سال سیزدهم،شماره‌های۴۷و۴۸،بهار و تابستان
۱۴۰۰
صص: ۳۵_۵۷

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

انتخاب ویژگی برای شناسایی نویسنده در متون کوتاه برخط فارسی

سمیه عارفی* محمد احسان بصیری امید روزمند*****

* کارشناس ارشد مهندسی فناوری اطلاعات، مؤسسه آموزش عالی صفاهان، اصفهان

** استادیار، دانشکده فنی و مهندسی، دانشگاه شهرکرد

*** استادیار، گروه مهندسی کامپیوتر، پردیس شهرضا، دانشگاه اصفهان

تاریخ دریافت: ۱۳۹۹/۰۵/۲۵ تاریخ پذیرش: ۱۳۹۹/۱۱/۲۳

نوع مقاله: پژوهشی

چکیده

رشد فزاینده‌ی استفاده از رسانه‌های اجتماعی و ارتباطات برخط به‌منظور بیان نظرات، تبادل عقاید و همچنین گسترش استفاده‌ی کاربران فارسی زبان از این ابزارها باعث افزایش متون فارسی در وب شده است. این رشد چشمگیر در کنار سوءاستفاده‌های ناشی از ناشناس بودن نویسنده‌ی نوشته‌ها نیاز به سامانه‌ی خودکار شناسایی نویسنده در این زبان را بیش از پیش آشکار می‌سازد. هدف از این پژوهش، بررسی ویژگی های مؤثر در شناسایی نویسندگان نظرات فارسی تولید شده توسط خریداران گوشی و همچنین ارزیابی روش‌های نظارتی و غیرنظارتی می‌باشد. عواملی که در این پژوهش بررسی می‌شود شامل ویژگی‌های لغوی، نگارشی، معنایی، ساختاری، دستوری، مختص متن و مختص شبکه‌های اجتماعی است. پس از استخراج ویژگی‌های مذکور، انتخاب ویژگی‌های برتر توسط چهار الگوریتم همبستگی ویژگی، نسبت بهره، OneR و تحلیل اجزای اصلی آزمایش می‌شود. در ادامه از الگوریتم‌های خوشه‌بندی مبتنی بر چگالی، K-means و EM برای خوشه‌بندی و الگوریتم‌های شبکه‌ی بیز، جنگل تصادفی و Bagging برای دسته‌بندی استفاده خواهد شد. ارزیابی الگوریتم‌های فوق بر روی نظرات فارسی مربوط به خریداران گوشی‌های سامسونگ نشان می‌دهد که بهترین تشخیص در بین الگوریتم‌های خوشه‌بندی با دقت ۵۹/۱۶٪ مربوط به الگوریتم EM روی ۱۵ ویژگی برتر انتخابی توسط OneR است درحالی‌که الگوریتم جنگل تصادفی به‌همراه نسبت بهره برای ۹۰ ویژگی با دقت ۷۹/۵۷٪ بهترین کارایی را در بین الگوریتم‌های دسته‌بندی دارد. همچنین مقایسه‌ی ویژگی‌ها نشان داد که ویژگی‌های نگارشی بیشترین تأثیر را در شناسایی نویسنده‌ی متون کوتاه داشته و پس از آن به ترتیب ویژگی‌های لغوی ، مختص متن، مختص شبکه‌های اجتماعی، ساختاری، دستوری و معنایی قرار گرفتند.

واژگان کلیدی: تحلیل متن، تحلیل سبک، استخراج ویژگی، انتخاب ویژگی و شناسایی نویسنده.

نویسنده مسئول: محمد احسان بصیری basiri@eng.sku.ac.ir

۱ مقدمه

بروز مسائل و مشکلات امنیتی و جرائم الکترونیکی از قبیل فعالیت‌های تروریستی، باج‌خواهی، حق‌السکوت، نامه‌های تهدید آمیز ناشناس، سرقت متون ادبی و علمی و مسائلی از این قبیل موجب ایجاد دلهره در کاربران به هنگام استفاده از ابزارهای شبکه‌های اجتماعی، نامه‌های الکترونیکی، پیام‌رسان‌های برخط و به‌طورکلی محیط‌های ارتباطی و اطلاعاتی شده است. امروزه شناسایی خودکار نویسنده‌ی این متون ناشناس به یک ابزار ارزشمند جهت حمایت از پژوهش‌های مربوط به حوزه‌ی جرم و جنایت و امنیت مبدل گردیده است.

با گسترش استفاده از اینترنت، ارتباطات برخط و شبکه‌های اجتماعی، خدمات مبتنی بر وب به‌عنوان یکی از مؤثرترین ابزارها جهت ارسال اطلاعات متنی و محتویات تولید شده توسط کاربر تبدیل شده‌اند. بیشتر این خدمات به کاربر اجازه می‌دهند که هویت واقعی خود را مخفی کرده، با هویت جعلی وارد فضای مجازی شده [۱] و به‌صورت ناشناس یا تحت نام مستعار با ارائه‌ی اطلاعات نادرست در وبلاگ‌ها، وب‌سایت‌ها و نامه‌های الکترونیکی مرتکب اقدامات خراب‌کارانه شوند [۲]؛ که این امر موجب برانگیختن رعب و وحشت و عدم اطمینان در میان کاربران می‌گردد. در چنین فضایی، مجرمان با مخفی کردن هویت اصلی خود و ارائه‌ی اطلاعات نادرست در مورد جنسیت، مکان، سن، سطح تحصیلات و ملیت خود، آزادانه به بیان عقاید و نظراتشان پرداخته و مرتکب اقدامات خرابکارانه می‌شوند [۱]. همچنین گروه‌های تروریستی و جنایی از نامه‌های الکترونیکی به‌عنوان یک کانال امن برای ارتباطات مخفی خود استفاده می‌کنند [۳] و نفوذگرها به‌منظور انجام فعالیت‌های غیرقانونی و ارتکاب جرائم زیربنایی مانند انتقال کرم، ویروس، تروجان و دیگر بدافزارهای قابل اجرا روی اینترنت، کاربران فضای مجازی را هدف قرار می‌دهند. نامه‌های تبلیغاتی، نامه‌های تهدیدآمیز، بدگویی‌های نژادی و هرزه‌نگاری رایج‌ترین نمونه‌های سوءاستفاده از نامه‌های الکترونیکی می‌باشند. علاوه‌براین، به‌دلیل بین المللی بودن جرائم سایبری، مسائل چندزبانگی نیز به یک چالش جدید برای تحلیل هویت نویسنده تبدیل شده است [۴]. به‌دلیل بروز چنین مسائلی، شناسایی مشخصات تولیدکنندگان محتوا و اهمیت حفاظت از این فضا هر روز پررنگ‌تر و نیاز به شناسایی خودکار نویسنده‌ی متون بیش از پیش آشکار می‌گردد.

تحلیل هویت نویسنده تا کنون در سه زمینه‌ی پژوهشی مختلف شامل شناسایی هویت نویسنده[۱]، تأیید هویت نویسنده[۲] و توصیف شخصیت نویسنده[۳] دسته‌بندی و مورد مطالعه قرارگرفته است [۴]. شناسایی هویت نویسنده، به مقایسه‌ی یک سند بی‌نام و نشان با اسناد موجود پرداخته و به بررسی این‌که آیا آن‌ها توسط یک نویسنده‌ی واحد تولید شده‌اند یا خیر می‌پردازد [۵]. مطالعاتی که تا کنون در این زمینه صورت گرفته به شناسایی راه‌هایی جهت شناسایی الگوهای ارتباطات تروریسم، کشف سرقت متون ادبی و علمی، حل مشاجرات در نوشته‌های ادبی و تاریخی یا نگارش موسیقی پرداخته است. تأیید هویت نویسنده، توسط مقایسه‌ی نمونه نگارش شخص مدعی با سایر نوشته‌های وی به اثبات این‌که آیا یک سند ناشناس توسط همان شخص نوشته شده است یا خیر می‌پردازد و تعیین خصوصیاتی مثل جنسیت، سن و سطح تحصیلات نویسنده‌ی یک سند ناشناس در حوزه‌ی توصیف شخصیت نویسنده قرار می‌گیرد [۴] و [۵] .

با این فرض، که هر نویسنده سبک نوشتاری مخصوص به خود را دارد و از کلمات خاصی استفاده می‌کند که نوشته‌اش را منحصربه‌فرد می‌سازد، استخراج ویژگی‌هایی از درون متون، گزینه‌ای مناسب جهت شناسایی نویسنده می‌باشد [۶]. انتخاب این ویژگی‌های متمایز کننده، با استفاده از برخی روش‌های آماری و یادگیری ماشین صورت می‌گیرد. این ویژگی‌ها که مطابق نظر پژوهشگران مشخص می‌شود، شامل ویژگی‌های لغوی، نحوی، معنایی، ساختاری، دستوری، مختص متن و مختص شبکه‌های اجتماعی می‌باشد. مطالعاتی که تا کنون در این زمینه صورت گرفته بیانگر این مطلب است که استفاده از ویژگی‌های سبک‌شناسی بهترین شیوه جهت شناسایی نویسنده‌ی متون می‌باشد [۵] و [۶].

تجزیه و تحلیل متون جهت شناسایی نویسنده‌ی ناشناس، از ابتدای قرن ۱۹ به‌صورت سنتی آغاز گردید و توزیع طول کلمات، از اولین ویژگی‌های مورد استفاده جهت شناسایی نویسنده‌ی متون در این دوره بوده است. با حضور رایانه، شناسایی خودکار نویسنده توسط تحلیل متون با استفاده از ویژگی‌هایی از سبک نگارش نویسنده میسر گردید. نخستین مطالعات صورت گرفته در این زمینه مربوط به استفاده از لغات دستوری با استفاده از طبقه‌بند بیز ساده بوده است که از این روش آماری جهت شناسایی هویت نویسندگان مقالات فدرالیست استفاده شد [۷] .

گروه دیگری از مطالعات که در سال‌های اخیر توجه بیشتری را به خود جلب کرده‌اند، با استفاده از ویژگی‌های سبک‌شناسی و با اعمال روش‌های یادگیری ماشین و متن‌کاوی به بررسی و ایجاد مدل‌هایی برای تعیین میزان تأثیر ویژگی‌های مختلف بر روی شناسایی نویسنده‌ی ناشناس پرداخته و با استفاده از روش‌های جدید به ارزیابی و انتخاب مناسب‌ترین ویژگی‌ها جهت حل مسئله‌ی شناسایی نویسنده‌ی متون پرداخته‌اند.

به‌دلیل تفاوت‌های بنیادی در قواعد دستوری زبان انگلیسی با زبان فارسی، مطالعات کم‌تری در زمینه‌ی طراحی ابزارهای ماشینی و محاسباتی مرتبط با زبان فارسی صورت گرفته [۲] که این امر

---

[۱] Authorship Identification

[۲] Authorship Verification

[۳] Authorship Characterization

موجب بروز چالش‌هایی در فاز پیش‌پردازش داده‌های متنی، از جمله ریشه‌یابی و برچسب‌گذاری اجزای کلام می‌گردد. علی‌رغم چنین کاستی‌هایی، نیاز به وجود سیستمی جامع و هوشمند برای کمک به مراجع قضایی و تحقیقاتی جهت شناسایی مجرمان سایبری و نویسنده‌ی مطالب بازنویسی شده و هرزنامه‌ها در سایت‌ها و شبکه‌های اجتماعی، ضرورت انجام پژوهش در زمینه‌ی شناسایی نویسنده‌ی متون کوتاه در زبان فارسی را آشکار می‌سازد.

با توجه به این‌که پژوهش‌های صورت گرفته در زبان فارسی در گذشته، اغلب بر روی متون ساختارمند و طولانی بوده و تنها از ویژگی‌های سنتی جهت شناسایی نویسنده‌ی متون استفاده شده است؛ در این پژوهش، تمرکز ما بر روی متون کوتاه برخط شامل نظرات فارسی تولید شده توسط خریداران گوشی می‌باشد که جهت دقیق‌تر شدن کار شناسایی نویسنده‌ی متون کوتاه، علاوه بر استفاده از ویژگی‌های سنتی متداول که در گذشته بر روی متون فارسی مورد استفاده قرار گرفته است، ویژگی‌های مختص شبکه‌های اجتماعی را نیز به مجموعه ویژگی‌ها افزودیم تا با استفاده از الگوریتم‌های انتخاب ویژگی مناسب بر روی ویژگی‌های استخراج شده و اعمال الگوریتم‌های خوشه‌بندی و دسته‌بندی، به مقایسه‌ی کارایی الگوریتم‌های نظارتی و غیرنظارتی پرداخته و همچنین ویژگی‌های مؤثر در شناسایی نویسندگان متون کوتاه را بر اساس سبک نگارش افراد مورد بررسی قرار دهیم.

به‌صورت خلاصه، نوآوری‌های پژوهش جاری به‌صورت زیر می‌باشد:

- ارائه‌ی مجموعه داده‌ی نظرات فارسی در رابطه با محصولات گوشی‌های سامسونگ، مربوط به سال‌های ۲۰۱۵ و ۲۰۱۶.
- پیشنهاد استفاده از ویژگی‌های مختص شبکه‌های اجتماعی در کنار سایر ویژگی‌های سنتی.
- مقایسه‌ی ویژگی‌های مختلف و انتخاب ترکیب مناسب آن‌ها برای شناسایی نویسنده‌ی متون کوتاه.

ادامه‌ی مقاله به شرح ذیل سازماندهی گردیده است. در بخش دوم، مبانی نظری و پیشینه‌ی پژوهش ارائه خواهد شد و در بخش سوم، روش پژوهش و ویژگی‌های سبک‌شناسی مورد استفاده جهت شناسایی نویسنده‌ی متون کوتاه بیان می‌گردد. بخش چهارم به پیاده‌سازی، بخش پنجم به تحلیل نتایج، بخش ششم به بحث اختصاص داده شده و درنهایت، بخش هفتم به نتیجه‌گیری پرداخته شده است.

۲ پیشینه پژوهش

شناسایی نویسنده با استفاده از ویژگی‌های سبک نوشتاری، شامل استخراج ویژگی‌های موجود در متن می‌باشد، به‌طوری‌که بتوان با استفاده از این ویژگی‌ها، بین متونی که توسط افراد مختلف نوشته شده تمایز قائل شد. استفاده از ویژگی‌های سبک‌شناسی جهت شناسایی نویسنده سابقه‌ای طولانی دارد. سیر تکاملی این حوزه را می‌توان در چهار دوره مورد بررسی قرار داد. نخستین پژوهش‌های صورت گرفته در این زمینه، به قرن ۱۹ میلادی باز می‌گردد که در آن زمان مندنهال [۸] به‌منظور تشخیص اشعار مورد تردید ویلیام شکسپیر و فرانسیس بیکن، از توزیع طول کلمات جهت شناسایی نویسنده‌ی متون استفاده نمود.

نخستین پژوهش‌های غیرسنتی در این زمینه، با کمک رایانه و با استفاده از مدل‌های احتمالاتی توسط موستلر و والاس در سال ۱۹۶۴ انجام شد. این مطالعات که با استفاده از طبقه‌بند بیز ساده و لغات دستوری بر روی مجموعه مقالات فدرالیست صورت گرفت، مؤثر بودن کلمات تابع[۱] در حل مسئله‌ی شناسایی نویسنده را به اثبات رساند [۷]. کرایگ در سال ۱۹۹۹ برای شناسایی نویسنده‌ی متون از تکیه‌کلام‌های افراد که در حقیقت همان عادات نوشتاری افراد می‌باشند، استفاده نمود. وی با استفاده از تجزیه و تحلیل نمایشنامه‌های میدلتون توماس و دیگران، رابطه‌ای که میان هویت و خصوصیات نگارش افراد وجود دارد را اثبات کرد [۹]. کوپل و همکارانش در سال ۲۰۰۴، ماشین بردار پشتیبان با هسته‌ی خطی را بر روی ویژگی‌های لغوی شامل ۲۵۰ کلمه‌ی پرتکرار استخراج شده از ۲۱ کتاب انگلیسی نوشته شده توسط ۱۰ نویسنده‌ی مختلف به‌کار بردند و با نادیده گرفتن نمونه‌های منفی، تعیین هویت نویسنده را به‌عنوان یک مسئله‌ی طبقه‌بندی تک کلاسه نشان دادند. آن‌ها متن را به بخش‌هایی با ۵۰۰ کلمه‌ی تقریباً مساوی تقسیم کرده و اختلاف بین سند نمونه تولید شده مشکوک و کاربران دیگر را سنجیده و به دقت ۹۵/۷٪ دست یافتند [۱۰].

