Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال اول، شماره دوم، پاییز و زمستان۱۴۰۲، صفحات ٦٦-٥٤

تشخیص ارائه‌دهندگان خدمات در پیام‌های فارسی تلگرام مبتنی بر روش‌های انتخاب ویژگی

فرزانه ابراهیمی [۱]، محمد علی زارع چاهوکی [۲*]، علی هاشمی [۳]

[۱] دانشجوی ارشد مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

[۲] استادیار دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

[۳] دکتری مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۱/۰۵/۰۹

تاریخ پذیرش:
۱۴۰۱/۱۱/۱۰

کلیدواژه‌ها:
دسته‌بندی، کاربران ارائه‌دهنده خدمات، انتخاب ویژگی، کاهش ویژگی، یادگیری ماشین

نویسنده‌مسئول:
chahooki@yazd.ac.ir

چکیده

پیام‌رسان تلگرام بستری مناسب برای کاربرانی است که به دنبال خرید محصول یا دریافت خدمات به‌صورت آنلاین هستند. در این پیام‌رسان، برای درخواست محصول، امکان دسترسی مستقیم به ارائه‌دهندگان کالا و خدمات وجود ندارد. لذا باید ابتدا در گروه‌های تلگرامی مرتبط عضو شد و درخواست خود را در تک تک گروه‌ها ثبت کرد و منتظر پاسخ ماند. این امر ضمن زمان‌بر بودن، با مشکلاتی همراه است. هدف از این پژوهش تشخیص کاربران ارائه‌دهنده خدمات به‌منظور ارتباط مستقیم و مؤثر با مشتریان است. بدین منظور از ایده دسته‌بندی پیام‌های فارسی منتشر شده در تلگرام استفاده شد. یکی از مشکلات دسته‌بندی این پیام‌ها، ابعاد بزرگ فضای ویژگی است که سبب کاهش دقت و افزایش زمان دسته‌بندی می‌شود. برای حل این مسئله از روش‌های انتخاب ویژگی استفاده شد. روش پیشنهادی این پژوهش، بر مبنای ترکیب روش‌های انتخاب ویژگی مبتنی بر فیلتر محلی و سراسری است. نوآوری این پژوهش در استفاده از روش‌های ترکیبی انتخاب ویژگی جهت دسته‌بندی خودکار پیام‌های فارسی تلگرام، به‌منظور شناسایی کاربران ارائه‌دهنده خدمات است. روش پیشنهادی، ضمن کاهش تعداد ویژگی‌ها و انتخاب ویژگی‌های مرتبط، سبب بهبود عملکرد دسته‌بندی و تشخیص کاربران ارائه‌دهنده خدمات می‌شود.

10.22034/ABMIR.2023.18780.1012

E-ISSN: 2821-2037

۱- مقدمه

با گسترش شبکه‌های اجتماعی[1]، تحولی بنیادین در نحوه ارتباط صاحبان کسب‌وکار با مشتریانشان به‌وجود آمد. در حال حاضر، هر فردی بدون توجه به‌اندازه کسب‌وکار و میزان بودجه‌ای که دارد، می‌تواند از فضای مجازی و شبکه‌های اجتماعی برای ارتباط با مشتریان، و معرفی محصولات و خدماتش به طیف وسیعی از مردم استفاده کند. تلگرام[2]، یکی از پیام‌رسان‌هایی است که بین ایرانیان محبوبیت زیادی کسب کرده است. صاحبان کسب‌وکارهای مختلف با ایجاد گروه و کانال‌های تلگرامی، روزانه میلیون‌ها پیام فارسی از جنس تبلیغ و معرفی محصول و خدمات منتشر می‌کنند که بستر مناسبی جهت درآمدزایی برای آن‌ها ایجاد کرده است.

شناسایی ارائه‌دهندگان کالا و خدمات در تلگرام سبب برقراری ارتباط مستقیم بین صاحبان کسب‌وکار و مشتریانشان می‌شود. به‌طور مثال کاربری که درخواست خرید یک قطعه خودرو را دارد می‌تواند فقط با نوشتن درخواست خود و یک‌بار ارسال آن، این تقاضا را به‌طور مستقیم برای مجموعه‌ای از ارائه‌دهندگان کالا که در حوزهٔ فروش قطعات خودرو فعالیت می‌کنند، ارسال کند. لذا دیگر نیاز به عضویت در تک‌تک کانال‌ها یا گروه‌های مرتبط نیست. این امر برای ارائه‌دهندگان کالا و خدمات در هر حوزه کاری دسترسی سریع فراهم می‌کند. همچنین زمان پاسخگویی به درخواست مشتریان را کاهش و احتمال پاسخ مثبت را افزایش می‌دهد. از طرفی با شناسایی کاربران ارائه‌دهنده خدمات، می‌توان کمپین‌هایی مخصوص هر حوزه کاری راه‌اندازی کرد که صاحبان کسب‌وکار در آن حوزه خاص، نظرات و تجربیات خود را بیان کنند. این کمپین‌ها کمک شایانی به افراد تازه‌وارد به هر حوزه می‌کند، زیرا می‌توانند در مدت‌زمان کمی از تجربیات پیش‌کسوتان باخبر شوند و مسیر هموارتری را برای ادامه کسب‌وکار خود ترسیم کنند. بسیاری از افراد برای راه‌اندازی یک فعالیت، خواهان اطلاعات درباره چالش‌های ورود به آن حوزه کاری هستند تا با آمادگی و بینش قوی‌تری ورود کرده و احتمال شکست خود را پایین بیاورند.

با شناسایی کاربران ارائه‌دهنده کالا و خدمات، این امکان وجود دارد تا بتوانند از فعالان هر حوزه، مشاوره بگیرند.

جهت تشخیص کاربران ارائه‌دهنده خدمات از روش دسته‌بندی[3] پیام‌های ارسالی آن‌ها در گروه‌های مختلف تلگرامی استفاده شد. عموماً دسته‌بندی خودکار متن بر مبنای ویژگی‌های استخراجی انجام می‌گیرد. یکی از مشکلات اساسی در دسته‌بندی، تعداد زیاد ویژگی‌های استخراجی است، چرا که اغلب این ویژگی‌ها نامربوط و زائد هستند و بر کارایی دسته‌بندی تأثیر منفی می‌گذارند [۱]. انتخاب ویژگی[4]، راه‌حلی مناسب برای کاهش ابعاد بزرگ فضای ویژگی و افزایش کارایی عملکرد دسته‌بندی است [۲]. هدف از انتخاب ویژگی، انتخاب بهینه‌ترین زیرمجموعه ویژگی از کل فضای ویژگی‌های اصلی متن موردنظر است [۳]، به‌طوری‌که بتوان ضمن کاهش تعداد ویژگی‌ها، به‌دقت دسته‌بندی قابل قبولی نیز دست‌یافت. با انتخاب ویژگی‌های مناسب، زمان آموزش و محاسبات در یادگیری ماشین کاهش پیدا کرده و دقت پیش‌گویی افزایش می‌یابد [٤].

این پژوهش روشی را جهت شناسایی کاربران ارائه‌دهنده خدمات در پیام‌رسان تلگرام ارائه می‌دهد. مبنای کار این روش، دسته‌بندی پیام‌های ارسالی کاربران در گروه‌های مختلف تلگرامی است. در گام اول پیام‌های ارسال‌شده توسط هر کاربر در گروه‌های مختلف تلگرامی استخراج می‌شود. در گام دوم با استفاده از پرکاربردترین روش‌های انتخاب ویژگی مبتنی بر فیلتر که به دودسته محلی و سراسری تقسیم می‌شوند، ویژگی‌های غیرضروری حذف می‌شوند. در این پژوهش ویژگی‌های استخراجی، همان کلمات موجود در پیام‌ها هستند. سپس هر یک از روش‌های محلی و سراسری باهم ترکیب و ویژگی‌های بهتر انتخاب و با استفاده از این ویژگی‌ها دسته‌بندی انجام می‌شود. نتایج این آزمایش‌ها نشان داد که ترکیب روش‌های تکی انتخاب ویژگی، در اکثر موارد می‌تواند سبب افزایش دقت نسبت به حالت بدون انتخاب ویژگی، در دسته‌بندی و شناسایی کاربران ارائه‌دهنده خدمات شود.

