فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Spring 2023, 4 (1), 51-61
DOI: 10.71856/impcs.2023.903623

Received: 14 Feb 2023
Accepted: 13 Apr 2023

مقاله پژوهشی

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

Zohre Karimi[1*], Hadith Haghiri[2]

1. Assistant Professor, Department of Engineering, Damghan University, Damghan, Iran.
Corresponding Author, z.karimi@du.ac.ir
2. B.A, Computer Engineering, Damghan University, Damghan, Iran. 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.

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران

عقیده‌کاوی نظرات دیجی‌کالا با استفاده از روش نیمه‌نظارتی مبتنی بر ماشین بردار پشتیبان

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

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

زهره کریمی[1*]، حدیث حقیری[2]

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

چکیده: رشد فراوان نظرات دیجیتال کاربران در مورد خدمات و محصولات منجر به توسعه روش‌های عقیده‌کاوی شده‌است. مدل‌های یادگیری ماشین نظارتی در این زمینه به نتایج خوبی دست‌یافته‌اند. 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 پیشنهادشده که از داده‌های بدون برچسب فراوان همراه با تعداد کمی دادهٔ برچسب‌دار در مرحله آموزش بهره‌می‌گیرد. با توجه به عملکرد مناسب روش ماشین بردار پشتیبان نظارتی جهت عقیده‌کاوی، به‌کارگیری روش نیمه‌نظارتی ماشین بردار پشتیبان پیشنهاد شده‌است. این روش در مقایسه با روش‌های موجود با چالش تقویت خطا مواجه‌نبوده و قادر به تخمین قطبیت نظراتی که در مرحله آموزش دیده‌نشده‌اند، نیز است. روش پیشنهادی روی مجموعه داده نظرات دیجی‌کالا مورد ارزیابی قرارگرفته و با الگوریتم ماشین بردار پشتیبان بر اساس ملاک‌های دقت، صحت، بازخوانی و F1 مقایسه شده‌است. نتایج به‌دست‌آمده 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 دیجی‌کالا.

۱. مقدمه

عقیده‌کاوی، نظرکاوی یا تحلیل احساس به مطالعه و کاربرد روش‌های محاسباتی مرتبط با متن از جمله پردازش زبان طبیعی و متن‌کاوی جهت مشخص‌کردن 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 قرارگرفته‌است. رویکردهای موجود در دو گروه رویکردهای مبتنی بر واژگان و رویکردهای مبتنی بر یادگیری ماشین قرارمی‌گیرند [۳]. روش‌های مبتنی بر واژگان از منابع زبانی شامل واژگان احساسی که دربرگیرندهٔ واژگان احساسی مثبت و منفی هستند استفاده‌می‌کنند 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 اعمال‌می‌کنند. اخیراً نیز روش‌هایی پیشنهادشده‌اند که روش‌های یادگیری ماشین و روش‌های مبتنی بر واژگان را با یکدیگر ترکیب‌می‌کنند.

الگوریتم‌های یادگیری ماشین عموماً در دو دستهٔ روش‌های نظارتی و بدون ناظر قرارمی‌گیرند. روش‌های نظارتی از داده‌های برچسب‌دار به‌عنوان داده‌های آموزشی استفاده‌می‌کنند و نیاز به دردسترس‌بودن داده‌های برچسب‌دار به میزان کافی دارند. هرچند در نظرکاوی برچسب داده‌ها به‌صورت دستی توسط خبره انسانی تعیین‌می‌شوند و لذا نیاز به صرف هزینه و زمان زیاد است. روش‌های بدون ناظر از اطلاعات برچسب داده‌ها استفاده‌نمی‌کنند 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 دادهٔ برچسب‌دار و تعداد زیادی داده بدون برچسب جهت آموزش مدل استفاده‌می‌کنند [۹, ۱۰]. این روش‌ها دانش موردنظر برای یادگیری مدل را از تعداد زیادی دادهٔ بدون برچسب در کنار تعداد اندکی دادهٔ برچسب‌دار دردسترس به‌دست‌می‌آوردند 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 انجام‌می‌شوند. هدف روش‌های مبتنی‌بر سند، پیش‌بینی قطبیت یا تمایل موجود در کل یک نظر است. به‌عبارت‌دیگر هر نظر به‌عنوان یک سند در نظر گرفته‌می‌شود. در روش‌های مبتنی‌بر جمله، قطبیت جملات موجود در هر نظر به صورت جداگانه پیش‌بینی‌می‌شود و در روش‌های مبتنی‌بر جنبه، ابتدا جنبه‌های موجود در نظرات استخراج‌شده و سپس تمایل مربوط به هر جنبه پیش‌بینی‌می‌شود.

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 برگراف، روش‌های هم‌آموزی[1] 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[2] جهت دسته‌بندی نیمه‌نظارتی نظرات استفاده‌کرده‌اند. روش خودآموز ابتدا یک دسته‌بند با ناظر را از داده‌های اندک برچسب‌دار در دسترس یادگرفته و از این دسته‌بند جهت برچسب‌گذاری داده‌های بدون برچسب استفاده‌می‌کند. سپس داده‌هایی را که با اطمینان بالایی درست دسته‌بندی‌کرده‌است به داده‌های برچسب‌دار موجود اضافه‌کرده و این روند را تکرارمی‌کند. 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 استفاده‌می‌کند، خطای خود را در هر مرحله تقویت‌می‌کند. «عسگریان، کاهانی و شریفی» [۱۶] نیز یک روش نیمه‌نظارتی مبتنی بر گراف را جهت استخراج یک منبع واژگان احساسی زبان فارسی به‌کاربرده‌اند. روش‌های نیمه‌نظارتی مبتنی‌بر گراف، چون بر مبنای ایجاد گراف روی داده‌های برچسب‌دار و بدون برچسب اولیه عمل‌می‌کنند، فقط قابلیت برچسب‌گذاری داده‌های بدون برچسبی را دارند که در مرحله آموزش در دسترس آن‌ها قرارگرفته‌است 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 دادهٔ 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 علاوه‌بر نظرات مرحله آموزش (غلبه بر محدودیت روش‌های نیمه‌نظارتی متداول مبتنی بر گراف).