با ظهور اینترنت و وب جهانی و درنهایت، با حضور رسانه‌های شبکه‌های اجتماعی و پیام‌رسانی تعاملی و پویا، تجزیه و تحلیل اسناد الکترونیکی و پیام‌های برخط به یک حوزه‌ی پژوهشی جدید در زمینه‌ی شناسایی نویسنده‌ی متون کوتاه مبدل گردید [۱۱].

از آن زمان تاکنون، با استفاده از مجموعه ویژگی‌های سبک‌شناسی و روش‌های طبقه‌بندی متفاوت، پژوهش‌های فراوانی بر روی مجموعه داده‌های مختلف شامل نامه‌های الکترونیکی، گروه‌های خبری، متون برخط، پست‌های شبکه‌های اجتماعی، مقالات، پیام‌های کوتاه، وبلاگ‌ها، وب‌سایت‌ها و کتاب‌ها صورت گرفته است که در ادامه، به بررسی ویژگی‌های استخراج شده از این متون و نتایج حاصل از اعمال الگوریتم‌های متفاوت بر روی این مجموعه داده‌ها می‌پردازیم.

۲,۱ مجموعه داده‌های مربوط به نامه‌های الکترونیکی

بیشترین پژوهش‌های انجام شده در این زمینه، به دسته‌بندی و شناسایی نویسنده‌ی نامه‌های الکترونیکی می‌پردازد. نخستین

[۱] Function words

پژوهش‌های انجام شده بر روی این مجموعه داده در سال ۱۹۹۳ توسط گوتنر و همکارانش صورت گرفت که با استفاده از فرهنگ واژگان به‌همراه شبکه عصبی[۱] دقت ۷۹/۱٪ را نتیجه گرفتند [۱۲]. دی ول در سال ۲۰۰۰ مدل دسته‌بندی ماشین بردار پشتیبان[۲] را روی مجموعه‌ای از ویژگی‌های ساختاری و نحوی به‌منظور تحلیل هویت نویسنده‌ی نامه‌های الکترونیکی ناشناس به‌کار برد و دریافت که با افزایش تعداد کلمات تابع از ۱۲۲ به ۳۲۰ کارایی دسته‌بند وخیم‌تر می‌گردد و نه تنها افزودن ویژگی‌های بیشتر جهت بهبود دقت لازم نیست، بلکه افزودن ویژگی‌های بی‌فایده ممکن است موجب کاهش دقت دسته‌بند گردد [۱۳].

در سال ۲۰۰۲، کورنی و همکارانش با استفاده از نامه‌های الکترونیکی به بررسی رابطه‌ی بین سبک نگارش افراد با سطح تحصیلاتشان پرداختند. آن‌ها ۲۵۳ نامه‌ی الکترونیکی ۲۰۰ تا ۵۰۰ کلمه‌ای مربوط به چهار کاربر را توسط ویژگی‌های سبک‌شناسی، ساختاری و کلمات تابع به‌همراه ماشین بردار پشتیبان آزمایش کرده و دقت ۷۰/۲ درصد را نتیجه گرفتند و دریافتند که با افزایش تعداد نویسندگان، کاهش طول اسناد و همچنین کاهش اندازه‌ی مجموعه‌ی آموزشی، دقت دسته‌بندی کاهش می‌یابد [۱۴].

اقبال و همکارانش در سال ۲۰۰۸ برای اولین بار از یک روش داده‌کاوی مبتکرانه به‌نام «چاپ‌نوشته[۳]» استفاده کردند. این روش شامل ترکیباتی از ویژگی‌های منحصربه‌فرد استخراج شده از متن است که به‌طور مکرر در نوشته‌های یک شخص اتفاق می‌افتد و همانند اثر انگشت بوده و نگارش فرد را از دیگران متمایز می‌کند [۱۵]. چن و عباسی نیز در همان سال به‌منظور شناسایی و کشف شباهت نگارش از چاپ‌نوشته در آزمایش‌هایشان استفاده کردند. آن‌ها به‌منظور انتخاب ویژگی‌های برتر از الگوریتم تحلیل اجزای اصلی[۴] استفاده کرده و جهت کشف شباهت اجزاء، یک موجودیت[۵] ناشناس را انتخاب نموده و آن را با تمام موجودیت‌های دیگر مقایسه کرده و یک رتبه محاسبه نمودند؛ اگر رتبه بالاتر از مقدار از پیش تعریف شده بود، موجودیت ناشناس در دسته‌ای با موجودیت تطبیق شده دسته‌بندی می‌شد [۱۶]. اقبال و همکارانش در ادامه‌ی تحقیقات خود در سال ۲۰۱۰، آزمایش‌های تجربی خود را بر روی ۲۰۰ نامه‌ی الکترونیکی ۶۲۸ تا ۱۳۴۲ کلمه‌ای با استفاده از ۲۹۲ ویژگی استخراج شده شامل ویژگی‌های لغوی، نحوی، خطاهای دستور زبانی و املایی و ویژگی‌های وابسته به محتوا را با دو روش متفاوت روی مجموعه داده مدیریت کردند و نرخ خطای ۱۷/۱٪ تا ۲۲/۴ درصد را نتیجه گرفتند [۱۷]. آن‌ها در همان سال در آزمایشی دیگر، با استفاده از الگوریتم‌های خوشه‌بندی K-means، EM و Bisecting K-Means و ترکیبی از چهار ویژگی ساختاری، لغوی، نحوی و مختص محتوا و استخراج چاپ‌نوشته‌ی مربوط به نویسندگان نامه‌های الکترونیکی، به ارزیابی ویژگی‌های سبک‌شناسی و تأثیر تعداد پیام‌های هر نویسنده پرداختند [۳]. آلیسون و گوتری در سال ۲۰۰۸ با استفاده از ماشین بردار پشتیبان و مجموعه ویژگی‌های نحوی، ۶۳ ویژگی مربوط به شمارش تعداد کلمات پرسشی زبان و ویژگی‌های n-تایی (۲-تایی و ۳-تایی) برای ۸ نویسنده با تعداد ۱۷۴ تا ۷۰۶ نامه‌ی الکترونیکی برای هر نویسنده و طول وبلاگ‌های ۱۰۰ تا ۶۰۰ کلمه، به شناسایی نویسنده یادداشت‌ها و وبلاگ‌ها پرداخته و میانگین دقت ۸۶/۷۴ درصد را کسب نمودند [۱۸].

چنگ و همکارانش در سال ۲۰۱۱ آزمایش‌های خود را بر روی نامه‌های الکترونیکی و مجموعه‌ی خبری رویترز با استفاده از ماشین بردار پشتیبان انجام دادند و دقت‌های ۸۲/۲۳٪ و ۷۶/۷۵٪ را به‌دست آوردند [۱۹]. در سال ۲۰۱۱ چن و همکارانش به بررسی شباهت نگارش نامه‌های الکترونیکی مربوط به ۴۰ نویسنده پرداختند و بدین منظور از ۱۵۰ ویژگی سبک‌شناسی شامل ۴۰ ویژگی لغوی، ۹ ویژگی ساختاری، ۷۶ ویژگی نحوی و ۲۵ ویژگی وابسته به محتوا برای تأیید هویت نویسنده استفاده، و برای بلاک‌های ۳۰ تا ۵۰ کلمه‌ای با استفاده از ماشین بردار پشتیبان برای ۱۰ نمونه دقت ۸۳/۹۰٪ و دقت ۸۸/۳۱٪ را برای ۱۵ نمونه نتیجه گرفتند [۲۰]. بروکاردو و همکارانش در سال ۲۰۱۵ تحقیقات خود را بر روی نامه‌های الکترونیکی و توئیت‌های ۴۰، ۲۸۰ و ۵۰۰ کلمه‌ای با استفاده از ویژگی سبک‌شناسی شامل ۵۲۸ ویژگی علامت‌محور، ۷۵ ویژگی کلمه‌محور، ۶۳۲ ویژگی نحوی، ۷ ویژگی وابسته به کاربرد و n-تایی‌های استخراج شده با استفاده از ماشین بردار پشتیبان انجام دادند و نرخ خطای متغیر از ۹/۹۸٪ تا ۴۵/۲۱٪ را نتیجه گرفتند [۵]. نیرخی و همکارانش در سال ۲۰۱۶ روش‌های غیرنظارتی خوشه‌بندی سلسله مراتبی[۶] شامل روش همجوشی[۷] و روش جداگر[۸] را به‌همراه رتبه‌بندی چندبعدی[۹] برای نگاشت متون به فضای دو بعدی روی نامه‌های الکترونیکی به‌منظور تعیین هویت نویسنده به‌کار بردند [۴].

---

[۱] Neural Network (NN)

[۲] Support Vector Machine (SVM)

[۳] write print

[۴] Principal Component Analysis (PCA)

[۵] entity

[۶] Hierarchial Clustering

[۷] Agglomerative way

[۸] Divisive way

[۹] Multimentional scaling

۲,۲ مجموعه داده‌های مربوط به رسانه‌های شبکه‌های اجتماعی

گسترش استفاده از رسانه‌های شبکه‌های اجتماعی در این سال‌ها، و بروز مشکلات و جرائم الکترونیکی موجب شد محققان کارهای پژوهشی خود را بر روی چنین مجموعه داده‌هایی متمرکز کنند. در سال ۲۰۱۰ جرمی و همکارانش به‌منظور شناسایی نویسندگان استاتوس‌های فیس‌بوک با استفاده از دسته‌بندهای بیز ساده و پرسپترون به ترتیب دقت‌های ۶۷/۷٪ و ۵۹٪ را نتیجه گرفتند [۲۱]. در همان سال، لایتون و همکارانش به‌منظور احراز هویت کاربران توئیتر، آزمایش‌های خود را بر روی ۱۲۰ توئیت با حدبیشتر طول ۱۴۰ علامت مربوط به ۵۰ کاربر، با استخراج n-تایی (۳-تایی) و با استفاده از الگوریتم K-means انجام داده و دقت ۷۰ درصد را به‌دست آوردند [۲۲]. لی و همکارانش در سال ۲۰۱۲، تکنیک انتخاب ویژگی غیرنظارتی را برای رسانه‌های اجتماعی پیشنهاد کردند که بر اساس مدل احتمالاتی ابتدا ویژگی‌های بالقوه اجتماعی برای هر مورد به‌دست آمده، سپس اهمیت هر ویژگی اندازه‌گیری شده و فضای ویژگی به‌صورت ماتریس چندبعدی تعریف می‌گردد. هرگاه ویژگی جدیدی دریافت شود، یک آزمون برای پذیرش یا رد ویژگی انجام می‌شود. اگر ویژگی‌ها پذیرفته شوند، مدل دوباره بهینه‌سازی می‌گردد و همچنین امکان حذف ویژگی‌هایی که قبلاً انتخاب شده بود نیز وجود دارد [۲۳].

در سال ۲۰۱۴، لی و همکارانش به‌منظور احراز هویت نویسنده پست‌های شبکه‌های اجتماعی، پست‌های ۳۰ کاربر فیس‌بوک که به‌طور میانگین از ۲۰/۶ کلمه تشکیل شده بود را مورد مطالعه قرار داده و از ۲۳۳ ویژگی سبک‌شناسی شامل ۵۰ ویژگی علامت‌محور، ۱۸ ویژگی کلمه‌محور، ۱۵۸ ویژگی نحوی، ۱ ویژگی ساختاری و ۶ ویژگی مختص شبکه‌های اجتماعی استفاده و آزمایش‌های خود را با روش‌های متفاوتی تکرار کردند و بهترین نرخ دقت ۷۹/۶٪ به هنگام ترکیب ویژگی‌های مختص شبکه‌های اجتماعی و سبک‌شناسی سنتی به‌دست آوردند. آن‌ها با استفاده از SVM Light دقت ۹/۷۸٪ را برای ویژگی‌های سبک‌شناسی سنتی و دقت ۶۹/۸٪ را برای ویژگی‌های مختص شبکه‌های اجتماعی به‌طور مجزا به‌دست آوردند و دریافتند که ترکیب این ویژگی‌ها دقت بالاتری را نتیجه می‌دهد. آن‌ها به‌طورکلی دریافتند که افزایش تعداد کاربران موجب کاهش دقت دسته‌بندی شده و ترکیب ویژگی‌های سبک‌شناسی و مختص شبکه‌های اجتماعی موجب افزایش دقت و استفاده از ویژگی‌های مجزا موجب کاهش دقت می‌گردد و همچنین تعداد ویژگی‌های بیشتر موجب انحراف استاندارد کمتر و دقت بیشتر می‌شود. علاوه‌براین، نتایج نشان داد که ویژگی‌های علامت‌محور جهت شناسایی نویسنده‌ی متون کوتاه معیار مطلوب‌تری نسبت به کلمه‌محور می‌باشد. همچنین برای مقایسه‌ی الگوریتم ماشین بردار پشتیبان با سایر الگوریتم‌ها، آزمایش‌های خود را با الگوریتم K-نزدیک‌ترین همسایه تکرار کرده و به دقت ۶۵/۵٪ دست یافتند [۲۴].

زوبیاگا و همکارانش در سال ۲۰۱۵ روشی جهت دسته‌بندی توئیت‌های جهان واقعی در چهار دسته‌ی اخبار، حوادث، یادداشت‌ها و یادبودها توسط تعیین مجموعه‌ی کوچکی از ویژگی‌های مستقل زبانی پیشنهاد کردند. بدین منظور از ویژگی‌هایی شامل عناصر لغوی، هش‌تگ‌ها، URLها، فهرست واژگان، علامات سؤال و تعجب استفاده کردند. چهار ویژگی که روش آن‌ها را جهت دسته‌بندی توئیت‌های جهان واقعی مناسب می‌سازند عبارتند از: مجموعه ویژگی‌های کوچک مورد نیاز که می‌تواند به‌طور واضح محاسبه گردد، بهبودبخشی قدرت تأثیر محتوا، هزینه‌ی محاسباتی خطی برای تعداد توئیت‌های تحلیل شده و بدون تغییر ماندن تعداد ویژگی‌های باقی مانده علی‌رغم تعداد نمونه‌ها. به‌هرحال، به‌دلیل این‌که ویژگی‌ها به‌طور اختصاصی طراحی شده بودند، کاربردی بودن آن‌ها جهت استفاده در زمینه‌های دیگر ممکن نبود [۲۵].

۲,۳ مجموعه داده‌های مربوط به وبلاگ‌ها، وب‌سایت‌ها، گروه‌های خبری، پیام‌های آنی و متون برخط

سایر مجموعه داده‌های مورد استفاده در پژوهش‌های اخیر شامل وبلاگ‌ها، وب‌سایت‌ها، گروه‌های خبری، پیام‌های آنی و متون برخط می‌باشد که این مجموعه داده‌ها در زبان‌های مختلف مورد بررسی قرار گرفته‌اند. در سال ۲۰۰۶، آنژلا با استفاده از دسته‌بند بیز ساده به‌همراه استخراج ۶۹ ویژگی سبک‌شناسی شامل ساختار جمله، علامت‌های خاص از پیش تعریف شده مربوط به احساسات، مخفف‌ها و تحلیل تکرار علائم بر روی چهار پیام آنی، یک چارچوب سامانه کشف نفوذ پیام آنی ایجاد کرده و بهترین دقت با میانگین ۶۸ درصد را نتیجه گرفت [۲۶].

در سال ۲۰۱۱، کانالز و همکارانش به‌منظور تأیید هویت نویسندگان متون از ۴۰ نمونه متن که به‌صورت برخط جمع‌آوری شده بودند استفاده کرده و یک دسته‌بند K-نزدیک‌ترین همسایه[1] را با ۸۲ ویژگی سبکی شامل ۶۲ ویژگی لغوی (۴۹ ویژگی علامت‌محور[2]، ۱۳ ویژگی کلمه‌محور) و ۲۰ ویژگی نحوی آموزش دادند و نرخ خطای ۲۰/۲۵٪ تا ۴/۱۸٪ را به‌دست آوردند و دریافتند که نه تنها ویژگی‌های نحوی لزوماً بیانگر سبک نگارشی افراد نبوده، بلکه این مجموعه ویژگی‌ها جهت تأیید هویت نویسنده کافی نیست و باید ویژگی‌های دیگری نیز به مجموعه ویژگی‌ها افزوده شود [۲۷].

