تاریخ دریافت مقاله: ۹۶/۱۰/۰۲
تاریخ پذیرش مقاله: ۹۷/۰۹/۰۴

روشی نوین برای شناسایی نویسنده متون با ترکیب الگوریتم‌های بهینه‌سازی توده ذرات و ماشین بردار پشتیبان

فرهاد سلیمانیان قره چپق*
استادیار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: farhad@iaurmia.ac.ir

محسن موتمن فر
کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی،ارومیه، ایران
پست الکترونیکی: motaman.mohsen@yahoo.com

مهدی وفادار
کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: v.mehdi1364@gmail.com

چکیده

تشخیص هوشمند نویسنده متون در زمینه‌های جرم‌شناسی دارای کاربردهای فراوانی می‌باشد اما به دلیل این‌که عوامل بسیاری در تشخیص نویسنده متون دخیل می‌باشند، نمی‌توان به‌طور دقیق و مطمئن نویسندگان متون را تشخیص داد. از این رو تحقیقات بسیاری در این زمینه از قرن نوزدهم صورت گرفته است اما تاکنون روشی با دقت ۱۰۰ درصد برای تمامی متون ارائه نشده است و محققان هر روزه روش‌های جدیدی را ارایه می‌دهند تا به دقت ۱۰۰ درصد نزدیک بشوند. در این مقاله برای حل این مسئله از ترکیب الگوریتم بهینه‌سازی توده ذرات و الگوریتم ماشین بردار پشتیبان استفاده شده است که از الگوریتم بهینه‌سازی توده ذرات به‌عنوان الگوریتم استخراج‌کننده ویژگی و از الگوریتم ماشین بردار پشتیبان به‌عنوان تشخیص‌دهنده نویسندگان استفاده شده است. ارزیابی برروی مجموعه داده Reuter_50_50 انجام شده است. در روش پیشنهادی پس از استخراج ویژگی‌ها، داده‌ها با نسبت ۸۰ به ۲۰ به داده‌های آموزشی و آزمایشی تقسیم می‌شوند. مجموعه داده‌های آموزشی به‌عنوان ورودی به الگوریتم ماشین بردار پشتیبان داده می‌شوند و پس از آموزش دیدن توسط ماشین بردار پشتیبان، مدل مناسب ساخته می‌شود و سپس مجموعه داده‌های آزمایشی براساس مدل ساخته شده، اعتبارسنجی می‌شوند. تقسیم‌بندی داده‌ها در حالت ۸۰ (آموزش) درصد و ۲۰ (آزمایشی) درصد بهترین نتیجه را داشته است. نتایج برمبنای معیارهای دقت و فراخوانی نشان می‌دهد که دقت تشخیص مدل پیشنهادی در مقایسه با مدل ماشین بردار پشتیبان بیشتر است.

واژه‌های کلیدی: الگوریتم بهینه‌سازی توده ذرات، الگوریتم ماشین بردار پشتیبان، تشخیص نویسنده متون، استخراج ویژگی

* نویسندهٔ مسئول

۱- مقدمه

باگسترش روزافزون متون و داده‌های الکترونیکی در فضای مجـازی، محققان برای جلوگیری از برخی سـوء استفاده‌ها در محیط‌های مجازی، به دنبال ایجاد ابزارهای امنیتی بیشـتری برای کاربران در جوامع مجازی هستند. بر همین اسـاس به‌منظور جلوگیری از دزدی‌های ادبی و علمی، محققان بسیاری به شناسایی افراد بر اساس سبک نگارش آن‌ها پرداخته‌اند. ایده شناسایی نویسنده، از مبحث طبقه‌بنـدی متون که خـود زیرشـاخه‌ای از علم پردازش زبان‌های طبیعی اسـت گرفته شده است[۱] و در آن سعی می‌شود تا با تجزیه و تحلیل واژگان، دستور زبان، مفهوم یـک جمله و با کمـک گرفتن از دانش مربـوط به واژگان، معنـای آن جملـه برای ماشـین قابل درک شـود [۲]. بر اساس فرضیه‌های مطرح شده توسط محققان، هر شخص الگوی خاصی در نحوه نگارش متون دارد که این موضوع به‌گونه‌ای همانند اثر انگشـت هر نویسنده‌ای عمل می‌کند. شناسـایی نویسنده از روی نثر، سبک و شیوه نوشتاری، یـا به‌عبـارت دیگر، ویژگی‌هـای نهفته در متون نوشـته شـده توسط نویسـندگان، یکی از مباحث جدید در زمینه هوش‌مصنوعی و پردازش زبان طبیعی به‌شمار می‌رود.

برای حل مسـئله شناسایی نویسـنده متون باید ابتدا ویژگی‌هـای نهفته داخل متن اسـتخراج گردند و سـپس براسـاس ویژگی‌های اسـتخراج شـده و بااسـتفاده از روش‌های موجود در حوزه یادگیری ماشین به شناسایی نویسـنده متون پرداخته می‌شـود. میزان دقت درمرحله اسـتخراج ویژگی‌ها تاثیر بسـزایی در نحوه تشـخیص نویسـنده دارد به‌طـوری که انتخاب درسـت ویژگی‌های اسـتخراج شـده باعث کاهش مدت زمان اجرای الگوریتم و افزایـش میزان بهینگی الگوریتم می‌گـردد. به‌طوری که اگر ویژگی‌های اضافی‌تر در تشـخیص نویسنده استفاده گردد به‌جای بهبود دقت تشـخیص، باعث کاهش دقت در تشخیص می‌گردد[۳].

در این مقاله برای تشخیص نویسنده متون از یک مدل ترکیبـی جدید برمبنای الگوریتم بهینه‌سـازی توده ذرات [۱۸] و ماشـین بردار پشتیبان [۱۹] استفاده شده است. در روش پیشنهادی از الگوریتم بهینه‌سازی توده ذرات برای انتخاب ویژگی و از ماشین بردار پشتیبان برای طبقه‌بندی اسـتفاده می‌شـود. انتخاب ویژگی موجب کاهش سربار فضا و زمان بـرای پیاده‌سـازی الگوریتم‌های طبقه‌بندی مانند ماشـین بردار پشتیبان خواهد شد. برای ساخت یک روش طبقه‌بندی مناسب نیاز به داده‌های آموزشی با کیفیت اسـت. این کیفیت با تعداد داده‌ها و ویژگی‌ها در مجموعه آمـوزش ارتباط دارد. انتخـاب ویژگی‌ها به عنوان یکی از روش‌های پیش‌پردازش داده می‌تواند باعث افزایش کیفیت مجموعه داده آموزش برای سـاخت مدل طبقه‌بندی گردد. انتخاب ویژگی‌ها دارای مزایای متعددی است. نخست زمان یادگیری در الگوریتم‌های طبقه‌بندی کاهش می‌یابد. و دیگر ایـن که با انتخاب ویژگی‌ها نیاز کمتـری به اندازه‌گیری و ذخیره‌سازی مقادیر ویژگی‌ها است.

