مقاله پژوهشی

یک چارچوب یادگیری نیمه‌نظارتی جهت دسته‌بندی دقیق موارد آزمون با بهره‌گیری از تعبیه‌های زبانی و ویژگی‌های معنایی متن

محمدحسین پروانه و مریم نورائی آباده

چکیده: با گسترش کاربرد هوش مصنوعی در مهندسی نرم‌افزار، استفاده از روش‌های هوشمند برای دسته‌بندی موارد آزمون به ضرورتی کلیدی تبدیل شده است. یکی از چالش‌های اصلی در این زمینه، وابستگی شدید مدل‌ها به داده‌های برچسب‌خورده است که تولید آن‌ها هزینه‌بر و زمان‌بر است. در این پژوهش، با هدف بررسی اثربخشی یادگیری نیمه‌نظارتی در چنین شرایطی، چارچوبی مبتنی بر pseudo-labeling طراحی شد تا داده‌های بدون برچسب را در فرآیند آموزش مدل ادغام کند و به بخش بدون‌نظارت وزن مناسبی در تابع خطا اختصاص دهد. برای ارزیابی، از مجموعه‌داده AG News شامل ۱۲۰،۰۰۰ نمونه آموزشی و ۷،۶۰۰ نمونه آزمایشی استفاده شد که از میان داده‌های آموزشی، ۲۰٪ (۲۴،۰۰۰ نمونه) به‌عنوان داده برچسب‌خورده و ۸۰٪ (۹۶،۰۰۰ نمونه) به‌عنوان داده بدون‌برچسب به کار رفت. استخراج ویژگی‌ها با مدل BERT-base انجام شد که بردارهای ۷۶۸ بعدی تولید می‌کند. نتایج بر اساس سنجه‌های صحت، دقت، فراخوانی و معیار-F1 نشان داد که روش نیمه‌نظارتی در مقایسه با ماشین بردار پشتیبان نظارتی، بهبود اندک اما معناداری در عملکرد ارائه می‌دهد. این یافته‌ها نشان می‌دهد که داده‌های بدون‌برچسب می‌توانند به‌طور مؤثر در بهبود مدل‌های یادگیری ماشین در شرایط کم‌داده به‌کار گرفته شوند.

کلیدواژه: **یادگیری نیمه‌نظارت‌شده، پردازش زبان طبیعی، یادگیری معنایی، S۳VM، SVM.**

۱- مقدمه

یادگیری ماشین به‌عنوان یکی از حوزه‌های کلیدی هوش مصنوعی، در سال‌های اخیر پیشرفت چشمگیری داشته و در بسیاری از مسائل واقعی، از جمله پردازش زبان طبیعی (NLP) و بینایی ماشین، نقشی اساسی ایفا کرده است [۱] و [۲]. با این حال، بیشتر الگوریتم‌های یادگیری ماشین برای دستیابی به عملکرد مطلوب، به حجم زیادی از داده‌های برچسب‌خورده نیاز دارند؛ داده‌هایی که گردآوری و برچسب‌گذاری آن‌ها فرآیندی پرهزینه و زمان‌بر است [۳]. در چنین شرایطی، یادگیری نیمه‌نظارتی به‌عنوان رویکردی میان یادگیری نظارتی و بدون‌نظارت مطرح می‌شود که با استفاده هم‌زمان از داده‌های برچسب‌خورده و بدون‌برچسب، می‌تواند عملکرد مدل را بهبود بخشد [۴] و [۵]. یکی از مدل‌های شناخته‌شده در این حوزه، ماشین بردار پشتیبان نیمه‌نظارتی (S۳VM) است که نخستین‌بار در ۲۰۰۳ م. توسط ژو معرفی شد [۶]. هر چند مطالعات اولیه اثربخشی این روش را نشان داده‌اند، اما بررسی جامع آن در کنار مدل‌های زبانی عمیق همچون BERT-base، RoBERTa و DeBERTa هنوز محدود است [۷] تا [۹]. این مدل‌ها با تکیه بر معماری ترانسفورمری، قابلیت استخراج تعبیه‌های متنی غنی و معنایی را دارند و می‌توانند به‌طور بالقوه دقت روش‌های نیمه‌نظارتی را افزایش دهند .در مقابل، ماشین بردار پشتیبان (SVM) به‌عنوان یک روش نظارتی کلاسیک، همچنان در بسیاری از مسائل دسته‌بندی متن، به‌ویژه در شرایط داده‌ی محدود، عملکردی قابل‌قبول ارائه می‌دهد [۱۰]. مقایسه‌ی مستقیم S۳VM و SVM، به‌ویژه در شرایطی که از یک چارچوب استخراج ویژگی مشترک مانند BERT-base استفاده می‌شود، می‌تواند دیدگاهی دقیق‌تر درباره‌ی توانایی یادگیری نیمه‌نظارتی در تعمیم الگوهای پنهان ارائه دهد.

پژوهش‌های اخیر [۱۱] و [۱۲] نشان داده‌اند که ترکیب یادگیری نیمه‌نظارتی با تعبیه‌های زبانی مبتنی بر معماری ترانسفورمری، در شرایط کمبود داده‌ی برچسب‌خورده منجر به بهبود معنادار عملکرد مدل‌ها می‌شود. بر همین اساس، نوآوری پژوهش حاضر در تمرکز بر کاربرد یادگیری نیمه‌نظارتی S۳VM برای دسته‌بندی دقیق موارد آزمون در شرایط محدودیت داده‌های برچسب‌خورده است. هدف اصلی این پژوهش، مقایسه‌ی نظام‌مند کارایی مدل‌های S۳VM و SVM با استفاده از ویژگی‌های استخراج‌شده از BERT-base بر روی مجموعه‌داده‌ی AG News است تا میزان تأثیر استفاده از داده‌های بدون‌برچسب بر عملکرد کلی مدل‌ها به‌طور تجربی سنجیده شود.

۲- مرور ادبیات

طبقه‌بندی متن یکی از حوزه‌های مهم پردازش زبان طبیعی[1] (NLP) است که در سال‌های اخیر با گسترش روش‌های یادگیری ماشین و یادگیری عمیق پیشرفت چشمگیری داشته است. این مسئله به‌طور گسترده در کاربردهایی مانند تحلیل احساسات، دسته‌بندی اخبار و فیلتر محتوای نامناسب مورد استفاده قرار می‌گیرد. دو رویکرد اصلی در این حوزه، یادگیری نظارت‌شده و یادگیری نیمه‌نظارت‌شده هستند که هر یک مزایا و محدودیت‌های خاص خود را دارند. در ادامه، ابتدا به معرفی روش‌های یادگیری ماشین در طبقه‌بندی متن، سپس نقش استخراج ویژگی با استفاده از BERT، و در نهایت مرور مطالعات پیشین پرداخته می‌شود.

در سال‌های اخیر، الگوریتم‌هایی مانند ماشین بردار پشتیبان (SVM)،

این مقاله در تاریخ ۲۰ آبان ماه ۱۴۰۴ دریافت و در تاریخ ۳ بهمن ماه ۱۴۰۴ بازنگری شد.

محمدحسین پروانه، گروه مهندسی کامپیوتر، واحد بین‌المللی اروند، دانشگاه آزاد اسلامی، آبادان، ایران، (email: ma.parvaneh@iau.ac.ir).

مریم نورائی آباده (نویسنده مسئول)، گروه مهندسی کامپیوتر، واحد بین‌المللی اروند، دانشگاه آزاد اسلامی، آبادان، ایران، (email: ma.nooraei@iau.ac.ir).

1. Natural Language Processing

شبکه‌های عصبی مصنوعی[1] (ANN) و مدل‌های یادگیری عمیق مانند BERT به‌طور گسترده برای طبقه‌بندی متن به‌کار گرفته شده‌اند [۱] تا [۴]. الگوریتم SVM به دلیل توانایی بالا در تفکیک داده‌های متنی و پایداری در برابر داده‌های با ابعاد بالا، یکی از پرکاربردترین روش‌های نظارت‌شده به شمار می‌رود [۱۳]. با این حال، عملکرد این مدل به میزان داده‌های برچسب‌خورده وابسته است و در شرایط کمبود داده‌های برچسب‌دار، دقت آن کاهش می‌یابد [۳]. یادگیری نیمه‌نظارت‌شده راهکاری برای رفع این محدودیت است که با ترکیب داده‌های برچسب‌خورده و بدون برچسب، عملکرد مدل را بهبود می‌دهد [۵]، [۱۴] و [۱۵]. یکی از روش‌های برجسته در این رویکرد، S۳VM است که با بهره‌گیری از داده‌های بدون برچسب، مرز تصمیم را به گونه‌ای تنظیم می‌کند که تعمیم‌پذیری مدل افزایش یابد [۴]. تحقیقات نشان داده‌اند که این رویکرد در شرایط داده محدود می‌تواند نسبت به SVM مزیت قابل توجهی داشته باشد.