[1] K Nearest Neighbor (KNN)

[2] Character-based

در سال ۲۰۰۹، لای جهت شناسایی نویسندگان وبلاگ‌ها با استفاده از طبقه‌بند بیز، دقت ۶۹ درصد را نتیجه گرفت [۲۸]. آلام و کومار در سال ۲۰۱۳ ویژگی‌های معنایی نوین را برای کمک به سیستم شناسایی نویسنده برای زبان عربی توسعه دادند. بدین منظور، از تجزیه کننده‌های مختص پردازش زبان طبیعی، واژگان، پردازش معنایی، انتساب نقش موضوعی، هیوریستیک‌های معنایی و تکنیک‌های یادگیری ماشین استفاده کردند و همچنین ماشین بردار پشتیبان را برای طبقه‌بندی داده‌های متنی به گروه‌های مختلف بر اساس سبک نگارش به‌همراه ویژگی‌های لغوی، نحوی، ساختاری، معنایی و مختص متن به‌کار بردند. آن‌ها آزمایش‌های خود را بر روی مقالات وب‌سایت عربی در حوزه‌ی تروریسم و نظم و قانون انجام دادند؛ بدین صورت که ابتدا هر ویژگی را به‌طور مجزا مورد بررسی قرار داده و پس از ترکیب تمام ویژگی‌ها به دقت ۹۸٪ دست یافتند [۲۹].

در همین سال آدامز و همکارانش، جهت بهینه‌سازی جست‌وجو برای افزایش دقت شناسایی، از ترکیب روش‌های الگوریتم ژنتیک[۱] و تکاملی[۲] استفاده کردند. سپس به‌منظور بهبود فرایند تطبیق، توسعه‌ی اکتشافی ژنتیک[۳] را معرفی کرده و آزمایش‌های خود را بر روی ویژگی‌های استخراج شده از نوشته‌های مربوط به وبلاگ خبری Huffingtonpost.com و CNN.com شامل ۱۷۰ ویژگی لغوی به‌همراه نشانه‌های نقطه‌گذاری، ۲۱ ویژگی ساختاری و ۱۲۲ کلمه‌ی تابع انجام داده و میانگین دقت‌های ۲۱/۵۰٪، ۴۳/۴۳٪ و ۵۷ درصد را برای آزمایش‌های حالت پایه، ترکیب روش‌های الگوریتم ژنتیک و تکاملی و توسعه‌ی اکتشافی ژنتیک به‌دست آوردند و دریافتند که این روش به‌طور موفق دقت تشخیص را افزایش می‌دهد درحالی‌که به‌صورت چشمگیر تعداد خصوصیات مورد نیاز جهت تشخیص را کاهش می‌دهد [۳۰].

در سال ۲۰۰۶، استاماتاتوس و هوواردس n-تایی[۴] با طول متغیر را بر روی مجموعه داده رویترز به‌کار بردند. آن‌ها از بهره‌ی اطلاعاتی[۵] برای انتخاب ویژگی و از ماشین بردار پشتیبان با هسته‌ی خطی به‌منظور دسته‌بندی استفاده کرده و به دقت ۷۳/۰۸٪ دست یافتند [۳۱].

در سال ۲۰۱۲ گونال و یوسال یک روش جدید احتمالاتی انتخاب ویژگی به‌نام انتخابگر ویژگی متمایز[۶] را به‌منظور انتخاب ویژگی پیشنهاد کردند و بدین منظور دسته‌بندهای ماشین بردار پشتیبان، شبکه عصبی و درخت تصمیم را بر روی چهار مجموعه داده رویترز، گروه‌های خبری، خدمات پیام کوتاه و نامه‌های الکترونیکی آزمایش کرده و به‌منظور اثبات کارایی روش پیشنهادی، با تکرار آزمایش‌ها دریافتند که روش پیشنهادیشان از لحاظ دقت دسته‌بندی و زمان پردازش، عملکردی رقابتی با سایر روش‌ها داشته است [۳۲].

۲,۴ مجموعه داده‌های مربوط به پژوهش‌های صورت گرفته در زبان فارسی

در سال‌های اخیر در زمینه‌ی شناسایی نویسنده در متون فارسی نیز تحقیقاتی بر روی مقالات، متون ادبی و پیام‌های الکترونیکی صورت گرفته است. در سال ۱۳۹۱، فرهمندپور و همکارانش از ویژگی‌های لغوی، نحوی، معنایی و وابسته به کاربرد جهت شناسایی هویت نویسنده استفاده کرده، کارایی این ویژگی‌ها و روش‌های K-نزدیک‌ترین همسایه، دلتا و الگوریتم ژنتیک را بر روی دو پایگاه داده شامل متون مربوط به دانشجویان دانشگاه بوعلی سینا و مقالات مربوط به ۸ نویسنده‌ی هم‌عصر، بررسی کرده و به دقت‌های ۵۰ تا ۱۰۰ درصد دست یافتند [۲]. یک سال بعد، زنگویی و شمس آباد، جهت شناسایی نویسندگان پیام‌های الکترونیکی مربوط به ۵۰ نفر از مشتریان بالقوه‌ی وب‌سایت آمازون، روش یادگیری ماشین به‌نام تجمیع هسته‌های وزن‌دار[۷] را پیشنهاد کردند. آن‌ها در ابتدا ویژگی‌های لغوی، نحوی، ساختاری و خاص متن را به‌صورت مجزا به‌کار برده و سپس با استفاده از یک روند تکاملی ویژگی‌ها را اضافه نموده و درنهایت آزمایش‌های خود را با ترکیبی از با استفاده از دسته‌بندهای ماشین بردار پشتیبان، تجمیع هسته‌های وزن‌دار، شبکه عصبی و $C_{4.5}$ به پایان رسانده و دقت‌های ۹۷/۶۹٪، ۷۸/۹۸٪، ۶۶/۹۶٪ و ۹۳/۳۶ درصد را نتیجه گرفتند [۳۳]. مرادی و بحرانی در سال ۱۳۹۴ با استفاده از ویژگی‌های سبک‌شناسی و روان‌شناختی به تشخیص جنسیت نویسنده پرداختند. بدین منظور مطالعات خود را بر روی متون ادبی فارسی (داستان و رمان) و نظرات کاربران در سایت هلوکیش با استفاده از ماشین بردار پشتیبان، درخت تصمیم و بیز ساده به انجام رساندند و دقت ۶/۷۱٪، ۶/۷۳٪ و ۶۶/۳ درصد برای پیکره متون و دقت ۵۸/۸٪، ۶۳٪ و ۵۳ درصد را برای پیکره نظرات نتیجه گرفتند [۱].

به‌منظور مقایسه‌ی نتایج پژوهش‌های صورت گرفته تا کنون، چند نمونه از آن‌ها در جداول ۱ و ۲ آورده شده است.

به‌طورکلی می‌توان گفت در سال‌های اخیر توجه بیشتری به شناسایی نویسنده در چارچوب کاربردهای عملی مثل بررسی نویسنده پیام‌های الکترونیکی، شناسایی الگوهای تروریسم، حل مشاجرات تاریخی و ادبی، موارد دادگاهی، کشف سرقت ادبی و

[۱] Genetic Algorithm (GA)
[۲] Evolutionary Algorithm (EA)
[۳] Genetic Heuristic Development (GHD)
[۴] n-gram
[۵] Information Gain (IG)
[۶] Distinguishing Feature Selector (DFS)
[۷] Weighted Kernel Fusion based on SVM-Parallel Hierarichal Gride Search (WKF)

علمی در مقالات دانشجویی و توصیف شخصیت نویسنده صورت گرفته است. بدین صورت که ابتدا ویژگی‌های برتر استخراج شده، سپس الگوریتم‌های انتخاب ویژگی مناسب بر روی این ویژگی‌ها اعمال شده و درنهایت با استفاده از الگوریتم‌های خوشه‌بندی یا دسته‌بندی مناسب، به شناسایی نویسندگان پرداخته شده است.

جدول ۱. نتایج مربوط به پژوهش‌های انجام شده در زبان‌های مختلف .

مرجع	مجموعه داده	ویژگی	روش	دقت
[۳]	نامه‌های الکترونیکی	لغوی، نحوی، ساختاری، وابسته به محتوا	EM, K-means, Bisecting K-Means	نرخ‌خطا ۰/۷۳-۰/۸۰ ۰/۷۳-۰/۸۸ ۰/۷۵-۰/۸۳
[۴]	نامه‌های الکترونیکی	لغوی	خوشه‌بندی سلسله مراتبی، رتبه‌بندی چندبعدی	
[۵]	نامه‌های الکترونیکی، توئیت‌ها	لغوی (n-تایی)، نحوی، ساختاری، وابسته به محتوا	ماشین بردار پشتیبان	۹۰/۰۲ ۷۸/۵۵
[۱۰]	۲۱ کتاب انگلیسی مربوط به ۱۰نویسنده	لغوی	ماشین بردار پشتیبان با هسته‌ی خطی	۹۵/۷
[۱۲]	نامه‌های الکترونیکی	لغوی (فرهنگ واژگان)	شبکه عصبی	۷۹/۱
[۱۳]	نامه‌های الکترونیکی	ساختاری، نحوی	ماشین بردار پشتیبان	
[۱۴]	۲۵۳ نامه‌ی الکترونیکی مربوط به ۴ کاربر	ساختاری، کلمات تابع	ماشین بردار پشتیبان	۷۰/۲
[۱۸]	۱۷۴ تا ۷۰۶ نامه‌ی الکترونیکی مربوط به ۸	لغوی(n-تایی و کلمه‌های پرسشی)، نحوی	ماشین بردار پشتیبان	۸۶/۷۴
[۲۰]	نامه‌های الکترونیکی مربوط به ۴۰ نویسنده	لغوی، نحوی، ساختاری، وابسته به محتوا	ماشین بردار پشتیبان	۸۳/۹۰ ۸۸/۳۱
[۲۲]	۱۲۰ توئیت مربوط به ۵۰ کاربر	لغوی(n-تایی)	k-نزدیک‌ترین همسایه	۷۰
[۲۴]	پست‌های ۳۰ کاربر فیس‌بوک	لغوی، نحوی، ساختاری، مختص شبکه‌های اجتماعی	SVM Light, k-نزدیک‌ترین همسایه	۷۸/۹ ۶۹/۸ ۶۵/۵
[۲۶]	پیام‌های آنی	ساختاری، علامت‌های خاص	بیز ساده	۶۸
[۲۷]	۴۰ نمونه متن جمع‌آوری شده به صورت برخط	لغوی، نحوی	k-نزدیک‌ترین همسایه	۹۵/۸۲ ۷۹/۷۵
[۲۹]	مقالات وب سایت عربی	لغوی، نحوی، ساختاری، معنایی، مختص متن	SVM Light	۹۸
[۳۰]	وبلاگ خبری Huffington post.com, CNN.com	لغوی، نشانه‌های نقطه‌گذاری، ساختاری، کلمات تابع	الگوریتم ژنتیک و تکاملی، توسعه‌ی اکتشافی ژنتیک	۴۳/۴۳ ۵۷
[۳۱]	مجموعه خبری رویترز	لغوی (n-تایی با طول متغیر)	ماشین بردار پشتیبان با هسته‌ی خطی + بهره‌ی اطلاعاتی	۷۳/۰۸

[۳۲]	مجموعه‌ی خبری رویترز، گروه‌های خبری، خدمات پیام کوتاه، نامه‌های الکترونیکی	لغوی (فرهنگ واژگان)	ماشین بردار پشتیبان، شبکه عصبی، درخت تصمیم	

جدول ۲. نتایج مربوط به پژوهش‌های انجام شده در زبان فارسی.

مرجع	مجموعه داده	ویژگی	روش	دقت
[۱]	متون ادبی فارسی (داستان و رمان فارسی)	لغوی، نحوی، ساختاری، دستوری، نشانه‌های روانی-زبانی	ماشین بردار پشتیبان، بیز ساده، درخت تصمیم	۷۱/۶ ۶۶/۳ ۷۳/۶
	نظرات کاربران در سایت هلو کیش			۵۸/۸ ۵۳ ۶۳
[۲]	متون مربوط به دانشجویان دانشگاه بوعلی سینا،	لغوی، نحوی، معنایی، وابسته به کاربرد	دلتا k-نزدیک‌ترین همسایه الگوریتم ژنتیک + دلتا الگوریتم ژنتیک + k-نزدیک‌ترین همسایه	۵۰ ۷۰ ۵۰ ۸۰
	مقالات مربوط به ۸ نویسنده‌ی هم‌عصر			۸۷ ۱۰۰ ۸۷/۵ ۱۰۰
[۳۳]	پیام‌های الکترونیکی مربوط به ۵۰ نفر از مشتریان بالقوه‌ی وب‌سایت آمازون	لغوی، نحوی، ساختاری، خاص متن	ماشین بردار پشتیبان، تجمیع هسته‌های وزن‌دار، شبکه عصبی، C۴.۵	۹۷/۶۹ ۹۸/۷۸ ۹۶/۶۶ ۹۳/۳۶
این پژوهش	نظرات فارسی مربوط به خریداران گوشی‌های سامسونگ	لغوی، نگارشی، معنایی، ساختاری، دستوری، مختص متن،	خوشه‌بندی مبتنی بر چگالی + تحلیل اجزای اصلی، خوشه‌بندی مبتنی بر	۴۸ ۵۵/۲۵
			مختص شبکه‌های اجتماعی	چگالی + همبستگی ویژگی، + K-means نسبت بهره، EM + OneR ۵۶/۵۵ ۵۹/۱۵
				شبکه‌ی بیز + همبستگی ویژگی، Bagging + OneR، جنگل‌تصادفی + تحلیل اجزای اصلی، جنگل‌تصادفی + نسبت بهره ۷۰/۶۷ ۷۶/۰۷ ۷۷/۸۷ ۷۹/۵۷

نتایج به‌دست آمده از تحقیقات گذشته گویای این مطلب است که تکنیک‌های انتخاب ویژگی سنتی که در تحقیقات گذشته جهت شناسایی نویسنده‌ی متون ساختارمند و طولانی مورد استفاده قرار گرفته و دقت بالایی داشته‌اند، در مورد متون کوتاه از دقت و کارایی کمتری برخوردارند و می‌توان با افزودن ویژگی‌های مختص شبکه‌های اجتماعی به مجموعه ویژگی‌های سنتی، تا حدی این مشکل را برطرف نمود.

با توجه به این‌که پژوهش‌های صورت گرفته در زبان فارسی در گذشته، اغلب بر روی متون ساختارمند و طولانی بوده که تنها از ویژگی‌های سنتی جهت شناسایی نویسنده‌ی متون استفاده شده است؛ در این پژوهش، تمرکز ما بر روی متون کوتاه برخط می‌باشد که جهت دقیق‌تر شدن کار شناسایی نویسنده‌ی متون کوتاه، علاوه بر استفاده از ویژگی‌های سنتی متداول که در گذشته بر روی متون فارسی مورد استفاده قرار گرفته است، ویژگی‌های مختص شبکه‌های اجتماعی را نیز به مجموعه ویژگی‌ها افزودیم تا ضمن بررسی تأثیر این ویژگی‌ها، به مقایسه‌ی کارایی الگوریتم‌های نظارتی و غیرنظارتی بپردازیم.