هدف از انتخاب ویژگی، جسـتجوی یـک مجموعه از ویژگی‌ها است که عملکرد طبقه‌بندی با روش‌های بانظارت مانند ماشـین بردار پشـتیبان را افزایش می‌دهد. در واقع باید ابعاد ویژگی کاهش یابد که هدف از این کار به‌دسـت آوردن بهترین و کوچکترین زیرمجموعه از ویژگی‌ها است. در روش انتخـاب ویژگی‌ها، ویژگی‌هایی که در طبقه‌بندی مفید هستند، از یک مجموعه کامل ویژگی انتخاب می‌شوند. انتخاب ویژگی در تشخیص نویسنده متون اهمیت بسزائی دارد، زیـرا تعداد زیادی ویژگی وجود دارند که بسـیاری از آن‌ها یا بی‌اسـتفاده هسـتند و یا این‌که بـار اطلاعاتی چندانی ندارند و فقط به منظور انسجام و شکل‌گیری متون بیان شـده‌اند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد.

ادامه مقاله به صورت زیر سـازماندهی شده است .در بخش دوم، کارهای قبلی در مورد تشخیص نویسنده متون بررسی شـده است. در بخش سـوم ویژگی‌های استفاده

شـــده در تشخیص نویســـنده متون، تحلیل شده است، در بخش چهارم روند طراحی و پیاده‌سازی روش پیشنهادی برای تشخیص نویسنده متون بیان شده است و در بخش پنجم به بررسی نتایج حاصله از روش پیشنهادی خواهیم پرداخت. و نهایتا در بخش ششم به نتیجه‌گیری و کارهای آتی خواهیم پرداخت.

۲-روش‌های پیشین

مســـائل موجود در حوزه پـــردازش زبان‌های طبیعی، از مباحثـــی می‌باشـــند که قابلیت حل شـــدن با روش‌های متن‌کاوی و یادگیری ماشـــین را دارا می‌باشند. تشخیص نویســـنده متون نیز به‌عنوان یکی از مسائل مطرح در این حوزه می‌باشـــد که تاکنون برای حل این مسئله روش‌های فراوانی ارائه شده است که تعدادی از این روش‌ها را به این صورت می‌توان بیان نمود:

محققان [٤] نخســـتین بار برای تشـــخیص نویسندگان متون از نمایشنامه‌های شکســـپیر استفاده نمودند که این تحقیق به‌عنوان اولین پژوهش در زمینه تشخیص نویسنده متون در قرن ۱۹ میلادی صورت گرفت و سپس در اواسط قرن بیستم مطالعات آماری در این زمینه توسط زیف[1] در ســـال ۱۹۳۲ [٥] و یول[2] در [٦] و [۷] انجام گردید. مطالعات دقیق‌تر توســـط موســـتلر و همکارانش در ســـال ۱۹٦٤ برروی پایگاه‌داده The Federalist Paper انجام شد که پس از این تحقیق محققان ســـعی در تشخیص نویسنده توسط روش‌های غیرسنتی نمودند[۸]. محققان[۹] برای شناسایی نویسنده متون موجود در اینترنت از الگوریتم‌های: بیزین ســـاده، ماشـــین بردار پشـــتیبان، نزدیک‌ترین همسایه و طبقه‌بندی حداقل مربعات منظم اســـتفاده نمودند. اینان از وب‌نوشت‌های ۱۰۰ هزار نویسنده مختلف در سطح اینترنت به‌عنوان مجموعه‌داده استفاده نمودند.

باتوجه به تاثیر ویژگی‌های استخراج شده در شناسایی نویســـندگان متون، محققان [۱۰] برای تشخیص نویسنده متون از ویژگی‌های موجود در «تابع کلمه» استفاده نمودند که این ویژگی‌ها شـــامل حروف تعریـــف، حروف اضافه، حروف ربط، فعل و اسم می‌باشـــد. این ویژگی ها باتوجه بـــه این‌که معنای متن را تغییر نمی‌دهند با تکرار زیادی در متن اســـتفاده می‌گردند. محققان دیگری برای افزایش دقت نویســـندگان متون از ویژگی‌های معنایی استفاده نمودند. ویژگی‌هـــای معنایی به تحلیل دقیق‌تـــر متن و ویژگی‌های موجود در سطوح بالاتر می‌پردازند. این ویژگی‌ها شامل: ۱) گـــراف وابســـتگی معنایی که خود شـــامل ویژگی‌های دودویـــی معنایی و روابط اصلاح معنایی اســـت، ۲) زمان و وجـــه فعل‌های به کار برده شـــده توســـط نویســـنده و شباهت‌های معنایی بین کلمات متن و همچنین شامل کلمات و عباراتی که جمله‌واره‌ها را به هم مربوط می‌کنند می‌باشد که این‌هـــا به‌عنوان حروف یا افزوده‌های ربطی شـــناخته می‌شوند[۱۱].

بـــرکاردو و همکارانش برای شناســـایی نویســـندگان پیام‌هـــای کوتاه برخـــط از الگوریتم n گـــرام و روش‌های یادگیری بانظارت اســـتفاده نمودند[۱۲]. برای شناســـایی نویســـنده متون عربـــی الگوریتم‌های مختلفـــی مبتنی بر الگوریتم بیزین ساده ارائه شده است[۱۳]. این الگوریتم‌ها شامل: الگوریتم چندجمله‌ای، الگوریتم چندجمله‌ای برنولی و الگوریتـــم چندجمله‌ای پواســـون می‌باشـــند کـــه از این الگوریتم‌ها برای شناســـایی ۱۰ نویسنده مختلف استفاده شده است و براساس نتایج حاصله، الگوریتم چندجمله‌ای برنولـــی دارای بهتریـــن عملکرد و دقـــت ۹۷,٤۳ درصدی می‌باشـــد. محققان دیگری از ۳۳ ویژگی استخراج شده از معیارهای شـــباهت جملهٔ متنی، شـــباهت مبتنی بر دانش، شباهت مبتنی بر پیکره، شباهت تعدادی، شباهت وابستگی نحوی و معیارهای ترجمه ماشـــینی به‌عنوان ویژگی‌های موثر در تشـــخیص نویســـنده و از الگوریتم رگرســـیون بردار پشتیبان برای محاسبه میزان شباهت در میان متون استفاده نمودند[۱٤].