۲-۱ کاربرد ویژگی‌های استخراج‌شده از BERT در طبقه‌بندی متن

یکی از چالش‌های اساسی در طبقه‌بندی متن، نمایش بهینه ویژگی‌ها است. در گذشته، روش‌هایی مانند Word2Vec و GloVe برای استخراج ویژگی‌ها به کار می‌رفتند، اما با ظهور مدل‌های یادگیری عمیق، این رویکردها به‌طور قابل توجهی ارتقا یافتند [۱۶] و [۱۷]. مدل BERT که بر پایه معماری ترنسفورمر طراحی شده است [۱]، با یادگیری پیش‌متناظر[2] بر روی حجم عظیمی از داده‌های متنی و سپس ریزتنظیم[3] برای وظایف خاص، توانایی بالایی در استخراج روابط معنایی کلمات و جملات دارد [۱۸]. مطالعات اخیر نشان داده‌اند که ترکیب ویژگی‌های استخراج‌شده از BERT با مدل‌هایی مانند SVM و S۳VM می‌تواند دقت و پایداری طبقه‌بندی را به شکل معناداری افزایش دهد، به‌ویژه در مجموعه‌داده‌هایی که دارای داده‌های برچسب‌خورده محدود هستند [۱۹] تا [۲۱].

۲-۲ مطالعات پیشین و مقایسه نتایج

تحقیقات متعددی به مقایسه عملکرد مدل‌های نظارت‌شده و نیمه‌نظارت‌شده در طبقه‌بندی متن پرداخته‌اند. به‌عنوان مثال، ژو و همکاران نشان دادند که S۳VM در بسیاری از مسائل طبقه‌بندی متنی عملکرد بهتری نسبت به SVM ارائه می‌دهد [۶] و [۲۲]. همچنین، یانگ و همکاران [۲۱] با استفاده از ویژگی‌های استخراج‌شده از شبکه‌های عصبی عمیق، اثبات کردند که ترکیب BERT و روش‌های نیمه‌نظارت‌شده می‌تواند باعث بهبود قابل توجه دقت مدل شود [۱۹]. چن و همکاران [۱۹] نیز در بررسی مجموعه‌داده‌های خبری دریافتند که استفاده از یادگیری نیمه‌نظارت‌شده، نسبت به یادگیری نظارت‌شده، دقت مدل‌ها را به‌طور معناداری افزایش می‌دهد [۱۱]. بر اساس این مطالعات و پژوهش‌های مشابه [۵]، [۶]، [۱۲] و [۲۳]، انتظار می‌رود ترکیب روش S۳VM با ویژگی‌های استخراج‌شده از BERT نسبت به مدل‌های کاملاً نظارت‌شده مانند SVM عملکرد بهتری داشته باشد، به‌ویژه در سناریوهایی که داده‌های برچسب‌دار محدود هستند [۱۴] و [۲۲].

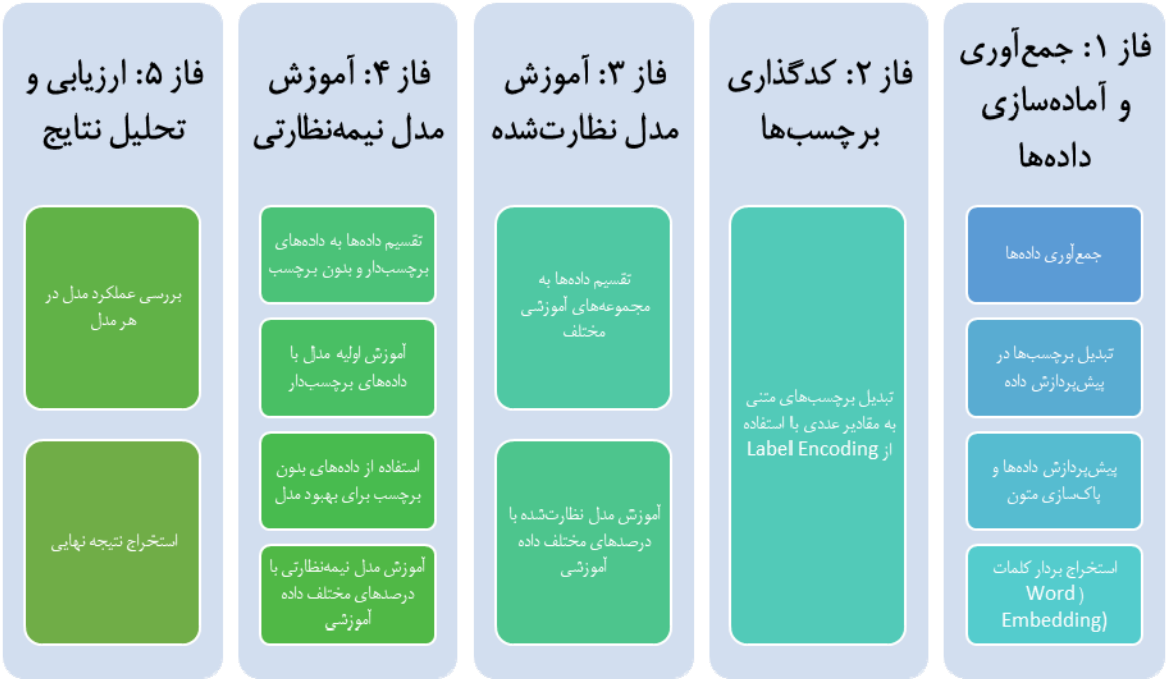
۳- مدل پیشنهادی

یک مدل پیشنهادی این پژوهش با هدف بررسی و مقایسه عملکرد دو رویکرد متفاوت یادگیری ماشین، یعنی یادگیری نظارت‌شده و نیمه‌نظارت‌شده، در مسئله طبقه‌بندی متون خبری طراحی و پیاده‌سازی شده است. نوآوری اصلی این مدل در ترکیب یک استخراج‌کننده ویژگی قدرتمند مبتنی بر معماری BERT با دو الگوریتم طبقه‌بندی SVM (نسخه نظارت‌شده) و S۳VM (نسخه نیمه‌نظارت‌شده) نهفته است. ایده کلی این است که با استفاده از BERT، ویژگی‌های متنی به بردارهایی با ابعاد بالا و معنای غنی تبدیل شوند و سپس این بردارها در اختیار دو نوع طبقه‌بند قرار گیرند تا عملکرد آن‌ها در شرایط مختلف حجم داده‌های برچسب‌خورده بررسی شود .BERT یک مدل قدرتمند پردازش زبان طبیعی است که می‌تواند روابط معنایی و نحوی میان کلمات را با دقت بالا درک کند [۱]. برخلاف روش‌های سنتی مانند TF-IDF یا Word2Vec که تنها وابستگی‌های محلی را در نظر می‌گیرند، BERT از مکانیسم توجه چندسر[4] استفاده کرده و متن را به‌صورت دوسویه تحلیل می‌کند [۱] و [۱۶]. در این مرحله، متن‌های ورودی ابتدا به توکن‌های عددی تبدیل می‌شوند و سپس به مدل BERT داده می‌شوند [۱]. این مدل با استفاده از لایه‌های ترنسفورمر، معنای هر کلمه را بر اساس بافت جمله استخراج کرده و برای هر جمله یک بردار عددی چندبعدی تولید می‌کند. این بردارها ویژگی‌های زبانی متون را در یک فضای برداری نمایش می‌دهند و به عنوان ورودی مدل‌های یادگیری ماشین مورد استفاده قرار می‌گیرند. استفاده از BERT به‌جای روش‌های سنتی موجب افزایش درک مدل از مفاهیم پیچیده زبانی شده و دقت دسته‌بندی را بهبود می‌بخشد [۱۹]. پس از استخراج ویژگی‌های متنی، داده‌های عددی به دو مدل مختلف یادگیری ماشین داده می‌شوند:

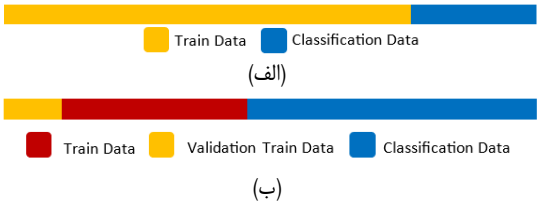
- ماشین بردار پشتیبان: این مدل که از یادگیری نظارت‌شده استفاده می‌کند، تنها از داده‌های برچسب‌گذاری‌شده برای یادگیری و دسته‌بندی نمونه‌های جدید استفاده می‌کند [۱۳]. SV سعی می‌کند مرز تصمیم بهینه‌ای را بین کلاس‌های مختلف داده ایجاد کند.
- ماشین بردار پشتیبان نیمه‌نظارتی:این مدل نسخه‌ای پیشرفته‌تر از SVM است که علاوه بر داده‌های برچسب‌دار، از داده‌های بدون برچسب نیز برای بهبود عملکرد دسته‌بندی استفاده می‌کند. S۳VM تلاش می‌کند تا مرز تصمیم‌گیری را به شکلی تنظیم کند که نه‌تنها داده‌های برچسب‌دار، بلکه داده‌های بدون برچسب نیز درست طبقه‌بندی شوند [۷].