۳ روش پژوهش

پژوهشگران به‌منظور ایجاد امنیت در فضای مجازی، به دنبال ابزارهایی برای جلوگیری از برخی سوءاستفاده‌ها در این فضا می‌باشند. بدین منظور جهت شناسایی نویسنده‌ی سندهای ناشناس شامل نامه‌های الکترونیکی، وبلاگ‌ها، وب‌سایت‌ها، متون برخط، پست‌های شبکه‌های اجتماعی، پیام‌های کوتاه، مقالات، گروه‌های خبری و کتاب‌ها روش‌های مختلفی مورد بررسی قرار گرفته که

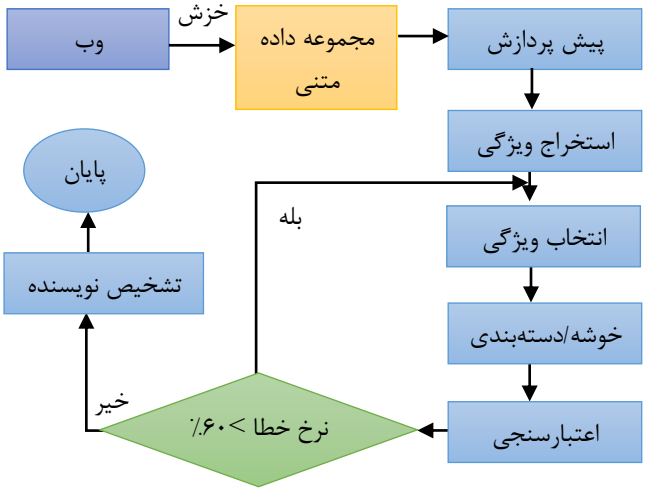
عموماً برمبنای الگوریتم‌های یادگیری ماشین بوده و شامل روش‌های یادگیری نظارتی و غیرنظارتی می‌باشند.

۳,۱ انتخاب نوع الگوریتم

روش‌های یادگیری نظارتی زمانی مناسب هستند که برچسب داده‌های آموزشی از قبل شناخته شده است. در بیشتر پژوهش‌های صورت گرفته تاکنون، الگوریتم‌های یادگیری نظارتی به‌طور عمده مورد استفاده قرار گرفته و تأثیرات ویژگی‌های سبک نگارشی متفاوت با استفاده از دسته‌بندهای ماشین بردار پشتیبان و درخت تصمیم مورد بررسی قرار گرفته‌اند [۳۰].

روش‌های یادگیری غیرنظارتی نیازی به برچسب‌گذاری داده از قبل ندارند، بلکه شناسایی بر پایه مقایسه شباهت بین یک سند ناشناخته مفروض در برابر اسناد شناخته شده با استفاده از ویژگی‌های متفاوت است، به‌طوری‌که سند ناشناخته در همان دسته‌ای قرار گیرد که توسط همان نویسنده نوشته شده است [۴]. به‌دلیل این‌که بیشتر متون در فضای وب به‌صورت ناشناس منتشر می‌گردند و اطلاعات اندکی در رابطه با نویسندگان حقیقی متون در دسترس می‌باشد، جهت طبقه‌بندی نویسندگان، روش‌های غیرنظارتی بر روی ویژگی‌های استخراج شده از نمونه‌ها اعمال شده و کار خوشه‌بندی آن‌ها به گروه‌های مختلف صورت می‌گیرد.

در این پژوهش با توجه به دردسترس بودن مجموعه داده‌ی برچسب‌گذاری شده برای نوشته‌های فارسی، هم از روش‌های نظارتی و هم از روش‌های غیرنظارتی استفاده خواهد شد و نتایج حاصل از ارزیابی‌ها با یکدیگر مقایسه می‌شوند. فرایند کلی شناسایی نویسنده‌ی متون در شکل ۱ بیان شده است.



شکل ۱. فرایند شناسایی نویسنده.

۳,۲ ویژگی‌های سبک‌شناسی

مطالعه‌ی ویژگی‌های سبک‌شناسی نشان می‌دهد که می‌توان با استفاده از سبک نگارشی افراد که نسبتاً ثابت می‌باشد، به شناسایی نویسنده‌ی متون پرداخت. این ویژگی‌ها، مربوط به سبک نگارشی نویسندگان بوده که از قبل توسط پژوهشگران تعریف شده‌اند. بر این اساس، هفت ویژگی شامل ویژگی‌های لغوی، نگارشی، ساختاری، دستوری، معنایی، ویژگی‌های مختص شبکه‌های اجتماعی و مختص متن جهت انجام این پژوهش مورد استفاده قرار گرفت. جداول ۳ تا ۹ شامل فهرست کلی ویژگی‌های استخراج شده از متن در این پژوهش می‌باشند.

- **ویژگی‌های نگارشی:** این ویژگی‌ها ساختار جمله را از دیدگاه نشانه‌های نقطه‌گذاری، تعریف مرزها و نوع جملات (پرسشی، تعجبی و خبری) توسط شکستن پاراگراف به جمله و جملات به نشانه‌های متفاوت، بیان کرده و شامل نشانه‌های نقطه‌گذاری و تعداد تکرار کلمات تابع می‌باشد [۵]. کلمات تابع، که به‌صورت ناخودآگاه توسط نویسنده تولید می‌شوند، مستقل از موضوع بوده و از نظر معنایی کم‌ارزش می‌باشند؛ اما به‌دلیل استفاده‌ی مکرر نویسنده از این کلمات و توزیع فراوان آن‌ها در متن، این ویژگی می‌تواند نتایج خوبی در تشخیص سبک نگارش افراد داشته باشد [۲].
- **ویژگی‌های معنایی:** این ویژگی‌ها که کلمات و عباراتی که جمله‌واره‌ها را به هم مربوط می‌کنند، به‌عنوان ۳ ویژگی مورد استفاده در این پژوهش می‌باشند که شامل افزوده‌های ربطی تشریحی، گسترشی و تفضیلی می‌باشند
- **ویژگی‌های ساختاری:** این ویژگی‌ها ترکیبی از جملات و پاراگراف‌ها و همچنین نحوه‌ی سازماندهی جملات درون پاراگراف‌ها و پاراگراف‌ها درون اسناد و به‌طورکلی، چیدمان متن توسط نویسنده می‌باشند.
- **ویژگی‌های لغوی:** این ویژگی‌ها به دو دسته‌ی علامت‌محور و کلمه‌محور دسته‌بندی می‌شوند. در ویژگی‌های علامت‌محور، متن به‌صورت رشته‌ای از علائم درنظر گرفته می‌شود. ویژگی‌های کلمه‌محور، متن را به‌صورت یک رشته از کلمات درنظر می‌گیرند و شامل ۱۸ ویژگی از جمله پرمایگی واژگان شامل توابع یول [۳۴]، هونور [۳۵]، برونت [۳۶]، سیشل [۳۷] و سیمسون [۳۸] می‌باشند که در روابط ۱ تا ۵ بیان گردیده‌اند.

$$\text{Yules K} = ۱۰^{۴}\left(-\frac{1}{N} + \sum_{i=1}^{V}\left(\frac{i}{N}\right)^{2}\right) \quad (۱)$$

$$\text{Honores R} = \frac{100\log_{10}^{N}}{1-\frac{count\ of\ HL}{V}} \quad (۲)$$

$$\text{Brunet w} = N^{(V)^{-\alpha}} \quad (۳)$$

$$\text{Sicheles S} = \frac{count\ of\ HD}{V} \quad (۴)$$

$$\text{Simpsons D} = \sum_{i=1}^{V} V_i \frac{i}{N}\frac{i-1}{N-1} \quad (۵)$$

که در این روابط، V تعداد کلمات یکتا، α پارامتر ثابت با مقدار ۰/۱۷، N تعداد کل کلمات، i مرتبه تکرار، HL تعداد کلمات با یک‌بار تکرار و HD تعداد کلمات با دو بار تکرار می‌باشد.

- **ویژگی‌های مختص شبکه‌های اجتماعی:** این ویژگی‌ها بیانگر حالت روحی نویسنده با استفاده از نمادها بوده و شامل نمادهای متن‌محور، نمادهای متفرقه، شکلک‌های مثبت، خنثی و منفی می‌باشند. نمادهای متن محور مانند :-) و موارد مشابه که نشانه حالت لبخند، خیلی‌شاد، اخم، عصبانی، ناراحت و گریان هستند در جدول ۱۰ و شکلک‌ها شامل علامت‌هایی با یونیکدهای معین، در جدول ۱۱ مشخص شده‌اند.
- **ویژگی‌های دستوری:** این ویژگی‌ها، مربوط به دستور زبان بوده و شامل کلماتی است که از لحاظ معنایی، معنای واژگانی کمی داشته و برای نشان دادن روابط دستوری بین کلمات مورد استفاده قرار می‌گیرند [۱].
- **ویژگی‌های مختص متن:** این ویژگی‌ها به‌منظور نمایش دقیق‌تر تفاوت سبک افراد در حوزه‌ی خاصی به‌کار می‌روند. انتخاب چنین ویژگی‌هایی وابسته به حوزه‌های کاربرد بوده و با توجه به زمینه متن انتخاب می‌گردند. در این پژوهش با توجه به مجموعه داده، انواع گوشی‌ها و تکیه کلام‌های افراد به‌عنوان ویژگی‌های منحصربه‌فرد در نظر گرفته شده است.

جدول ۳. مجموعه ویژگی‌های نگارشی.

ویژگی‌های نگارشی	
F۸۱	تعداد کل علائم نقطه‌گذاری ÷ تعداد کل علائم
F۸۲	تعداد کاما ÷ تعداد کل علائم
F۸۳	تعداد نقطه ÷ تعداد کل علائم
F۸۴	تعداد دونقطه ÷ تعداد کل علائم
F۸۵	تعداد سمیکلون ÷ تعداد کل علائم
F۸۶	تعداد علامت سؤال ÷ تعداد کل علائم
F۸۷	تعداد علامت تعجب ÷ تعداد کل علائم
F۸۸	تعداد علامت سؤال سه‌تایی ÷ تعداد کل علائم
F۸۹	تعداد علامت تعجب سه‌تایی ÷ تعداد کل علائم
F۹۰	تعداد سه‌نقطه ÷ تعداد کل علائم
F۹۱	تعداد علامت کاما سه‌تایی ÷ تعداد کل علائم
F۹۲	تعداد علامت نقل قول ÷ تعداد کل علائم
F۹۳	تعداد کلمات تابع ÷ تعداد کل کلمات

جدول ۴. مجموعه ویژگی‌های معنایی.

ویژگی‌های معنایی	
F۱۱۸	افزوده‌های ربطی تشریحی ÷ تعداد کل کلمات
F۱۱۹	افزوده‌های ربطی گسترشی ÷ عداد کل کلمات
F۱۲۰	افزوده‌های ربطی تفضیلی ÷ تعداد کل کلمات

جدول ۵. مجموعه ویژگی‌های ساختاری.

ویژگی‌های ساختاری	
F۹۴	تعداد کل خطوط ÷ تعداد کل علائم
F۹۵	تعداد کل خطوط ÷ تعداد کل کلمات
F۹۶	تعداد کل جملات ÷ تعداد کل علائم
F۹۷	تعداد کل جملات ÷ تعداد کل کلمات
F۹۸	تعداد علامت enter ÷ تعداد کل علائم
F۹۹	تعداد جملات بدون نقطه پایانی

جدول۶. مجموعه ویژگی‌های لغوی.

ویژگی	توضیح ویژگی
	ویژگی‌های علامت‌محور
$F_{۱}$	تعداد آدرس‌های URL
$F_{۲}$	تعداد کل اعداد ÷ تعداد کل علائم
$F_{۳}$	تعداد کل حروف بزرگ انگلیسی ÷ تعداد کل علائم
$F_{۴}$	تعداد کل حروف کوچک انگلیسی ÷ تعداد کل علائم
$F_{۵}$	تعداد کل حروف انگلیسی ÷ تعداد کل علائم
$F_{۶}$	تعداد کل حروف فارسی و انگلیسی ÷ تعداد کل علائم
$F_{۷}$	تعداد کل حروف فارسی ÷ تعداد کل علائم
$F_{۸}$	تعداد کل کلمات ÷ تعداد کل علائم
$F_{۹}$	تعداد کل علامت‌های خاص ÷ تعداد کل علائم
$F_{۱۰}$-$F_{۲۴}$	تعداد هر یک از علامت‌های خاص ÷ تعداد کل علائم خاص
$F_{۲۵}$-$F_{۶۰}$	تعداد هر یک از حروف انگلیسی (A-Z , a-z) ÷ تعداد کل علائم
$F_{۶۱}$	تعداد علامت فاصله ÷ تعداد کل علائم
$F_{۶۲}$	تعداد علامت نیم‌فاصله ÷ تعداد کل علائم
	ویژگی‌های کلمه‌محور
$F_{۶۳}$	تعداد کلمات فارسی ÷ تعداد کل کلمات
$F_{۶۴}$	تعداد کلمات انگلیسی ÷ تعداد کل کلمات
$F_{۶۵}$-$F_{۶۷}$	تعداد کلمات کوتاه (۱،۲و۳ حرفی) ÷ تعداد کل کلمات
$F_{۶۸}$-$F_{۷۰}$	تعداد کلمات طولانی (۶،۷ و بیشتر از ۸ حرفی) ÷ تعداد کل کلمات
$F_{۷۱}$	تعداد کلمات با یک بار تکرار ÷ تعداد کل کلمات
$F_{۷۲}$	تعداد کلمات با دو بار تکرار ÷ عداد کل کلمات
$F_{۷۳}$	تعداد کلمات یکتا ÷ تعداد کل کلمات
$F_{۷۴}$-$F_{۷۵}$	تعداد کلمات پرتکرار (۷ و بیشتر از۸) ÷ تعداد کل کلمات
$F_{۷۶}$	معیار Sichel
$F_{۷۷}$	معیار Honore

$F_{۷۸}$	معیار Brunet
$F_{۷۹}$	معیار Yule
$F_{۸۰}$	معیار Simpson

جدول ۷. مجموعه ویژگی‌های مختص شبکه‌های اجتماعی.

ویژگی‌های مختص شبکه‌های اجتماعی	
F۱۲۱	تعداد کل نمادهای متن‌محور ÷ تعداد کل علائم
F۱۲۲	تعداد کل شکلک‌ها ÷ تعداد کل علائم
F۱۲۳	تعداد کل نمادهای متفرقه ÷ تعداد کل علائم
F۱۲۴	تعداد کل شکلک‌های مثبت ÷ تعداد کل علائم
F۱۲۵	تعداد کل شکلک‌های خنثی ÷ تعداد کل علائم
F۱۲۶	تعداد کل شکلک‌های منفی ÷ تعداد کل علائم

جدول ۸. مجموعه ویژگی‌های دستوری.

ویژگی‌های دستوری	
F۱۰۰	تعداد اسامی ÷ تعداد کل کلمات
F۱۰۱	تعداد حروف اضافه ÷ تعداد کل کلمات
F۱۰۲	تعداد حروف ربط ÷ تعداد کل کلمات
F۱۰۳	تعداد حروف ندا ÷ تعداد کل کلمات
F۱۰۴	تعداد ضمایر فاعلی ÷ تعداد کل کلمات
F۱۰۵	تعداد ضمایر مفعولی ÷ تعداد کل کلمات
F۱۰۶	تعداد ضمایر اشاره ÷ تعداد کل کلمات
F۱۰۷	تعداد ضمایر پرسشی ÷ تعداد کل کلمات
F۱۰۸	تعداد کلمات مربوط به قطعیت ÷ تعداد کل کلمات
F۱۰۹	تعداد کلمات مربوط به شک و تردید ÷ تعداد کل کلمات
F۱۱۰	تعداد کلمات مربوط به رنگ‌ها ÷ تعداد کل کلمات
F۱۱۱	تعداد ضمایر ÷ تعداد کل کلمات
F۱۱۲	تعداد صفات ÷ تعداد کل کلمات
F۱۱۳	تعداد قیود ÷ تعداد کل کلمات
F۱۱۴	تعداد گروه اسمی
F۱۱۵	تعداد گروه فعلی
F۱۱۶	تعداد گروه قیدی
F۱۱۷	تعداد گروه صفتی یا مسندی

جدول ۹. مجموعه ویژگی‌های مختص متن.

ویژگی‌های مختص متن	
F۱۲۷	تعداد کلمات مربوط به متن ÷ تعداد کل کلمات
F۱۲۸	تعداد گوشی‌های سری A ÷ کلمات مربوط به متن
F۱۲۹	تعداد گوشی‌های سری C ÷ کلمات مربوط به متن
F۱۳۰	تعداد گوشی‌های سری E ÷ کلمات مربوط به متن
F۱۳۱	تعداد گوشی‌های سری J و G ÷ کلمات مربوط به متن
F۱۳۲	تعداد گوشی‌های سری S ÷ کلمات مربوط به متن
F۱۳۳	تعداد گوشی‌های سری Note ÷ کلمات مربوط به متن
F۱۳۴	تعداد تکیه کلام‌ها ÷ تعداد کل کلمات

جدول ۱۰. شکلک‌ها.