بدین‌منظور، پیشنهادمی‌شود که روش S3VM جهت نظرکاوی عقاید فارسی به‌کاررود. این روش که تعمیم‌یافته ماشین بردار پشتیبان است از همان ابتدا مرز تصمیم را با استفاده از داده‌های برچسب‌دار و بدون برچسب با یکدیگر یادمی‌گیرد، لذا در معرض چالش تقویت خطا نیست و محدودیتی برای برچسب‌گذاری داده‌های بدون برچسب ندارد.

در ادامه، در ابتدا پیشینهٔ پژوهش بررسی‌شده، 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 انجام‌شده‌است [۱۷, ۱۸, ۱۹, ۲۰]. در این بخش، پژوهش‌هایی که در دو منظر زبان و رویکرد با این مقاله مشترک هستند، مورد بررسی قرارمی‌گیرند. ابتدا پژوهش‌هایی که به‌صورت خاص روی زبان فارسی متمرکزشده‌اند مورد بررسی قرارگرفته و سپس مروری بر رویکردهای نظرکاوی نیمه‌نظارتی صورت‌می‌گیرد. درنهایت هم پژوهش‌های اندکی که مبتنی بر رویکرد نیمه‌نظارتی در عقیده‌کاوی زبان فارسی هستند به‌صورت خاص مورد توجه قرارمی‌گیرند.

پردازش متون در زبان فارسی با چالش تعامل با ویژگی‌های خاص این زبان مواجه است. به‌کارگیری 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 ایجادشده هستند [۲۳, ۲۴, ۲۵, ۲۶, ۲۷, ۲۸, ۲۹].(« باقری و سرائی » روی مسئلهٔ انتخاب ویژگی در عقیده‌کاوی متون فارسی متمرکزشده و یک معیار بهبودیافته اطلاعات متقابل را برای انتخاب ویژگی پیشنهادکرده‌اند [۲۱]. در نهایت، دسته‌بند بیزین ساده برای دسته‌بندی نظرات استفاده‌شده‌است. «وزیری‌پور، گیرود کریر و زاپالا» روند تغییر نظرات توئیت‌های فارسی در زمینه عناوین سیاسی را بررسی-کرده‌اند. آن‌ها ابتدا از الگوریتم خوشه‌بندی براون[۳] [۱۸] برای انتخاب ویژگی استفاده‌کرده و سپس دسته‌بند SVM را اعمال‌کرده‌اند [۳۰]. پژوهش‌هایی نیز در عقیده‌کاوی نظرات فارسی با استفاده از روش‌های یادگیری عمیق انجام‌شده‌است. « گنبدی و رنجبر» از مدل زبانی «برت» جهت دسته‌بندی تمایلات موجود در نظرات فارسی در مورد برخی خوردوهای ایرانی استفاده‌کرده‌اند [۳۱]. « دستغیب، کلینی و راستی» [۳۲] ترکیبی از روش‌های یادگیری متناظر ساختاری[۴] و شبکه‌های عصبی کانولوشنال را جهت دسته‌بندی نظرات فارسی اعمال‌کرده‌اند. دستیابی به دقت مناسب در روش‌های یادگیری عمیق نیاز به دردسترس‌بودن 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 پیشنهادکرده‌اند [۳۳].

در ادامه، روش‌های نیمه‌نظارتی عقیده‌کاوی بررسی‌می‌شوند. «انند و نارم » [۱۲] یک روش نیمه‌نظارتی جهت نظرکاوی مبتنی‌بر جنبه، روی نقد فیلم‌ها انجام‌داده‌اند. در این پژوهش، استخراج جنبه‌ها به روش نیمه‌نظارتی مبتنی‌بر گراف صورت‌می‌گیرد. «حسن خان، قمر، بشیر» [۱۱] از یک روش مبتنی‌بر واژگان، بهره‌گرفته و با استفاده از ملاک‌های بهرهٔ اطلاعات[۵] و شباهت کسینوسی امتیازهای کلمات احساسی موجود در واژگان را بهبوددادهاند. در نهایت، الگوریتم SVM جهت تعیین تمایل موجود در نظر در این پژوهش اعمال‌شده‌است. برچسب لازم برای مجموعه آموزشی SVM از طریق اعمال بهبودهایی روی منبع واژگان احساسی تأمین‌شده‌است. در پژوهشی دیگر، از دسته‌بندهای شبکه بیزین نیمه‌نظارتی جهت تعیین عینی یا ذهنی بودن[۶] نظرات استفاده‌شده‌است [۱۴]. «انصاری، احمد، دجا و ساکسنا» [۳۴] نیز یک رویکرد نیمه‌نظارتی مبتنی‌بر گراف جهت استخراج جنبه‌ها پیشنهادکردند. «رن، کاجی، یوشیناگا وکیتسورگاوا » [۳۵] نیز یک روش نیمه‌نظارتی مبتنی‌بر گراف را جهت عقیده‌کاوی به‌کارگرفته‌اند. آن‌ها روی انتخاب مناسب نمونه‌های برچسب‌دار متمرکز-شده و تأثیر ملاک‌های گوناگون از جمله درجهٔ رأس‌ها و مقدار رتبهٔ صفحه[۷] را با یکدیگر مقایسه‌کرده‌اند. یک روش مبتنی‌بر شبکه عصبی گراف ناهمگن نیمه‌نظارتی نیز برای دسته‌بندی نظرات پیشنهاد شده است [۳۶]. این روش، در مقایسه با روش‌های نیمه‌نظارتی ترارسان[۸] که فقط می‌توانند برچسب داده‌های بدون برچسب به‌کاررفته در مرحله آموزش را تخمین‌بزنند، قابلیت دسته‌بندی نظرات بدون برچسبی که در مرحلهٔ آموزش دیده‌نشده‌اند را نیز دارد. مطالعهٔ جامعی در مورد اعمال انواع روش‌های نیمه‌نظارتی روی پیام‌های موجود در توییتر نیز انجام‌شده‌است [۳۷]. پژوهش‌های مورد اشاره در این پاراگراف نظرات انگلیسی را تحلیل‌کرده‌اند.