برای مقایسه عملکرد این دو مدل، مجموعه‌ای از داده‌های آزمون با تعداد محدودی نمونه برچسب‌گذاری‌شده در نظر گرفته شده است. نتایج نشان داده‌اند که S۳VM نسبت به SVM عملکرد بهتری داشته است، زیرا با استفاده از داده‌های بدون برچسب توانسته است مرز دسته‌بندی را بهینه‌تر تنظیم کند و دقت کلی را افزایش دهد [۴] و [۶]. یکی از چالش‌های اصلی در دسته‌بندی موارد آزمون، کمبود داده‌های برچسب‌گذاری‌شده است [۳]. برچسب‌گذاری دستی حجم بالایی از داده‌ها نیازمند زمان و هزینه زیادی است. برای رفع این مشکل، مدل S۳VM به شکلی طراحی شده است که بتواند از داده‌های بدون برچسب نیز برای یادگیری استفاده کند [۵] . در این روش، داده‌های بدون برچسب ابتدا وارد مدل می‌شوند و مدل تلاش می‌کند تا آن‌ها را در یکی از دسته‌های از

1. Artificial Neural Network
2. Pre-Training
3. Fine-Tuning
4. Multi-Head Attention



شکل ۱: معماری روش پیشنهادی.



شکل ۲: تفاوت شیوه بخش‌بندی داده‌ها در دو روش (الف) نظارت‌شده و (ب) نیمه‌نظارت‌شده.

پیش‌تعریف‌شده قرار دهد [۵]. فرض اصلی S۳VM این است که داده‌های بدون برچسب اطلاعات ارزشمندی درباره ساختار کلی داده‌ها ارائه می‌دهند و می‌توانند به یافتن مرز تصمیم بهینه‌تر کمک کنند [۷] و [۱۵]. این فرایند باعث می‌شود که مدل در مواجهه با نمونه‌های جدید، عملکرد بهتری داشته باشد و وابستگی آن به داده‌های برچسب‌دار کاهش یابد [۲۱].

برای ارزیابی مدل‌ها، از سه معیار صحت[1]، F۱-score و ضریب کاپا[2] استفاده شده است. نتایج نشان داد که مدل S۳VM به دلیل بهره‌گیری از داده‌های بدون برچسب، در تمامی این معیارها عملکرد بهتری نسبت به SVM داشته است. در شکل ۱ روندنمای روش پیشنهادی نشان داده شده است.

۳-۱ داده‌های مورد استفاده

در این پژوهش از مجموعه داده AG_News که شامل چهار دسته خبری (ورزشی، سیاسی، اقتصادی و علمی) است، استفاده شده است این مجموعه شامل ۱۲۰,۰۰۰ نمونه آموزشی و ۷,۶۰۰ نمونه تست می‌باشد. تقسیم‌بندی داده‌ها به صورت زیر انجام شده است:

- تعداد۲۰٪ از داده‌های برچسب‌خورده برای آموزش مدل‌های نظارت‌شده و نیمه‌نظارت‌شده
- تعداد ۸۰٪ از داده‌های بدون برچسب که فقط در روش نیمه‌نظارت‌شده استفاده شده است.

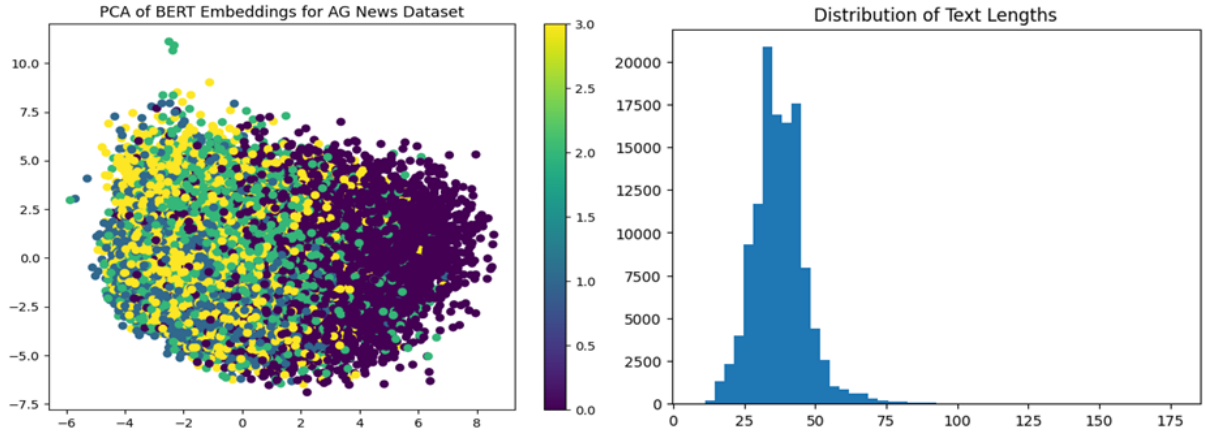
در شکل ۲، تفاوت شیوه بخش‌بندی داده‌ها در دو روش نظارت‌شده و نیمه‌نظارت‌شده نشان داده شده است.

۳-۲ پیش‌پردازش داده‌ها

پس از انجام مراحل اولیه پاک‌سازی متن، از مدل BERT برای استخراج ویژگی‌های معنایی و نحوی استفاده شده است. BERT یک مدل از پیش‌آموزش‌دیده‌شده مبتنی بر شبکه‌های عصبی ترنسفورمری است که قادر به درک عمیق‌تر متن در مقایسه با روش‌های سنتی مانند TF-IDF یا Word2Vec است [۱۷]. این مدل با استفاده از نمایش دوسویه، معنای هر کلمه را در بافت جمله در نظر می‌گیرد که باعث افزایش دقت در پردازش زبان طبیعی می‌شود [۱۷]. در این مرحله، هر جمله ورودی ابتدا به توکن‌های عددی تبدیل شده و پس از آن به مدل BERT داده می‌شود [۱]. BERT با استفاده از لایه‌های ترنسفورمر چندسری، اطلاعات وابستگی کلمات را استخراج کرده و یک بردار عددی پرمعنا برای هر جمله تولید می‌کند. این بردارها نمایانگر ویژگی‌های زبانی متن بوده و برای مدل‌های یادگیری ماشین یا شبکه‌های عصبی قابل استفاده هستند [۱۹]. در نهایت، بردارهای استخراج‌شده به‌عنوان ورودی مدل دسته‌بندی استفاده می‌شوند. این ویژگی‌ها که شامل اطلاعات نحوی، معنایی و حتی روابط بین کلمات در متن هستند، دقت دسته‌بندی موارد آزمون را بهبود می‌بخشند. استفاده از BERT به‌جای روش‌های سنتی، موجب افزایش درک مدل از مفاهیم پیچیده زبانی شده و وابستگی به داده‌های برچسب‌گذاری‌شده را کاهش می‌دهد، که در یادگیری نیمه‌نظارتی اهمیت ویژه‌ای دارد [۱۷]. شکل ۳ دو نمایش از مجموعه‌داده AG_News را نشان می‌دهد.

هر نمونه، بعد از پردازش، به یک بردار با ابعاد ۷۶۸ تبدیل شده است که برای مدل‌های یادگیری ماشین قابل استفاده است .مقدار ۷۶۸ در واقع ابعاد embedding پیش‌فرض BERT-base است و این هایپرپارامتر هیچگونه تغییری در این تحقیق نداشته است.

