بررسی جامع ترکیب روش‌های محلی و سراسری انتخاب ویژگی برای شناسایی درخواست در تلگرام

زهرا خلیفه‌زاده و محمدعلی زارع‌چاهوکی*

گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه یزد، ایران

چکیده

تلگرام سرویس پیام‌رسان متن‌بازی مبتنی بر رایانش ابری است. تلگرام به دلایلی همچون پشتیبانی از زبان‌ها، امکان ایجاد گروه و کانال با تعداد کاربران متعدد، به پیام‌رسانی محبوب و پرکاربرد تبدیل شد. داده‌های متنی زیادی که در گروه‌های تلگرامی وجود دارد حاوی دانش پنهانی هستند. استخراج این دانش‌ها، نظیر درخواست‌های موجود در پیام‌های کاربران می‌تواند سودمند باشد. لذا با شناسایی درخواست‌ها می‌توان به نیازهای کاربران پاسخ داد و به دسترسی سریع آن‌ها به خواسته‌هایشان کمک کرد که این امر موجب توسعه کسب‌وکار کاربران می‌شود. با توجه به ابعاد بالای فضای ویژگی‌ها در داده‌های متنی، کاهش ویژگی‌ها از طریق انتخاب ویژگی ضرورت می‌یابد. از روش‌های انتخاب ویژگی، دو روش مبتنی برفیلتر محلی و سراسری انتخاب شد. با بررسی و ترکیب پرکاربردترین آن‌ها به زیرمجموعه بهینه‌ای از ویژگی‌های بااهمیت دست یافتیم. این روش ترکیبی، با کاهش بهینه ویژگی‌ها سبب افزایش دقت در شناسایی درخواست، افزایش کارایی دسته‌بندی متن، کاهش زمان آموزش و محاسبات شد.

واژگان کلیدی: انتخاب ویژگی، متن‌کاوی، دقت دسته‌بندی، یادگیری ماشین

A General Investigation on the Combination of Local and Global Feature Selection Methods for Request Identification on Telegram

Zahra Khalifeh zadeh & Mohammad Ali Zare-Chahooki*

Department of Computer Engineering, Faculty of Engineering, Yazd University, Iran

Abstract

Nowadays, the use of various messaging services is expanding worldwide with the rapid development of Internet technologies. Telegram is a cloud-based open-source text messaging service. According to the US Securities and Exchange Commission and based on the statistics given for October 2019 to present, 300 million people worldwide used telegram per month. Telegram users are more concentrated in countries such as Iran, Venezuela, Nigeria, Kenya, Russia, and Ukraine. This messenger has become a popular and extensively used messenger because it supports various languages and provides diverse services such as creating groups and channels with a large number of users and members. There is a large amount of contextual data on telegram groups containing hidden knowledge; the extraction of this knowledge can be beneficial. The requests on telegram users' messages are examples of this sort of data with hidden knowledge. Hence, identifying requests can respond to users' needs and help them fulfill their desires immediately; this drives users' business development. The authors identified these requests in a telegram search engine named the Idekav system of Yazd University. Then, the authors created opportunities to earn money by sending these requests to the business owners who were able to respond to them. Given the high dimensions of feature space in contextual data, it is necessary to reduce attributes using feature selection.

In the present study, the appropriate features were selected for Persian text classification and request identification. Among the feature selection methods, two local and global filter-based methods were chosen. By general investigation and combining the most extensively used filter-based FS methods,

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۸/۲/۲۲ • تاریخ پذیرش: ۱۳۹۹/۱۲/۱۳ • تاریخ انتشار: ۱۴۰۱/۷/۷ • نوع مطالعه: پژوهشی

an optimal subset of important features was obtained. This hybrid feature selection method resulted in increased request identification accuracy, improved Persian text classification efficiency, and reduced training time and computation by optimizing the feature reduction. Of course, it is noteworthy that the classification accuracy is reduced in some methods; however, this value is negligible compared to the feature reduction value. Incorporating the concept of opinion mining into the analysis of emotions and questions can be a method to identify positive or negative demand in social networks. Therefore, the requests in the Persian telegram messages can be identified using opinion mining researches. For experiments in the present article, a dataset called Persian is used, which is extracted from the Idekav system. The selection of suitable features to increase model accuracy in request identification is an important part of this research. The support vector machine was employed to calculate accuracy. Given the acceptable results of the SVM, its various kernels were also calculated. Micro-averaging and macro-averaging criteria were also used for evaluation. Model inputs include many optimal feature subsets. Furthermore, feature selection methods have been proposed to produce suitable features for each model for increasing the accuracy of the model. Afterward, among all the features investigated, appropriate features have been selected for each of the applied feature selection models. For a more precise explanation, the main innovations of the present study are as follows:

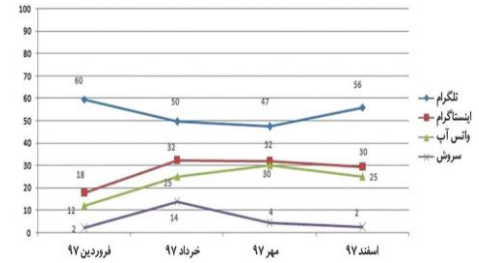
- **Use of the most common filters based on local and global feature selection methods to find the optimal feature set.**
- **Use of hybrid methods to create suitable features for predictive models of accuracy in Persian text classification and their application in identifying requests in Persian messages on telegram.**
- **Selecting suitable features to increase accuracy and reduce computational time for each of the models under consideration. In this regard, in addition to picking an efficient algorithm, it is attempted to provide a method for making more appropriate choices.**
- **Evaluation and testing of the proposed models for a large set of Persian data and many different features.**

Keywords: Feature Selection, Text mining, Classification Accuracy, Machine Learning

۱- مقدمه

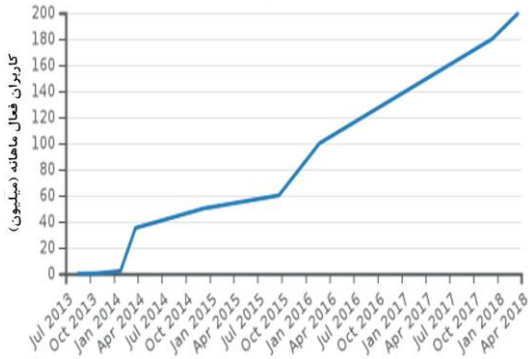
امروزه با پیشرفت سریع فناوری‌های اینترنتی، استفاده از پیام‌رسان‌های مختلف در سراسر جهان گسترش یافته است. تعداد کاربران این پیام‌رسان‌ها در کشورهای مختلف به‌دلیل ارائه خدمات متنوع متفاوت است. پیام‌رسان تلگرام ازجمله این پیام‌رسان‌ها است که در کشورهایی مانند ونزوئلا، نیجریه، کنیا، روسیه، اوکراین و ایران بیشتر از کشورهای دیگر استفاده می‌شود [1]. در ایران به‌دلایل مختلفی ازجمله پشتیبانی از زبان فارسی به پیام‌رسانی محبوب و پرکاربرد تبدیل شده است که حاوی اطلاعات مفید و ارزشمندی است. طبق آمارهای اخیر، شصت درصد مردم ایران از تلگرام به‌عنوان پیام‌رسان استفاده می‌کنند [2]. در اکتبر ۲۰۱۳، تلگرام یکصدهزار کاربر فعال روزانه داشت. در مارس ۲۰۱۸، کاربران فعال ماهانه تلگرام به دویست میلیون نفر رسید. طبق آمار کمیسیون بورس و اوراق بهادار ایالات‌متحده از اکتبر ۲۰۱۹، تعداد کاربران تلگرام به‌صورت ماهانه سیصد میلیون نفر در سراسر جهان است [3].

به‌دلیل وجود مخاطبان زیاد در تلگرام، فرصت‌های زیادی برای کسب درآمد از این طریق برای افراد زیادی به‌وجود آمده است. مهم‌ترین دلیل کسب درآمد حتی در کانالی با کمترین عضو، این است که این کانال‌ها نیاز واقعی افراد را رفع می‌کنند. به همین دلیل اعضای واقعی دارند و همین اعضا مشتریان وفاداری می‌شوند که سبب توسعهٔ کسب‌وکار می‌شوند.



(شکل-۱): میزان استفاده مردم ایران از شبکه‌های اجتماعی در سال ۱۳۹۷ [2]

(Figure-1): The extent of Iranian people's use of social media in 1397 [2]



(شکل-۲): میزان استفاده تلگرام از آگوست ۲۰۱۳ تا مارس ۲۰۱۸ [3]

(Figure-2): Telegram timeline (August 2013 - March 2018) [3]

به این منظور ابتدا باید این نیازها یا درخواست‌ها را پیدا کرد. در تلگرام میلیون‌ها پیام فارسی وجود دارد که به‌صورت درخواست ردوبدل می‌شوند. پیام‌هایی که از جنس درخواست هستند برای بسیاری از کسب‌وکارها