«عسگریان، کاهانی و شریفی» یک منبع واژگان حسی فارسی را با استفاده از الگوریتم نیمه‌نظارتی مبتنی‌بر مدل مخفی مارکوف و بهره‌گیری از منبع زبانی نگاشت قطبیت SentiWordNet موجود به زبان انگلیسی ایجادکرده‌اند[۱۶]. روش یادگیری عمیق بدون ناظر بازنمایی رمزگذار دوطرفه از مبدل‌ها[۹] که با کمک مجموعه داده کوچکی از شایعات فارسی تنظیم‌شده و با یک مدل یادگیری باناظر ترکیب‌شده‌است جهت اعتبارسنجی شایعات در زبان فارسی به‌کاررفته‌است [۳۸]. هرچند پژوهش‌های ذکرشده دقیقاً در راستای کاربرد مورد نظر این مقاله نیستند اما مؤید اهمیت یادگیری نیمه‌نظارتی در عقیده‌کاوی متون فارسی هستند. دسته‌بند مبتنی‌بر قانون مدل مخفی مارکوف در قالب یادگیری نیمه‌نظارتی خودآموز جهت تعیین قطبیت نظرات موجود در دیجی‌کالا به‌کاررفته‌است [۱۵]. روش خودآموز ابتدا با استفاده از تعداد کم داده‌های برچسب‌دار موجود، مدل را یادمی‌گیرد و سپس احتمال تعلق داده‌های بدون برچسب آموزشی به هر دسته را با استفاده از مدل یادگرفته‌شده، تخمین‌می‌زند. در تکرار بعد، داده‌هایی که با احتمال بالایی متعلق به هر دسته هستند به داده‌های آموزشی اضافه‌می‌شوند و این گام تا برچسب‌گذاری تمام داده‌ها ادامه‌می‌یابد. این الگوریتم در معرض دو چالش قراردارد [۱۰]: چالش تقویت خطا بدین معنا که اگر برچسب‌های تخمینی اشتباه مشخص‌شده‌باشد در تکرارهای بعد، آموزش با استفاده از داده‌های خطادار انجام‌می‌شود و (۲) ناکافی بودن داده بدین معنا که مدل اولیه با تعداد کمی از داده‌ها که برای تعیین مرز تصمیم کافی نیستند، آموزش‌داده‌می‌شود. «دهداربهبهانی، شاکری و فایلی» نیز از یک رویکرد نیمه‌نظارتی جهت عقیده‌کاوی استفاده‌کرده‌اند. آن‌ها از منابع زبانی غنی زبان انگلیسی بهره‌برده و یک شبکه معنایی وزن‌دار چندزبانه

ناهمگن[10] ایجادکرده و با استفاده از الگوریتم نیمه‌نظارتی قدم‌زدن تصادفی دسته‌بندی را انجام‌داده‌اند. الگوریتم قدم‌زدن تصادفی[11] الگوریتمی ترارسان بوده و قابلیت برچسب‌گذاری داده‌های بدون برچسبی که در مرحله آموزش وجودنداشته‌اند، 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 داشته‌باشد مدّنظر 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 داده‌می‌شود. هدف مقاله پرداختن به این مسئله است که آیا می‌توان با تعداد کمتری نظر برچسب‌گذاری‌شده به زبان فارسی در مرحله آموزش مدل به کارایی حداقل برابر یا حتی بهتر از روش‌های باناظری که از همان تعداد داده برچسب‌دار استفاده‌می‌کنند، رسید؟ مدلی مدنظر است که در نهایت، قابلیت تخمین نظرات جدید را داشته‌باشد و در معرض خطای ناشی از تعداد کم داده‌های برچسب‌دار نباشد.

نتایج اعمال روش پیشنهادی روی مجموعه دادهٔ حاوی نظرات دیجی‌کالا ارزیابی‌شده‌است. این مجموعه داده حاوی ۱۰۰۰۰ نظر بدون برچسب را دیجی‌نکست، مرکز نوآوری و سرمایه‌گذاری دیجی کالا منتشرکرده‌است. زیرمجموعه‌ای از این نظرات به‌صورت دستی برچسب‌گذاری‌شده و ارزیابی روش پیشنهادی روی این زیرمجموعه داده انجام‌شده‌است. نمونه‌ای از این نظرات در جدول ۱ آمده‌است.

ابتدا نظرات تبدیل به بردارهای عددی شده و سپس الگوریتم یادگیری ماشین نیمه‌نظارتی مدنظر روی آن‌ها آموزش‌داده‌می‌شود. تعدادی داده برچسب‌دار به‌صورت تصادفی از بین داده‌ها انتخاب‌شده و همراه با تعدادی داده بدون‌برچسب در مرحله آموزش استفاده‌می‌شوند. پیاده‌سازی با استفاده از زبان پایتون انجام‌شده‌است. کتابخانه‌های سکلرن[12] و سمی‌سوپروایزد[13] از کتابخانه‌های اصلی هستند که مورد استفاده قرارگرفته‌اند.