[1] Social network
[2] Telegram
[3] classification
[4] Feature selection

در گام سوم خروجی ویژگی‌های روش ترکیبی گام دوم، به‌عنوان ورودی برای روش انتخاب ویژگی یادگیرمبنا استفاده می‌شود. با این روش بار دیگر ویژگی‌های کم‌اهمیت حذف می‌شود. نتایج آزمایش‌ها نشان می‌دهد که استفاده از روش ترکیبی فیلتر ویادگیر مبنا عملکرد بهتری نسبت به روش‌های فیلتر دارد.

در ادامه در بخش دوم به بررسی مطالعات پیشین پرداخته می‌شود. روش پیشنهادی در بخش سوم به‌تفصیل بیان می‌شود. نتایج استفاده از روش پیشنهادی در بخش چهارم ارائه و در بخش پنجم نیز به بیان خلاصه کارهای انجام‌شده و نتیجه‌گیری پرداخته می‌شود.

۲- مطالعات پیشین

تکنیک‌های انتخاب ویژگی به‌طورکلی به سه دسته فیلتر[1]، یادگیر مبنا[2] و یادگیر حاصل[3] طبقه‌بندی می‌شوند. روش‌های فیلتر، مستقل از الگوریتم یادگیری، ویژگی‌ها را بر اساس یک مرحله پیش‌پردازش انتخاب می‌کنند. در این روش ابتدا به تک‌تک ویژگی‌ها امتیازی داده می‌شود و بر اساس این امتیاز مرتب می‌شوند و ویژگی‌هایی با بالاترین امتیاز انتخاب می‌شوند. درنهایت این مجموعه به‌عنوان ورودی به الگوریتم دسته‌بندی داده می‌شود.

روش‌های فیلتر شامل دودسته محلی[4] و سراسری[5] می‌شوند. روش‌های انتخاب ویژگی سراسری معروف عبارت‌اند از: فراوانی سند[6] (DF)، بهره اطلاعات[7] (IG)، شاخص‌های جینی[8] (GI) و انتخاب‌گر ویژگی متمایز[9] (DFS). از روش‌های انتخاب ویژگی محلی پرکاربرد می‌توان نسبت شانس[10] (OR) و ضریب همبستگی[11] (CC) را نام برد. روش فیلتر از الگوریتم یادگیری خارجی برای ارزیابی عملکرد ویژگی‌های انتخاب‌شده استفاده می‌کند [۱].

در روش‌های یادگیر مبنا، زیرمجموعه‌ای از ویژگی‌ها بر اساس یک الگوریتم یادگیری به‌گونه‌ای انتخاب می‌شوند که سبب بیشینه شدن کارایی آن روش یادگیری شود. روش‌های یادگیرمبنا تقریباً شبیه به روش‌های فیلتر عمل می‌کنند با این تفاوت که از یک الگوریتم طبقه‌بندی از پیش تعریف‌شده به‌جای اندازه‌گیری مستقل برای ارزیابی زیرمجموعه ویژگی‌ها استفاده می‌کنند. روش‌های یادگیرمبنا در مقایسه با روش‌های فیلتر، عملکرد بهتری دارند، اما زمانی که تعداد ویژگی‌ها بسیار زیاد شود، ازنظر محاسباتی گران‌تر می‌شوند. دو روش انتخاب متوالی روبه‌جلو (SFS) و حذف متوالی به عقب (SBS)، دو روش متداول حریصانه هستند که اغلب برای انتخاب ویژگی استفاده می‌شوند. [۵].

در روش‌های یادگیرحاصل، انتخاب ویژگی و یادگیری مدل به‌صورت هم‌زمان انجام می‌شود. روش‌های یادگیرحاصل زمان محاسبه برای دسته‌بندی مجدد زیرمجموعه‌های مختلف درروش یادگیرمبنا را کاهش می‌دهند [٦]. درمجموع روش‌های فیلتر محاسبات ساده‌تر و زمان پردازش کوتاه‌تری دارند. لذا جهت کار با مجموعه داده‌های با ابعاد بالا مناسب هستند. اما روش‌های یادگیرمبنا باوجود کارایی قابل‌قبول، ازنظر محاسباتی پیچیده‌تر و پرهزینه‌تر هستند [۵]. ازاین‌رو روش‌های فیلتر در مقایسه با روش‌های یادگیرمبنا سریع‌تر عمل می‌کنند. ضعف روش‌های یادگیرحاصل در نادیده گرفتن ارتباط بین ویژگی‌ها است که سبب مشکلاتی در انتخاب ویژگی‌های نهایی می‌شود [۷]. آقای یوسال ترکیب یک روش انتخاب ویژگی سراسری مبتنی بر فیلتر و یک روش انتخاب ویژگی محلی را برای بهبود انتخاب ویژگی، استفاده کرده است که درواقع به دنبال ایجاد مجموعه‌ای از ویژگی‌هایی هستند که تقریباً به‌طور یکسان تمام کلاس‌ها را در برگیرد. نتایج تجزیه‌وتحلیل تجربی نشان می‌دهد که این روش ترکیبی، عملکرد دسته‌بندی بهتری نسبت به عملکرد تکی روش‌های انتخاب ویژگی سراسری دارد؛ اما این روش در برخورد با مجموعه داده‌های نامتعادل دارای تعداد زیادی کلاس، مشکل دارد [۷].

پژوهش آقای اگنی‌هتری طرح انتخاب ویژگی متغیر سراسری (VGFSS)را که ترکیبی از یک روش سراسری و روش نسبت شانس است، جهت طبقه‌بندی خودکار اسناد متنی ارائه می‌دهد. این

[1] Filter
[2] Wrapper
[3] Embed
[4] Local
[5] Global
[6] Document Frequency (DF)
[7] Information Gain (IG)
[8] Gini Index (GI)
[9] Distinguishing Feature Selector (DFS)
[10] Odd Ratio (OR)
[11] Correlation Coefficient (CC)

روش تعداد متغیری از ویژگی‌ها را از هر کلاس بر اساس توزیع عبارات در کلاس‌ها انتخاب می‌کند که این امر تضمین می‌کند حداقل تعداد اصطلاحات از هر کلاس انتخاب می‌شود [۸].