هویدی و همکارانش برای تولید ویژگی‌های موردنیاز

1- Zipf
2- Yule

در شناسایی نویسندگان متون از سطح N گرام کلمات و سطح N گرام نویسه‌ها و برای شناسایی نویسندگان متون از الگوریتم بیزین ساده و ماشین بردار پشتیبان استفاده نمودند [۱۵]. محققان [۱۶] از الگوریتم‌های ماشین بردار پشتیبان و دسته‌بندی فازی برای شناسایی نویسندگان متون استفاده نموده‌اند. این محققان ابتدا ویژگی‌های مربوط به هر نویسنده را استخراج نموده و در مرحله بعدی براساس ویژگی‌های استخراج شده نویسندگان متون را شناسایی نموده‌اند. براساس نتایج ارائه شده توسط این محققان دقت پیش‌بینی الگوریتم ماشین بردار پشتیبان بیشتر از الگوریتم دسته‌بندی فازی می‌باشد و در نهایت دقت الگوریتم روش ترکیبی بیشتر از الگوریتم ماشین بردار پشتیبان می‌باشد که دارای دقت ۷۶ درصدی می‌باشد. برای تشخیص نویسندگان متون ترکی پژوهشگران[۱۷] روش نوینی را ارائه داده‌اند. این پژوهشگران ابتدا از میان ۳۵ ویژگی کلی، ۲۲ ویژگی را توسط الگوریتم جستجوی رتبه‌بندی انتخاب نموده و سپس ویژگی‌های انتخاب شده را به الگوریتم‌های یادگیری ماشین ارائه نموده‌اند که براساس نتایج حاصله الگوریتم چندجمله‌ای دارای دقت ۸۰ درصدی می‌باشد.

۳-ویژگی‌های استخراج شده برای شناسایی متون

استخراج ویژگی‌های موجود در درون متون به‌عنوان اصلی‌ترین مرحله در مسئله تشخیص نویسنده متون می‌باشد. ویژگی‌های انتخاب شده جهت استخراج، تاثیر بسزایی در نحوه تشخیص نویسنده متون دارد به‌طوری که اگر ویژگی‌های استخراج شده بهینه نباشد باعث کاهش دقت روش پیشنهادی خواهد شد. در این تحقیق نیز تعدادی از ویژگی‌های موجود در متون براساس سعی و خطا انتخاب و به‌عنوان ورودی به روش‌پیشنهادی ارائه شده است که این ویژگی‌ها در جدول(۱) بیان شده است [۱۲]. ویژگی‌های موجود در جدول(۱)، برمبنای آزمایش برروی مجموعه داده‌های تشخیص نویسنده متون گردآوری و استخراج شده‌اند [۱۲]. این ویژگی‌ها در تشخیص نویسنده اسناد بسیار مهم هستند. چون بیشتر نویسندگان از دسته‌بندی‌های جدول(۱) در متون خود استفاده می‌کنند، لذا دسته‌بندی متون برمبنای آن‌ها به الگوریتم‌های طبقه‌بندی کمک می‌کنند که به دنبال ویژگی‌های اصلی باشند و برمنبای آن‌ها نوع ردهٔ متون را تشخیص دهند.

براساس جدول(۱) ویژگی‌های استخراج شده شامل ۵ دسته از ویژگی‌ها می‌باشد که هر کدام از ویژگی‌ها نیز دارای تعدادی زیرمجموعه می‌باشد. ۵ دسته اصلی شامل: ویژگی‌های لغوی، ویژگی‌های ساختاری، ویژگی‌های نحوی، کلمات دستوری، ویژگی‌های روانشناختی و زبانی می‌باشد [۱۲].

۴-روش پیشنهادی

با افزایش اهمیت مسئله تشخیص نویسنده متون در حوزه‌های مختلف، این امر به عنوان یک امر ضروری و پراهمیت در حوزه پردازش زبان‌های طبیعی مطرح می‌گرددکه نیازمند ارائه روش‌های جدیدی می‌باشد. لذا در این مقاله روش جدیدی مبتنی بر الگوریتم بهینه‌سازی توده ذرات و الگوریتم ماشین بردار پشتیبان ارائه شده است. باعنایت براین‌که ویژگی‌های استفاده شده در مجموعه‌داده تاثیر بسزایی در میزان دقت تشخیص دارند لذا در روش پیشنهادی از الگوریتم بهینه‌سازی توده ذرات به‌عنوان الگوریتم استخراج کننده ویژگی‌ها استفاده شده است.

الگوریتم بهینه‌سازی توده ذرات [۱۸] یک الگوریتم مبتنی بر جمعیت است که در آن ذرات یک ازدحام (جمعیت) را تشکیل می‌دهند. جمعیت موجود در فضای مسئله حرکت می‌کند و براساس تجربیات فردی خود و تجربیات جمعی سعی می‌کنند تا راه حل بهینه در فضای جستجو را بیابند. الگوریتم بهینه‌سازی توده ذرات به عنوان یک الگوریتم بهینه‌سازی، یک جستجوی مبتنی بر جمعیت را فراهم می‌کند که در آن هر ذره با گذشت زمان موقعیت خود را تغییر می‌دهد. در الگوریتم بهینه‌سازی توده ذرات، ذرات در یک فضای جستجوی چند بعدی از راه‌حل‌های ممکن مسئله،

جدول۱: ویژگی‌های استفاده شده برای تعیین نویسنده متون [۱۲]

دسته بندی	نوع ویژگی	توضیحات
ویژگی‌های لغوی	ویژگی‌های مبتنی بر نویسه	تعداد کل نویسه‌ها تعداد کل حروف / تعداد کل حروف کل تعداد نویسه‌های بزرگ / تعداد کل نویسه‌ها تعداد کل نویسه‌های عددی / تعداد کل نویسه‌ها تعداد کل نویسه‌های فضای خالی / تعداد کل نویسه‌ها تعداد کل فضاهای Tab / تعداد کل نویسه‌ها همه حروف نویسه‌های ویژه
	ویژگی مبتنی بر کلمه	تعداد کل کلمات تعداد کل کلمات مختصر / کل تعداد کلمات تعداد کل کلمات طویل / تعداد کل کلمات تعداد کل نویسه‌ها در کلمه / تعداد کل نویسه‌ها طول متوسط کلمات طول متوسط جملات به شکل نویسه طول متوسط جملات در قالب کلمات تعداد کل کلمات مختلف / کل تعداد کلمات تعداد کلمات که فقط یک بار در متن ظاهر میشود / تعداد کل کلمات تعداد کلمات که فقط دو بار در متن ظاهر میشود / تعداد کل کلمات
ویژگی‌های ساختاری	–	مجموع خطوط تعداد کل جمله‌ها تعداد پاراگراف‌ها تعداد جملات در پاراگراف تعداد نویسه‌ها در پاراگراف جداکننده بین پاراگراف‌ها تعداد جملات شروع شده با حروف بزرگ / تعداد عبارات تعداد جملات شروع شده با حروف کوچک / تعداد جملات میانگین تعداد کلمات در هر جمله تعداد خطوط خالی/ تعداد خطوط
ویژگی‌های نحوی	–	فراوانی علائم نشانه‌گذاری مانند و . ؟ ! : ؛ ' ( )
کلمات دستوری	–	تعداد اسناد / تعداد کل کلمات تعداد ضمایر / تعداد کل کلمات تعداد افعال کمکی / تعداد کل کلمات تعداد پیوندها تعداد قیدها / تعداد کل کلمات تعداد کلمات برای پاسخ کوتاه / تعداد کل کلمات
ویژگی‌های روانشناختی و زبانی	–	منفی احساسات مثبت احساسات منفی اضطراب خشم غمگینی بینش شک یقین محدودیت توافق