روش به‌کاررفته برای یادگیری نیمه‌نظارتی، S3VM، الگوریتم توسعه‌یافته ماشین بردار پشتیبان است که قابلیت تعامل با تمام چالش‌های بیان‌شده در این مقاله را دارد. به‌صورت همزمان از داده‌های برچسب‌دار و بدون برچسب یادمی‌گیرد و مدل به‌دست‌آمده 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 انجام‌شده‌است.

$$Precision = \frac{TP}{TP + FP} \quad (۱)$$

$$Recall = \frac{TP}{TP + FN} \quad (۲)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۳)$$

$$F1 = \frac{2 \times Precision \times recall}{Precision + recall} \quad (۴)$$

TP، TN، FP و FN بر اساس مقادیر ماتریس اغتشاش[14] به‌صورت زیر مشخص‌می‌شود:

Actual	Predicted	
	Negative	*Positive*
Negative	TN	FP
Positive	FN	TP

از آن‌جا که داده‌های برچسب‌دار انتخاب‌شده در نتایج به‌دست‌آمده تأثیرگذار هستند، روش پیشنهادی (و نیز سایر روش‌هایی که برای مقایسه انتخاب‌شده‌اند) ده مرتبه روی داده‌های برچسب‌دار متفاوت که به‌صورت تصادفی انتخاب‌شده‌اند، اجرامی‌شود و میانگین نتایج این اجراها گزارش‌می‌شود تا نتایج پایدار و قابل‌استنادی داشته‌باشیم. روش پیشنهادی با الگوریتم باناظری که از همان تعداد داده‌های برچسب‌دار برای آموزش استفاده‌می‌کند، مقایسه‌می‌شود. نتایج این مقایسه، تأثیر به‌کارگیری داده‌های بدون برچسب در مرحله آموزش را نشان‌می‌دهد. علاوه بر آن، روش پیشنهادی با روش نیمه‌نظارتی خودآموز مقایسه‌می‌شود.

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 وظیفهٔ[15] تعیین قطبیت در عقیده‌کاوی، یادگیری تابعی است که بتواند مثبت یا منفی بودن تمایل موجود در نظر را پیش‌بینی‌کند. 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 دیجی‌کالا

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

یادگیری مدل استفاده‌می‌شود. مراحل روش پیشنهادی در شکل (۱) نشان‌داده‌شده‌است. ابتدا گام‌های پیش‌پردازشی روی داده اعمال‌شده تا بردار مربوط به هر نظر استخراج‌شود. در این مرحله از دو منبع واژگان

احساسی موجود برای زبان فارسی استفاده‌شده‌است. سپس الگوریتم دسته‌بندی S3VM با استفاده از ترکیبی از داده‌های برچسب‌دار و بدون برچسب یادگرفته‌می‌شود. در ادامه جزئیات این مراحل بیان‌شده‌است.

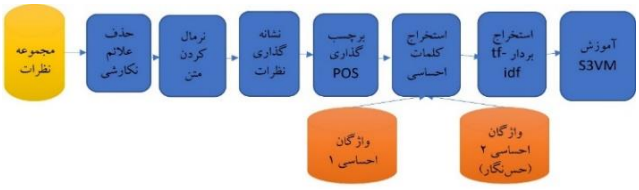
۱،۴. گام پیش پردازش

ابتدا بعد از حذف علائم نگارشی و نرمال‌سازی متن، جداسازی واژگان[۱۶] متن انجام‌شده و کلمات موجود در متن استخراج‌می‌شود. همچنین برچسب‌گذاری اجزا کلام[۱۷] نیز انجام‌می‌شود. 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[۱۸]. منبع واژگان دیگر، حس‌نگار است که شامل تعداد زیادی از لغات با بار احساسی مثبت و منفی با در نظرگرفتن 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) محاسبه‌می‌شود:

$$tf-idf_{i,j} = tf_{i,j} \times \log(\frac{n}{df_i}) \qquad (۵)$$

n تعداد کل نظرات است و $tf_{i,j}$ فراوانی عبارت i در سند j را نشان‌می‌دهد و df_i نشان‌دهنده تعداد اسناد شامل عبارت i است. منظور از عبارت در اینجا کلمات احساسی استخراج‌شده‌است که همان ویژگی‌های بردارها را تشکیل‌می‌دهند. بدین ترتیب، هر نظر r_i به صورت بردار $x_i = (x_{i1}, \dots, x_{iD})$ بازنمایی‌می‌شود که در آن $x_{ij} = tf-idf_{i,j}$ و D برابر تعداد کل عبارات احساسی موجود در نظرات است. بردارهای به‌دست‌آمده نهایتاً نرمال می‌شوند تا مقادیری بین ۰ و ۱ داشته‌باشند.