آقای ملو و همکارانش با مقایسه و ارزیابی تجربی دو روش انتخاب ویژگی محلی و سراسری بر روی مجموعه داده‌های طبقه‌بندی چند برچسبی مسطح و سلسله مراتبی رایج (که معمولاً در محک زدن عملکرد طبقه‌بندی کننده‌های چند برچسبی استفاده می‌شوند) نشان می‌دهند که انتخاب ویژگی محلی عملکرد بهتری نسبت به انتخاب ویژگی سراسری در این نوع داده‌ها دارند. [۹]

در برخی پژوهش‌ها، روش‌های محلی برای تعداد ویژگی‌های کم و روش‌های سراسری برای تعداد ویژگی‌های زیاد نتایج بهتری داشته‌اند [۵]. آقای کرست و همکارانش از روش‌های فیلتر به‌صورت مستقل و ترکیبی استفاده کردند [۱۰]. اندازه‌گیری دقیق روابط ویژگی‌های نامزد، ویژگی‌ها و دسته‌بندی‌های انتخاب‌شده در فرآیند انتخاب، به‌ویژه برای داده‌های با ابعاد بالا و اندازه نمونه کوچک، یک کار چالش‌برانگیز است. برای این منظور در پژوهش آقای وی و همکارانش معیار جدیدی به نام اهمیت ویژگی پویا (DFI) و همچنین الگوریتم انتخاب ویژگی متناظر آن بانام انتخاب ویژگی مبتنی بر اهمیت ویژگی پویا (DGIFS) پیشنهادشده است. این الگوریتم ترکیبی از روش انتخاب ویژگی مبتنی بر اهمیت ویژگی پویا با روش‌های فیلتر است. هدف آن‌ها رسیدن به‌دقت دسته‌بندی بالاتر با تعداد کمتری از ویژگی‌ها است. [۱۱].

تکنیک کیسه کلمات اغلب برای ارائه سند در طبقه‌بندی متن استفاده می‌شود. بااین‌حال، برای مجموعه بزرگی از اسناد که ابعاد بردار کیسه کلمات بسیار بزرگ است، دسته‌بندی متن به دلیل داده‌های پراکنده، برازش بیش‌ازحد، و ویژگی‌های نامربوط، به چالشی جدی تبدیل می‌شود. روش انتخاب ویژگی فیلتر با حذف ویژگی‌های نامربوط از بردار کیسه کلمات، تعداد ویژگی‌ها را کاهش می‌دهد. آقای نام و همکارانش نقاط ضعف و نقاط قوت دو رویکرد انتخاب ویژگی فیلتر که رویکرد مبتنی بر فرکانس و رویکرد مبتنی بر خوشه است را تحلیل کردند، سپس از روش‌های انتخاب ویژگی فیلتر ترکیبی برای بهبود طبقه‌بندی متن با تعداد ویژگی‌های زیاد استفاده کردند. این روش، ترکیبی از رویکرد مبتنی بر فرکانس و رویکرد مبتنی بر خوشه است [۱۲].

آقای اسیف و همکارانش یک روش انتخاب ویژگی جدید مبتنی بر بهینه‌سازی ازدحام ذرات تطبیقی با وزن خود اینرسی (SIW-APSO) برای افزایش عملکرد سیستم‌های طبقه‌بندی متن پیشنهاد می‌کنند. این روش به دلیل صلاحیت جستجوی بالا و توانایی یافتن زیرمجموعه ویژگی‌ها به‌طور کارآمد، دارای پدیده همگرایی سریع است. تجزیه‌وتحلیل تجربی این مقاله نشان می‌دهد که روش پیشنهادی نسبت به الگوریتم‌های پیشرفته موجود در مجموعه داده‌های رویترز-۲۱۵۷۸ عملکرد بهتری دارد [۱۳].

آقای افروسیدینیز و همکارانش از ۱۲ روش تکی و ۶ روش گروهی برای انتخاب ویژگی جهت طبقه‌بندی داده‌های محیطی، به‌ویژه در حوزه مدل‌سازی توزیع گونه‌ها استفاده می‌کنند. این روش‌ها هر ۳ دسته روش‌های انتخاب ویژگی فیلتر، یادگیرمبنا، و یادگیرحاصل را شامل می‌شوند. آزمایش آن‌ها بر روی ۸ مجموعه داده‌های محیطی نشان می‌دهد که توضیحات افزودنی (SHAP) و اهمیت جایگشت مؤثرترین روش‌های فردی هستند که هر دو در دسته روش‌های یادگیرمبنا قرار می‌گیرد. به‌طورکلی، روش‌های فیلتر عملکرد ضعیفی دارند و روش‌های یادگیرحاصل دراین‌بین قرار می‌گیرند. از ۶ روش گروهی در نظر گرفته‌شده، Reciprocal Ranking، بهترین عملکرد را دارد [۱۴].

۲ ـ ۱ روش‌های انتخاب ویژگی مورداستفاده در این پژوهش

روش‌های انتخاب ویژگی محلی مورداستفاده در این پژوهش شامل نسبت شانس و ضریب همبستگی است که در ادامه به توضیح عملکرد این روش‌ها پرداخته می‌شود. معیار نسبت شانس، عضویت در کلاس خاصی را با صورت کسر و عدم عضویت را با مخرج کسر اندازه‌گیری می‌کند. امتیاز عضویت و عدم عضویت، با تقسیم به یکدیگر نرمال می‌شوند [۱۰]. طبق رابطه (۱) به‌صورت مجزا امتیاز نسبت شانس هر اصطلاح برای تمام کلاس‌ها محاسبه و بالاترین امتیاز را برای اختصاص یک برچسب کلاس به‌اصطلاح انتخاب می‌شود. برای تعیین برچسب، مقدار امتیاز مطلق در نظر گرفته می‌شود.

$$OR(t|C_i) = \log \frac{P(t|C_i)[1-P(t|\bar{C}_i)]}{[1-P(t|C_i)]P(t|\bar{C}_i)} \quad (۱)$$

$P(t|C_i)$ احتمال وجود اصطلاح[1] t به‌شرط کلاس C_i است. $P(t|\bar{C}_i)$ احتمال شرطی اصطلاح t برای تمام کلاس‌ها به‌جز کلاس C_i است.

روش ضریب همبستگی، شایستگی یک زیرمجموعه از اصطلاحات را با در نظر گرفتن قدرت پیش‌بینی تکی هر اصطلاح و درجه افزونگی بین آن‌ها، ارزیابی می‌کند. زیرمجموعه مناسب از ویژگی‌ها، همبستگی بالایی در کلاس و همبستگی پایینی بین کلاس‌های مختلف دارد. در رابطه (۲) ضریب همبستگی CC برای اصطلاح t در یک کلاس C_i برای مجموعه‌ای با N نمونه تعریف‌شده است [۳].

$$CC\left(t,C_i\right) = \frac{\sqrt{N}\begin{bmatrix} P\left(t,C_i\right)P\left(\bar{t},\bar{C}_i\right) \\ -P\left(t,\bar{C}_i\right)P\left(\bar{t},C_i\right)\end{bmatrix}}{\sqrt{P(t)P(\bar{t})P(C_i)P(\bar{C}_i)}} \quad (۲)$$

N تعداد کل اسناد را نشان می‌دهد. $P(t|C_i)$ احتمال وجود اصطلاح t به‌شرط کلاس C_i است. $P(\bar{t}|\bar{C}_i)$ احتمال عدم وجود اصطلاح t برای تمام کلاس‌ها به‌جز کلاس C_i است. $P(\bar{t}|C_i)$ احتمال عدم وجود اصطلاح t به‌شرط کلاس C_i است. $P(t)$ احتمال وجود اصطلاح t و $P(\bar{t})$ احتمال عدم وجود اصطلاح t را نشان می‌دهد. $P(C_i)$ احتمال وجود کلاس C_i و $P(\bar{C}_i)$ احتمال عدم وجود کلاس C_i را نشان می‌دهد.