حرکت می‌کننـد. در این فضا یک معیـار ارزیابی تعریف می‌شود و سـنجش کیفیت راه‌حل‌های مسئله از طریق آن صورت می‌پذیرد. تغییـر حالت هر ذره در یک گروه تحت تاثیر تجربه‌های خود و یا دانش همسایگانش بوده و رفتار جسـتجوی یک ذره در گروه تحت تاثیر ذرات دیگر است. همین رفتار ساده باعث پیدا شدن ناحیه‌های بهینه از فضای

جســـتجو می‌گردد. بنابراین در الگوریتم بهینه‌سازی توده ذرات هـــر ذره به محض پیدا کردن موقعیت بهینه آن را به نحو مناســـبی به اطلاع سایر ذرات می‌رساند و هر ذره بر اساس مقادیر به‌دست آمده برای تابع هزینه با یک احتمال معیـــن تصمیم می‌گیـــرد تا از ذرات دیگر پیـــروی نماید و جستجو در فضای مســـئله با استفاده از دانش قبلی ذرات انجام گیرد. این عمل باعث می‌شود تمام ذرات بیش از حد به یکدیگر نزدیک نشـــوند و به طـــور موثری از عهده حل مسائل بهینه‌سازی پیوسته برآیند.

در هر تکـــرار الگوریتم، هـــر ذره موقعیت بعدی خود را بـــا توجه به دو مقـــدار تغییر می‌دهد، نخســـت بهترین موقعیتـــی که خود آن ذره تاکنون داشـــته اســـت (pbest) و دیگـــری بهتریـــن موقعیتی که تاکنون توســـط کل ذرات جمعیت به‌وجود آمده است و در واقع بهترین pbest در کل جمعیت می‌باشـــد(gbest). برمبنای مفهوم ریاضی، pbest برای هر ذره در واقع حافظه بیولوژیکی آن ذره محســـوب می‌شود. gbest همان دانش عمومی جمعیت است و وقتی کـــه افراد موقعیت خود را براســـاس gbest تغییر می‌دهند در واقع تلاش می‌کنند که ســـطح دانش خود را به ســـطح دانش جمعیت برسانند. از نظر مفهومی، بهترین ذره گروه همـــه ذرات گروه را به یکدیگر مرتبـــط می‌نماید. الگوریتم بهینه‌ســـازی توده ذرات الهام گرفته شده از حرکت دسته جمعی ذرات می‌باشـــد که در ســـال ۱۹۹۵ پیشـــنهاد شده اســـت[۱۸]. در این الگوریتم برای حل مســـئله از دو معیار ســـرعت و موقعیت ذرات استفاده می‌شود به‌طوری که در هر مرحله، ذرات براساس سرعت خاصی به سمت موقعیت بهینه حرکت می‌کنند.

در این الگوریتم برای بیان ســـرعت و موقعیت ذرات از فرمول‌های(۱)و(۲) استفاده می‌شود.

$$v_{i,j}(t+1)=wv_{i,j}(t)+c_1 r_1(t)(pbest_{i,j}(t)x_{i,j}(t))+ c_2 r_2(t)(gbest_i(t)-x_{i,j}(t)) \quad (۱)$$

در فرمول(۱)، w به‌عنوان ضریب اینرسی، ضرایب c_2 و c_1 به‌عنوان ضرایب یادگیری و ضرایب r_2 و r_1 به‌عنوان اعداد تصادفی استفاده شده است. متغیر pbest بهترین جوابی اســـت که تا به حال توسط ذره اُم پیدا شده است و متغییر gbest نیز بهترین جوابی است که توسط کل ذرات موجود در جامعه، پیشنهاد شده‌است.

$$x_i(t+1)=x_i(t)+(v_i(t+1)) \quad (۲)$$

در فرمـــول(۲) متغییر x_i بیانگر موقعیت فعلی ذره جاری و v_i بیانگر ســـرعت ذره جاری می‌باشد که توسط فرمول(۱) تولید می‌گردد.

در روش پیشـــنهادی پس از مشـــخص شده ویژگی‌ها توسط الگوریتم بهینه‌سازی توده ذرات، از الگوریتم ماشین بردار پشتیبان برای تشخیص نویسنده متون استفاده شده است. الگوریتم ماشین بردار پشتیبان در سال ۱۹۹۲ توسط واپنیک معرفی شـــده و بر پایه نظریـــه یادگیری آماری بنا گردیـــده اســـت[۱۹]. در ایـــن الگوریتم داده‌ها براســـاس بردارهای پشـــتیبان به دســـته‌های مختلفی تقســـیم‌بندی می‌شـــوند به‌طوری که با تغییر بردارهای پشـــتیبان نتایج الگوریتم تغییر پیدا می‌کند. ماشین بردار پشتیبان [۱۹] یک رده‌بند دودویی اســـت که دو رده را با استفاده از یک خط مرزی از هم جدا می‌کند. ماشین بردار پشتیبان قادر است به‌طور همزمان خطای رده‌بندی تجربی داده‌ها را کاهش و تفکیک‌پذیری رده را با اســـتفاده از تغییر شکل‌های مختلف افزایـــش دهد. این مزیت ماشـــین بردار پشـــتیبان را قادر می‌ســـازد تا با داده با ابعاد زیاد و یـــا رده‌هایی با فضای توزیـــع چندبعدی بهتر عمل کند. در تقســـیم خطی داده‌ها، هدف دست‌یابی به تابعی است که تعیین‌کننده ابرصفحه‌ای با بیشترین حاشیه می‌باشـــد. با حداکثر شدن حاشیه این ابرصفحه، تفکیک بین طبقات حداکثر می‌گردد. فرض کنید که $S=\{xi,yi\}$ یک نمونه آموزشی است که از دو رده $yi=\pm1$ و هر رده از $xi,i=1,2,..,m$ ویژگی تشکیل شده است.