Positive	:-) :) :o) :] :۳ :c) :> =] ۸) =) :} :^) :-d :d ۸-d ۸d x-d xd X-d xd =-d =d =-۳ =۳ b^d :-)) ^_^ (^_^)/ (^^)/(^o^) >^_^< (^.^) (^_^.) (^_^) (^^) ;;) ^_^ (#^.^#)(y) ;d :* :") (*^۰^*) ;) :') !(^^)! (*^^)v (^^)v (^_^)v
Negative	>:[:-(:(:-c :c :-< -_,- :< :-[:[:{ :-|| :@ qq (>_<)(>_<)> -____- </۳.___. :]x \m/ :| (/_;) (t_t) (;_;)(;_:) (;o;) (:_;) (tot) -___- (-.-) (-_-) x_x --" -__- :/ (*_*) (*_*; (+_+) (@_@) :-/ -_- (~o~) (~_~) :-| :'(._.
Neutral	:o :p

جدول ۱۱. یونیکد نمادهای احساسی.

Face positive	\U+۱F۶۰۰\U+۱F۶۰۱\U+۱F۶۰۲\U+۱F۹۲۳\U+۱F۶۰۳\U+۱F۶۰۴\U+۱F۶۰۵\U+۱F۶۰۶\U+۱F۶۰۹\U+۱F۶۰A\U+۱F۶۰B\U+۱F۶۰E\U+۱F۶۰D\U+۱F۶۱۸\U+۱F۶۱۷\U+۱F۶۱۹\U+۱F۶۱A\U+F۲۶۳A\U+۱F۶۴۲\U+۱F۹۱۷\U+۱F۹۲۹
Face neutral	\U+۱F۹۱۴\U+۱F۹۲۸\U+۱F۶۱۰\U+۱F۶۱۱\U+۱F۶۳۶\U+۱F۶۴۴\U+۱F۶۰F\U+۱F۶۲۳\U+۱F۶۲۵\U+۱F۶۲E\U+۱F۹۱۰\U+۱F۶۲F\U+۱F۶۲A\U+۱F۶۲B\U+۱F۶۳۴\U+۱F۶۰C\U+۱F۶۱B\U+۱F۶۱C\U+۱F۶۱D\U+۱F۹۲۴\U+۱F۶۱۲\U+۱F۶۱۳\U+۱F۶۱۴\U+۱F۶۱۵\U+۱F۶۴۳\U+۱F۹۱۱\U+۱F۶۳۲
Face negative	\U+۲F۶۳۹\U+۱F۶۴۱\U+۱F۶۱۶\U+۱F۶۱E\U+۱F۶۱F\U+۱F۶۲۴\U+۱F۶۲۲\U+۱F۶۲D\U+۱F۶۲۶\U+۱F۶۲۷\U+۱F۶۲۸\U+۱F۶۲۹\U+۱F۹۲F\U+۱F۶۲C\U+۱F۶۳۰\U+۱F۶۳۱\U+۱F۶۳۳\U+۱F۹۲A\U+۱F۶۳۵\U+۱F۶۲۱\U+۱F۶۲۰\U+۱F۹۲C

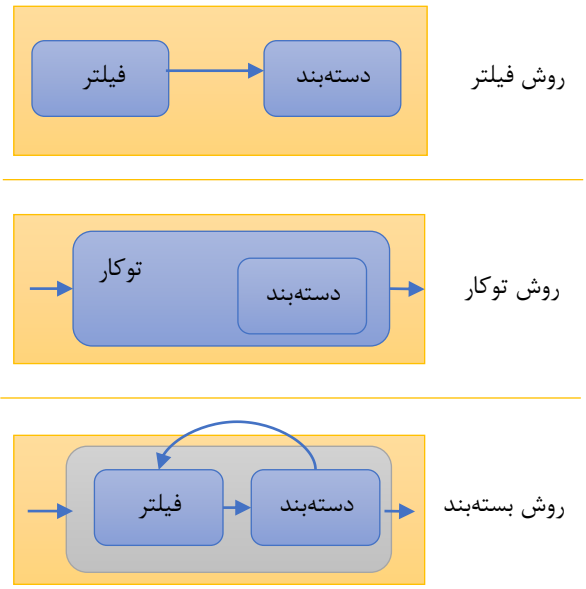
۳,۳ انتخاب ویژگی‌ها

در بسیاری موارد پس از مرحله اسـتخراج ویژگـی، بـا داده‌هـایی بـا ابعاد زیاد مواجه هستیم که کارایی الگوریتم یادگیری را کاهش داده و بـه مشـکل بـیش‌بـرازش[۱] منجـر مـی‌گردنـد. بنـابراین، انتخـاب

[۱] Overfitting

ویژگی‌های مرتبط و ضروری در مرحله‌ی پیش‌پردازش از اهمیتی بنیادین برخوردار است. در این مرحله، با انتخاب زیرمجموعه‌ای از ویژگی‌های اولیه، ابعاد داده‌ها کاهش می‌یابد [۳۹]. برخلاف روش‌های مبتنی بر استخراج ویژگی، این نوع روش‌ها معنای اصلی ویژگی‌ها را بعد از کاهش حفظ می‌کنند. الگوریتم‌های انتخاب ویژگی همان‌گونه که در شکل ۲ نشان داده شده است به سه دسته‌ی فیلتر، بسته‌بند[۱] و توکار[۲] تقسیم می‌شوند [۳۹]. در روش اول، الگوریتم مستقل از دسته‌بند است و در روش‌های دوم و سوم همان‌گونه که در شکل مشخص شده است، برای انتخاب ویژگی با دسته‌بند در تعامل است.

چهار الگوریتم انتخاب ویژگی که در این پژوهش مورد استفاده قرار گرفته شامل همبستگی ویژگی[۳] با تابع ارزیاب بر اساس همبستگی با فیلد هدف، نرخ بهره[۴] با تابع ارزیاب بر اساس آنتروپی با فیلد هدف، OneR با تابع ارزیاب شامل یک دسته‌بند ساده و الگوریتم تحلیل اجزای اصلی با تابع ارزیاب بر اساس کوواریانس می‌باشند.



شکل ۲. انواع الگوریتم‌های انتخاب ویژگی.

الگوریتم همبستگی ویژگی یک الگوریتم از نوع فیلتر است که بر اساس میزان همبستگی بین ویژگی‌ها کار می‌کند. در این الگوریتم، بر اساس رابطه‌ی ۶ ابتدا ضریب همبستگی پیرسون برای دو ویژگی X و Y به‌صورت زیر محاسبه می‌شود:

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\sigma^2(X)+\sigma^2(Y)}} \qquad (۶)$$

که در این رابطه cov کوواریانس و σ واریانس است. سپس بر اساس مقدار به‌دست آمده که عددی بین ۱- و ۱ است میزان همبستگی و درنتیجه تکراری بودن ویژگی‌ها مشخص می‌شود.

الگوریتم نرخ بهره نسخه‌ی نرمال شده‌ی الگوریتم بهره‌ی اطلاعاتی است. برای نرمال‌سازی در این الگوریتم، مقدار بهره‌ی اطلاعاتی بر آنتروپی ویژگی نسبت به دسته تقسیم می‌شود. این کار با استفاده از روابط زیر انجام می‌شود.

$$\text{Gain Ratio (C,A)} = \frac{H(C)-H(C|A)}{H(A)} \qquad (۷)$$

در این رابطه C دسته، A ویژگی و H تابع آنتروپی است که از رابطه‌ی زیر محاسبه می‌گردد.

$$H(S) = -\sum_{j=1}^{j=m} p_j \log_2 p_j \qquad (۸)$$

در این رابطه p احتمالی است که برای آن یک مقدار از فضای نمونه‌ها (S) رخ می‌دهد.

الگوریتم OneR یک الگوریتم ساده است که به ازای هر ویژگی یک قانون ایجاد کرده و بر اساس میزان خطای هر قانون آن‌ها را ارزیابی می‌کند. سپس قانونی که کمترین نرخ خطا را داشته باشد انتخاب می‌کند. این روش یک الگوریتم توکار است.

الگوریتم تحلیل اجزای اصلی بر اساس انتخاب ویژگی‌هایی که به بهترین شکل پخشی داده در یک مجموعه داده را نشان می‌دهند، کار می‌کند. این الگوریتم ویژگی‌هایی که با هم همپوشانی زیادی دارند را ادغام می‌کند. این الگوریتم نیز یک روش فیلتر می‌باشد.

۳,۴ خوشه‌بندی

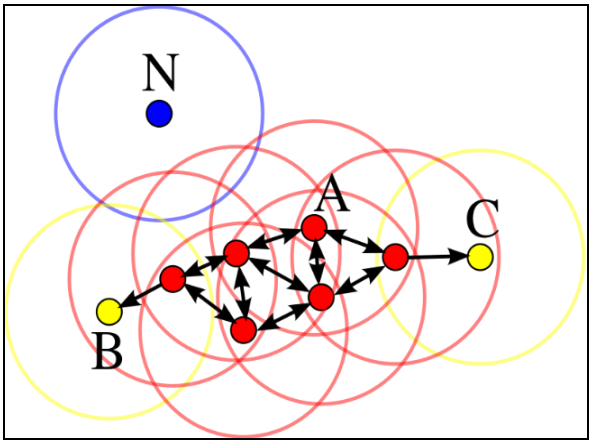
هدف خوشه‌بندی یافتن یک ساختار درون یک مجموعه از داده‌های بدون برچسب بوده و خوشه‌ها شامل مجموعه‌ای از داده‌ها می‌باشد که شباهت بیشتری به یکدیگر داشته باشند. در خوشه‌بندی سعی می‌شود تا داده‌ها به خوشه‌هایی تقسیم شوند که شباهت درون هر خوشه بیشینه و شباهت داده‌های خوشه‌های متفاوت کمینه شود.

الگوریتم‌های خوشه‌بندی اعمال شده در این پژوهش شامل الگوریتم K-means، EM و خوشه‌بندی مبتنی بر چگالی[۵] می‌باشد. به این دلیل که الگوریتم‌های K-means و EM الگوریتم‌های شناخته شده‌تری می‌باشند در این قسمت فقط الگوریتم DBSCAN مختصراً توضیح داده خواهد شد.

همان گونه که در شکل ۳ دیده می‌شود، الگوریتم DBSCAN بر اساس دسته‌بندی داده‌ها به سه دسته‌ی هسته، دسترسی‌پذیر و نویز تقسیم می‌شوند.

---

[۱] Wrapper
[۲] Embeded
[۳] Correlation Attribute (CA)
[۴] Gain Ratio
[۵] Density-based spatial clustering of applications with noise (DBSCAN)



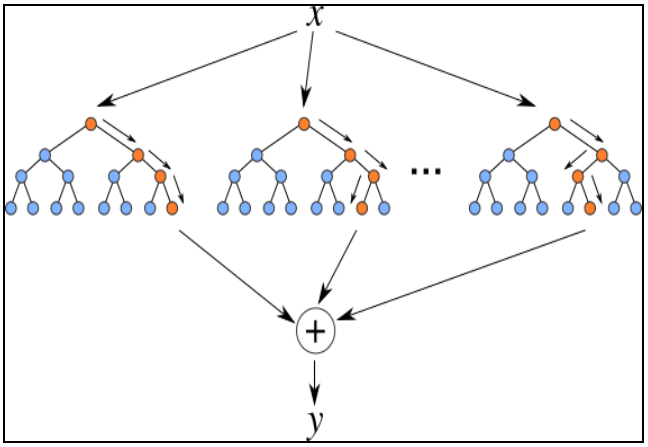
شکل ۳. توصیف الگوریتم DBSCAN. نقاط A و سایر نقاط قرمز هسته و نقاط زرد دسترسی‌پذیر از نقاط هسته می‌باشند. نقطه‌ی N نویز است.

در این الگوریتم هر نمونه که با یک نقطه نشان داده می‌شود، به شرطی که به تعداد یک متغیر از پیش مشخص شده (*minPts*) در یک فاصله‌ی از پیش مشخص (ε) در اطراف آن قرار داشته باشند (نقاط دسترسی‌پذیر)، هسته تلقی می‌شود و در غیر این صورت نویز هستند. به‌عنوان مثال در شکل ۳ اگر $minPts=۴$ باشد، نقطه A و سایر نقاط قرمز هسته هستند زیرا ناحیه‌ی اطراف آن‌ها با شعاع دارای حداقل ۴ نقطه است. به این دلیل که تمام این نقاط از همدیگر ε قابل دسترسی هستند، همگی در یک خوشه قرار می‌گیرند. نقاط B و C اگرچه هسته نیستند اما نقاط دسترسی پذیر هستند و درنتیجه در یک خوشه با هسته قرار می‌گیرند. نهایتاً نقطه‌ی N نویز است زیرا نه هسته است و نه نقطه دسترسی‌پذیر.

۳,۵ دسته‌بندی

هدف دسته‌بندی تمایز بین داده‌های برچسب خورده درون تعداد دسته‌های از پیش معین می‌باشد. معمولاً از روش‌های یادگیری ماشین برای دسته‌بندی استفاده می‌شود [۳۹]. الگوریتم‌های دسته‌بندی اعمال شده در این پژوهش شامل الگوریتم‌های شبکه‌ی بیز، جنگل تصادفی[۱] و Bagging می‌باشد. این سه الگوریتم از سه خانواده‌ی مختلف انتخاب شده‌اند و در نحوه‌ی محاسبه‌ی دسته‌ها تفاوت ساختاری دارند.

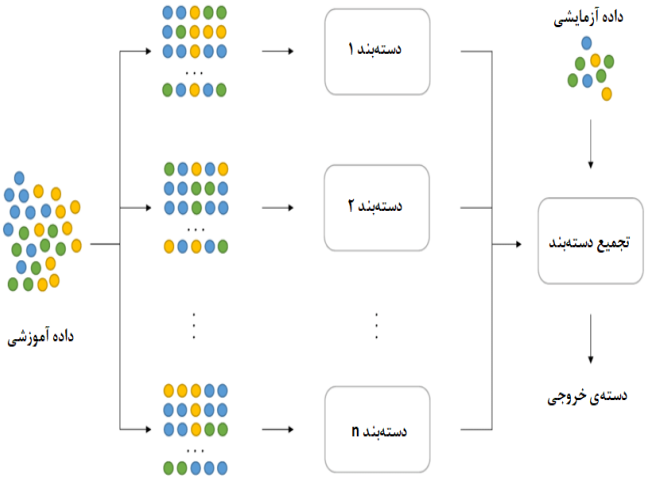
شبکه‌ی بیز یک نوع مدل گرافیکی احتمالی است که از استنتاج بیز برای محاسبه‌ی احتمالات استفاده می‌کند. هدف شبکه‌ی بیز مدل کردن وابستگی شرطی و درنتیجه علیت با استفاده از یال‌ها در یک گراف جهت‌دار می‌باشد. جنگل‌های تصادفی یا جنگل‌های تصمیم‌گیری تصادفی یک روش یادگیری گروهی[۲] برای دسته‌بندی است که با ساختن تعدادی درخت تصمیم‌گیری در زمان آموزش دسته‌ای را که میانه دسته‌ها باشد یا میانگین پیش‌بینی درخت‌ها (در مواردی که دسته‌ها عددی هستند) را به‌عنوان خروجی برمی‌گرداند. نمای کلی این الگوریتم در شکل ۴ آورده شده است.



شکل ۴. نمای کلی الگوریتم جنگل تصادفی. خروجی (y) بر اساس تجمیع خروجی درخت‌های تصادفی با ورودی x انتخاب می‌شود.

الگوریتم Bagging یک الگوریتم تخمین گروهی است که تعدادی دسته‌بند را روی زیرمجموعه‌های تصادفی از داده‌ی آموزشی آموزش داده و پیش‌بینی آن‌ها را با روش‌هایی مثل رأی‌گیری تجمیع می‌کند. نمای کلی این الگوریتم در شکل ۵ آورده شده است.

همان‌گونه که از مقایسه‌ی شکل‌های ۴ و ۵ مشخص است، تفاوت دو الگوریتم جنگل تصادفی و Bagging در این است که در الگوریتم Bagging داده‌ی آموزشی به چند بخش تصادفی تقسیم می‌شود و هر دسته‌بند روی بخشی از داده‌ی آموزشی آموزش می‌بیند.