روش‌های انتخاب ویژگی سراسری این پژوهش شامل فراوانی سند، بهره اطلاعات، شاخص‌های جینی و انتخاب‌گر ویژگی متمایز است که در ادامه به معرفی این روش‌ها پرداخته می‌شود. روش بهره اطلاعات، طبق رابطه (۳) سهم وجود یا عدم وجود اصطلاح برای دسته‌بندی صحیح اسناد متنی را محاسبه می‌کند. روش بهره اطلاعات یک استراتژی نظارت‌شده است. یعنی از برچسب اسناد در انتخاب اصطلاحات برای حفظ آن‌ها استفاده می‌کند [۵].

$$\begin{aligned} IG(t) &= -\sum_{i=1}^{M} P(C_i)\, log\, P(C_i) \\ &+ P(t)\sum_{i=1}^{M} P(C_i|t)\, log\, P(C_i|t) \\ &+ P(\bar{t})\sum_{i=1}^{M} P(C_i|\bar{t})\, log\, P(C_i|\bar{t}) \end{aligned} \quad (۳)$$

M تعداد کلاس‌ها، $P(C_i)$ احتمال کلاس C_i، $P(t)$ احتمال وجود اصطلاح t و $P(\bar{t})$ احتمال عدم وجود اصطلاح t را نشان می‌دهد. $P(C_i|t)$ احتمال کلاس C_i به‌شرط وجود اصطلاح t است. $P(C_i|\bar{t})$ احتمال کلاس C_i به‌شرط عدم وجود اصطلاح t است. طریقه محاسبه روش شاخص جینی در رابطه (٤) بیان‌شده است. این روش به شکل بهبودیافته در دسته‌بند درخت تصمیم استفاده می‌شود [۱۰].

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2\, P(C_i|t)^2 \quad (٤)$$

M تعداد کلاس‌ها، $P(t|C_i)$ احتمال وجود اصطلاح t به‌شرط کلاس C_i است. $P(C_i|t)$ احتمال کاس C_i به‌شرط وجود اصطلاح t است.

روش انتخاب ویژگی مبتنی بر فیلتر زمانی مطلوب است که امتیاز بالایی را به اصطلاحات متمایز و امتیاز پایینی را به موارد نامربوط اختصاص دهد. یکی از روش‌های انتخاب ویژگی مناسب برای دسته‌بندی متن، انتخاب‌گر ویژگی متمایز است [۱۰]. این روش اصطلاحات متمایز را انتخاب می‌کند [۸]. طریقه محاسبه این روش در رابطه (۵) قابل‌مشاهده است.

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1} \quad (۵)$$

$P(t|C_i)$ احتمال وجود کلاس C_i به‌شرط وجود اصطلاح t است. $P(\bar{t}|C_i)$ احتمال عدم وجود اصطلاح t به‌شرط کلاس C_i است. $P(t|\bar{C}_i)$ احتمال شرطی اصطلاح t برای تمام کلاس‌ها به‌جز کلاس C_iاست.

فراوانی سند مربوط به یک اصطلاح، تعداد اسناد آموزشی است که اصطلاح موردنظر در آن اسناد وجود دارد. از این روش به دلیل

[1] Term

هزینه کم محاسبات، در دسته‌بندی متن بیشتر استفاده می‌شود [۸]. رویکرد روش فراوانی سند این است که در دسته‌بندی متن، اصطلاحات نادر غیر آموزنده هستند و باید در هنگام انتخاب ویژگی حذف شوند. نحوه محاسبه این روش در رابطه (٦) نشان داده‌شده است. N تعداد کل اسناد، $P(t)$ احتمال رخداد اصطلاح t را نشان می‌دهد.

$$DF(t) = N \cdot P(t) \quad (٦)$$

۳- روش تحقیق

برای شناسایی کاربران ارائه‌دهنده خدمات، از دسته‌بندی پیام‌های ارسالی آن‌ها در گروه‌های تلگرامی استفاده شد. روش پیشنهادی در چندین مرحله انجام شد که به‌طور خلاصه شامل جمع‌آوری اطلاعات، پیش‌پردازش، انتخاب ویژگی، اجرا روش‌های یادگیری ماشین و ارزیابی نتایج می‌شود. در ادامه جزئیات هرکدام از مراحل به‌تفصیل بیان می‌شود.

۳-۱ جمع‌آوری داده‌ها

در گام نخست یک‌میلیون پیام منتشرشده در گروه‌های مختلف تلگرامی به همراه شناسه ارسال‌کننده آن‌ها، از سامانه ایده‌کاو[1] که یک موتور جست‌وجوی تلگرام است، جمع‌آوری شد. سپس این داده‌ها برای نمایه‌سازی به الستیک سرچ[2] منتقل شدند. الستیک سرچ یک موتور جستجوگر متن است. در گام بعد با جست‌وجو در نمایه ساخته‌شده مبتنی بر شناسه تعلق‌گرفته به هر کاربر، مجموع پیام‌های ارسالی هر کاربر، جمع‌آوری و مجدداً، نمایه جدیدی ایجاد شد. در نمایه جدید هر کاربر با مجموع پیام‌های ارسالی خود، به‌عنوان یک سند لحاظ می‌شود. تعداد اسناد ایجادشده در این مرحله به ۳۰ هزار می‌رسد.

در گام بعد، دسته‌بندی در قالب برچسب‌زنی دستی بر روی اسناد انجام شد. فرایند برچسب‌زنی به این صورت است که، اگر مجموع پیام‌های ارسالی یک کاربر، از نوع تبلیغات و معرفی خدمات باشد به آن سند برچسب کلاس مثبت تعلق می‌گیرد و آن کاربر به‌عنوان ارائه‌دهنده خدمات شناخته می‌شود. در غیر این صورت، سند برچسب کلاس منفی می‌گیرد. در طی این پژوهش پیام‌های مربوط به ۳۰ هزار کاربر تلگرام برچسب‌زنی شد. از این تعداد ۹،۵۰٦ کاربر دارای برچسب مثبت، ۱۷،۸۲۸ کاربر دارای برچسب منفی و ۲،٦٦٦ کاربر دارای برچسب نامشخص بودند. ابتدا داده‌های با برچسب‌های نامشخص کنار گذاشته شدند. لذا ۲۷،۳۳٤ کاربر با برچسب مثبت و منفی داریم. داده‌های به‌دست‌آمده را به دو قسمت تقسیم می‌کنیم و مراحل آماده‌سازی داده‌ها را بر روی این تقسیم‌بندی انجام می‌دهیم. ۸۰ درصد اول داده‌ها را به‌عنوان مجموعه آموزشی و ۲۰ درصد بعدی را به‌عنوان مجموعه آزمون در نظر می‌گیریم. در داده‌های آموزش و آزمون، نمونه‌های با برچسب منفی بیشتر از نمونه‌های با برچسب مثبت است.

در این پژوهش جهت دسته‌بندی از چهار روش ماشین بردار پشتیبان[3] (*SVM*)، بیزین ساده[4] (*NB*)، شبکه عصبی[5] (*MLP*) و درخت تصمیم[6] (*DT*) استفاده کردیم. همچنین از معیار میانگین میکرو[7] (*Micro-F1*)، برای ارزیابی نتایج استفاده شد.

۳-۲ انتخاب ویژگی

دسته‌بندی به‌صورت دستی، بسیار زمان‌بر است، لذا برای تشخیص کاربران ارائه‌دهنده خدمات، از روش‌های دسته‌بندی خودکار اسناد استفاده شد. روش‌های دسته‌بندی بر مبنای ویژگی‌های استخراجی از متن، عمل می‌کنند. در این پژوهش ویژگی‌های استخراجی، همان کلمات موجود در پیام‌ها هستند. لذا در گام بعد بر روی همه اسناد جداسازی کلمات[8] انجام شد و درنهایت برای هر کلمه، فرکانس تکرار آن در کل اسناد محاسبه شد. بعد از حذف کلمات توقف[9]، برای رسیدن به ویژگی‌های بهتر، کلماتی که فرکانس بسیار بالا یا بسیار پایین داشتند حذف شدند زیرا کلماتی که نادر و یا بسیار پرتکرار هستند معمولاً ارزش بالایی در دسته‌بندی ندارند. سپس کلمات باقیمانده، به‌عنوان ویژگی و واژگان موجود در کیسه کلمات[10] در نظر گرفته شدند.