شکل ۱: مراحل الگوریتم پیشنهادی

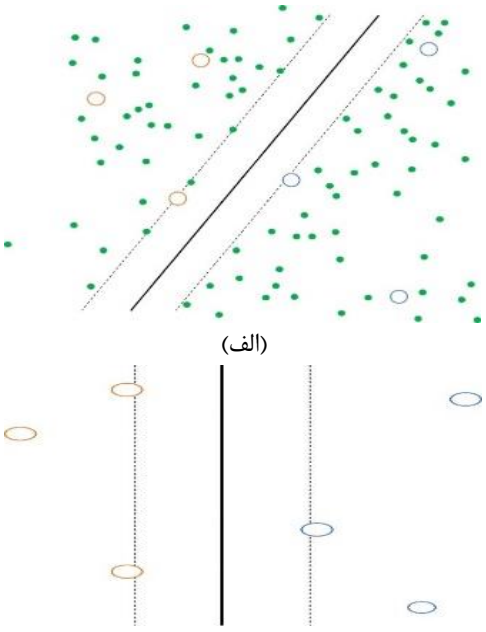
۲،۴. مدل ماشین بردار پشتیبان نظارتی

بعد از استخراج بردارهای عددی نظرات، الگوریتم S3VM به داده‌ها اعمال‌می‌شود تا مرز تصمیم مناسب برای تمایز دو قطبیت مثبت و منفی یادگرفته‌شود. الگوریتم S3VM توسعه یافتهٔ SVM جهت بهره‌گیری از داده‌های بدون برچسب در مرحله آموزش است [۹]. الگوریتم SVM یک مدل یادگیری نظارتی شناخته‌شده مبتنی‌بر یادگیری ماشین آماری است که بر مبنای این فرض عمل‌می‌کند که مرز تصمیم دسته‌ها در نواحی با چگالی پایین قراردارد. شهود استفاده از داده‌های بدون برچسب در مرحلهٔ آموزش در شکل (۲) نشان‌داده شده‌است. همان‌طور که در شکل مشخص است، داده‌های برچسب‌دار دو شکل الف و ب یکی هستند اما مرز تصمیمی که با کمک داده‌های بدون برچسب یادگرفته‌است بهتر از مرز تصمیمی است که در یادگیری آن فقط از داده‌های برچسب‌دار استفاده‌شده‌است. داده‌های بدون برچسبی که خارج از حاشیه[۱۹] 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 قرارمی‌گیرند جریمه‌می شوند. به عبارت دیگر، مرز تصمیم باید در ناحیهٔ با چگالی پایین مجموعه داده باشد به طوری که تعداد کمی از نمونه‌های بدون برچسب به آن نزدیک باشد. با در نظرگرفتن 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[۲۰] ($n = l + u$)، مسئلهٔ بهینه‌سازی S3VM به صورت رابطه (۶) است [۹]:

$$\min_{w,b} \sum_{i=1}^{l} max\,(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b), 0) + \lambda_1\|\mathbf{w}\|^2 + \lambda_2 \sum_{j=l+1}^{l+u} max\,(1 - |\mathbf{w}^T\mathbf{x_j} + b|, 0)] \qquad (۶)$$

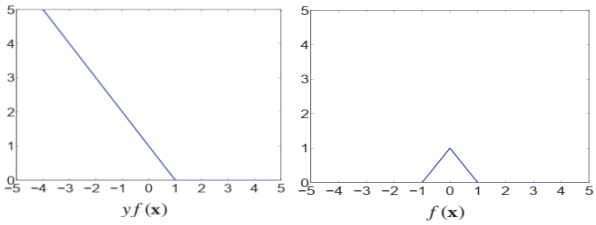
جملهٔ اول، تابع هزینه هینگ[۲۱] است که نمونه‌های برچسب‌داری را که برچسب تخمینی آن‌ها برابر برچسب واقعی آن‌ها نیست جریمه می‌کند. جمله دوم، جمله منظم‌ساز[۲۲] برای جلوگیری از بیش‌برازش است. اولین دو جمله، جملاتی هستند که مسئلهٔ بهینه سازی SVM را شکل‌می‌دهند و جملهٔ سوم مخصوص یادگیری نیمه‌نظارتی است. این جمله تابع هزینهٔ کلاه[۲۳] را محاسبه‌می‌کند که داده‌های بدون برچسب را در صورتی‌که در بین دو حاشیه قرارگیرد، جریمه‌می‌کند. توابع هزینهٔ هینگ و کلاه در شکل (۳) نشان‌داده‌شده‌اند. λ_2 و λ_1 پارامترهایی هستند که اهمیت جملات مربوطه را نسبت به یکدیگر نشان‌می‌دهند. همانند SVM، S3VM نیز می‌تواند با بهره‌گیری از توابع هسته[۲۴]، دسته‌هایی را که به صورت غیرخطی از یکدیگر جدا شده‌اند از همدیگر تمیز دهد.



(الف)

(ب)

شکل ۲: مقایسه (الف) SVM و (ب) S3VM. نقاط آبی و قرمز رنگ داده‌های آموزشی برچسب‌دار و نقاط سبز رنگ،داده‌های بدون برچسبی هستند که در مرحله آموزش S3VM استفاده‌شده‌اند [9]



(ب) (الف)

شکل ۳: (الف) تابع هزینه کلاه و (ب) تابع هزینه هینگ [9]

۵. تجزیه و تحلیل یافته‌ها

در این بخش، آزمایش هایی که جهت ارزیابی روش پیشنهادی ارائه‌شده‌اند، شرح‌داده‌شده و نتایج آن‌ها آمده‌است. مجموعه داده استفاده شده در بخش روش پژوهش شرح‌داده شد. زیر مجموعه‌ای از مجموعه دادهٔ اصلی به‌نحوی انتخاب شده‌است که تعداد نظرات مثبت و منفی در داده‌های برچسب‌دار و نیز بدون برچسب یکسان باشد. تعداد کلمات استخراج‌شده بر اساس گام‌های تشریح شده در مرحله پیش‌پردازش ۷۶۹۶ کلمه است.

ابتدا ۱۰۰۰ دادهٔ بدون برچسب درآموزش مدل S3VM به‌کاررفته و نتایج پیش‌بینی تمایل موجود در نظر روی ۲۰۰ دادهٔ تست مقایسه‌شده‌است. الگوریتم‌های SVM و S3VM با هستهٔ گاوسی، $k(x_i, x_j) = exp(-\gamma \left| |x_i - x_j|^2 \right)$، اعمال‌شده‌اند. پارامترهای الگوریتم‌ها با استفاده از جستجوی توری[۲۵] مشخص شده و مقادیر در نظرگرفته‌شده برای هر پارامتر در جدول ۲ آمده‌است. مقدار صحت، بازخوانی،F1 و دقت دو مدل SVM و S3VM به‌ازای تعداد یکسانی دادهٔ برچسب‌دار با یکدیگر مقایسه‌شده و میانگین و انحراف معیار ملاک‌های نامبرده به‌ازای ۱۰ بار اجرا در جدول ۳ نشان‌داده شده‌است. همان‌طور که انتظارمی‌رود با افزایش تعداد داده‌های برچسب‌دار، کارایی هر دو الگوریتم بهبود پیدامی‌کند اما از آن‌جا که S3VM از داده‌های بدون برچسب نیز در مرحله آموزش بهره‌می‌گیرد عملکرد بهتری در مقایسه با S3VM دارد.