در روش پیشنهادی به‌منظور نگاشت داده‌ها به‌صورت خطی از تابع پایه شـــعاعی استفاده شـــده است. تابع پایه شعاعی برای ایجاد ماشین‌هایی با انواع مختلف از سطوح غیرخطی در فضای داده‌ها، ضرب‌های داخلی تولید می‌کند [۱۹]. معمولاً تابع پایه شعاعی برای طبقه‌بندی و پیش‌بینی

عملکرد بهتری دارد. برای پیاده‌سازی این تابع از فرمول(۳) استفاده شده است.

$$k(x_i,x_j)=\exp(-\alpha\ |x_i-x_j|^2) \quad ,\alpha>0 \qquad (۳)$$

در فرمول(۳) پارامتر x_i و x_j بیانگر بردار iام و jام می‌باشد و پارامتر الفا بیانگر گاما می‌باشد که پارامتر گاما مشخص کننده شکل مرز تصمیم می‌باشد. ماشین بردار پشتیبان در مرحله آموزش، مرز تصمیم‌گیری را به گونه‌ای انتخاب می‌نماید که حداقل فاصله آن با هر یک از دسته‌های مورد نظر را بیشینه کند. این نوع انتخاب باعث می‌شود که تصمیم‌گیری در شرایط نوفه‌دار بهتر انجام شود و نمونه‌های بین دو رده با دقت بیشتری تشخیص داده شوند.

برای بیان بهتر روش پیشنهادی نحوه عملکرد این الگوریتم در شکل(۱) نمایش داده شده است.

همان‌گونه که شکل(۱) نشان می‌دهد، پس از آن‌که بهینه‌ترین ویژگی‌ها از میان ویژگی‌های استخراج شده توسط الگوریتم بهینه‌سازی توده ذرات انتخاب گردید، مجموعه‌داده تولید شده توسط ویژگی‌های استخراج شده به دو مجموعه‌داده آزمایش و آموزش تقسیم‌بندی می‌شوند که مجموعه‌داده آموزش به‌عنوان ورودی به الگوریتم ماشین بردار پشتیبان داده می‌شود و پس از آموزش دیدن ماشین بردار پشتیبان، نویسنده متون موجود در مجموعه‌داده آزمایشی براساس ماشین بردار پشتیبان آموزش دیده، تعیین می‌گردند و در نهایت براساس دو معیار ارزیابی دقت و دوباره فراخوانی مورد ارزیابی قرار می‌گیرند. انتخاب ویژگی‌ها توسط الگوریتم بهینه‌سازی توده ذرات برمبنای ویژگی‌های مطرح شده در متن اسناد مانند شمارش تعداد کلمات کلیدی انجام می‌گیرد. به هر ویژگی یک وزن اختصاص داده می‌شود و برمبنای وزن داده شده به ویژگی‌ها، ذرات جستجو در فضای جستجو را شروع می‌کنند و همسایه‌های نزدیک به نقطه انتخاب شده را جستجو می‌کنند. لذا ویژگی‌هایی انتخاب می‌شوند که نزدیکی و تشابه بیشتری به یکدیگر دارند.

۵-ارزیابی و نتایج

در این مقاله برای ارزیابی روش پیشنهادی از چهار معیار ارزیابی: تعداد متونی که نویسنده آن‌ها درست تشخیص داده شده است، تعداد متونی که نویسنده آن‌ها درست تشخیص داده نشده است، درصد دقت و درصد فراخوانی از مجموعه‌داده Reuter_50_50 به‌عنوان مجموعه‌داده آموزشی و آزمایشی با نسبت ۸۰ به ۲۰ درصد استفاده شده است. در بیشتر موارد برای ارزیابی طبقه‌بندی و پیش‌بینی نمونه‌ها از ۸۰درصد داده‌ها برای آموزش و از ۲۰درصد داده‌ها برای آزمایش استفاده می‌شود. در مرحله آموزش، مدل ساخته می‌شود؛ که مدل برمبنای قوانین و قواعد الگوریتم طبقه‌بندی و استفاده از شرط‌ها ایجاد می‌شود و سپس برمبنای مدل ساخته شده در مرحله آموزش از ۲۰درصد نمونه برای آزمایش استفاده می‌شود. نتایج حاصله از روش پیشنهادی براساس دو معیار دقت و فراخوانی [۲۰] مورد ارزیابی قرار گرفته است.

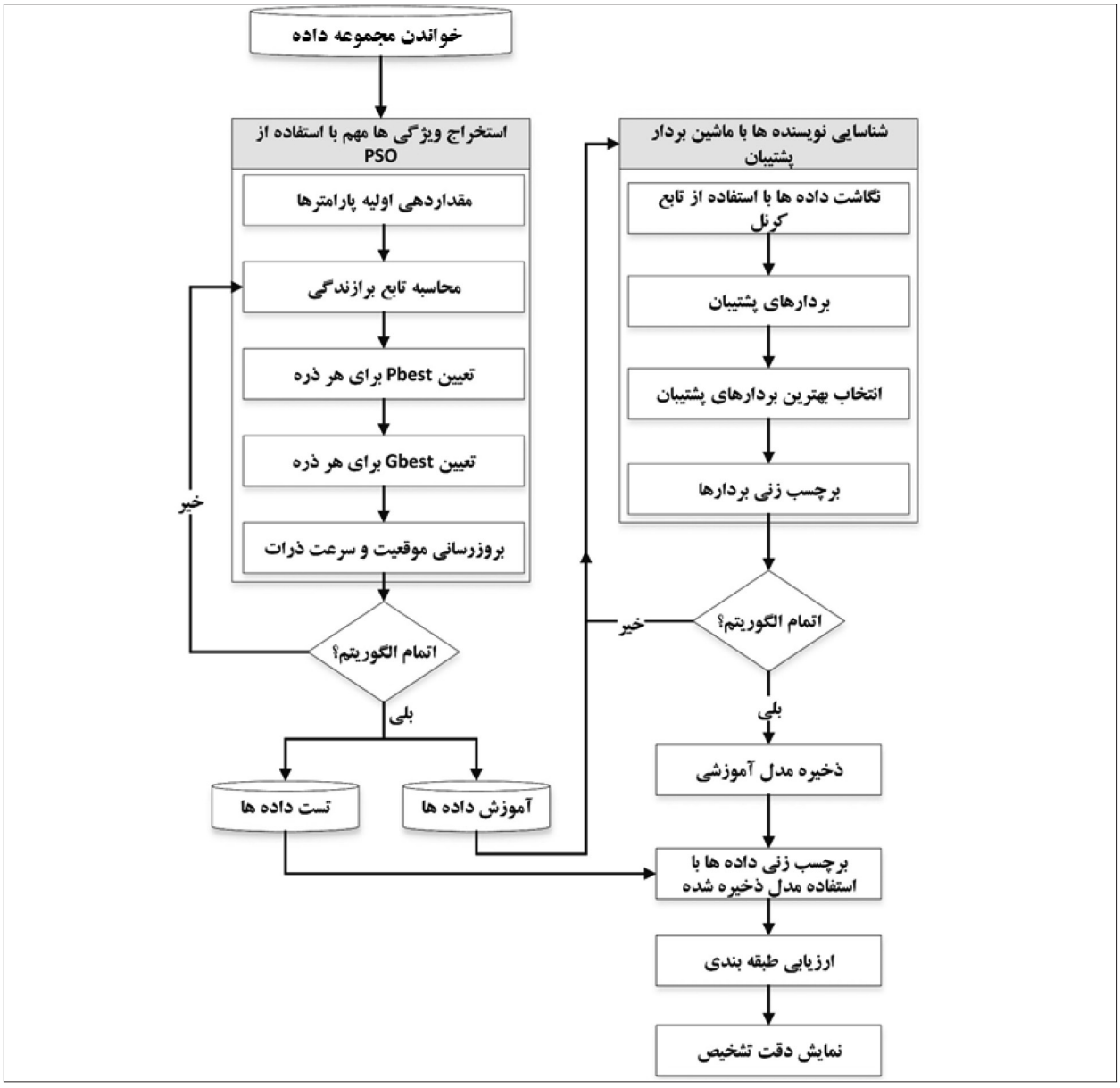
$$Precision = \frac{True\ positive}{True\ positive + False\ Positive} \qquad (٤)$$