[1] Idekav.com
[2] Elastic search
[3] Support Vector Machine (SVM)
[4] Naive Bayes (NB)
[5] Multi-Layer Perceptron (MLP)
[6] Decision Tree (DT)
[7] Micro-averaging
[8] Tokenization
[9] Stop words
[10] Bag Of Words (BOW)

بعد از مراحل پیش‌پردازش تعداد کل ویژگی‌های استخراج‌شده از متن اصلی برابر ۷۴۳۵ عدد است. مجموعه ویژگی‌های به‌دست‌آمده از مرحله پیش‌پردازش را با $F=\{F_1, F_2, \cdots, F_N\}$ نشان داده و آن را به‌عنوان مجموعه‌ای از تمام ویژگی‌های اولیه تعریف می‌کنیم که در این مجموعه $N=$ ۷۴۳۵ است.

دسته‌بندی اسناد با استفاده از این تعداد ویژگی، زمان آموزش مدل را در یادگیری ماشین، بسیار طولانی می‌کند. لذا برای کاهش این زمان و افزایش دقت دسته‌بندی، نیاز به اعمال روش‌های انتخاب ویژگی بر روی ویژگی‌های اولیه است تا ضمن حذف ویژگی‌های زائد، زمان دسته‌بندی اسناد را کاهش و دقت دسته‌بندی را افزایش دهد. برای اعمال روش‌های انتخاب ویژگی نیاز به ماتریس کلاس_اصطلاح است. ماتریس کلاس_اصطلاح ماتریسی است که سطرهای آن نماینده دو کلاس مثبت و منفی و ستون‌های آن، ویژگی‌های اولیه هستند. درایه a_{ij} از این ماتریس، تعداد اسناد کلاس j ام که ویژگی i ام در آن وجود داشته را نشان می‌دهد. با استفاده از این ماتریس می‌توان هر یک از روش‌های انتخاب ویژگی محلی و سراسری را اجرا کرد.

انتخاب ویژگی با روش‌های محلی و سراسری

در این مرحله با استفاده از روش‌های انتخاب ویژگی محلی CC و OR و روش‌های انتخاب ویژگی سراسری GI، IG، DF و DFS به ویژگی‌های هر کلاس C_i موجود در مجموعه F یک امتیاز داده می‌شود. این امتیازها نشان‌دهنده تفاوت میان ویژگی‌ها در یک مجموعه داده است. در مرحله بعد همه ویژگی‌ها با توجه به امتیازی که در مرحله انتخاب ویژگی کسب کرده‌اند به ترتیب نزولی مرتب می‌شود. سپس در هر یک از روش‌های انتخاب ویژگی محلی، L ویژگی با بالاترین امتیاز به‌عنوان ویژگی‌های محلی و در هر یک از روش‌های انتخاب ویژگی سراسری، G ویژگی با بالاترین امتیاز به‌عنوان ویژگی‌های سراسری نهایی انتخاب می‌شوند. مقدار L و G، یک تعداد تعیین‌شده است که به‌صورت تجربی حاصل‌شده است. در هر روش محلی، $F\ Local \subseteq F$ به‌عنوان مجموعه‌ای از ویژگی‌های انتخاب‌شده محلی، شامل تعداد L ویژگی است. در هر روش سراسری، $F\ Global \subseteq F$ را به‌عنوان مجموعه‌ای از ویژگی‌های انتخاب‌شده سراسری تعریف می‌کنیم که شامل تعداد G ویژگی است.

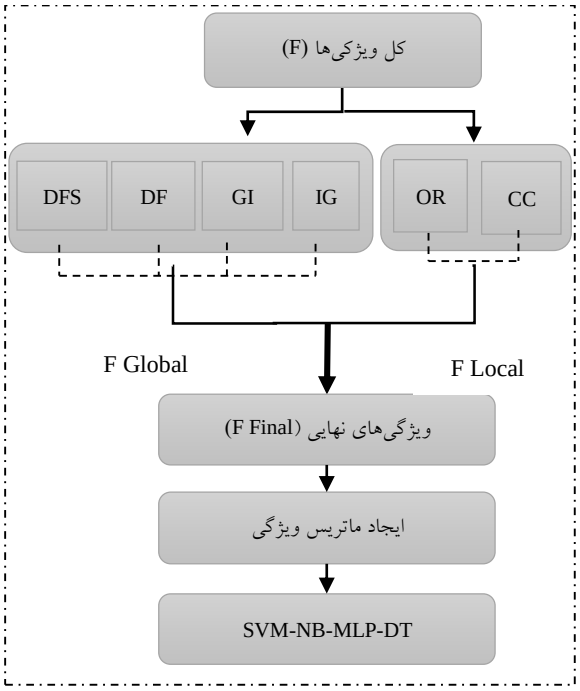
ترکیب مجموعه ویژگی‌ها

از دو مرحله قبل، دو مجموعه ویژگی به دست آمد که هرکدام مقدار تعیین‌شده‌ای دارند. در این مرحله با توجه به شکل (۱) مجموعه ویژگی خروجی مرتب‌شده از هر یک از روش‌های محلی را با مجموعه ویژگی خروجی مرتب‌شده از هر یک از روش‌های سراسری ترکیب می‌کنیم. از این دو مجموعه، ویژگی‌هایی انتخاب می‌شوند که امتیازات بالاتری داشته باشند. درنهایت یک مجموعه ویژگی نهایی $F\ Final$ به دست می‌آید.

$F\ Final \subseteq \{F\ Local \cup F\ Global\}$ است. مجموعه ویژگی نهایی $F\ Final$ مقدار تعیین‌شده‌ای دارد و در آن بهترین ویژگی‌ها با مقدار بالاتر انتخاب‌شده‌اند. بعد از انتخاب ویژگی در هر روش، نوبت به ساخت بردار ویژگی می‌رسد. بردار ویژگی مورداستفاده برای روش‌های یادگیری به‌صورت ورودی است و هر ویژگی حضور یا عدم حضور یک کلمه را نشان می‌دهد. مجموع بردارهای ویژگی به‌عنوان ماتریس در نظر گرفته می‌شود. لذا برای هر ترکیب، ماتریس ویژگی مربوط به آن روش ساخته می‌شود. در گام بعد، عملکرد روش‌های انتخاب ویژگی در دسته‌بندی، توسط روش‌های یادگیری ماشین ارزیابی می‌شود.

انتخاب ویژگی با ترکیب روش‌های فیلتر و یادگیرمبنا

در گام بعد به‌منظور افزایش دقت، از ترکیب روش‌های فیلتر و یادگیرمبنا استفاده می‌شود. بدین منظور قبل از این‌که ویژگی‌های اصلی به روش یادگیرمبنا داده شود، باید از فیلتر روش پیشنهادی اول عبور کنند. در گام بعد ویژگی‌های مناسب انتخاب‌شده توسط روش‌های فیلتر، به‌عنوان ویژگی‌های ورودی، به روش یادگیرمبنا داده می‌شوند. با ترکیب روش‌های انتخاب ویژگی فیلتر و روش یادگیرمبنا (*SFS*) و بررسی نتایج، مشاهده می‌شود که استفاده از این روش سبب کاهش چشمگیر ویژگی‌ها و افزایش دقت در دسته‌بندی می‌شود.