شکل ۵. نمای کلی الگوریتم Bagging. خروجی (y) بر اساس تجمیع خروجی دسته‌بندهایی که هر یک بر روی بخشی از مجموعه داده‌ها آموزش داده شده‌اند، انتخاب می‌شود.

۴ پیاده‌سازی

[۱] Random forest

[۲] Ensemble learning

۴,۱ مجموعه داده

همان‌طور که در شکل ۱ مشاهده می‌شود، جمع‌آوری اسناد الکترونیکی موجود در محیط وب اولین قدم جهت شناسایی نویسنده می‌باشد. مجموعه داده‌ی استخراج شده در این پژوهش، شامل ۱۰۰۰ نظر فارسی در رابطه با محصولات گوشی‌های سامسونگ است که مربوط به سال‌های ۲۰۱۵ و ۲۰۱۶ می‌باشد. در این مجموعه، نظرات مربوط به ۹ نویسنده به‌همراه شناسه‌ی کاربری آن‌ها آورده شده است. متوسط نظرات برای نویسندگان ۱۲۵ نظر با تعداد ۱ تا ۱۹ جمله‌ی ۹ تا ۵۰۵ کلمه‌ای بوده است. نمونه‌ای از رکوردهای این مجموعه داده، مربوط به دو کاربر در جدول ۱۲ قابل مشاهده می‌باشد.

از آنجایی که این متون غیرساخت‌یافته بوده و استخراج دانش از چنین مجموعه‌ای در ابتدای امر کاری غیرممکن می‌باشد، باید با اعمال فرایند متن‌کاوی به‌صورتی تبدیل شود که برای الگوریتم قابل درک باشد. گام اول در چنین فرایندی معمولاً پیش‌پردازش است؛ در این فرایند سندهای متن خام به‌عنوان ورودی دریافت شده و خروجی آن مجموعه‌ای از کلمات است که در مدل فضای بردار[۱] (VS) مورد استفاده قرار می‌گیرد.

۴,۲ فرایند پیش‌پردازش متون

فرایند کلی پیش‌پردازش صورت گرفته در این پژوهش که در شکل ۶ آمده است، تنها جهت استخراج ویژگی‌های دستوری بوده و به‌دلیل این‌که هدف اصلی عدم تغییر در سبک نگارش نویسندگان می‌باشد، استخراج سایر ویژگی‌ها نیازی به مرحله پیش‌پردازش نداشته و همان متون اولیه مورد استفاده قرار گرفته است.

به‌دلیل این‌که در استخراج ویژگی‌های دستوری، هدف نهایی برچسب‌گذاری کلمات فارسی می‌باشد، تمامی نشانه‌ها و کلمات غیرفارسی حذف گردید و تنها کلمات فارسی موجود در متن باقی ماند. سپس، متن حاصل توسط ابزارهای پردازش زبان طبیعی موجود که مختص زبان فارسی می‌باشد نرمال‌سازی گشت و در گام بعد، ویژگی‌های سبکی از آن‌ها استخراج گردید. از ابزارهای پردازش زبان طبیعی رایج که در زمینه‌ی پیش‌پردازش متون فارسی می‌توان از آن‌ها بهره جست، می‌توان به کتابخانه هضم و فردوس‌نت[۲] اشاره نمود که در این پژوهش نیز نرمال‌سازی متون با کمک این ابزارها صورت گرفت. فرایند نرمال‌سازی صورت گرفته توسط این نرم‌افزارها شامل مراحل ذیل می‌باشد:

- اصلاح کدینگ: جایگزین کردن حروف غیرفارسی که بیش از یک یونیکد دارند، با حروف معادل فارسی.
- حذف اعراب، تنوین، تشدید و علائمی از این قبیل.
- اصلاح نیم‌فاصله: استفاده از نیم‌فاصله جهت تعیین مرزهای کلمات مرکب که می‌توانند به سه شکل متفاوت چسبیده، با نیم‌فاصله و یا فاصله نوشته شوند.
- تبدیل کلمات عامیانه به رسمی: جهت از بین بردن ابهام در کلمات و اصطلاحات عامیانه و محاوره‌ای که پردازش زبان طبیعی را با مشکل مواجه می‌سازند.
- واحدسازی: استفاده از فاصله جهت تعیین مرزهای کلمات.
- برچسب‌زنی اجزای کلام: انتساب برچسب‌های واژگانی به کلمات و نشانه‌های تشکیل دهنده‌ی یک متن جهت نشان دادن نقش کلمات و نشانه‌ها در جمله. پس از طی مراحل اولیه‌ی نرمال‌سازی، جهت تعیین نوع کلمات از قبیل اسامی، افعال، حروف، ضمایر و صفات نوبت برچسب‌گذاری می‌باشد که با استفاده از روش‌های مبتنی بر قاعده[۳]، صورت می‌گیرد و برچسب مناسب با استفاده از قواعد دستوری و زبان‌شناسی انتخاب می‌شود.
- تجزیه‌ی جملات به اجزای تشکیل دهنده‌ی خود شامل گروه‌های اسمی، فعلی، قیدی و ....
- درخت نحوی با استفاده از ساختار لغات، موقعیت و ترتیب لغات در جمله، حروف یا عبارات قبل و بعد از آن‌ها و نوع لغات، ایجاد می‌شود. درواقع، این عملیات با توجه به ریخت‌شناسی (مطالعه ساختار و حالت‌های مختلف یک کلمه) در دستور زبان فارسی صورت می‌گیرد.

جدول ۱۲. نمونه‌ای از رکوردهای مجموعه داده.

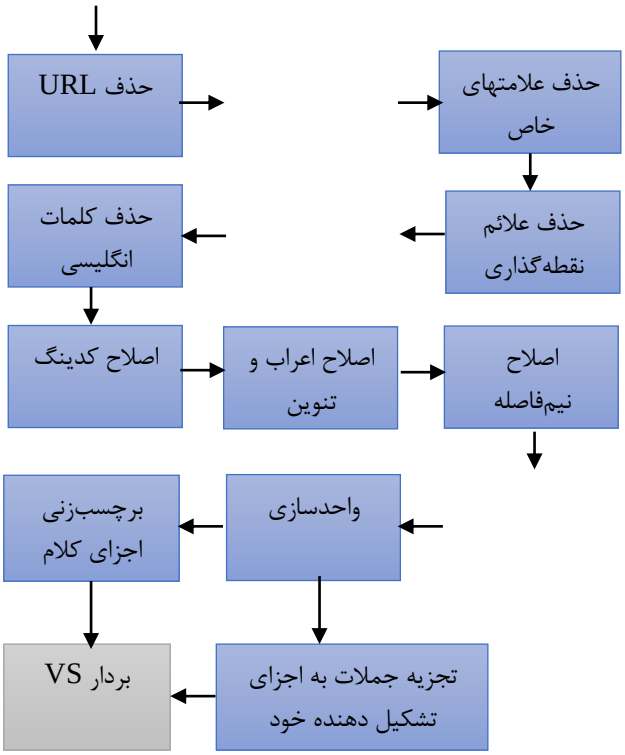
شناسه کاربر	نظرات
۹۵۴۴۶	ببین الان بهترین مدلی که به ایران وارد شده همون مدل C یا CD هستش که من مدل دو سیم دارم الان سه ماه دارم استفاده میکنم هیچ ایرادی نتونستم بهش بگیرم مصرف باطریش هم معمولیه و نسبت به سخت افزار و ... خیلی طبیعیه هیچ مشکلی در آنتن دهی نداره .
۹۵۴۴۶	اگه بودجه براتون مهم نیس ۱۰۰ درصد نوت ۵ بهترینه; باطری نوت ۵ از نوت ۴ هم بهتره
۹۵۴۴۶	سلام بله خیلی بهتر از نسل های قبلی هستش مولیتی تسکینگ هم با آپدیتی که اخیرا داده به مراتب بهتر شده اما بی شک میگم باز سرعت اجرای برنامه ها فوق العادست
۱۰۴۳۷۹	مشکل رم گوشی های سامسونگ فقط تو نوت ۵ و

[۱] Vector Space

[۲] ابزارهای پردازش متون زبان فارسی، آزمایشگاه فناوری وب دانشگاه فردوسی مشهد، ۱۳۹۱.

[۳] Rule-based

	گلکسی اس ۶ بود الان اس ۷ و اس ۷ اج چنین مشکلی ندارن میتونید ی تحقیق بکنید و خودتون از نزدیک ببینید گوشی جی ۷ هم با اینکه رمش ۱,۵ هیتش کوچکترین لگ و تاخیری نداره
۱۰۴۳۷۹	مقایسه سی ۷ و آ ۷ ۲۰۱۶ حتما ببینید..; http://www.aparat.com/v/vFnJW
۱۰۴۳۷۹	من الان چند ماهه این گوشی دستمه مدل فورجی; خدایش تاحالا هیچ مشکلی باش نداشتم و دارم ازش لذت میبرم



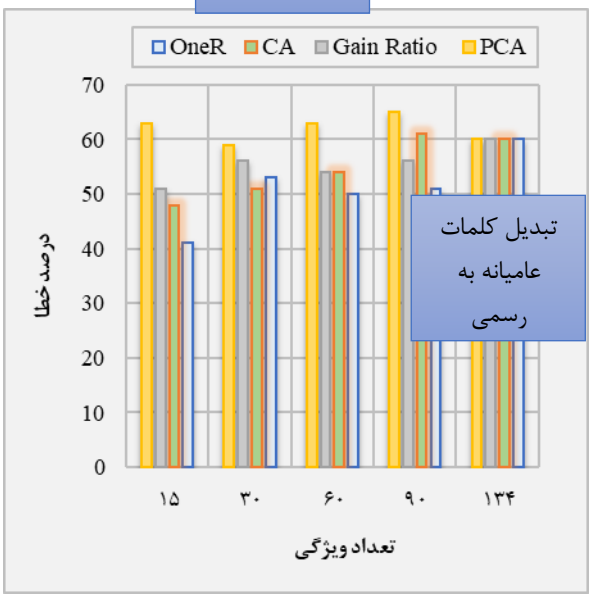
شکل ۶. فرایند پیش‌پردازش متون فارسی.

۴,۳ تکمیل فرایند پیاده‌سازی

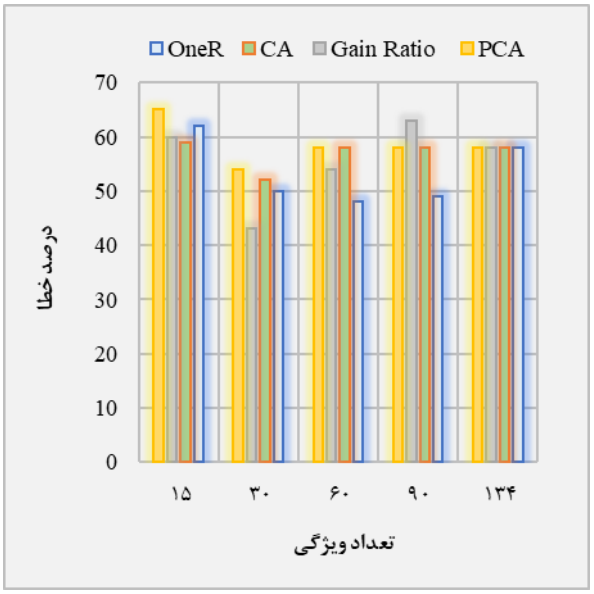
مرحله‌ی بعد شامل استخراج ویژگی‌ها است که در آن برای هر سند متنی یک بردار ۱۳۴ بعدی شامل ویژگی‌های لغوی، نگارشی، ساختاری، دستوری، معنایی، ویژگی‌های مختص شبکه‌های اجتماعی و مختص متن جهت بازنمایی مقادیر تولید شد؛ به‌طوری‌که این ویژگی‌ها دارای تمام یا بخش اعظمی از اطلاعات موجود در متون اولیه بودند. در ادامه، فرایند انتخاب ویژگی، خوشه‌بندی و دسته‌بندی بر روی بردار ویژگی‌ها با استفاده از الگوریتم‌های ذکر شده در بخش‌های ۳,۳، ۳,۴ و ۳,۵ صورت گرفت.

۵ تحلیل نتایج

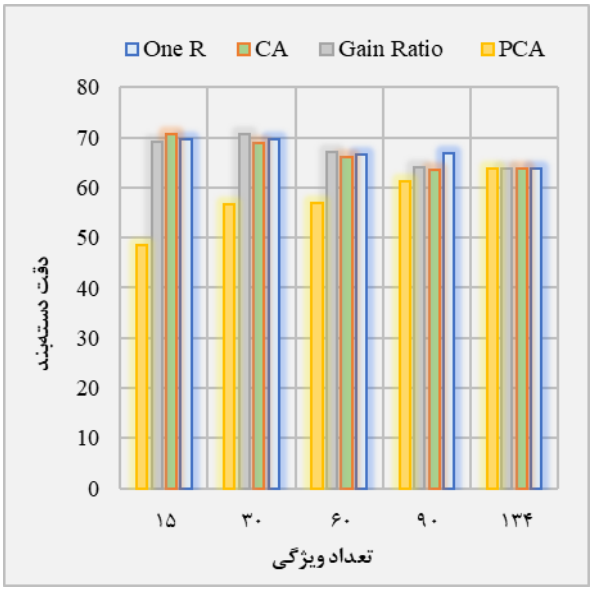
در این پژوهش، به‌منظور اعمال الگوریتم‌های انتخاب ویژگی بر روی ویژگی‌های استخراج شده و همچنین خوشه‌بندی و دسته‌بندی با استفاده از ویژگی‌های انتخاب شده، از بسته نرم‌افزاری وکا استفاده شد. جهت انتخاب ۱۵، ۳۰، ۶۰، ۹۰ و ۱۳۴ ویژگی برتر جهت شناسایی نویسنده، در مرحله‌ی رتبه‌بندی ویژگی‌ها، الگوریتم‌های نرخ بهره، الگوریتم تحلیل اجزای اصلی، همبستگی ویژگی و OneR بر روی مجموعه داده اعمال و سپس با استفاده از الگوریتم‌های EM، K-means و DBSCAN کار خوشه‌بندی نمونه‌ها صورت گرفت. سپس برای مقایسه‌ی نتایج در حالت نظارت شده و غیرنظارتی، دسته‌بندی نمونه‌ها با استفاده از بیز، جنگل تصادفی و Bagging انجام الگوریتم‌های مذکور، از درصد خطای به‌دست آمده برای خوشه‌بندی و دقت (تعداد ی دسته‌بندی شده) برای دسته‌بندی استفاده رسی‌های مربوط به روش‌های غیرنظارتی در شکل‌های ۷ تا ۹ و روش‌های نظارتی در شکل‌های ۱۰ تا ۱۲ قابل



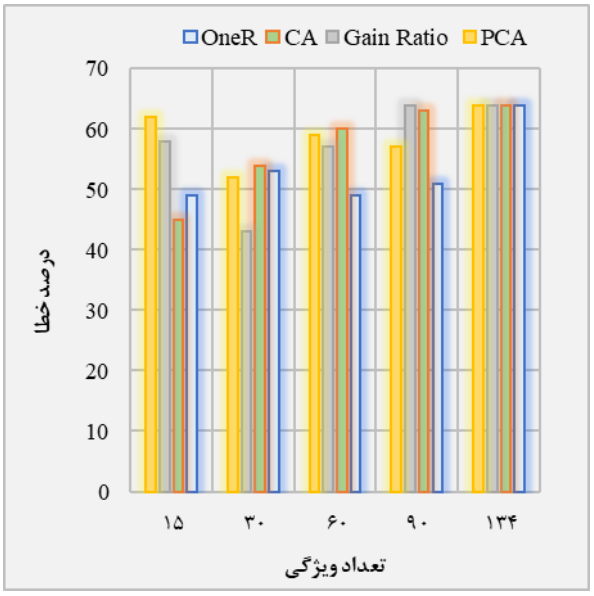
شکل۷. درصد خطای حاصل از اعمال الگوریتم EM .