1. Accuracy
2. Cohen's Kappa



(ب) (الف)

شکل ۳: پیش‌پردازش داده‌ها، (الف) PCA تعبیه‌سازی BERT روی مجموعه‌داده و (ب) توزیع طول متن.

۳-۳ مدل یادگیری نیمه‌نظارت‌شده

در این پژوهش، برای بهره‌گیری از داده‌های بدون برچسب در فرآیند دسته‌بندی، از ماشین بردار پشتیبان نیمه‌نظارتی (S۳VM) استفاده شده است [۴]. این مدل نسخه‌ای توسعه‌یافته از ماشین بردار پشتیبان (SVM) است که علاوه بر داده‌های برچسب‌گذاری‌شده، می‌تواند از داده‌های بدون برچسب نیز برای بهبود دقت دسته‌بندی استفاده کند [۷]. در این روش، برای استفاده از داده‌های بدون برچسب، از الگوریتم شبه‌برچسب‌زنی[1] بهره گرفته شده است. در شبه‌برچسب‌زنی، ابتدا مدل با داده‌های برچسب‌خورده آموزش اولیه می‌بیند. سپس این مدل اولیه برای پیش‌بینی برچسب داده‌های بدون برچسب به‌کار می‌رود، و پیش‌بینی‌های با اعتماد بالا به‌عنوان شبه‌برچسب[2] در نظر گرفته می‌شوند. این نمونه‌های دارای برچسب شبهی سپس به داده‌های آموزشی افزوده شده و در تکرارهای بعدی آموزش مجدداً مدل را بهبود می‌دهند، به‌طوری که مدل بتواند ساختار پنهان داده‌ها را بهتر بیاموزد و مرز تصمیم‌گیری دقیق‌تری بیابد [۲۴] و [۲۵].

در یادگیری نظارتی، مدل فقط از نمونه‌های دارای برچسب یاد می‌گیرد، اما S۳VM با افزودن نمونه‌های دارای شبه‌برچسب، تلاش می‌کند تا مرز تصمیم بهینه‌ای ایجاد کند که داده‌های مشاهده‌نشده را نیز به درستی طبقه‌بندی کند [۷]. S۳VM با این فرض کار می‌کند که داده‌های بدون برچسب به‌ویژه پس از برچسب‌گذاری با شبه‌برچسب می‌توانند اطلاعات ارزشمندی درباره ساختار توزیع داده‌ها ارائه دهند [۱۵]. این مدل سعی می‌کند فاصله بین کلاس‌ها را حداکثر کند، در حالی که هم‌زمان از داده‌های بدون برچسب برای تنظیم بهتر مرز تصمیم‌گیری بهره می‌برد [۴]. به این ترتیب، نقاط بدون برچسب که به‌طور طبیعی در فضاهای خالی بین کلاس‌ها قرار دارند، نقش کلیدی در تعیین مرز دسته‌بندی ایفا می‌کنند. این ویژگی باعث می‌شود مدل نه‌تنها برای داده‌های دارای برچسب، بلکه برای داده‌های جدید و برچسب‌گذاری‌نشده نیز تعمیم‌پذیری بالاتری داشته باشد.

استفاده از S۳VM در دسته‌بندی موارد آزمون، به مدل کمک می‌کند تا از حداقل داده‌های برچسب‌دار حداکثر بهره را ببرد [۲۱] و [۲۶]. این روش به‌ویژه در شرایطی که برچسب‌گذاری دستی داده‌ها زمان‌بر و پرهزینه است، عملکرد مناسبی دارد. به دلیل استفاده از داده‌های بدون برچسب و شبه برچسب، مدل می‌تواند به درک بهتری از ساختار توزیع داده‌ها دست یابد و در نهایت دقت بالاتری را نسبت به روش‌های کاملاً نظارتی ارائه دهد مدل ماشین بردار پشتیبان نیمه‌نظارت‌شده برای بهره‌گیری از داده‌های بدون برچسب استفاده شده است [۴]. در این مدل، تابع هزینه به‌صورت زیر تعریف می‌شود

$$\min_{w,b}\left(\frac{1}{2}\|W\|^2\right)+C\sum_{i=1}^{l}\max\left(0,1-y_i f(X_i)\right)+ C_u\sum_{j=l+1}^{l+u}\phi\left(f(X_j)\right) \quad (1)$$

که در آن l تعداد داده‌های برچسب‌خورده، u تعداد داده‌های بدون برچسب، C و C_u ضرایب تنظیم‌کننده برای داده‌های برچسب‌خورده و بدون برچسب و $\phi\left(f(X_j)\right)$ تابعی است که میزان اطمینان مدل را نسبت به نمونه‌های بدون برچسب تنظیم می‌کند.

٤- تحلیل و ارزیابی

٤-١ مدل مقایسه‌شده

مدل ماشین بردار پشتیبان (SVM) به عنوان یکی از مدل‌های رایج و مؤثر در یادگیری نظارت‌شده در این پژوهش استفاده شده است. SVM یک الگوریتم طبقه‌بندی است که تلاش می‌کند تا به صورت یک هایپرپلین (ابرصفحه) برای تفکیک داده‌ها بین کلاس‌های مختلف عمل کند. این مدل با تمرکز بر حداکثر کردن حاشیه (Margin) بین داده‌های مختلف در فضای ویژگی‌ها، قادر به انجام پیش‌بینی‌های دقیق حتی در مواردی با داده‌های پیچیده و غیرخطی است. در این پژوهش، SVM به عنوان معیار اصلی برای مقایسه با مدل‌های نیمه‌نظارتی قرار گرفته و عملکرد آن در شرایط مختلف داده‌های برچسب‌خورده بررسی شده است. تابع تصمیم SVM به‌صورت زیر تعریف می‌شود

$$f(x)=\sum_{i=1}^{n}a_i y_i K(X_i,X)+b \quad (2)$$

که در آن a_i ضرایب لاگرانژ، y_i برچسب داده‌های آموزشی، b مقدار بایاس مدل و $K(X_i,X)$ تابع کرنل است که در این پژوهش از کرنل خطی و کرنل RBF استفاده شده است.

1. Pseudo-Labeling
2. Pseudo-Label

٤-۲ معیارهای ارزیابی عملکرد

برای مقایسه عملکرد مدل‌های یادگیری نظارت‌شده و نیمه‌نظارت‌شده، از معیارهای زیر استفاده شده است.

صحّت

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

که در آن، TP تعداد نمونه‌های درست مثبت، TN تعداد نمونه‌های درست منفی، FP نمونه‌های منفی که اشتباه مثبت تشخیص داده شده‌اند و FN نمونه‌های مثبت که اشتباه منفی تشخیص داده شده‌اند را نشان می‌دهند.

میانگین دقت وزنی[1]

$$F1-Score = 2 \times \frac{Precision + Recall}{Precision \times Recall} \quad (۴)$$

که دقّت[2] و بازخوانی[3] به صورت زیر تعریف می‌شوند

$$Precision = \frac{TP}{TP + FP} \quad (۵)$$

$$Recall = \frac{TP}{TP + FN} \quad (۶)$$

معیار کاپا

$$Kappa = \frac{P_. - P_e}{1 - P_e} \quad (۷)$$

که در آن، $P_.$ میزان توافق مشاهده‌شده و P_e میزان توافق تصادفی است.

٤-۳ تنظیمات آزمایش و اجرای مدل‌ها

برای بهینه‌سازی عملکرد مدل‌ها و انتخاب مقادیر بهینه ابرپارامترها، از روش جستجوی شبکه‌ای[4] استفاده شد. در این روش، برای هر ابرپارامتر مانند C در مدل SVM (که میزان جریمه اشتباه طبقه‌بندی را تعیین می‌کند) و gamma در کرنلRBF، یک بازه معقول از مقادیر ممکن تعریف می‌شود. به‌طور مشابه، در مدل S۳VM لاوه بر C، پارامتر Cu که اهمیت داده‌های بدون برچسب را در تابع هزینه تنظیم می‌کند نیز در جستجوی شبکه‌ای لحاظ گردید. در فرآیند جستجوی شبکه‌ای، برای هر ترکیب ممکن از این مقادیر تعریف‌شده، مدل آموزش داده می‌شود و عملکرد آن با استفاده از معیارهای ارزیابی تعیین‌شده) مانند صحّت یا امتیاز-F۱ سنجیده می‌شود. سپس بهترین ترکیب، یعنی آن مجموعه از مقادیر که بالاترین عملکرد را نشان داده است، به‌عنوان پارامترهای نهایی انتخاب می‌شود. این رویکرد باعث می‌شود که انتخاب پارامترها بر اساس ارزیابی سیستماتیک و کمی انجام شود، نه صرفاً انتخاب دستی یا پیش‌فرض، و در نتیجه به بهبود دقت و پایداری مدل‌ها کمک می‌کند. آزمایش‌ها روی مجموعه داده‌های AG_News انجام شده است [۲۷]که شامل اخبار دسته‌بندی‌شده در چهار گروه مختلف است و به طور معمول در کارهای مربوط به پردازش زبان طبیعی استفاده می‌شود. برای ارزیابی پایدارتر عملکرد مدل‌ها، آزمایش‌ها ۵ بار با اجرای مجدد انجام شده است، که این کار به کاهش تاثیر تصادفی و افزایش صحّت نتایج کمک می‌کند. این روش‌های ارزیابی دقیق و تنظیم پارامترها به پژوهش اعتبار علمی بیشتری می‌بخشند.