شکل (۱): روند اجرایی روش پیشنهادی

۳-۳ ارزیابی

در این پژوهش عملکرد روش‌های یادگیری ماشین با تعداد متفاوتی از ویژگی‌ها، بر روی نمونه‌های ارزیابی، موردبررسی قرار گرفت و در هر روش دسته‌بندی، بهینه‌ترین تعداد ویژگی انتخاب شد. مجموعه بهینه از ویژگی‌ها، مجموعه‌ای است که با تعداد ویژگی کمتر، بیش‌ترین افزایش دقت را در دسته‌بندی حاصل کند. برای انتخاب زیرمجموعه بهینه، روند کار بدین‌صورت است که در هر مرحله تعداد ۵۰، ۱۰۰، ۲۰۰، تا ۴۰۰ ویژگی انتخاب می‌شود. سپس این ویژگی‌ها در یادگیرنده‌های متفاوت قرارگرفته و ارزیابی می‌شوند. به عبارتی تعداد ویژگی‌های بهینه را با استفاده از دادگان آموزش به دست می‌آوریم.

بعد از به دست آوردن مجموعه ویژگی بهینه مربوط به هر روش دسته‌بندی، در گام بعد، عملکرد روش‌های متفاوت دسته‌بندی بر روی‌داده‌های آزمون با این مجموعه ویژگی بهینه، موردبررسی قرار گرفت که نتایج آن در بخش ۳ بیان‌شده است.

٤- نتایج

در این بخش ابتدا نتایج استفاده از روش‌های انتخاب ویژگی تکی و ترکیبی فیلتر محلی و سراسری در دسته‌بندهای مختلف بر روی‌داده‌های آموزشی بررسی می‌شود. با توجه به نتایج به‌دست‌آمده، مجموعه ویژگی بهینه انتخاب می‌شود. در گام بعد، نتایج حاصل از هر دسته‌بند با استفاده از مجموعه بهینه، بر روی‌داده‌های آزمون موردبررسی قرار می‌گیرد.

٤-١ ارزیابی روش‌های انتخاب ویژگی تکی فیلتر محلی و سراسری

استفاده از روش‌های انتخاب ویژگی تکی بر روی‌داده‌های آموزشی، در بیشتر موارد سبب افزایش دقت شد. طبق نتایج ذکرشده در جدول (۱) روش انتخاب ویژگی *CC* در میان همه روش‌های انتخاب ویژگی تکی محلی و سراسری، بهترین عملکرد را داشت به‌طوری‌که در تمامی روش‌های دسته‌بندی، بالاترین میزان دقت را حاصل کرد. دسته‌بندهای ماشین بردار پشتیبان و درخت تصمیم عملکرد بهتر و درصد افزایش دقت بیشتری نسبت به سایر روش‌ها داشته‌اند.

جدول (۱): تعداد ویژگی‌های بهینه در بهترین روش‌های تکی انتخاب ویژگی محلی و سراسری در معیار Micro_F1

روش دسته‌بندی	روش انتخاب ویژگی بهتر	تعداد ویژگی	دقت با انتخاب ویژگی	دقت بدون انتخاب ویژگی	درصد کاهش ویژگی	درصد افزایش دقت
SVM	CC	٤٠٠	۰/۸۳۵	۰/۷۰۵	۹٤/٦۲	۱۸/۵۱
DT	CC	٤٠٠	۰/۸۰۳	۰/۷۰۲	۹٤/٦۲	۱٤/۳۰
NB	CC	٤٠٠	۰/۸۳۵	۰/۷۵۵	۹٤/٦۲	۱۰/۵۳
MLP	CC	۳۰۰	۰/۸۲	۰/۷۹٦	۹۵/۹٦	۲/۹۵

در دسته‌بند ماشین بردار پشتیبان، روش انتخاب ویژگی *CC* مقدار دقت بالاتری با تعداد ٤۰۰ ویژگی داشت. دقت این روش ۰/۸۳۵ است. مقدار دقت بدون کاهش ویژگی برای این دسته‌بند ۰/۷۰ بود. بنابراین استفاده از روش‌های کاهش ویژگی دقت را تا ۱۸/۵۱

درصد افزایش داده است. دسته‌بند درخت تصمیم برای روش انتخاب ویژگی *CC* مقدار دقت بالاتری با ۴۰۰ ویژگی داشت. دقت این روش ۰/۸۰ است. مقدار دقت بدون کاهش ویژگی برای این دسته‌بند ۰/۷۰۲ بود. بنابراین استفاده از روش‌های کاهش ویژگی دقت را تا ۱۴/۳۰ درصد افزایش داده است که در مقایسه با دسته‌بند ماشین بردار پشتیبان درصد کمتری است.

۴-۲ ارزیابی ترکیب روش‌های انتخاب ویژگی محلی و سراسری

در این بخش، عملکرد ترکیبی روش‌های انتخاب ویژگی محلی و سراسری بر روی‌داده‌های آموزشی مقایسه می‌شود. روش‌های موردبررسی را با تعداد ویژگی‌های متفاوت اجرا کرده و نتایج را مقایسه کردیم. طبق نتایج جدول (۲) استفاده از روش‌های انتخاب ویژگی ترکیبی در اکثر موارد سبب افزایش دقت شد. میزان افزایش دقت با استفاده از روش‌های کاهش ویژگی نسبت به دقت‌های بدون کاهش ویژگی به ترتیب مربوط به دسته‌بندهای ماشین بردار پشتیبان، درخت تصمیم و در آخر شبکه عصبی است، که از این میان میزان درصد افزایش دقت در دسته‌بند ماشین بردار پشتیبان قابل‌ملاحظه است.

همچنین میزان افزایش دقت با روش‌های انتخاب ویژگی ترکیبی شامل *CC* بهتر است. در دسته‌بند ماشین بردار پشتیبان روش انتخاب ویژگی *CC_GI* میزان دقت بالاتری داشت. میزان دقت این روش ۰/۸۲ است. میزان دقت، بدون کاهش ویژگی برای این دسته‌بند ۰/۷۲ است. بنابراین استفاده از روش‌های کاهش ویژگی دقت را تا ۱۶ درصد افزایش داده است. در بخش قبل برای روش‌های تکی مشاهده شد که دقت روش *CC* به‌تنهایی نیز دارای مقادیر بالایی بود. دسته‌بند ماشین بردار پشتیبان در مقایسه با دسته‌بندهای دیگر، نسبت به مقدارهای بدون کاهش ویژگی دارای درصد افزایش دقت بیشتری برای روش‌های ترکیبی است. دسته‌بند درخت تصمیم برای روش انتخاب ویژگی ترکیبی *CC_DFS* مقدار دقت بالاتری با ۴۰۰ ویژگی داشت. مقدار دقت این روش ۰/۸۰ است. درحالی‌که مقدار دقت بدون کاهش ویژگی برای این دسته‌بند ۰/۷۰ است. بنابراین در درخت تصمیم استفاده از روش‌های کاهش ویژگی دقت را تا ۱۴ درصد افزایش داده است که در مقایسه با دسته‌بند ماشین بردار پشتیبان درصد کمتری است.

جدول (۲): تعداد ویژگی‌های بهینه در بهترین ترکیب روش‌های انتخاب ویژگی محلی و سراسری در معیار Micro_F1

روش دسته‌بندی	روش انتخاب ویژگی بهتر	تعداد ویژگی بهینه	دقت با انتخاب ویژگی	دقت بدون انتخاب ویژگی	درصد کاهش ویژگی	درصد افزایش دقت
SVM	CC_GI	۴۰۰	۰/۸۲۲	۰/۷۰۵	۹۴/۶۲	۱۶/۵۹
DT	CC_DFS	۴۰۰	۰/۸۰۵	۰/۷۰۲	۹۴/۶۲	۱۴/۷۴
NB	CC_DFS	۴۰۰	۰/۸۳۱	۰/۷۵۵	۹۴/۶۲	۱۰/۰۶
MLP	CC_GI	۴۰۰	۰/۸۲۳	۰/۷۹۶	۹۴/۶۲	۳/۳۸

پس از بررسی نتایج مختلف، تعداد ویژگی بهینه برای داده‌های این پژوهش، ۴۰۰ ویژگی در نظر گرفته شد. لذا در مرحله بعد یادگیرنده‌های مختلف را بر روی‌داده‌های آزمون با زیرمجموعه ویژگی بهینه به‌دست‌آمده، موردسنجش قرار می‌دهیم.