جذابیت دارند. ما در سامانه ایده‌کاو دانشگاه یزد (موتور جستجوگر تلگرام) پیام‌های از جنس درخواست را شناسایی کرده‌ایم. با تشخیص این درخواست‌ها و ارسال آن‌ها به صاحبان مشاغل مرتبط با درخواست، هم به درخواست کاربر تلگرام پاسخ مناسب داده می‌شود و هم صاحبان مشاغل مشتریان خود را از این طریق شناسایی می‌کنند. در این راستا ما نیز هزینه‌ای را بابت شناسایی و ارسال درخواست، از صاحبان مشاغل دریافت می‌کنیم؛ لذا شناسایی درخواست یک بحث مهم است که باید به آن پرداخته شود. موضوع شناسایی و پاسخ به درخواست‌ها همیشه مطرح بوده، ولی امروزه با گسترش شبکه‌های اجتماعی و دنیای دیجیتال اهمیت بیشتری یافته‌اند، امروزه هر فرد می‌تواند به‌راحتی و حتی با استفاده از دستگاه تلفن همراه خود یک درخواست را ایجاد کرده و پاسخی را دریافت کند. اگر این پاسخ متناسب با درخواست باشد، درواقع کار بازاریابی را با صرفه‌جویی در زمان و هزینه انجام داده‌ایم؛ از طرفی گسترش و محبوبیت شبکه‌های اجتماعی در بین افراد، جوامع مجازی را به منابع ارزشمندی از اطلاعات سیاسی، اجتماعی و تجاری تبدیل کرده است. این اطلاعات نمود افکار و احساسات افراد است، که کاوش آن‌ها دانش ارزشمندی را در حوزه‌های گوناگونی نظیر مدیریت ارتباط با مشتری، پی گیری افکار عمومی، عقیده‌کاوی و فیلترینگ متن فراهم می‌کند. تحلیل این حجم از اطلاعات غیر ساخت‌یافته نیازمند روش‌های بهینه متن‌کاوی و پردازش زبان طبیعی است. گاهی افراد یک جامعه دارای وجوه و ارزش‌های مشترکی هستند که این باعث تأثیر گذاشتن افراد بر روی‌هم و به‌وجودآمدن شبکه‌های اجتماعی مشترک با موضوعی خاص می‌شود. شبکه‌های اجتماعی در بسیاری از مشاغل به‌عنوان ابزاری جهت ارائه خدمات و تعامل با مشتریان تبدیل شده است؛ لذا دانش مستخرج از شبکه‌های اجتماعی مانند فیس‌بوک، تلگرام و دیگر شبکه‌های اجتماعی برای شرکت‌های پژوهشی بازاریابی، سازمان‌های افکار عمومی و دیگر واحدهای متن‌کاوی ارزش زیادی دارند. هدف عقیده‌کاوی (تحلیل احساسات، نظرکاوی، هوش مصنوعی احساسات) این است که رایانه را قادر سازیم تا احساسات را تشخیص داده و بیان کند [4-10]. درنظرگرفتن مفهوم عقیده‌کاوی در تحلیل احساسات و سؤالات، می‌تواند روشی برای شناسایی درخواست مثبت یا منفی در شبکه‌های اجتماعی باشد، لذا با استفاده از پژوهش‌های عقیده‌کاوی می‌توان به شناسایی درخواست های موجود در پیام‌های فارسی تلگرام پرداخت. در این پژوهش‌ها نیز ما با داده‌های متنی سروکار داریم که برای شناسایی عقاید و درخواست‌های پنهان در آن‌ها نیاز به پردازش و دسته‌بندی متن را ضروری می‌سازد.

قدمت مطالعات دسته‌بندی متن به سال ۱۹۶۰ بازمی‌گردد، اما مطالعات اساسی در این زمینه از سال ۱۹۹۰ اهمیت یافته و تا به امروز ادامه داشته است [11]. در دنیای دیجیتال امروز فناوری‌های اینترنتی به‌سرعت در حال افزایش و رایج‌شدن هستند. هم‌زمان با این رشد سریع و استفاده از دستگاه‌های الکترونیکی و برنامه‌های کاربردی مدرن، میزان اسناد الکترونیکی در جهان افزایش می‌یابد؛ بنابراین سازمان‌دهی سلسله مراتبی این اسناد دیجیتال ضرورت پیدا می‌کند [11 , 12]؛ اما استفاده از نیروی انسانی برای دسته‌بندی متن‌ها بسیار زمان‌بر و پرهزینه است و محدودیت کاربری را ایجاد می‌کند؛ درنتیجه، دسته‌بندی خودکار متن به‌عنوان ابزاری مناسب برای سازمان‌دهی این حجم بالای اسناد بسیار موردتوجه قرارگرفته است. ابعاد بالای فضای ویژگی مشکلی اساسی در دسته‌بندی متن است [13 , 14]. این ویژگی‌ها اغلب نامربوط و زائد هستند و بر کارایی دسته‌بندی تأثیر منفی می‌گذارند؛ ازاین‌رو انتخاب ویژگی، راه‌حلی مناسب برای کاهش ابعاد بزرگ فضای ویژگی و افزایش کارایی دسته‌بندی متن است [12-17]. مسأله انتخاب ویژگی، بخش مهمی از یادگیری ماشین [18 , 19] است که زمان آموزش و زمان محاسباتی را کاهش داده و کیفیت پیش گویی را بهبود می‌دهد [15]، [18]، [20]. هدف از انتخاب ویژگی، انتخاب بهینه‌ترین زیرمجموعه ویژگی [17]، [21] از کل فضای ویژگی‌های اصلی مسأله موردنظر است، به‌طوری‌که ضمن کاهش ابعاد به دقت دسته‌بندی مطلوبی نیز دست یافت.