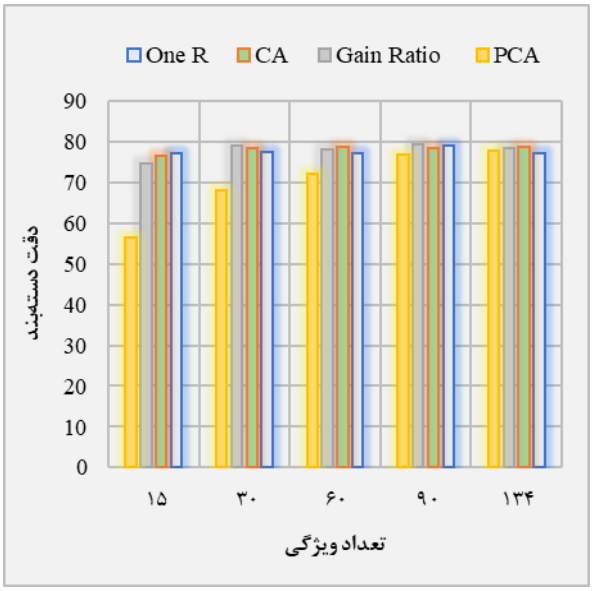
شکل ۸. درصد خطای حاصل از اعمال الگوریتم K-means.



شکل ۱۰. دقت دسته‌بندی مربوط به الگوریتم شبکه‌ی بیز.

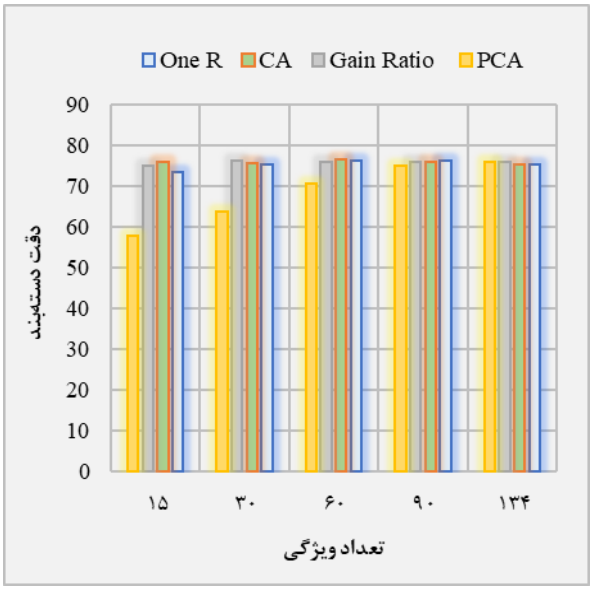


شکل ۹. درصد خطای حاصل از اعمال الگوریتم DBSCAN.



شکل ۱۱. دقت دسته‌بندی مربوط به الگوریتم جنگل تصادفی.

با دقت در نمودارهای شکل‌های ۷ تا ۹ ملاحظه می‌شود که در روش غیرنظارتی، بهترین نتیجه مربوط به زمانی است که ۱۵ ویژگی برتر توسط الگوریتم انتخاب ویژگی OneR انتخاب شده و سپس توسط الگوریتم خوشه‌بندی EM به خوشه‌های مجزا تقسیم‌بندی شده است. همان‌طور که در شکل ۷ مشاهده می‌شود با کاهش ابعاد ویژگی‌ها از ۱۳۴ به ۱۵ ویژگی، نرخ خطا کاهش یافته و درصد تشخیص درست افزایش می‌یابد. به‌طوری‌که در بهترین حالت الگوریتم‌های غیرنظارتی به دقت ۵۹/۱۶٪ می‌رسند. از این مشاهده می‌توان چنین نتیجه گرفت که افزودن ویژگی‌های بی‌فایده نه تنها منجر به بهبود دقت نمی‌شود، بلکه ممکن است موجب کاهش دقت دسته‌بند نیز گردد.



شکل ۱۲. دقت دسته‌بندی مربوط به الگوریتم Bagging.

همچنین با دقت در نمودارهای شکل‌های ۱۰ تا ۱۲ مشاهده می‌شود که در روش نظارتی، بهترین عملکرد مربوط به الگوریتم جنگل تصادفی به‌همراه نسبت بهره برای ۹۰ ویژگی می‌باشد که دقت ۷۹/۵۷ درصد کسب گردیده است.

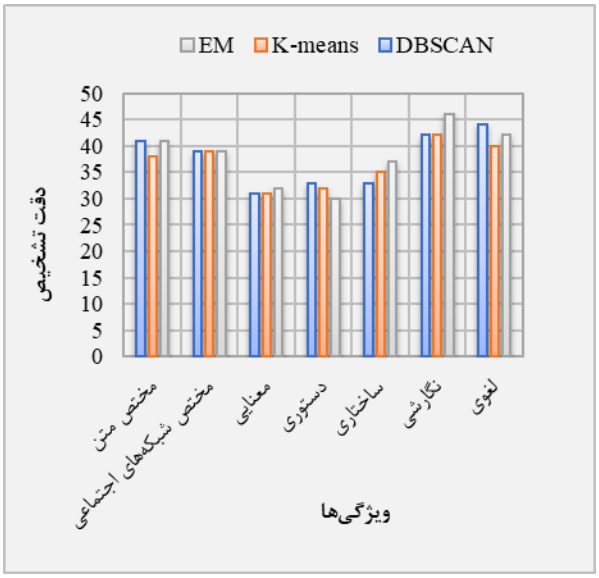
خلاصه‌ی نتیجه‌ی ارزیابی‌ها شامل ترتیب قرار گرفتن ۱۵ ویژگی برتر انتخاب شده با استفاده از الگوریتم انتخاب ویژگی OneR، به‌همراه گروه مربوط به هر ویژگی، در جدول ۱۳ قابل مشاهده می‌باشد. بررسی دقیق‌تر این ۱۵ ویژگی برتر شامل ویژگی‌های F۳، F۶۴، F۶۳ ، F۶۵، F۴۴، F۹۰، F۸۳، F۴۱، F۵، F۱۳۱، F۸۰، F۱۳۲، F۹۵، F۸۱ و F۱۱۵ بیانگر این مطلب است که کدام‌یک از ویژگی‌های سبک‌شناسی که در گذشته برای متون طولانی و ساختارمند مورد استفاده قرار گرفته و کار شناسایی نویسنده را با دقت بالایی انجام داده‌اند، درمورد متون کوتاه نیز می‌توانند به همان اندازه مؤثر واقع شوند.

جدول ۱۳. بررسی ۱۵ ویژگی برتر.

با توجه به جدول ۱۳ با مقایسه‌ی ویژگی‌ها مشاهده می‌شود که در این پژوهش، ویژگی‌های لغوی (علامت‌محور و کلمه‌محور) بیشترین تأثیر را در شناسایی نویسنده‌ی متون کوتاه داشته و پس از آن‌ها به ترتیب ویژگی‌های نگارشی، مختص متن، ساختاری و دستوری قرار گرفته‌اند.

به‌دلیل این‌که مجموعه داده‌ی مورد استفاده در این پژوهش مربوط به نظرات خریداران گوشی می‌باشد، حروف s ، j و g در جدول ۱۳ مربوط به حروف به‌کار رفته در مدل‌های گوشی بوده که افراد نظرات خود را درمورد آن‌ها بیان نموده‌اند. این حروف جزء ویژگی‌های لغوی (علامت‌محور) بوده و به دلیل پرتکرار بودن این حروف در مجموعه داده، موجب بالارفتن دقت ویژگی لغوی در میان سایر ویژگی‌ها گردیده‌اند.

در ادامه به‌منظور بررسی تأثیر ویژگی‌های سبک‌شناسی، هر یک از ویژگی‌های نگارشی، لغوی، معنایی، دستوری، ساختاری، مختص متن و مختص شبکه‌های اجتماعی به‌طور مجزا در نظر گرفته شد و سپس الگوریتم‌های خوشه‌بندی بر روی داده‌های موجود اعمال شد. با توجه به شکل ۱۳ می‌توان نتیجه گرفت که ویژگی‌های نگارشی بیشترین تأثیر را در شناسایی نویسنده‌ی متون کوتاه داشته و پس از آن به ترتیب ویژگی‌های لغوی ، مختص متن، مختص شبکه‌های اجتماعی، ساختاری، دستوری و معنایی قرار گرفته‌اند.



شکل ۱۳. بررسی ویژگی‌های سبک‌شناسی به‌طور مجزا.

۶ بحث

ویژگی	شرح ویژگی	گروه ویژگی
F۳	تعدادکل حروف بزرگ انگلیسی ÷ تعدادکل علائم	لغوی-علامت‌محور
F۶۴	تعداد کلمات انگلیسی ÷ تعداد کل کلمات	لغوی- کلمه محور
F۶۳	تعداد کلمات فارسی ÷ تعداد کل کلمات	لغوی- کلمه محور
F۶۵	تعداد کلمات کوتاه ۱ حرفی ÷ تعداد کل کلمات	لغوی- کلمه محور
F۴۴	تعداد حروف انگلیسی J و j ÷ تعداد کل علائم	لغوی-علامت‌محور
F۹۰	تعداد سه‌نقطه ÷ تعداد کل علائم	نگارشی
F۸۳	تعداد نقطه ÷ تعداد کل علائم	نگارشی
F۴۱	تعداد حروف انگلیسی G و g ÷ تعداد کل علائم	لغوی-علامت‌محور
F۵	تعداد کل حروف انگلیسی ÷ تعداد کل علائم	لغوی-علامت‌محور
F۱۳۱	تعداد گوشی‌های سری J و G ÷ کلمات مختص متن	مختص متن
F۸۰	معیار Simpson	لغوی-کلمه محور
F۱۳۲	تعداد گوشی‌های سری S ÷ کلمات مربوط به متن	مختص متن
F۹۵	تعداد کل خطوط ÷ تعداد کل کلمات	ساختاری
F۸۱	تعداد کل علائم نقطه‌گذاری ÷ تعداد کل علائم	نگارشی
F۱۱۵	تعداد گروه فعلی	دستوری

۶,۱ مقایسه‌ی روش‌های نظارتی و غیرنظارتی

همان‌گونه که از مقایسه‌ی شکل ۷ و شکل ۱۱ مشخص است، در بهترین حالت الگوریتم‌های غیرنظارتی به دقت ۵۹/۱۶٪ (درصد خطای ۴۰/۸۴٪) می‌رسند، درحالی‌که الگوریتم‌های نظارتی در بهترین حالت به دقت ۷۹/۵۷٪ می‌رسند. همچنین بدترین حالت برای الگوریتم‌های نظارتی فقط کمی بدتر از بهترین حالت برای الگوریتم‌های غیرنظارتی است. این نشان‌دهنده‌ی کارایی بالاتر الگوریتم‌های نظارتی می‌باشد. این نتیجه از پیش قابل پیش‌بینی بود زیرا الگوریتم‌های نظارتی با در اختیار داشتن برچسب نمونه‌ها و با استفاده از یادگیری بهتر می‌توانند نمونه‌های دیده نشده را برچسب‌گذاری نمایند.

به‌طورکلی، بهترین نتایج حاصل از اعمال الگوریتم‌های مذکور که در جدول ۱۴ نیز آمده است به‌صورت زیر می‌باشد. در روش نظارتی الگوریتم‌های Bagging، شبکه‌ی بیز و جنگل تصادفی به‌منظور دسته‌بندی و از الگوریتم‌های همبستگی ویژگی، نسبت بهره، OneR و تحلیل اجزای اصلی به‌منظور انتخاب ویژگی‌های برتر استفاده گردید که با استفاده از الگوریتم‌های Bagging به‌همراه OneR دقت ۷۶/۰۷٪، شبکه‌ی بیز به‌همراه همبستگی ویژگی دقت ۷۰/۶۷٪، جنگل تصادفی به‌همراه نسبت بهره دقت ۷۹/۵۷٪ و جنگل تصادفی به‌همراه تحلیل اجزای اصلی دقت ۷۷/۸۷ درصد کسب گردید.

جدول ۱۴. بررسی دقت طبقه‌بندی الگوریتم‌های متفاوت، بر روی ویژگی‌های استخراج شده از متون کوتاه فارسی.

مجموعه داده	ویژگی	روش	دقت
نظرات فارسی مربوط به خریداران گوشی	لغوی، نگارشی، معنایی، ساختاری، دستوری، مختص متن، مختص شبکه‌های اجتماعی	خوشه‌بندی مبتنی بر چگالی + تحلیل اجزای اصلی،	۴۸
		خوشه‌بندی مبتنی بر چگالی + همبستگی ویژگی،	۵۵/۲۵
		K-means + نسبت بهره،	۵۶/۵۵
		EM + OneR	۵۹/۱۵
		شبکه‌ی بیز + همبستگی ویژگی،	۷۰/۶۷
		Bagging + OneR,	۷۶/۰۷
		جنگل‌تصادفی + تحلیل اجزای اصلی،	۷۷/۸۷
		جنگل‌تصادفی + نسبت بهره	۷۹/۵۷

در روش غیرنظارتی از الگوریتم‌های مبتنی بر چگالی، K-means و EM به‌منظور خوشه‌بندی و از الگوریتم‌های همبستگی ویژگی، نسبت بهره، OneR و تحلیل اجزای اصلی به‌منظور انتخاب ویژگی‌های برتر استفاده گردید که با استفاده از الگوریتم‌های مبتنی بر چگالی به‌همراه همبستگی ویژگی دقت ۵۵/۲۵٪، K-means به‌همراه نسبت بهره دقت ۵۶/۵۵٪، EM به‌همراه OneR دقت ۱۵/۵۹٪ و خوشه‌بندی مبتنی بر چگالی به‌همراه تحلیل اجزای اصلی دقت ۴۸ درصد کسب گردید.

۶,۲ بررسی ویژگی‌های سبک‌شناسی

با توجه به این‌که مجموعه داده‌ی مورد استفاده در این پژوهش، متون کوتاه شامل نظرات مربوط به خریداران گوشی می‌باشد و این متون کوتاه به‌صورت تعاملی بوده و از سبک نگارشی رسمی پیروی نمی‌کنند و همچنین افراد برای غلط‌های املایی و دستور زبانی اهمیتی قائل نمی‌شوند و به‌طور کلی در قالب و ساختار و چیدمان متن، با اسناد متنی معمولی تفاوت چشمگیری دارند چنین خصوصیاتی موجب کاهش دقت در ویژگی‌های سبک‌شناسی سنتی شده و کار شناسایی نویسنده‌ی متون کوتاه به کمک این ویژگی‌ها را با مشکل مواجه می‌سازد.

بررسی نتایج حاصل از اعمال روش‌های متفاوت بر روی مجموعه داده‌ی متون کوتاه برخط در این پژوهش نشان داد که ویژگی‌های نگارشی بیشترین تأثیر را در شناسایی نویسنده‌ی متون کوتاه داشته و پس از آن به ترتیب ویژگی‌های لغوی، مختص متن، مختص شبکه‌های اجتماعی، ساختاری، دستوری و معنایی قرار می‌گیرند.

به‌دلیل این‌که ویژگی‌های لغوی متن را به‌صورت رشته‌ای از کلمات و علائم درنظر گرفته و ویژگی‌های نگارشی شامل نشانه‌های نقطه‌گذاری و تعداد تکرار کلمات تابع می‌باشند؛ خصوصیات این دو ویژگی موجب شده که سبک نگارشی، غلط‌های املایی و دستور زبانی، محاوره‌ای یا رسمی بودن متن، ساختار و سازماندهی متن، تأثیری در عملکرد این دو ویژگی نداشته و نسبت به سایر ویژگی‌ها از دقت بالاتری برخوردار باشند.

ویژگی‌های مختص متن که با توجه به محتوای متن، مدل‌های گوشی و تکیه کلام‌های افراد درنظر گرفته شده‌اند عملکرد مناسبی در شناسایی نویسنده داشته‌اند. البته این ویژگی‌ها نیز تحت تأثیر محاوره‌ای بودن متن و همچنین غلط‌های املایی موجود در متن قرار گرفته و نسبت به دو ویژگی نگارشی و لغوی از دقت پایین‌تری برخوردار می‌باشند.