٤-٤ یافته‌ها

نتایج تجربی نشان می‌دهد که مدل نیمه‌نظارت‌شده S۳VM نسبت به مدل نظارت‌شده SVM عملکرد بهتری از خود نشان می‌دهد، به‌ویژه در شرایطی که تنها درصد کمی از داده‌ها برچسب‌خورده هستند (مانند استفاده از ۲۰٪ داده‌های برچسب‌خورده در این پژوهش). مدل S۳VM با بهره‌گیری از داده‌های بدون برچسب، قادر است الگوهای پنهان در داده را بهتر شناسایی کرده و از ساختار درونی داده‌ها برای بهبود مرز تصمیم بهره گیرد. این مزیت به‌ویژه در معیارهایی مانند صحّت، میانگین دقت وزنی و معیار کاپا قابل مشاهده است، جایی که در اغلب آزمایش‌ها مقادیر بهتری نسبت به مدل SVM ثبت کرده است. بنابراین می‌توان نتیجه گرفت که در مسائل با داده‌های محدود، بهره‌گیری از رویکردهای نیمه‌نظارت‌شده می‌تواند تأثیر خوبی در بهبود عملکرد مدل یادگیری ماشین داشته باشد.

در ادامه، نتایج به‌دست‌آمده از مقایسه مدل‌های یادگیری ماشین در زمینه طبقه‌بندی داده‌های AG_News ارائه می‌شود. این نتایج شامل مقایسه دقت، کارایی و میزان بهبود مدل‌های S۳VM نیمه‌نظارتی) و SVM (نظارت‌شده) در شرایط مختلف داده‌های برچسب‌خورده و بدون برچسب است.

٤-٤-۱ مقایسه عملکرد مدل‌ها

نتایج آزمایش‌ها نشان داد که در شرایطی که تنها ۲۰٪ از داده‌ها دارای برچسب بودند، مدل نیمه‌نظارتی S۳VM عملکردی نسبتاً بهتر اما نزدیک به مدل نظارتی SVM ارائه داد. هرچند انتظار می‌رفت با بهره‌گیری از داده‌های بدون برچسب، بهبود بیشتری در معیارهایی چون امتیاز-F۱ و بازخوانی نسبت به مدل صرفاً نظارتی حاصل شود ، اما این افزایش محسوس نبود.

به‌طور خاص، مدل نیمه‌نظارتی S۳VM همان‌طور که انتظار می‌رفت، با افزایش پیوسته داده‌های بدون برچسب (و کاهش داده‌های برچسب‌خورده) به تدریج عملکرد بهتر و برتری بیشتری از خود نشان داد، اما میزان برتری آن نسبت به مدل نظارتی SVM همچنان اندک بود.

یکی از دلایل این رفتار می‌تواند به ساختار پیچیده و توزیع غیریکسان داده‌ها بازگردد که ممکن است فرآیند پویای یادگیری نیمه‌نظارتی را محدود کرده باشد. همچنین، استفاده از ویژگی‌های استخراج‌شده توسط مدل زبانی BERT علی‌رغم مزیت‌های آن در نمایش معنایی داده‌ها منجر به ایجاد بردارهای ویژگی بسیار غنی و متراکم شده است که در صورت نبود تنظیمات دقیق، می‌تواند تأثیر مثبت داده‌های بدون برچسب را خنثی کند. علاوه بر این، انتخاب بهینه پارامترهای حساس در مدل‌های نیمه‌نظارتی (نظیر وزن‌دهی به نمونه‌های بدون برچسب در تابع هزینه) نقش مهمی در عملکرد دارد؛ در غیر این صورت، ممکن است عملکرد مدل نیمه‌نظارتی با مدل صرفاً نظارتی هم‌سطح شود.

٤-٤-۲ تأثیر تعداد داده‌های برچسب‌خورده

آزمایش‌ها روی مجموعه داده AG_News نشان داد که هرچه درصد داده‌های برچسب‌خورده افزایش یابد، مدل نظارتی SVM عملکرد بهتری دارد. اما در حجم‌های کم داده‌های برچسب‌خورده، مدل نیمه‌نظارتی S۳VM تا حدی توانست از داده‌های بدون برچسب برای بهبود عملکرد

1. Weighted F1-Score
2. Precision
3. Recall
4. Grid Search

جدول ۱: مقایسه عملکرد روش‌های نیمه‌نظارتی و نظارت‌شده بر اساس معیارها و کرنل‌های مختلف.

نظارت‌شده										
هسته	هسته تابع پایه شعاعی (RBF Kernel)					هسته خطی (Linear Kernel)				
درصد داده‌های در دسترس	۲۰	۲۵	۳۰	۳۵	۴۰	۲۰	۲۵	۳۰	۳۵	۴۰
معیار ها ↓										
صحّت	۰/۹۰۲	۰/۹۰۷	۰/۹۰۸	۰/۹۱۱	۰/۹۱۳	۰/۸۹۱	۰/۹۰۱	۰/۹۰۲	۰/۹۰۳	۰/۹۰۵
بازخوانی	۰/۹۰۲	۰/۹۰۶	۰/۹۰۸	۰/۹۱۰	۰/۹۱۳	۰/۸۹۰	۰/۹۳۰	۰/۹۰۲	۰/۹۰۳	۰/۹۰۴
امتیاز F۱	۰/۹۰۳	۰/۹۰۷	۰/۹۰۸	۰/۹۱۱	۰/۹۱۲	۰/۸۹۲	۰/۹۰۱	۰/۹۰۱	۰/۹۰۳	۰/۹۰۵
SVM										
صحّت	۰/۸۸۰۵		۰/۸۸۵۶		۰/۸۹۵۶	۰/۹۰۲۵			۰/۹۰۶۳	
بازخوانی	۰/۸۸۰۸		۰/۸۸۵۹		۰/۸۹۵۹	۰/۹۰۲۷			۰/۹۰۲۶	
امتیاز F۱	۰/۸۸۰۵		۰/۸۸۵۶		۰/۸۹۵۶	۰/۹۰۲۵			۰/۹۰۶۳	
SVM										

نیمه‌نظارت‌شده										
هسته	هسته تابع پایه شعاعی (RBF Kernel)					هسته خطی (Linear Kernel)				
درصد داده‌های در دسترس	۲۰	۲۵	۳۰	۳۵	۴۰	۲۰	۲۵	۳۰	۳۵	۴۰
معیار ها ↓										
صحّت	۰/۸۷۲	۰/۸۹۴	۰/۹۰۲	۰/۹۱۲	۰/۹۲۳	۰/۸۵۱	۰/۸۷۳	۰/۸۷۸	۰/۸۸۴	۰/۸۸۶
بازخوانی	۰/۸۷۴	۰/۸۹۴	۰/۹۰۳	۰/۹۱۲	۰/۹۲۳	۰/۸۵۵	۰/۸۷۳	۰/۸۷۸	۰/۸۸۴	۰/۸۸۶
امتیاز F۱	۰/۸۷۲	۰/۸۹۴	۰/۹۰۲	۰/۹۱۲	۰/۹۲۰	۰/۸۵۱	۰/۸۷۳	۰/۸۷۲	۰/۸۸۴	۰/۸۸۷
گسترش برچسب Label Spreading										
صحّت	۰/۷۷۰	۰/۷۷۹	۰/۷۹۳	۰/۸۰۲	۰/۸۱۰	۰/۸۵۴	۰/۸۶۴	۰/۸۷۱	۰/۸۷۶	۰/۸۸
بازخوانی	۰/۶۹۳	۰/۷۱۲	۰/۷۴۰	۰/۷۷۳	۰/۷۸۶	۰/۸۵۵	۰/۸۶۴	۰/۸۷۲	۰/۸۷۶	۰/۸۸
امتیاز F۱	۰/۶۵۴	۰/۷۰۷	۰/۷۲۸	۰/۷۶۷	۰/۷۷۹	۰/۸۵۴	۰/۸۶۴	۰/۸۷۱	۰/۸۷۶	۰/۸۸
انتشار برچسب Label Propagation										
صحّت	۰/۸۸۵۲		۰/۸۹۰۲		۰/۸۹۲۴	۰/۸۹۹۶			۰/۹۰۰۸	
بازخوانی	۰/۸۸۴۹		۰/۸۸۹۷		۰/۸۹۲۴	۰/۸۹۲۶			۰/۹۰۰۴	
امتیاز F۱	۰/۸۸۵۲		۰/۸۹۰۲		۰/۸۹۲۲	۰/۸۹۹۶			۰/۹۰۰۸	
S3VM										

بهره ببرد، هرچند که میزان این بهبود کمتر از مقدار مورد انتظار بود.