همچنین روش S3VM با روش خودآموز مقایسه شده‌است. جهت مقایسه بهتر، دسته‌بند پایه در روش خودآموز SVM در نظرگرفته شده‌است. نتایج مقایسه به‌ازای دو مقدار ۰,۸ و ۰,۹ برای حدآستانه به ازای تعداد متفاوت ۱۰ و ۳۰ عدد داده‌های برچسب‌دار در شکل ۴ آمده‌است. همان‌طور که در نتیجه مشخص است S3VM عملکرد بهتری در مقایسه با روش خودآموزی دارد که دلیل اصلی آن، این است که SVM اولیه‌ای که درخودآموز با تعداد کمی دادهٔ برچسب‌دار آموزش داده‌می‌شود، کارایی لازم برای تخمین برچسب داده‌های بدون برچسب را ندارد و لذا نمی‌تواند به‌خوبی S3VM که از همان ابتدا داده‌های بدون برچسب را در آموزش واردمی‌کند، عمل‌کند.

در آزمایش بعد، تاثیر تعداد داده‌های بدون برچسب در عملکرد S3VM بررسی‌شده‌است و نتایج الگوریتم به‌ازای تعداد متفاوت داده‌های بدون برچسب در شکل ۳ نشان‌داده‌شده‌است. با افزایش داده‌های بدون برچسب تاحدّی، نتایج بهبود می‌یابد و از حدی به بعد تغییر چندانی حاصل‌نمی‌شود.

۶. نتیجه‌گیری

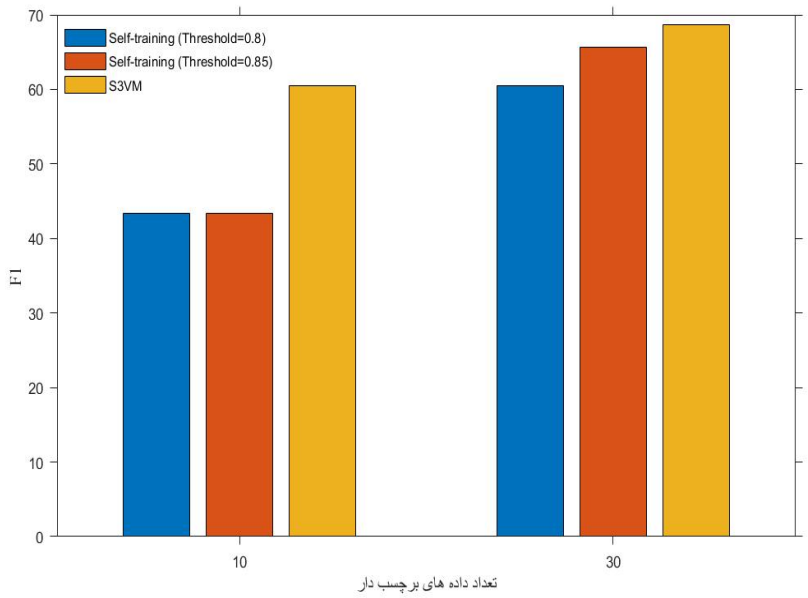
با رشد به‌اشتراک‌گذاری نظرات افراد در رسانه‌های اجتماعی در سال‌های اخیر، ضرورت پردازش خودکار این نظرات به‌وجودآمده‌است. پردازش دقیق نظرات با استفاده از روش‌های یادگیری ماشین نظارتی نیاز به دردسترس‌بودن داده‌های برچسب‌گذاری‌شدهٔ فراوان دارد که عملی هزینه‌بر و مستلزم صرف زمان زیاد است. درحالی‌که نظرات بدون برچسب به‌وفور در دسترس هستند. در این مقاله، روشی نیمه‌نظارتی جهت پیش‌بینی تمایلات موجود در نظرات توسعه‌داده‌شده‌است. بعد از گام‌های پیش‌پردازشی با کمک منابع واژگان احساسی و ویژگی‌های آماری این واژگان، بردارهای نظرات استخراج‌شده و با استفاده از دسته‌بند نیمه‌نظارتی ماشین بردار پشتیبان دسته‌بندی‌می‌شوند. مدل به‌دست‌آمده روی مجموعه دادهٔ نظرات دیجی کالا ارزیابی‌شد و نتایج حاکی از عملکرد بهتر مدل در مقایسه با الگوریتم ماشین بردار پشتیبانی است که فقط از داده‌های برچسب‌دار برای یادگیری مدل استفاده‌می‌کند. جهت ادامه پژوهش، اعمال روش‌های یادگیری نیمه‌نظارتی روی بردارهای تعبیه استخراج‌شده از متن پیشنهادمی‌شود.