ویژگی‌های مختص شبکه‌های اجتماعی که با استفاده از شکلک‌های مثبت، خنثی و منفی و نمادهای متن‌محور، بیانگر حالت روحی نویسنده می‌باشند نیز عملکرد مناسبی در شناسایی نویسنده‌ی

متون کوتاه داشته و سبک نگارشی افراد و سایر خصوصیات مربوط به متون کوتاه تأثیر چندانی در دقت تشخیص آن‌ها نداشته‌اند.
ویژگی‌های ساختاری تحت تأثیر خصوصیات متون محاوره‌ای قرار گرفته و درنتیجه نسبت به سایر ویژگی‌ها، تأثیر کمتری در شناسایی نویسنده‌ی متون کوتاه داشته‌اند.
به دلیل این‌که ویژگی‌های دستوری تحت تأثیر غلط‌های املایی و دستور زبانی و همچنین سبک نگارشی افراد قرار می‌گیرند و همچنین فرایند استخراج این ویژگی‌ها با کمک ابزارهای پردازش زبان طبیعی صورت گرفته، عدم رعایت قواعد نگارشی و دستور زبانی، کار برچسب‌گذاری مجموعه داده را با مشکل مواجه ساخته و برچسب‌های به‌دست آمده با کمک این ابزارها، از دقت بالایی برخوردار نبوده و همین امر موجب کاهش دقت نتایج حاصل از اعمال الگوریتم‌ها با استفاده از ویژگی‌های دستوری شده است و این ویژگی با دقت ۳۲/۲۳ درصد، عملکرد ضعیفی را در شناسایی نویسنده‌ی متون کوتاه نشان داده است.
عامیانه بودن متون کوتاه و غلط‌های املایی و دستور زبانی موجب شده است که ویژگی‌های معنایی نسبت به تمامی ویژگی‌ها از کمترین دقت تشخیص برخوردار باشد.

۷ نتیجه‌گیری

در پژوهش حاضر شناسایی نویسندگان نظرات فارسی با استفاده از الگوریتم‌های مختلف خوشه‌بندی و دسته‌بندی، همچنین با به‌کارگیری الگوریتم‌های انتخاب ویژگی مورد بررسی قرار گرفت. یافته‌های این پژوهش نشان می‌دهد که با کاهش ابعاد ویژگی‌ها درصد تشخیص درست افزایش می‌یابد به‌طوری‌که بهترین نتیجه مربوط به زمانی است که ۱۵ ویژگی برتر توسط الگوریتم انتخاب ویژگی OneR انتخاب شدند. از میان این ۱۵ ویژگی ویژگی‌های لغوی بیشترین تأثیر را در شناسایی نویسنده‌ی متون کوتاه داشته و پس از آن‌ها به ترتیب ویژگی‌های نگارشی، مختص متن، ساختاری و دستوری قرار گرفتند.
نتایج حاصل از آزمودن الگوریتم‌های نظارتی و غیرنظارتی نشان می‌دهد که بهترین دقت تشخیص در بین الگوریتم‌های خوشه‌بندی مربوط به الگوریتم EM روی ۱۵ ویژگی برتر انتخابی توسط OneR می‌باشد درحالی‌که الگوریتم جنگل تصادفی به‌همراه نسبت بهره برای ۹۰ ویژگی بالاترین کارایی را در بین الگوریتم‌های دسته‌بندی دارد. همچنین با توجه به نتایج حاصل از مقایسه‌ی الگوریتم‌های خوشه‌بندی و دسته‌بندی مشخص شد که الگوریتم‌های نظارتی در بهترین حالت به میزان ۲۰/۴۱٪ نسبت به الگوریتم‌های غیرنظارتی برتری دارند.
این پژوهش نشان می‌دهد با وجود آن‌که به کمک ویژگی‌های سبک‌شناسی سنتی نیز می‌توان به شناسایی نویسنده‌ی متون کوتاه پرداخت، اما تکنیک‌های انتخاب ویژگی سنتی که در تحقیقات گذشته جهت شناسایی نویسنده‌ی متون ساختارمند و طولانی مورد استفاده قرار گرفته و دقت بالایی داشتند در مورد متون برخط دقت و کارایی کمتری دارند. اسناد برخط، تعاملی‌تر بوده، سبک نگارش رسمی کمتری داشته و الگوی واژگان در آن‌ها ثابت نیست و همچنین نسبت به اسناد متنی عادی در سبک نگارش، قالب، ساختار و چیدمان متن متفاوت بوده و این‌که افراد برای غلط‌های املایی و گرامری اهمیتی قائل نمی‌شوند، یافتن الگوی نگارشی نویسنده را با مشکل مواجه می‌سازد. کوتاه بودن طول این متون نیز موجب کاهش دقت دسته‌بندی و افزایش نرخ خطا می‌گردد.
نتایج حاصل از این پژوهش گویای این مطلب است که به‌دلیل این‌که ویژگی‌های دستوری وابسته به سبک نگارشی افراد بوده و فرایند استخراج این ویژگی‌ها با کمک ابزارهای پردازش زبان طبیعی صورت می‌گیرد، برچسب‌های به‌دست آمده با استفاده از این ابزارها از دقت کافی برخوردار نبوده و درنهایت، نتایج حاصل از اعمال الگوریتم‌ها از دقت بالایی برخوردار نمی‌باشند. ویژگی‌های معنایی و مختص متن تحت تأثیر محاوره‌ای بودن متن قرار گرفته و همین امر به کاهش دقت تشخیص منجر می‌شود. در میان این ویژگی‌ها، ویژگی‌های لغوی و نگارشی که وابستگی چندانی به ساختار و قالب متن ندارند و محاوره‌ای و رسمی بودن متون در عملکرد آن‌ها تأثیری نمی‌گذارد، نسبت به سایر ویژگی‌ها نتیجه‌ی بهتری از خود نشان می‌دهند.
با توجه به یافته‌های این پژوهش می‌توان جمع‌بندی کرد که درصورتی که نوشته‌هایی از نویسندگان وب در دسترس باشد، مثلا زمانی که هدف شناسایی کاربری از بین کاربران موجود در یک وب‌سایت از روی نوشته‌های موجود است و سابقه‌ی نگارش افراد وجود دارد، بهتر است از الگوریتم‌های دسته‌بندی استفاده شود؛ درحالی‌که اگر هدف یافتن نوشته‌های یک نویسنده بر اساس تشابه نوشته‌ها باشد، لازم است از الگوریتم‌های خوشه‌بندی استفاده شود.
به‌عنوان ادامه‌ی راه برای این پژوهش می‌توان استفاده از الگوریتم‌های فراابتکاری که در سایر حوزه‌ها کارایی خوبی نشان داده‌اند را برای انتخاب ویژگی پیشنهاد داد. همچنین به‌کارگیری روابط موجود در سایت‌های اجتماعی می‌تواند جهت شناسایی بهتر نویسندگان به‌عنوان مسیری برای کارهای آتی پیشنهاد شود.

مراجع

[۱] مرادی، مهدی و بحرانی، محمد، "تشخیص خودکار جنسیت نویسنده در متون فارسی"، فصل‌نامه پردازش علائم و داده‌ها، شماره ۴، پیاپی ۲۶، صفحات ۸۳-۹۴، ۱۳۹۴.

[۲] فرهمندپور، زینب، نیک‌مهر، هومان، منصوری زاده، محرم و طبیب زاده قمصری، امید، "یک سیستم نوین هوشمند تشخیص هویت نویسنده فارسی زبان بر اساس سبک نوشتاری-مقاله برگزیده هفدهمین کنفرانس

ملی انجمن کامپیوتر ایران"، *مجله محاسبات نرم*، شماره دوم، صفحات ۲۶-۳۵، ۱۳۹۱.

[۳] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *Digit. Investig.*, vol. ۷, no. ۱–۲, pp. ۵۶–۶۴, ۲۰۱۰.

[۴] S. Nirkhi, R. V Dharaskar, and V. M. Thakare, "Authorship Verification of Online Messages for Forensic Investigation," *Procedia Comput. Sci.*, vol. ۷۸, pp. ۶۴۰–۶۴۵, ۲۰۱۶, doi: https://doi.org/۱۰,۱۰۱۶/j.procs.۲۰۱۶,۰۲,۱۱۱.

[۵] M. L. Brocardo, I. Traore, and I. Woungang, "Authorship verification of e-mail and tweet messages applied for continuous authentication," *J. Comput. Syst. Sci.*, vol. ۸۱, no. ۸, pp. ۱۴۲۹–۱۴۴۰, ۲۰۱۵.

[۶] Y. Yiming and P. Jan O., "A Comparative Study on Feature Selection in Text Categorization," *Proceeding ICML '۹۷ Proc. Fourteenth Int. Conf. Mach. Learn.*, vol. ۵۳, no. ۹, pp. ۴۱۲–۴۲۰, ۱۹۹۷.

[۷] M. Frederick and L. Wallace David, "Inference and Disputed Authorship: The Federalist. Reading, Addison." Wessley Publishing Company. Republié sous le titre Applied Bayesian and …, ۱۹۸۴.

[۸] T. C. Mendenhall, "The Characteristic Curves of Composition," *Science (۸۰-.).*, vol. ۹, no. ۲۱۴, pp. ۲۳۷–۲۴۹, Dec. ۱۸۸۷, [Online]. Available:http://www.jstor.org/stable/۱۷۶۴۶۰۴.

[۹] H. Craig, "Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?," *Lit. Linguist. Comput.*, vol. ۱۴, no. ۱, pp. ۱۰۳–۱۱۳, ۱۹۹۹.

[۱۰] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proceedings of the twenty-first international conference on Machine learning*, ۲۰۰۴, p. ۶۲.

[۱۱] E. Villar-Rodriguez, J. Del Ser, M. N. Bilbao, and S. Salcedo-Sanz, "A feature selection method for author identification in interactive communications based on supervised learning and language typicality," *Eng. Appl. Artif. Intell.*, vol. ۵۶, pp. ۱۷۵–۱۸۴, ۲۰۱۶, doi: https://doi.org/۱۰,۱۰۱۶/j.engappai.۲۰۱۶,۰۹,۰۰۴

[۱۲] P. Geutner, U. Bodenhausen, and A. Waibel, "Flexibility through incremental learning: Neural networks for text categorization," in *Proceedings of WCNN-۹۳, World Congress on Neural Networks*, ۱۹۹۳, pp. ۲۴–۲۷.

[۱۳] O. De Vel, "Mining e-mail authorship," ۲۰۰۰.

[۱۴] M. Corney, O. De Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *۱۸th Annual Computer Security Applications Conference, ۲۰۰۲. Proceedings.*, ۲۰۰۲, pp. ۲۸۲–۲۸۹.

[۱۵] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics," *Digit. Investig.*, vol. ۵, pp. S۴۲–S۵۱, ۲۰۰۸.

[۱۶] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. ۲۶, no. ۲, pp. ۱–۲۹, ۲۰۰۸.

[۱۷] F. Iqbal, L. A. Khan, B. C. M. Fung, and M. Debbabi, "E-Mail Authorship Verification for Forensic Investigation," in *Proceedings of the ۲۰۱۰ ACM Symposium on Applied Computing*, ۲۰۱۰, pp. ۱۵۹۱–۱۵۹۸, doi: ۱۰,۱۱۴۵/۱۷۷۴۰۸۸,۱۷۷۴۴۲۸.

[۱۸] B. Allison and L. Guthrie, "Authorship Attribution of E-Mail: Comparing Classifiers over a New Corpus for Evaluation.," ۲۰۰۸.

[۱۹] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digit. Investig.*, vol. ۸, no. ۱, pp. ۷۸–۸۸, ۲۰۱۱.

[۲۰] X. Chen, P. Hao, R. Chandramouli, and K. P. Subbalakshmi, "Authorship similarity

[۲۹] H. Alam and A. Kumar, "Multi-lingual author identification and linguistic feature extraction—A machine learning approach," in *۲۰۱۳ IEEE International Conference on Technologies for Homeland Security (HST)*, ۲۰۱۳, pp. ۳۸۶–۳۸۹.

[۳۰] J. Adams, H. Williams, J. Carter, and G. Dozier, "Genetic Heuristic Development: Feature selection for author identification," in *۲۰۱۳ IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, ۲۰۱۳, pp. ۳۶–۴۱.

[۳۱] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in *International conference on artificial intelligence: Methodology, systems, and applications*, ۲۰۰۶, pp. ۷۷–۸۶.

[۳۲] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Syst.*, vol. ۳۶, pp. ۲۲۶–۲۳۵, ۲۰۱۲, doi: https://doi.org/۱۰,۱۰۱۶/j.knosys.۲۰۱۲,۰۶,۰۰۵

[۳۳] زنگویی، سمیرا، نعمتی شمس‌آباد، حسنعلی "شناسایی نویسندگان پیام های الکترونیکی از طریق واکاوی نوع و سبک نگارش آن ها مبتنی بر روش های یادگیری ماشین (WKF based on SVM-PHGS)"، *پردازش و مدیریت اطلاعات (علوم و فناوری اطلاعات)*، شماره ۲، دوره ۲۹، صفحات ۴۷۶-۴۵۳، ۱۳۹۲.

[۳۴] G. U. Yule, "The statistical study of literary vocabulary. Cambridge, Cambridge [Eng.]." University Press. Journal of the Royal Statistical Society, ۱۹۴۴.

[۳۵] A. Honoré, "Some simple measures of richness of vocabulary," *Assoc. Lit. Linguist. Comput. Bull.*, vol. ۷, no. ۲, pp. ۱۷۲–۱۷۷, ۱۹۷۹.

[۳۶] E. Brunet, *Le Vocabulaire de Jean Giraudoux: structure et évolution : statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française*. Slatkine, ۱۹۷۸.

[۳۷] H. S. Sichel, "On a Distribution Law for Word Frequencies," *J. Am. Stat. Assoc.*, vol. ۷۰, no. ۳۵۱a, pp. ۵۴۲–۵۴۷, ۱۹۷۵, doi:

detection from email messages," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, ۲۰۱۱, pp. ۳۷۵–۳۸۶.

[۲۱] J. Keeshin, Z. Galant, and D. Kravitz, "Machine Learning and Feature Based Approaches to Gender Classification of Facebook Statuses." ۲۰۱۰.

[۲۲] R. Layton, P. Watters, and R. Dazeley, "Authorship Attribution for Twitter in ۱۴۰ Characters or Less," in *۲۰۱۰ Second Cybercrime and Trustworthy Computing Workshop*, Jul. ۲۰۱۰, pp. ۱–۸, doi: ۱۰,۱۱۰۹/CTC.۲۰۱۰,۱۷.

[۲۳] C. Li, A. Sun, and A. Datta, "Twevent: Segment-Based Event Detection from Tweets," in *Proceedings of the ۲۱st ACM International Conference on Information and Knowledge Management*, ۲۰۱۲, pp. ۱۵۵–۱۶۴, doi: ۱۰,۱۱۴۵/۲۳۹۶۷۶۱,۲۳۹۶۷۸۵.

[۲۴] J. S. Li, J. V Monaco, L.-C. Chen, and C. C. Tappert, "Authorship authentication using short messages from social networking sites," in *۲۰۱۴ IEEE ۱۱th International Conference on e-Business Engineering*, ۲۰۱۴, pp. ۳۱۴–۳۱۹.

[۲۵] A. Zubiaga, D. Spina, R. Martínez, and V. Fresno, "Real- time classification of twitter trends," *J. Assoc. Inf. Sci. Technol.*, vol. ۶۶, no. ۳, pp. ۴۶۲–۴۷۳, ۲۰۱۵.

[۲۶] A. Orebaugh, "An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation," in *Proceedings ۴۰th Annual ۲۰۰۶ International Carnahan Conference on Security Technology*, ۲۰۰۶, pp. ۱۶۰–۱۷۲.

[۲۷] O. Canales *et al.*, "A stylometry system for authenticating students taking online tests," *P. Student-Faculty Res. Day, Ed., CSIS. Pace Univ.*, ۲۰۱۱.

[۲۸] C.-Y. Lai, "Author Gender Analysis'," *Final Proj. from I*, vol. ۲۵۶, ۲۰۰۹.

۱۰,۱۰۸۰/۰۱۶۲۱۴۵۹,۱۹۷۵,۱۰۴۸۲۴۶۹.

[۳۸] E. H. SIMPSON, “Measurement of Diversity,” *Nature*, vol. ۱۶۳, no. ۴۱۴۸, p. ۶۸۸, ۱۹۴۹, doi: ۱۰,۱۰۳۸/۱۶۳۶۸۸a۰.

[۳۹] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, “A novel ACO–GA hybrid algorithm for feature selection in protein function prediction,” *Expert Syst. Appl.*, vol. ۳۶, no. ۱۰, pp. ۱۲۰۸۶–۱۲۰۹۴, ۲۰۰۹, doi:https://doi.org/۱۰,۱۰۱۶/j.eswa.۲۰۰۹,۰۴,۰۲۳.