٤-٤-٣ تحلیل میزان تأثیر ویژگی‌های استخراج‌شده

از آنجا که در این پژوهش به جای Word2Vec از BERT برای استخراج ویژگی‌های متنی استفاده شده است، تحلیل نتایج نشان می‌دهد که تأثیر ویژگی‌های استخراج‌شده بر عملکرد مدل‌های نظارتی و نیمه‌نظارتی بسیار مهم بوده است. ویژگی‌های حاصل از BERT باعث افزایش دقت کلی هر دو مدل شدند، اما تأثیر آن بر بهبود S۳VM نسبت به SVM محدودتر بود. در شرایطی که هر دو مدل ۲۰، ۲۵، ۳۰، ۳۵ و ۴۰ درصدِ داده‌ها را در اختیار داشتند، نتایج عملکرد آنها در جدول زیر ثبت شده است.

اما همانطور که در جدول ۱ قابل مشاهده است می‌توان برتری نسبی یادگیری نیمه نظارت شده در دسته بندی موارد آزمون دید. وقتی درصد کمی از داده‌ها برچسب دارند SVM فقط از این مقدار کم یاد می‌گیرد. اما گسترش برچسب[1] یا S۳VM از ساختار توزیع کل داده‌ها استفاده کرده و مرز تصمیم بهتری می‌سازد، حتی اگر برچسب‌ها کم باشند.

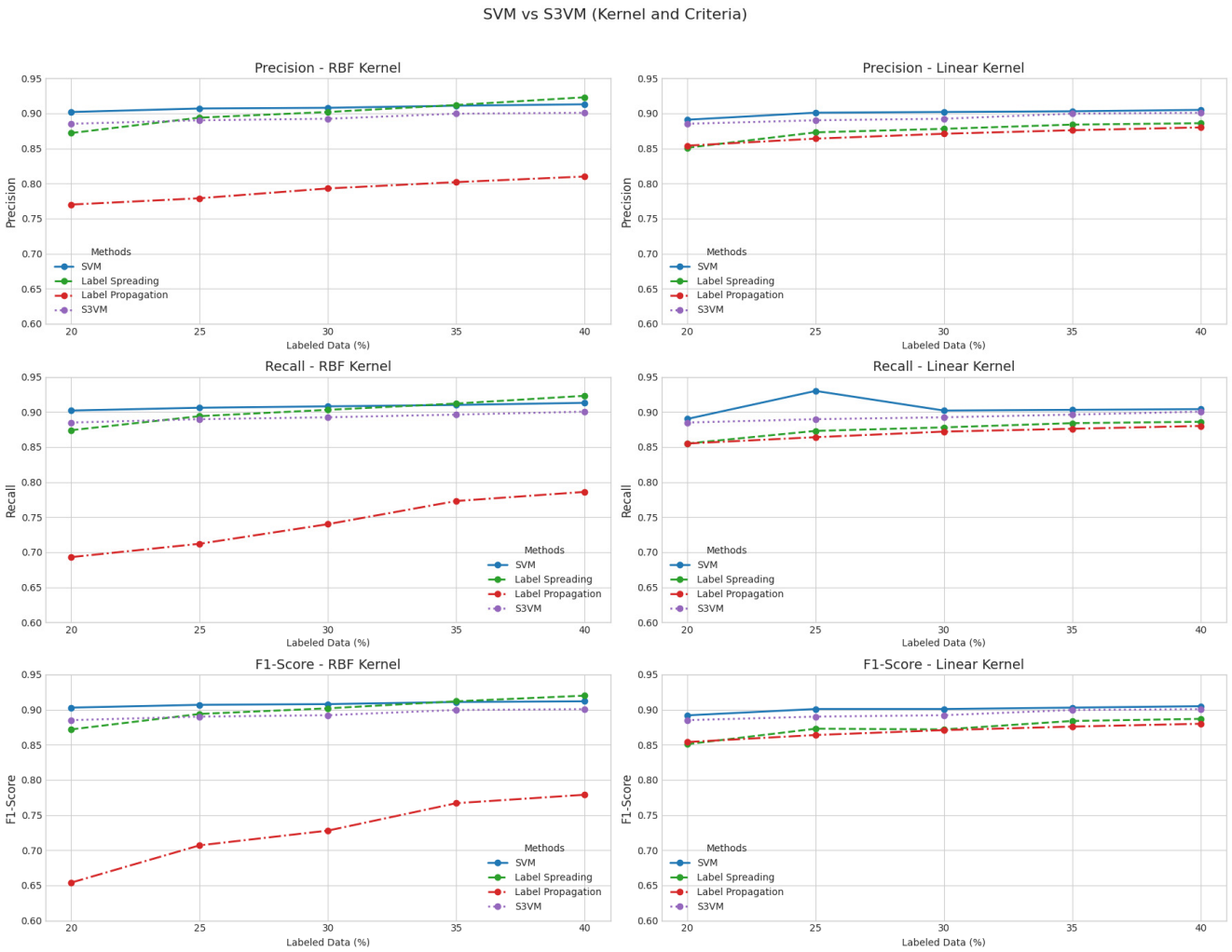
شکل ۴، مقایسه‌ای جامع بین روش‌های نظارتی و نیمه‌نظارتی در دو کرنل مختلف RBF و خطی را برای سه معیار اصلی یادگیری ماشین (صحّت، بازیابی و امتیاز-F۱) ارائه می‌دهد. در کرنل RBF، روش‌های نیمه‌نظارتی مانند گسترش برچسب و S۳VM عملکردی بسیار نزدیک به SVM دارند و حتی در برخی نقاط با داده‌های کمتر، عملکرد بالاتری را نشان می‌دهند، که بیانگر توانایی آن‌ها در بهره‌برداری بهتر از داده‌های بدون برچسب است. در کرنل خطی نیز تفاوت عملکرد روش‌ها کمتر است، اما باز هم روش‌های نیمه‌نظارتی خصوصاً انتشار برچسب[2] در معیار بازیابی، عملکرد پایدارتری نسبت به SVM نشان می‌دهد. این تحلیل نشان می‌دهد که استفاده از روش‌های نیمه‌نظارتی، به‌ویژه در شرایطی با داده‌های برچسب‌خورده محدود، می‌تواند منجر به بهبود دقت مدل شود، به‌ویژه در ترکیب با کرنل‌های غیرخطی مانند RBF.