از فرمول(٤) برای به‌دست آوردن دقت روش پیشنهادی و از فرمول(۵) برای به‌دست آوردن فراخوانی استفاده شده است.

$$Recall = \frac{True\ positive}{True\ positive + False\ Negative} \qquad (۵)$$

معیار دقت شامل پارامترهای درست مثبت(TP)و کاذب مثبت (FP) است. در این معیار نمونه‌های کاذب مثبت هم به عنوان مثبت درست در دقت تشخیص تاثیرگذار هستند. معیار فراخوانی شامل پارامترهای درست مثبت و کاذب منفی (FN) است. در این معیار نمونه‌های کاذب منفی هم در دقت تاثیرگذار هستند. نتایج حاصله از مقایسه روش پیشنهادی در اشکال(۲)، (۳) و جدول(۲) نمایش داده شده است.

براساس شکل(۲) که روش پیشنهادی براساس مجموعه‌داده آزمایشی با الگوریتم ماشین بردار پشتیبان

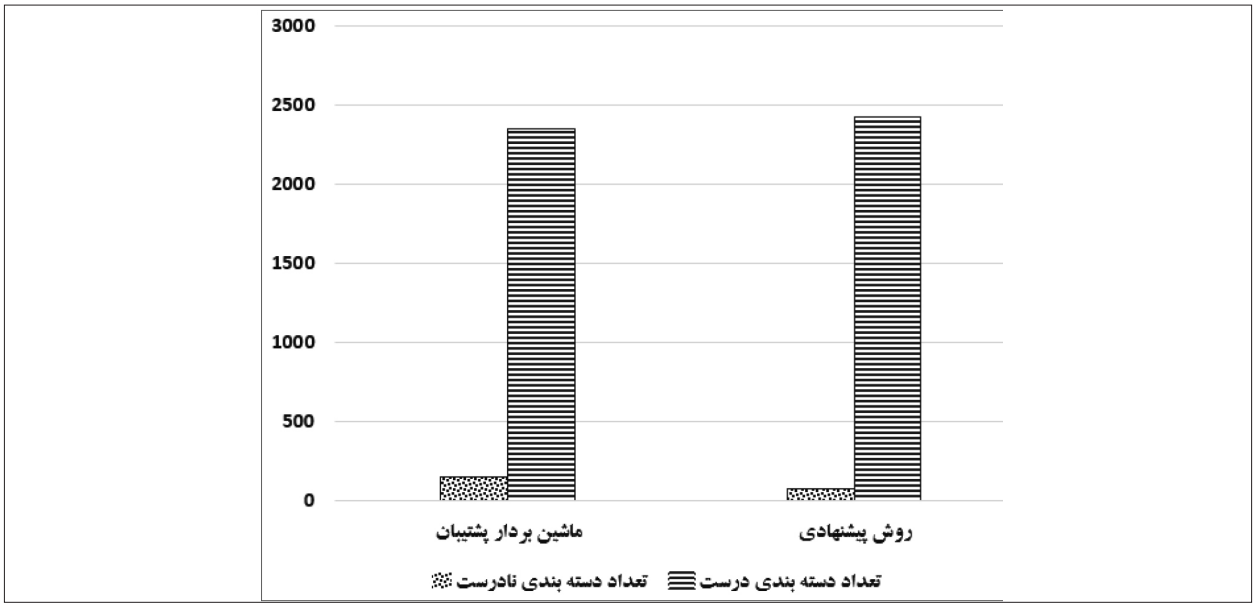


شکل۱: نحوه عملکرد روش پیشنهادی

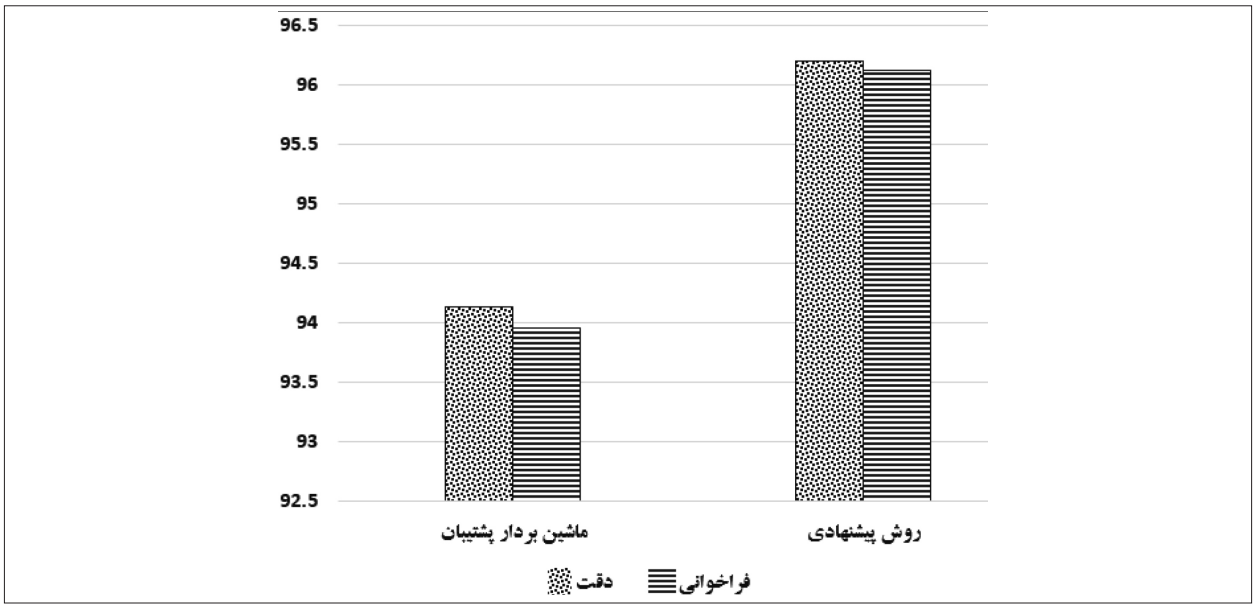
مورد ارزیابی قرار گرفته اســت، می‌تــوان گفت که روش پیشــنهادی فقط نویســنده ۷۷ متن را به‌درستی تشخیص نداده است؛ این درحالی می‌باشد که الگوریتم ماشین بردار پشتیبان نویســنده ۱۵۱ متن را به‌درستی تشخیص نداده اســت. لذا می‌توان بیان نمود که عملکرد روش پیشنهادی بهینه‌تر از الگوریتم ماشین بردار پشتیبان می‌باشد. تعداد دســته‌بندی درست در روش پیشــنهادی و ماشین بردار پشتیبان به ترتیب برابر با ۲۴۲۳ و ۲۳۴۹ می‌باشد.