در جدول (۳) و (۴) مقادیر نتایج ارزیابی یادگیرنده‌های مختلف بر روی‌داده‌های آزمون قابل‌مشاهده است که به‌طور میانگین در روش‌های انتخاب ویژگی متفاوت، ترکیب روش‌های انتخاب ویژگی مبتنی بر فیلتر محلی و سراسری، عملکرد دسته‌بندی بهتری نسبت به روش‌های انتخاب ویژگی تکی داشته است. این ترکیب‌ها، با تولید زیرمجموعهٔ بهینه از تمامی ویژگی‌های مهم و مؤثر، ضمن کاهش ابعاد ویژگی‌ها، سبب افزایش دقت شدند.

جدول (۳): دقت دسته‌بند NB و SVM با استفاده از روش‌های انتخاب ویژگی در معیار Micro_F1

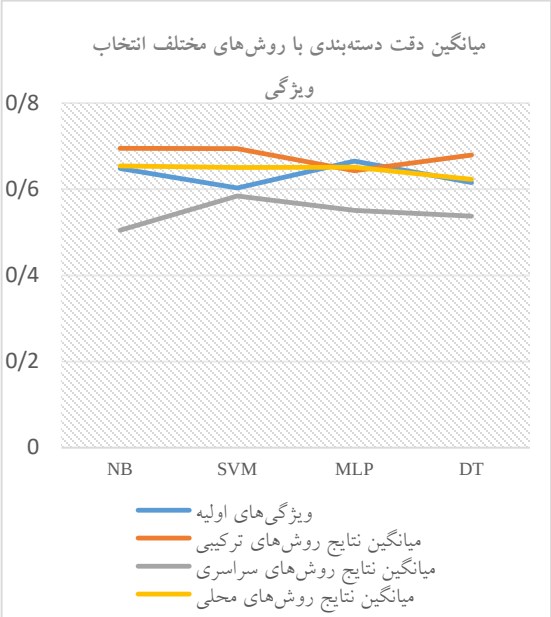
روش انتخاب ویژگی	NB	SVM
No FS	۰/۶۴۸۵	۰/۶۰۳۲
OR_DFS	۰/۷۰۵۳	۰/۷۰۲۵
OR_IG	۰/۵۷۵۶	۰/۶۰۲۵
OR_DF	۰/۷۲۳	۰/۷۰۵
OR_GI	۰/۷۰۳۵	۰/۷۰۲۲
CC_DFS	۰/۷۱۲۵	۰/۷۱۱۴
CC_IG	۰/۶۹۴۳	۰/۶۹۵۸

CC_DF	۰/۷۲۵٤	۰/۷۱٦٥
CC_GI	۰/۷۱۹٥	۰/۷۱٦٥
DFS	۰/٤۲٥٦	۰/٥۹٦٤
IG	۰/٤۰۲٥	۰/٥۳۲۸
GI	۰/٥۹٤	۰/٦۰۹٥
DF	۰/٥۹٦٤	۰/٥۹۸۸
OR	۰/٥٦٤۲	۰/٥۸٦٤
CC	۰/۷٤٤٥	۰/۷۱٥۲

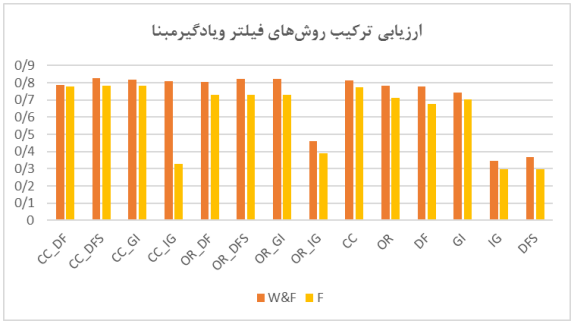
جدول (٤): دقت دسته‌بند MLP و DT با استفاده از روش‌های انتخاب ویژگی در معیار Micro_F1

روش انتخاب ویژگی	MLP	DT
No FS	۰/٦٦٥	۰/٦۱٥٤
OR_DFS	۰/۷۰۳٤	۰/۷۰۲۹
OR_IG	۰/٤۷۲٥	۰/٥٦۳٥
OR_DF	۰/٦۹٤	۰/۷۱۳٥
OR_GI	۰/۷۰۱۳	۰/٦۹۷۷
CC_DFS	۰/۷۱٤٥	۰/٦۹٦۳
CC_IG	۰/۳۹٦۸	۰/٦٦۸٥
CC_DF	۰/۷۳٦۸	۰/٦۹۹٤
CC_GI	۰/۷۲٦۱	۰/٦۹٤۲
DFS	۰/٥٤۸۲	۰/٤۸٥٦
IG	۰/٥۰۳٥	۰/٤۹٦٤
GI	۰/٥۹٦٤	۰/٥۹۹٤
DF	۰/٥٥٦	۰/٥٦۸
OR	۰/٥٦۷۳	۰/٥٦۲٥
CC	۰/۷۳٥٦	۰/٦۸۳٤

شکل (۲): میانگین نتایج روش‌های انتخاب ویژگی متفاوت در انواع دسته‌بندها را نشان می‌دهد. نتایج حاصل، بیانگر بیشترین تأثیر روش ترکیبی انتخاب ویژگی به ترتیب بر روی دسته‌بندهای ماشین بردار پشتیبان، بیزین ساده درخت تصمیم بوده است. در دسته‌بند شبکه عصبی نیز باوجود کاهش دقت نسبت به حالت بدون انتخاب ویژگی، عملکرد این دسته‌بند با کاهش تعداد ویژگی‌ها از ۷000 ویژگی به ٤۰۰ ویژگی سنجیده شده است که این کاهش ویژگی سبب کاهش چشم‌گیر زمان پردازش می‌شود و نتیجه بااهمیتی است. همچنین در میان روش‌های انتخاب ویژگی، عملکرد روش‌های ترکیبی به‌طور میانگین بهتر از روش‌های محلی و سراسری است. در میان روش‌های تکی، روش‌های محلی بهتر از روش‌های سراسری در دسته‌بندی عمل کردند.



شکل (۲): مقایسه نتایج دسته‌بندی با روش‌های انتخاب ویژگی فیلتر تکی و ترکیبی



شکل (۳): مقایسه نتایج دسته‌بندی با استفاده از روش فیلتر و روش ترکیبی فیلتر و یادگیرمبنا با ۵۰ ویژگی در دسته‌بند NB و معیار Micro-F1

٤-۳ ارزیابی ترکیب روش‌های انتخاب ویژگی فیلتر و یادگیرمبنا

همان‌طور که در شکل (۳) مشاهده می‌شود، ترکیب روش‌های فیلتر تکی و ترکیبی با روش یادگیرمبنا *SFS* دارای دقت بالاتری است. بالاترین دقت مربوط به ترکیب روش *SFS* با روش‌های فیلتری است که شامل ضریب همبستگی یا *CC* هستند. همچنین از میان روش‌های تکی، ترکیب *SFS* با *CC* یا ضریب همبستگی که از

روش‌های محلی است دارای دقت بالاتری است. از میان روش‌های ترکیبی، ترکیب این روش یادگیرمبنا با روش *CC_DFS* دارای دقت بالاتری است.