جدول ۲: مقادیر ابرپارامترها

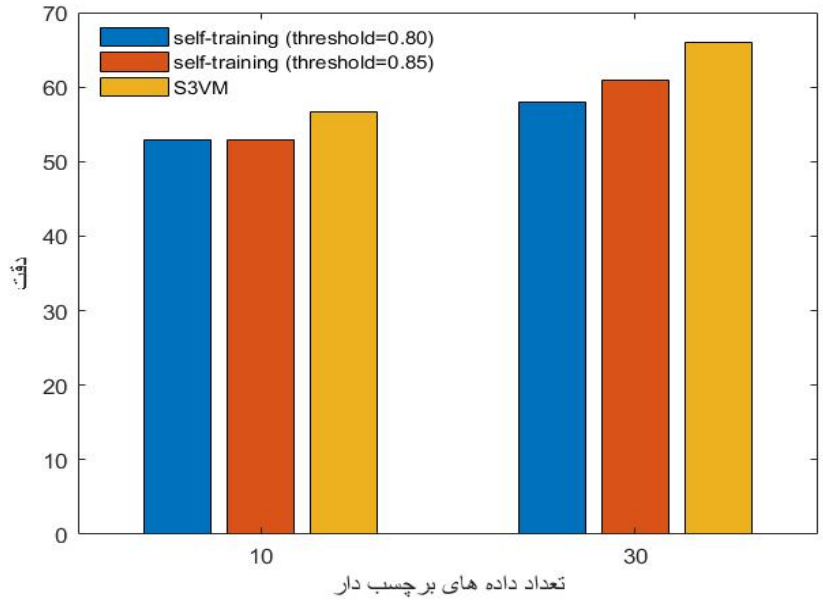
الگوریتم	پارامتر	مقادیر ممکن
SVM, S3VM	λ_1	{۰,۰,۰۰۰۰۰۱ ,۰,۰۰۰۰۱ ,۰,۰۰۰۱ , ۰,۰۰۱،۰,۱،۰,۵،۱,۳،۱۰،۸۰،۱۰۰،۵۰۰،۱۰۰۰،۱۰۰۰۰ ،۱۰۰۰۰۰،۱۰۰۰۰۰۰}
S3VM	λ_2	{۰,۹،۱،۱،۱}

جدول ۳. مقایسه نتایج الگوریتم S3VM با SVM به‌ازای مقادیر مختلف داده‌های برچسب‌دار

تعداد داده های برچسب دار	S3VM				SVM			
	دقت (٪)	بازخوانی (٪)	صحت (٪)	F1 (٪)	دقت (٪)	بازخوانی (٪)	صحت (٪)	F1 (٪)
۱۰	۵۶,۷۰ (۰,۰۷۱۵)	۵۶,۷۰ (۰,۰۷۱۵)	۶۴,۸۳ (۰,۱۰۳۰)	۶۰,۴۹ (۰,۰۸۹۸)	۵۳,۱۰ (۰,۰۳۸۵)	۵۳,۱۰ (۰,۰۳۸۵)	۳۵,۹۸ (۰,۱۵۱۰)	۴۲,۸۹ (۰,۱۰۴۰)
۲۰	۶۳,۳۰ (۰,۰۷۵۸)	۶۳,۳۰ (۰,۰۷۵۸)	۶۸,۵۱ (۰,۱۰۳۶)	۶۵,۸۰ (۰,۱۰۷۸)	۵۳,۷۵ (۰,۰۵۸۱)	۵۳,۷۵ (۰,۰۵۸۱)	۴۸,۲۷ (۰,۲۱۴۰)	۵۰,۸۶ (۰,۱۱۴۴)
۳۰	۶۶,۰۵ (۰,۰۴۲۰)	۶۶,۰۵ (۰,۰۴۲۰)	۷۱,۴۳ (۰,۰۵۴۴)	۶۸,۶۳ (۰,۰۵۶۶)	۵۹,۷۵ (۰,۵۷۱)	۵۹,۷۵ (۰,۵۷۱)	۶۰,۵۵ (۰,۱۸۱۹)	۶۰,۱۵ (۰,۱۱۰۱)
۴۰	۶۸,۳۹ (۰,۰۵۹۴)	۶۸,۳۹ (۰,۰۵۹۴)	۰۷.۷۴ (۰,۰۴۱۵)	۷۱,۱۲ (۰,۰۷۷۱)	۶۲,۹۵ (۰,۰۳۹۷)	۶۲,۹۵ (۰,۰۳۹۷)	۷۰,۷۳ (۰,۰۳۱۴)	۶۶,۶۱ (۰,۰۷۳۸)
۵۰	۶۹,۰۵ (۰,۰۵۸۲)	۶۹,۰۵ (۰,۰۵۸۲)	۷۵,۲۸ (۰,۰۲۳۵)	۷۲,۰۳ (۰,۰۷۹۷)	۶۷,۵۹ (۰,۰۶۰۲)	۶۷,۵۹ (۰,۰۶۰۲)	۷۳,۱۵ (۰,۰۳۷۰)	۷۰,۲۶ (۰,۰۸۸۱)
۶۰	۶۷,۶۵ (۰,۰۵۵۹)	۶۷,۶۵ (۰,۰۵۵۹)	۷۵,۴۵ (۰,۰۵۰۴)	۷۱,۴۳ (۰,۰۷۵۵)	۶۹,۰۵ (۰,۰۴۳۸)	۶۹,۰۵ (۰,۰۴۳۸)	۷۳,۶۴ (۰,۰۲۳۶)	۷۱,۲۷ (۰,۰۵۸۵)
۷۰	۷۲,۷۰ (۰,۰۴۲۶)	۷۲,۷۰ (۰,۰۴۲۶)	۷۷,۲۱ (۰,۰۲۶۵)	۷۴,۸۹ (۰,۰۵۰۷)	۶۸,۳۹ (۰,۰۶۹۱)	۶۸,۳۹ (۰,۰۶۹۱)	۷۳,۵۵ (۰,۳۲۶)	۷۰,۸۸ (۰,۰۹۶۵)
۸۰	۷۳,۱۵ (۰,۰۴۷۶)	۷۳,۱۵ (۰,۰۴۷۶)	۷۶,۲۸ (۰,۰۳۱۳)	۷۴,۶۸ (۰,۰۶۱۸)	۷۱,۴۰ (۰,۰۴۰۳)	۷۱,۴۰ (۰,۰۴۰۳)	۷۶,۰۲ (۰,۰۲۷۴)	۷۳,۶۴ (۰,۰۵۲۶)
۹۰	۷۵,۱۵ (۰,۰۵۵۳)	۷۵,۱۵ (۰,۰۵۵۳)	۷۸,۸۵ (۰,۰۳۱۲)	۷۶,۹۶ (۰,۰۶۶۳)	۷۱,۹۰ (۰,۰۴۰۴)	۷۱,۹۰ (۰,۰۴۰۴)	۷۴,۲۸ (۰,۰۳۹۲)	۷۳,۰۷ (۰,۰۴۵۹)
۱۰۰	۷۵,۶۵ (۰,۰۵۹۴)	۷۵,۶۵ (۰,۰۵۹۴)	۷۹,۰۵ (۰,۰۳۴۰)	۷۷,۳۱ (۰,۰۷۴۴)	۷۲,۶۵ (۰,۰۴۶۵)	۷۲,۶۵ (۰,۰۴۶۵)	۷۸,۱۴ (۰,۰۳۵۹)	۷۵,۳۰ (۰,۰۵۲۷)