۵- بحث

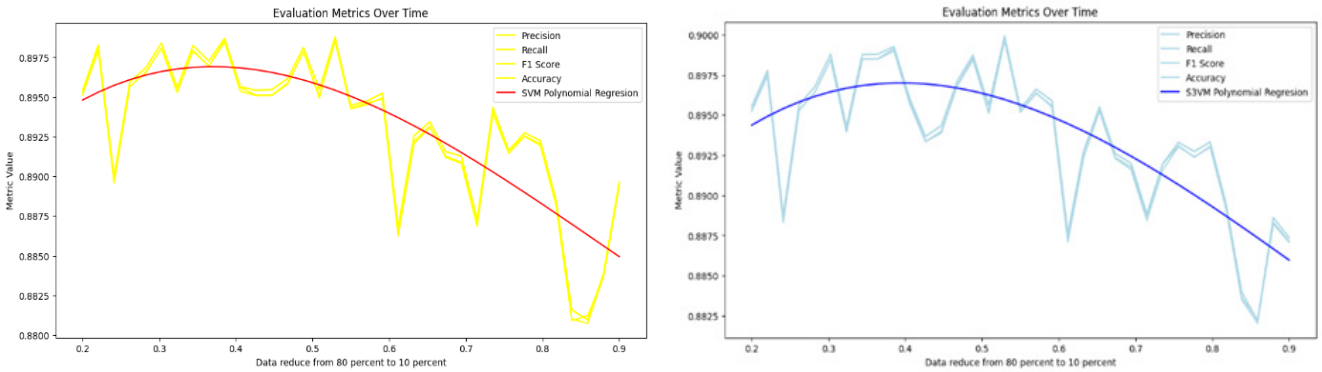
نتایج این پژوهش نشان داد که مدل نیمه‌نظارتی S۳VM در شرایطی که تنها ۲۰٪ از داده‌ها برچسب‌خورده است، عملکردی بهتر اما نزدیک به مدل نظارتی SVM داشت. برخلاف انتظار اولیه، تفاوت عملکرد این دو مدل اختلاف زیادی نداشت. این یافته با برخی مطالعات پیشین که مزایای

1. Label Spreading

2. Label Propagation



شکل ۴: مقایسه کارایی بر اساس معیارها.

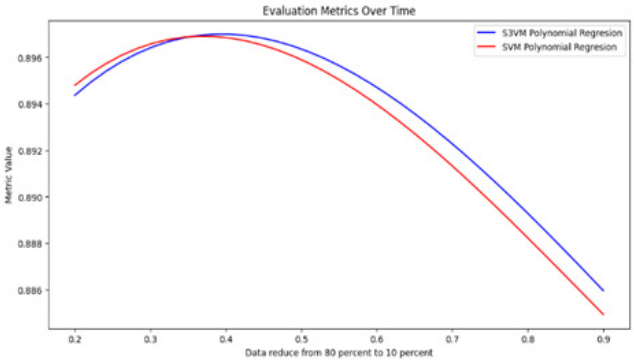


شکل ۵: خروجی هردو مدل به‌صورت جداگانه.

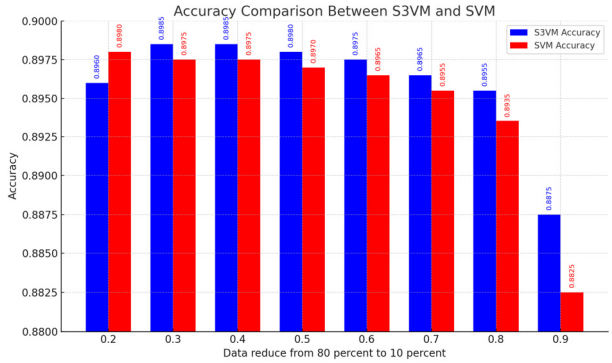
یادگیری نیمه‌نظارتی را در شرایط کمبود داده‌های برچسب‌خورده برجسته کرده‌اند، در تناقض است. برای مثال، مطالعاتی که در حوزه پردازش زبان طبیعی (NLP) انجام شده‌اند، نشان داده‌اند که در بسیاری از موارد، مدل‌های نیمه‌نظارتی می‌توانند با بهره‌گیری از داده‌های بدون برچسب، دقت بالاتری نسبت به مدل‌های کاملاً نظارتی کسب کنند. با این حال، در این پژوهش، مشاهده شد که این مزیت در مدل S۳VM چندان آشکار نبود.

در این تحقیق برای بررسی دقیق‌تر و درک بهتر تفاوت عملکرد دو مدل، آزمون را توسعه دادیم زیرا تنها بررسی عملکرد با حجم داده ۲۰ درصدی دید محدودی به ما می‌دهد. از این رو تصمیم گرفته شد که مدل را در یک حلقه تکرار قرار دهیم و حجم داده ها را در هر مرحله از تکرار کاهش دهیم به این صورت که مدل در هر مرحله دسترسی کمتری به داده ها دارد. در مرحله اول ۹۰ درصد از داده ها در اختیار مدل است و در مرحله آخر تنها ۱۰ درصد از داده ها باقی مانده‌اند و در پایان هر مرحله نیز عملکرد مدل ثبت می‌شود. نمایش صحّت عملکرد مدل‌ها در شکل ۵ مشاهده می‌شود.

در شکل ۶ به وضوح تفاوت عملکردی دو مدل قابل تشخیص است. در ابتدا، زمانی که حجم داده بیشتری در اختیار مدل‌ها قرار داشت، مدل مبتنی بر یادگیری نظارتی که با رنگ قرمز ثبت شده است عملکرد بهتری از خود نشان داد. اما با کاهش حجم داده، مدل نیمه‌نظارتی S۳VM افت



شکل ۶: نمایش هر دو عملکرد در یک پلات جهت مقایسه.



شکل ۷: نمایش تفاوت عملکرد در نمودار میله‌ای.

عملکرد کمتری را تجربه کرد و به برتری نسبی دست یافت. این نتیجه تا حدود زیادی پیش بینی ما در ابتدای تحقیق را تایید می‌کند. شکل ۷ عملکرد مدل‌های نظارتی و نیمه‌نظارتی را در چندین اجرا متوالی و بر اساس معیارهای مختلف (صحّت، بازیابی و امتیاز-F۱) مقایسه می‌کند. نوسانات در منحنی‌ها، میزان پایداری هر روش را نشان می‌دهد و مشخص می‌سازد که روش‌های نیمه‌نظارتی مانند S۳VM و گسترش برچسب در حضور داده‌های برچسب‌خورده محدود، نه تنها عملکرد رقابتی دارند بلکه در برخی اجراها از مدل نظارتی SVM نیز فراتر می‌روند، که حاکی از توانایی آن‌ها در تعمیم الگوهای نهفته در داده‌ها است.

بر اساس شکل ۷ می‌توان تفاوت عملکرد دو مدل مذکور را در هر مرحله از تکرار آموزش مقایسه کرد. عملکرد دو مدل S۳VM (نیمه‌نظارتی) و SVM (نظارتی) را در شرایط کاهش تدریجی داده‌های آموزشی از ۸۰٪ تا ۱۰٪ نمایش می‌دهد.همان‌گونه که مشاهده می‌شود مدل S۳VM در اکثر سطوح کاهش داده، دقت بالاتری نسبت به SVM ارائه داده است. این موضوع نشان می‌دهد که الگوریتم‌های نیمه‌نظارتی توانایی بهره‌گیری مؤثر از داده‌های بدون برچسب را دارند و در سناریوهای داده‌ی محدود عملکرد بهتری ارائه می‌دهند. در مقابل، عملکرد SVM با کاهش داده‌ها به‌شدت افت می‌کند که تأییدکننده‌ی نیاز شدید این الگوریتم به داده‌های برچسب‌خورده است. این پژوهش دارای چندین نقطه قوت مهم است:

- مقایسه‌ی سیستماتیک: یکی از مهم‌ترین نقاط قوت، مقایسه‌ی دقیق بین روش‌های نظارتی و نیمه‌نظارتی در شرایطی است که داده‌های برچسب‌خورده محدود هستند. این مقایسه، درک بهتری از کارایی روش‌ها در سناریوهای واقعی فراهم می‌سازد.
- تحلیل آماری: استفاده از آزمون‌های آماری برای بررسی معناداری تفاوت عملکرد مدل‌ها موجب شد نتایج پژوهش تنها به مقایسه‌ی سطحی محدود نشود، بلکه از نظر آماری نیز اعتبار یابد.

با این حال، پژوهش حاضر با محدودیت‌هایی همراه است که مسیرهای آتی را روشن می‌سازد:

- در فرآیند استخراج ویژگی از BERT استفاده شد، اما سایر روش‌های مبتنی بر شبکه‌های عصبی مانند (GPT-embeddings) مورد ارزیابی قرار نگرفتند.
- تنها چند الگوریتم نیمه‌نظارتی پایه مانند S۳VM بررسی شدند و روش‌های ترکیبی یا پیشرفته‌تر مانند الگوریتم‌های مبتنی بر گراف (انتشار برچسب) یا Self-training لحاظ نشدند.
- پژوهش به مجموعه‌داده‌ی AG_News محدود شد و تعمیم‌پذیری نتایج به زبان‌ها یا دامنه‌های دیگر بررسی نشد. استفاده از چندین مجموعه‌داده متنوع می‌تواند به تعمیم بهتر کمک کند.