در شــکل(۳)، روش پیشــنهادی براســاس دو معیار فراخوانی و دقت با الگوریتم ماشــین بردار پشتیبان مورد مقایسه قرارگرفته است. براساس این نتایج می‌توان گفت که روش پیشنهادی دارای عملکرد بهینه‌تری نسبت به ماشین بردار پشــتیبان می‌باشد. به دلیل این‌که در مدل پیشنهادی از الگوریتم بهینه‌ســازی توده ذرات برای انتخاب ویژگی استفاده شده است.

در جدول(۲) روش پیشــنهادی براســاس چهار معیار تعریف شــده با الگوریتم ماشــین بردار پشــتیبان مورد ارزیابی قرار گرفته شده است که نتایج حاصله از این جدول



شکل۲: مقایسه روش پیشنهادی با الگوریتم ماشین بردار پشتیبان



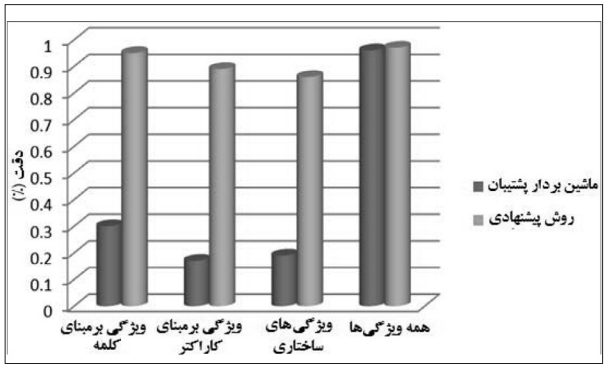
شکل۳: مقایسه روش پیشنهادی با الگوریتم ماشین بردار پشتیبان

جدول۲: مقایسه روش پیشنهادی با الگوریتم ماشین بردار پشتیبان

مدل‌ها	تعداد دسته‌بندی نادرست	تعداد دسته‌بندی درست	دقت	فراخوانی
ماشین بردار پشتیبان	۱۵۱	۲۳۴۹	۹۴,۱۳	۹۳,۹۶
روش پیشنهادی	۷۷	۲۴۲۳	۹۶,۲۰	۹۶,۱۲

بیانگر عملکرد مناسب روش پیشنهادی می‌باشد. جدول(۲)، نشان می‌دهد که مقدار فراخوانی در روش پیشنهادی برابر با ۹٦,۱۲ درصد است و در ماشین بردار پشتیبان برابر با ۹۳,۹٦ درصد است. همچنین مقدار دقت در مدل پیشنهادی و ماشین بردار پشتیبان به‌ترتیب برابر با ۹٦,۲۰ درصد و ۹٤,۱۳ درصد است.

در ایـــن تحقیق از ســـه نوع ویژگی: مبتنـــی بر کلمات، مبتنی بر نویســـه و ویژگی‌های ســـاختاری استفاده شده است. به‌منظور تعیین میزان تاثیرگذاری ویژگی‌های مطرح شـــده، هرکـــدام از ویژگی‌ها به‌صـــورت جداگانه به‌عنوان مجموعه‌داده درنظر گرفته شده‌اند که نتایج حاصله از این



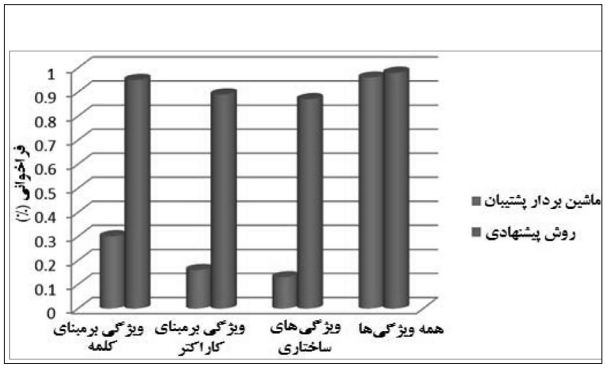
شکل۴: مقایسه روش پیشنهادی با روش‌های مطرح‌شده در مجموعه‌داده آزمایشی با تفکیک ویژگی‌ها براساس دقت

ارزیابی در اشکال (٤) و (٥) نمایش داده شده است.

در شکل(٥)، درصد معیار فراخوانی برمبنای دسته‌های مختلف نشان داده شده است. روش پیشنهادی در مقایسه با ماشین بردار پشتیبان درصد فراخوانی بیشتری دارد. مقدار فراخوانی در روش پیشنهادی برمبنای همه ویژگی‌ها نزدیک به ۱۰۰ درصد است. روش پیشنهادی در همه دسته‌ها دارای درصد فراخوانی بیشتری است و لذا در روش پیشنهادی استفاده از انتخاب ویژگی تاثیرگذار بوده است.

براساس مقادیر نشان داده شده در اشکال(٤) و (٥) می‌توان بیان نمود که ویژگی‌های مبتنی بر کلمه دارای بیشترین تاثیر در شناسایی نویسنده می‌باشند. درجدول(٣) روش پیشنهادی با روش‌های مطرح شده در بخش پیشینه تحقیق مورد بررسی قرار گرفته است. نتایج این جدول بیانگر عملکرد بهینه روش پیشنهادی می‌باشد. روش پیشنهادی با الگوریتم چندجمله‌ای برنولی [١٣]، بیزین ساده[١٥]، ماشین بردار پشتیبان [١٥]، الگوریتم چندجمله‌ای برمبنای جستجوی رتبهای [١٦] و ماشین بردار پشتیبان فازی [١٧] مقایسه شده است.