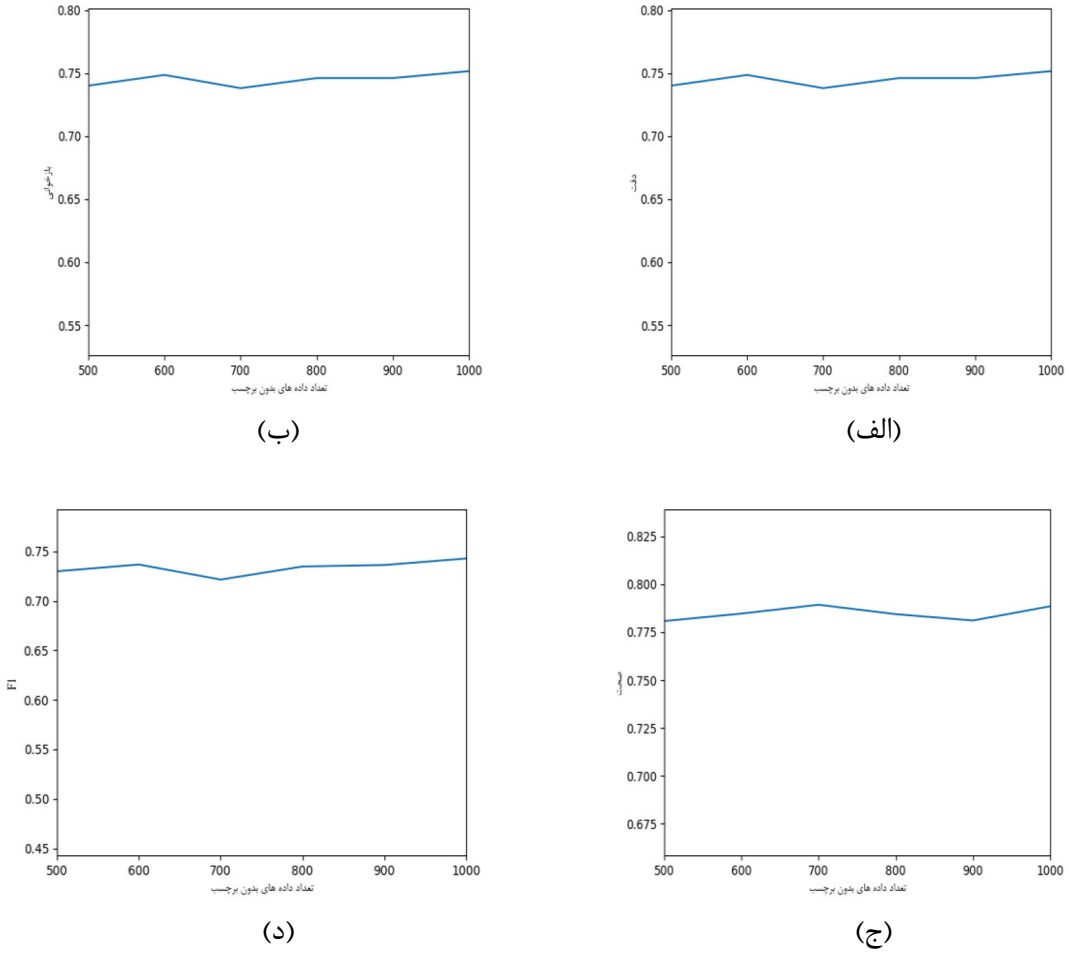


(الف)



(ب)

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



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

سپاسگزاری

این پژوهش تحت حمایت داده‌های متن باز دیجی کالا است.

References

[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. Montejo-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 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.

[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," ournal 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," 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.

پی‌نوشت

۱ Co-training

۲ Self-training

۳ Brown

۴ structural correspondence learning

۵ information gain

۶ subjective or objective

۷ Pagerank

۸ Transductive

۹ Bidirectional Encoder Representations from Transformers

۱۰ heterogeneous, multilingual and weighted semantic network

۱۱ Random Walk

۱۲ sklearn

۱۳ semisupervised

۱۴ Confusion

۱۵ task

۱۶ tokenize

۱۷ Part-of-speech tagging

۱۸ 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%AA)

۱۹ margin

۲۰ ترتیب بیان‌شده برای برچسب‌دار بودن یا نبودن داده‌ها اهمیت ندارد و صرفا جهت سادگی توضیح مسئله داده های اول برچسب‌دار فرض‌شده‌اند.

۲۱ hinge loss

۲۲ regularization term

۲۳ hat loss

۲۴ kernel function

۲۵ Grid Search