٦- نتیجه گیری

در این پژوهش با بهره‌گیری از یادگیری نیمه‌نظارتی، تلاش شد نقش داده‌های بدون برچسب در بهبود عملکرد مدل‌های دسته‌بندی، به‌ویژه در شرایطی که حجم داده‌ها یا میزان داده‌های برچسب‌خورده اندک است، مورد بررسی قرار گیرد. نتایج نشان داد که روش‌های S۳VM و گسترش برچسب با کرنل RBF در این شرایط عملکردی بهتر و نزدیک به روش‌های نظارتی مانند SVM ارائه دادند. در مقابل، روش انتشار برچسب با کرنل RBF عملکرد ضعیف‌تری نشان داد که بیانگر حساسیت آن به کمبود داده‌های برچسب‌خورده است. با این حال، استفاده از کرنل‌های غیرخطی مانند RBF موجب افزایش دقت و پایداری روش‌های نیمه‌نظارتی شد.

یافته‌ها نشان می‌دهد که بهره‌گیری هوشمندانه از داده‌های بدون برچسب می‌تواند ضمن کاهش هزینه‌های برچسب‌گذاری، کیفیت مدل‌های یادگیری ماشین را در شرایط محدودیت داده بهبود بخشد؛ چراکه در بسیاری از پروژه‌ها و تحقیقات، حجم داده‌ها یا داده‌های برچسب‌خورده محدود است. از نظر کاربردی، نتایج پژوهش حاضر تأیید می‌کند که یادگیری نیمه‌نظارتی می‌تواند رویکردی ارزشمند برای سناریوهای واقعی باشد، به‌ویژه زمانی که داده‌های برچسب‌خورده کافی در دسترس نیستند. با این حال، نتایج نشان دادند که این روش‌ها در همه‌ی شرایط جایگزین کامل یادگیری نظارتی نیستند و در صورت وجود داده‌های کافی، حتی ممکن است عملکرد ضعیف‌تری نسبت به روش‌های نظارتی داشته باشند. بنابراین، مدل‌های نیمه‌نظارتی نیازمند بهینه‌سازی و ترکیب با سایر روش‌های یادگیری برای دستیابی به عملکرد پایدارتر هستند. برای تحقیقات آینده، بررسی تأثیر روش‌های مختلف استخراج ویژگی مانند Word2Vec یا FastText بر یادگیری نیمه‌نظارتی و به‌کارگیری مدل‌های نیمه‌نظارتی پیشرفته‌تر مانند الگوریتم‌های گراف‌محور یا شبه‌برچسب‌زنی پیشنهادی می‌شود. همچنین، آزمون مدل‌ها بر روی مجموعه‌داده‌های متنوع برای ارزیابی میزان تعمیم‌پذیری و بهره‌گیری از یادگیری تقویتی و مقایسه‌ی نتایج با روش‌های موجود مطلوب است.

مراجع

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc The 16th Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* vol. 1, pp. 4171-4186, 16 pp., New Orleans, LA, USA, 1-6 Jun. 2018.

[2] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. of the 28th Annual Conf.*

[19] H. Chen, W. Han, and S. Poria, "SAT: Improving semi-supervised text classification with simple instance-adaptive self-training," in *Proc. of the Findings of the Association for Computational Linguistics*, pp. 6141-6146, 2022,
[20] E. Kotei and R. J. I. Thirunavukarasu, "A systematic review of transformer-based pre-trained language models through self-supervised learning," *Information*, vol. 14, no. 3, Article ID: 187, Mar. 2023.
[21] W. Yang, R. Zhang, J. Chen, and J. Sheng, "Calibrating Pseudo-Labeling with Class Distribution for Semi-supervised Text Classification," in *Proc. of the 2025 Conf. on Empirical Methods in Natural Language Processing*, pp. 13026-13039, Suzhou, China, 4-9 Nov. 2025.
[22] I. Sirbu, R. -A. Popovici, C. Caragea, □. Trău□an-Matu, and T. Rebedea, "MultiMatch: Multihead consistency regularization matching for semi-supervised text classification," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pp. 2792-2808, Suzhou, China, 4-9 Nov. 2025.
[23] J. M. Duarte and L. A. Berton, "A review of semi-supervised learning for text classification," *Artificial Intelligence Review,* vol. 56, pp. 9401-9469, 2023.
[24] K. Sohn, *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. 34th Conf. on Neural Information Processing Systems*, pp. 596-608, Vancouver, Canada, 6-12 Dec. 2020.
[25] A. Hatefi, X. -S. Vu, M. Bhuyan, and F. Drewes, "The efficiency of pre-training with objective masking in pseudo labeling for semi-supervised text classification," 2025.
[26] S. Cheng, W. Chen, W. Liu, and H. Qu, "Improving lightweight semi-supervised text classification via teacher intervention," *Applied Soft Computing*, vol. 184, pt. B, Article ID: 113844, Dec. 2025.
[27] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. of the 18th Int. Conf. on Neural Information Processing Systems*, pp. 529-536, Vancouver, Canada, 13-16, Dec. 2004.

محمدحسین پروانه تحصیلات خود را در مقاطع کارشناسی مهندسی ماشین آلات و کارشناسی ارشد نرم‌افزار بترتیب در دانشگاه علوم و فنون و دانشگاه آزاد اسلامی واحد بین‌المللی اروند به پایان رسانده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: تست نرم‌افزار، محاسبات نرم و کاربردهای آن در مهندسی نرم‌افزار.

مریم نورائی آباده در سال ۱۳۸۰ برای تحصیل در رشته مهندسی کامپیوتر در دانشگاه اصفهان پذیرفته شد. وی در سال ۱۳۹۴مدرک دکتری سیستم‌های نرم‌افزاری را از واحد علوم و تحقیقات تهران اخذ نمود. دکتر نورائی از سال ۱۳۸۷ در دانشکده مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد بین‌المللی اروند مشغول به فعالیت گردید و اینک نیز عضو هیأت علمی این دانشکده می‌باشد. زمینه‌های علمی مورد علاقه نامبرده متنوع بوده و شامل موضوعاتی مانند ایده‌های نو در استفاده از هوش مصنوعی در مهندسی نرم‌افزار، آزمون نرم‌افزار، محاسبات نرم و کاربردهای آن در مهندسی نرم‌افزار است.

on Neural Information Processing Systems, pp. 560-567, Montreal, Canada, 8-13 Dec. 2015.
[3] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. of the 16th Int. Conf. on Machine Learning*, pp. 200-209, Bled, Slovenia, 27-30 Jun. 1999.
[4] K. Bennett and A. J. A. i. N. I. p. s. Demiriz, "Semi-supervised support vector machines," in *Proc. of the 12th Int. Conf. on Neural Information Processing Systems*, pp. 368-374, Denver, CO, USA, 30 Nov.-5 Dec. 1998.
[5] D. -H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning*, p. 896, Atlanta, GA, USA, 2013.
[6] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," *in Proc. ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, vol. 3, pp. 58-65, Washington, DC, USA, 2003
[7] O. Chapelle, B. Schölkopf, and A. Zien, "A discussion of semi-supervised learning and transduction," in *Semi-Supervised Learning*: MIT Press, 2006, pp. 473-478.
[8] Y. Liu, *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019.
[9] P. He, X. Liu, J. Gao, and W. J. a. p. a. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," 2020.
[10] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2: Short Papers, pp. 302-308, Baltimore, MD, USA, 23-24 Jun 2014,
[11] Q. Xie, M. -T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687-10698, Seattle, WA, USA, 13-19 Jun. 2020.
[12] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proc. of the 32nd Int. Conf. on Neural Information Processing Systems*, pp. 3239-3250, Montréal, Canada, 3-8 Dec. 2018.
[13] F. Pedregosa, et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
[14] D. Berthelot, *et al.*, "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. of the 33rd Int. Conf. on Neural Information Processing Systems,* pp. 5049-5059*,* Montréal, Canada, 8-14 Dec. 2019.
[15] M. N. J. A. S. E. Abadeh, "Knowledge-enhanced software refinement: leveraging reinforcement learning for search-based quality engineering," *Automated Software Engineering*, vol. 31, Article ID: 57, 2024.
[16] A. Radford, *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[17] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15-18, Florence, Italy, 28 Jul.-2 Aug. 2019.
[18] G. Quétant, P. Molchanov, and S. J. Voloshynovskiy, TwinTURBO: Semi-Supervised Fine-Tuning of Foundation Models via Mutual Information Decompositions for Downstream Task and Latent Spaces, arXiv preprint arXiv:2503.07851, 2025.