جدول(٣)، نشان می‌دهد که درصد دقت روش پیشنهادی برابر ٩٦,٠٢ است که در مقایسه با مدل‌های [١٥]، ماشین بردار پشتیبان [١٥]، الگوریتم چندجمله‌ای برمبنای جستجوی رتبه‌ای [١٦] و ماشین بردار پشتیبان فازی [١٧] دقت بیشتری دارد. همچنین درصد دقت الگوریتم چندجمله‌ای برنولی [١٣] در مقایسه با همه مدل‌ها بیشتر است. روش پیشنهادی با غلبه بر چهار مدل توانسته



شکل۵: مقایسه روش پیشنهادی با روش‌های مطرح‌شده در مجموعه‌داده آزمایشی با تفکیک ویژگی‌ها براساس فراخوانی

جدول ۳: مقایسه روش پیشنهادی با روش‌های مطرح شده در بخش پیشینه تحقیق

دقت	روش‌ها
٩٧,٠٤	الگوریتم نیوی بیز چندجمله‌ای برنولی [١٣]
٧١,٨٥	بیزین ساده [١٥]
٦٢,٩٦	ماشین بردار پشتیبان [١٥]
٨٠,٠٠	الگوریتم نیوی بیز چندجمله‌ای برمبنای جستجوی رتبه‌ای [١٦]
٧٦,٠٠	ماشین بردار پشتیبان فازی [١٧]
٩٦,٠٢	روش پیشنهادی

است دقت بیشتری داشته باشد و لذا الگوریتم مناسبی برای شناسایی نویسنده متون باشد. دلیل افزایش دقت روش پیشنهادی در مقایسه با چهار مدل دیگر استفاده از الگوریتم بهینه‌سازی توده ذرات برای انتخاب ویژگی است. به عبارتی الگوریتم بهینه‌سازی توده ذرات با انتخاب ویژگی، بهترین متون را شناسایی و انتخاب می‌کند و سپس آن‌ها را به ماشین بردار پشتیبان می‌دهد و ماشین بردار پشتیبان با داده‌های دقیق‌تر و با نوفهٔ کمتر مواجه است.

۶-نتیجه گیری

باگسترش روزافزون متون الکترونیکی در سطح فضای مجازی، برای جلوگیری از وقوع تخلفات، افزایش امنیت در این فضا و کنترل بر محتوای تولید شده توسط کاربران باید روش‌های برای شناسایی نویسندگان متون نوشته‌شده ارائه گردد. از این رو در این مقاله سعی شده است تا با ارائه یک روش ترکیبی، میزان دقت را در

تشـــخیص نویسندگان متون افزایش داد. در روش ترکیبی از الگوریتـــم بهینه‌ســـازی توده ذرات به عنـــوان الگوریتم استخراج‌کننده ویژگی‌های متون موجود در مجموعه‌داده، از الگوریتم ماشین بردار پشتیبان به‌عنوان شناسایی‌کننده نویســـنده متـــون و از متون موجـــود در Reuter_50_50 به‌عنوان مجموعه‌داده اســـتفاده شده است. براساس نتایج حاصله از ارزیابی روش پیشـــنهادی، الگوریتم پیشنهادی دارای درصد دقت ۹۶،۲ و درصد فراخوانی ۹۶،۱۲ می‌باشد این درحالی اســـت که میزان درصد دقت و درصد دوباره فراخوانی برای الگوریتم ماشـــین بردار پشـــتیبان برابر با ۹٤،۱۳ و ۹۳،۹٦ می‌باشد. براساس این نتایج می‌توان گفت که روش پیشـــنهادی دارای عملکرد بهینه‌تری می‌باشد از این رو می‌توان از روش پیشنهادی در سایر مسائل مطرح در حوزه هوش مصنوعی استفاده نمود.

مراجع

1- D. D. Lewis, A new Benchmark Collection for Text Categorization Research, The Journal of Machine Learning Research, 5: 361-397, 2004.

2- J.Allen, Natural language Understanding, Benjamin/ Cummings Publishing Company, 1987.

3- L.S.Oliveira, A Methodology for Feature Selection Using Multi-objective Genetic Algorithms for Handwritten Digit String, International Journal of Pattern Recognition and Artificial intelligence, 17 (6), pp: 903-929, 2003.

4- T.C.Mendenhall, the Characteristic Curves of Composition, Science, IX, 237–249, 1887.

5- G.K. Zipf, Selected Studies of the Principle of Relative Frequency in Language, Cambridge, MA: Harvard University Press, 1932.

6- G.U. Yule, On Sentence-Length as a Statistical Characteristic of Style in Prose, With Application to Two Cases of Disputed Authorship, Biometrika, 363-390, 1938.

7- G.U. Yule, The Statistical Study of Literary Vocabulary, Cambridge University Press, 1944.

8-F. Mosteller and D.L.Wallace, Inference and Disputed Authorship: The Federalist, Reading, MA: Addison-Wesley, 1964.

9-A.Narayanan, H.Paskov, N.Z.Gong, J.Bethencourt, On the Feasibility of Internet-Scale Author Identification, Draft. A Version of This Paper Will Appear At IEEE S&P 2012.

10-M.R.Davarpanah, M. SanjiAramideh, Farsi lexical analysis and stop word list, Library Hi Tech, Vol. 27 Issue 3, pp.435-449, 2009.

11-S.Argamon, C.Whitelaw, P.Chase, S.R.Hota, N.Garg, S.Levitan, Stylistic text classification using functional lexical features, Journal of the American Society for Information Science and Technology, 58(6), 802–822, 2007.

12-M.L.Brocardo,I.Traore, Sh.Saad, I.Woungang, Authorship Verification for Short Messages Using Stylometry, Conference on Computer, Information and Telecommunication Systems (CITS 2013), Piraeus-Athens, Greece, 2013.

13-A.S.Altheneyan and M.B. Menai, Naıve Bayes Classifiers for Authorship Attribution of Arabic Texts, Journal of King Saud University – Computer and Information Sciences 26, 473–484, 2014.

14-T. T. ZHU And M. LAN, ECNUCS: Measuring Short Text Semantic Equivalence Using Multiple Similarity Measurements, Second Joint Conference On Lexical And Computational Semantics (*SEM), pp. 124-131, 2013.

15-F. Howedi, M.Mohd, Text Classification for Authorship Attribution Using Naive Bayes Classifier With Limited Training Data, Computer Engineering And Intelligent Systems, Vol.5, No.4, 2014.

16-M. Sudheep Elayidom, Ch.Jose, A.Puthussery, N.K Sasi, TEXT CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION ANALYSIS, Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September 2013.

17-T.TAŞ, A.K. GORUR, Author Identification for Turkish Texts, Journal of Arts and Sciences Say: 7, May›s 2007.

18-J.Kennedy, RC.Eberhart. Particle swarm optimization, In: Proceedings of the IEEE conference on neural networks, Perth,Australia; pp. 1942–8, 1995.

19-C.Cortes and V.Vapnik, “Support-Vector Networks”, Machine Learning, Vol. 20, Issue 3, pp. 273-297, 1995.

20-V. RAGHAVAN and G. JUNG, A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance, ACM Transactions on Information Systems, Vol. 7, No. 3, Pages 205-229, July 1989.