Processing and Communication Systems*, vol. 2, no. 2, pp. 9-16, 2021.
- [3] L. Yue, C. Weitong, L. Xue, Z. Wanli and Y. Minghao, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, pp. 617-663, 2019.
- [4] T. P.D., "Thumbs up or thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," *40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424, 2002.
- [5] X. Ding, B. Liu and P. S. Yu, "A Holistic Lexicon-based Approach to Opinion Mining," *Proceedings of the International Conference on Web Search and Web Data*, pp. 231-240, 2008.
- [6] Z. Karimi and K. Nasiri, "Sentiment Analysis of Digikala Opinions using Adaptive Neuro-Fuzzy Inference System," *In Proceeding of 4th International Conference on Soft Computing*, pp. 1035-1043, 2021.
- [7] M. S. Sabuj, Z. Afrin and K. M. A. Hasan, "Opinion mining using support vector machine with web based diverse data," *International Conference on Pattern Recognition and Machine Intelligence*, pp. 673-678, 2017.
- [8] M. R. Saleh, M. Teresa Martín-Valdivia, A. Montejor-Ráez and L. A. Ureña López, "Experiments with SVM to classify opinions in different domains," *Expert Systems with Applications*, pp. 14799-14804, 2011.
- [9] X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning," *Synthesis lectures on artificial intelligence and machine learning* 3, no. 1, pp. 1-130, 2009.
- [10] Z. Karimi and S. Shiry Ghidary, "Semi-Supervised Metric Learning in Stratified Spaces via Intergrating Local Constraints and Information-theoretic non-local Constraints," *Neurocomputing* 312, pp. 165-176, 2018.
- [11] F. Hassan Khan, U. Qamar and S. Bashir, "A Semi-Supervised Approach to Sentiment Analysis using Revised Sentiment Strength based on SentiWordNet," *Knowledge and information Systems*, pp. 851-872., 2017.
- [12] D. Anand and D. Naorem, "Semi-Supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering," *Procedia Computer Science*, pp. 86-93, 2016.
- [13] Y. He and D. Zhou, "Self-training from labeled features for sentiment analysis," *Information Processing & Management*, pp. 606-616, 2011.
- [14] J. Ortigosa-Hernández, J. Diego Rodríguez, L. Alzate, M. Lucania, I. Inza and J. A. Lozano, "Approaching Sentiment Analysis by using Semi-Supervised Learning

- IEICE TRANSACTIONS on Information and Systems, pp. 790-797, 2014.
- [36] T. Yang, L. Hu, C. Shi, H. Ji, X. Li and L. Nie, "HGAT: Heterogeneous Graph Attention Networks for Semi-Supervised Short Text Classification," ACM Transactions on Information Systems (TOIS), pp. 1-29, 2021.
- [37] N. F. F. D. Silva, L. F. Coletta and E. R. Hruschka, "A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning," ACM Computing Surveys (CSUR), pp. 1-26, 2016.
- [38] Z. Jahanbakhsh-Nagadeh, M.-R. Feizi-Derakhshi and A. Sharifi, "A Semi-Supervised Model for Persian Rumor Verification based on Content Information," Multimed Tools Applications 80, p. 35267-35295, 2021.
- [31] L. Gonbadi and N. Ranjbar, "Sentiment Analysis of People's opinion about Iranian National," Intelligent Multimedia Processing and Communication Systems, vol. 3, no. 4, pp. 51-60, 2023.
- [32] M. B. Dastgheib, S. Koleini and F. Rasti, "The Application of Deep Learning in Persian Documents Sentiment Analysis," International Journal of Information Science and Management (IJISM), pp. 1-15, 2020.
- [33] R. Ghasemi, S. A. Ashrafi Asli and S. Momtazi, "Deep Persian sentiment analysis: Cross-lingual training for low-resource languages," Journal of Information Science 48, pp. 449-462, 2022.
- [34] G. Ansari, C. Saxena, T. Ahmad and M. Doja, "Aspect Term Extraction using Graph-based Semi-Supervised Learning," Procedia Computer Science, vol. 167, pp. 2080-2090, 2020.
- [35] Y. Ren, N. Kaji, N. Yoshinaga and M. Kitsuregawa, "Sentiment Classification in Under-resourced Languages using Graph-based Semi-Supervised Learning Methods,"

### پی‌نوشت

<sup>۱۸</sup> [http://dataheart.ir/article/3460/%D9%84%D8%BA%D8%AA-%D9%86%D8%A7%D9%85%D9%87-%D9%81%D8%A7%D8%B1%D8%B3%DB%8C-%D8%A8%D8%B1%D8%A7%DB%8C-%D9%86%D8%B8%D8%B1%DA%A9%D8%A7%D9%88%DB%8C-\(%D8%B9%D9%82%DB%8C%D8%AF%D9%87-%DA%A9%D8%A7%D9%88%DB%8C-%DB%8C%D8%A7-%D8%AA%D8%AD%D9%84%DB%8C%D9%84-%D8%A7%D8%AD%D8%B3%D8%A7%D8%B3%D8%A7%D8%A](http://dataheart.ir/article/3460/%D9%84%D8%BA%D8%AA-%D9%86%D8%A7%D9%85%D9%87-%D9%81%D8%A7%D8%B1%D8%B3%DB%8C-%D8%A8%D8%B1%D8%A7%DB%8C-%D9%86%D8%B8%D8%B1%DA%A9%D8%A7%D9%88%DB%8C-(%D8%B9%D9%82%DB%8C%D8%AF%D9%87-%DA%A9%D8%A7%D9%88%DB%8C-%DB%8C%D8%A7-%D8%AA%D8%AD%D9%84%DB%8C%D9%84-%D8%A7%D8%AD%D8%B3%D8%A7%D8%B3%D8%A7%D8%A)

<sup>۱۹</sup> margin

<sup>۲۰</sup> ترتیب بیان شده برای برجسب‌دار بودن یا نبودن داده‌ها اهمیت ندارد و صرفاً جهت سادگی توضیح مسئله داده‌های اول برجسب‌دار فرض شده‌اند.

<sup>۲۱</sup> hinge loss

<sup>۲۲</sup> regularization term

<sup>۲۳</sup> hat loss

<sup>۲۴</sup> kernel function

<sup>۲۵</sup> Grid Search

<sup>۱</sup> Co-training

<sup>۲</sup> Self-training

<sup>۳</sup> Brown

<sup>۴</sup> structural correspondence learning

<sup>۵</sup> information gain

<sup>۶</sup> subjective or objective

<sup>۷</sup> Pagerank

<sup>۸</sup> Transductive

<sup>۹</sup> Bidirectional Encoder Representations from Transformers

<sup>۱۰</sup> heterogeneous, multilingual and weighted semantic network

<sup>۱۱</sup> Random Walk

<sup>۱۲</sup> sklearn

<sup>۱۳</sup> semisupervised

<sup>۱۴</sup> Confusion

<sup>۱۵</sup> task

<sup>۱۶</sup> tokenize

<sup>۱۷</sup> Part-of-speech tagging