۵- بحث و نتیجه‌گیری

هدف از این پژوهش شناسایی کاربران ارائه‌دهنده خدمات در تلگرام با استفاده از دسته‌بندی پیام‌های ارسالی آن‌ها در گروه‌های مختلف تلگرامی است. برای شروع کار ابتدا پیام‌های ارسال‌شده توسط هر کاربر در گروه‌های مختلف تلگرامی از سامانه ایده‌کاو استخراج شد. سپس برچسب‌زنی دستی انجام شد. دسته‌بندی دستی اسناد، کاری زمان‌بر و پرهزینه است. لذا استفاده از روش‌های دسته‌بندی خودکار متن، امری ضروری است. روش‌های دسته‌بندی خودکار بر مبنای ویژگی‌های استخراجی از متن، عمل می‌کنند. در این پژوهش ویژگی‌های استخراجی، همان کلمات موجود در پیام‌ها هستند. اما ابعاد بزرگ فضای ویژگی یکی از مشکلات اصلی در دسته‌بندی خودکار است. لذا با استفاده از روش‌های انتخاب ویژگی سعی در کاهش ویژگی‌ها و انتخاب ویژگی‌های مرتبط شد.

در مرحله اول با استفاده از پرکاربردترین روش‌های انتخاب ویژگی مبتنی بر فیلتر که به دودسته محلی و سراسری تقسیم می‌شوند ویژگی‌های غیرضروری حذف شد. در مرحله دوم، هر یک از این روش‌های محلی و سراسری باهم ترکیب و ویژگی‌های بهتر انتخاب شد. نتایج این آزمایش‌ها نشان داد که ترکیب روش‌های تکی انتخاب ویژگی، در اکثر موارد می‌تواند سبب افزایش دقت نسبت به حالت بدون انتخاب ویژگی، در شناسایی کاربران ارائه‌دهنده خدمات شود.

در مرحله سوم خروجی ویژگی‌های روش ترکیبی مرحله دوم به‌عنوان ورودی برای روش انتخاب ویژگی یادگیرمبنا استفاده شد. با این روش بار دیگر ویژگی‌های کم‌اهمیت حذف می‌شوند. این مراحل با تعداد ویژگی‌های مختلف انجام شد و درمجموع نتایج نشان داد که روش ترکیبی فیلتر ویادگیر مبنا عملکرد بهتری نسبت به روش‌های فیلتر دارد.

به‌طور میانگین، از میان پرکاربردترین روش‌های سراسری که در این پژوهش استفاده‌شده است، روش *DF* یا فرکانس سند و *GI* جهت دسته‌بندی پیام‌های تلگرام دارای دقت بالاتری بودند. از میان پرکاربردترین روش‌های محلی، روش *CC* یا ضریب همبستگی دارای دقت بالاتری بود. در روش‌های ترکیبی که با استفاده ترکیب پرکاربردترین روش‌های تکی استفاده‌شده در این پژوهش به دست آمد، روش *CC_DF* دقت بالاتری را در دسته‌بندی پیام‌های تلگرام نشان داد. درنهایت بالاترین دقت مربوط به ترکیب روش *SFS* با روش‌های فیلتری است که شامل ضریب همبستگی یا *CC* هستند. به‌طورکلی استفاده از روش‌های انتخاب ویژگی جهت دسته‌بندی پیام‌های تلگرام به‌منظور شناسایی کاربران ارائه‌دهنده خدمت، ضمن کاهش قابل‌توجه ویژگی‌ها و انتخاب ویژگی‌های مرتبط، سبب بالا رفتن دقت عملکرد روش‌های یادگیری ماشین و کاهش زمان آموزش مدل‌های یادگیری می‌شود.

References

[1] X. Deng, Y. Li, J. Weng, and J. Zhang, (2019), Feature selection for text classification: A review, Multimedia Tools & Applications, vol. 78, no. 3, pp..113-115

[2] M. Nekkaa and D. Boughaci, (2016), Hybrid harmony search combined with stochastic local search for feature selection, Neural Processing Letters, vol.44, no. 1, pp. 199-220.

[3] G. BİRİCİK, B. Dri, and A. C. SÖNMEZ, (2012), Abstract feature extraction for text classification, Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20,no. Sup. 1, pp. 1137-1159.

[4] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and O. A. Alomari, (2017), Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering, Expert Systems with Applications, vol.84, pp. 24-36.

[5] D. Ö. Şahin and E. Kılıç, (2019), "Two new feature selection metrics for text classification", Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, vol. 60, no. 2, pp. 162-171.

[6] G. Chandrashekar and F. Sahin, (2014), "A survey on feature selection methods, Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28.

[7] A. K. Uysal, (2016), An improved global feature selection scheme for text classification, Expert systems with Applications, vol. 43, pp. 82-92.

[8] D. Agnihotri, K. Verma, and P. Tripathi, (2017), Variable global feature selection scheme for automatic classification of text documents, Expert Systems with Applications, vol. 81, pp. 268-281.

[9] A. Melo and H. Paulheim, (2019), Local and global feature selection for multilabel classification with binary relevance, Artificial intelligence review, vol. 51, no.1, pp. 33-60.
[10] A. Kursat. Uysal, (2018), "On two-stage feature selection methods for text classification", IEEE Accebss, vol. 6, pp.43233-43251.
[11] G. Wei, J. Zhao, Y. Feng, A. He, and J. Yu, (2020), A novel hybrid feature selection method based on dynamic feature importance, Applied Soft Computing, vol.93, p. 106337.
[12] L. N. H. Nam and H. B. Quoc, (2017), The hybrid filter feature selection methods for improving high-dimensional text categorization, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 25, no. 02, pp.235.
[13] Asif, M., Nagra, A. A., Ahmad, M. B., & Masood, K, (2022),. Feature selection empowered by self-inertia weight adaptive particle swarm optimization for text classification. Applied Artificial Intelligence, vol. 36, no. 01, p.2004345.
[14] Effrosynidis, D., & Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. Ecological Informatics, vol. 61, p.101224.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 2, Autumn & Winter 2023, pp.54-66

service providers Identification in Telegram Persian messages based on feature selection methods

Farzaneh Ebrahimi [1], Mohammad Ali Zare Chahooki [2*], Ali Hashemi[3]

[1]Senior student of Computer Engineering, Yazd University

[2]Assistant Professor, Faculty of Computer Engineering, Yazd University

[3]PhD in Computer Engineering, Yazd University

Article Information

Original Research Paper

Received:
2022 July 31

Accepted:
2023 January 30

Keywords:
classification, Service Provider Users, Feature Selection, Feature Reduction, Machine Learning.

Corresponding Author*:
chahooki@yazd.ac.ir

Abstract

Telegram Messenger is a suitable platform for users who are looking to buy a product or receive services online. In these messengers, it is not possible to have direct access to the providers of goods and services, and in order to request the product, one must first become a member of the related groups and channels telegram. The purpose of this study is to directly identify users of service providers using the classification of Persian messages published in Telegram. One of the problems with categorizing these messages is the large size of the feature space, which reduces accuracy and increases classification time. Feature selection methods were used to solve this problem. The proposed method of this research is based on a combination of feature selection methods based on local and global filters. In this regard, in the first step, using the most widely used methods for selecting local and global filter feature, related features are selected. In the second step, a combination of local and global filtering methods is used to identify better features and increase classification accuracy. The innovation of this research is in using the combined methods of feature selection for automatic classification of Telegram Persian messages, in order to identify the users of the service provider. The proposed method, while reducing the number of features and selecting related features, improves the performance of classification and Identification of service providers.

: 10.22034/ABMIR.2023.18780.1012

E-ISSN: 2821-2037