در پژوهش‌ها برای مسأله انتخاب ویژگی، راه‌حل‌ها و الگوریتم‌های فراوانی ارائه شده است که بعضی از آن‌ها قدمت سی یا چهل‌ساله دارند. مشکل برخی از آن‌ها بار زیاد محاسباتی بود. برای حل این مشکل، بسیاری از پژوهش‌گران به انتخاب ویژگی پرداخته‌اند تا سرعت و دقت دسته‌بندی‌ها [11]، [17]، [22] در مسائل داده‌کاوی بیشتر شود؛ لذا مطالعات متعددی در مورد ترکیب روش‌های مختلف انتخاب ویژگی برای دسته‌بندی متن ارائه شده است. این روش‌ها به‌طورکلی به سه دسته فیلتر[1]، یادگیرمبنا[2] و یادگیرحاصل[1] طبقه‌بندی شده‌اند

[1] Filter
[2] Wrapper

[11]، [12]، [16]، [18]، [23]. روش‌های فیلتر، مستقل از الگوریتم یادگیری، ویژگی‌ها را بر اساس امتیازات مرتب می‌کنند و به‌عنوان ورودی به الگوریتم دسته‌بندی می‌دهند [11]، [12]، [18]. روش‌های یادگیرمبنا، زیرمجموعه‌ای از ویژگی‌هایی را بر اساس یک الگوریتم یادگیری انتخاب می‌کنند که باعث بیشینه‌شدن کارایی آن روش یادگیری شود [11]، [12]، [18]. روش‌های یادگیرحاصل، انتخاب ویژگی و یادگیری مدل را به‌صورت هم‌زمان انجام می‌دهند [11]، [12]، [14]، [18] و برای کاهش زمان محاسبه در روش‌های یادگیرمبنا هستند [24]. روش‌های فیلتر ازنظر محاسباتی خیلی ساده و سریع هستند [11]، [12] و به‌راحتی برای مجموعه داده‌های با ابعاد بالا به کار می‌روند. همچنین روش فیلتر نسبت شانس، مقدار زیادی از ویژگی‌های منفی [12]، [16] را تولید می‌کند؛ اما روش‌های یادگیرمبنا برخلاف کارایی قابل‌قبول درعمل زمان‌گیر [11] بوده و نسبت به روش فیلتر ازنظر محاسباتی پیچیده‌تر و پرهزینه‌تر است. ازاین‌رو روش‌های فیلتر در مقایسه با روش‌های یادگیرمبنا سریع‌تر هستند [11]. در روش یادگیرحاصل نیز ارتباط بین ویژگی‌ها نادیده گرفته می‌شود که باعث مشکلاتی در انتخاب نهایی ویژگی‌ها می‌شود. روش‌های یادگیرمبنا و یادگیرحاصل نیاز به تعامل طبقه‌بندی[2] مکرر دارند که می‌تواند زمان اجرا را افزایش دهد [12] ولی فیلترها در هنگام ساخت مجموعه ویژگی، هیچ تعامل طبقه‌بندی کننده‌ای نیاز ندارند [12]. با توجه به این دلایل، روش‌های مبتنی بر فیلتر نسبت به یادگیرمبنا و یادگیرحاصل کارآمدتر هستند و برای دسته‌بندی، بیشتر استفاده می شوند. ما نیز در این پژوهش از روش‌های مبتنی بر فیلتر استفاده کردیم. روش‌های فیلتر به دو دسته روش‌های محلی[3] و سراسری[4] تقسیم شده‌اند.

با توجه به نتایج پژوهش‌های اخیر که به برخی از آن‌ها اشاره شد، انتخاب ویژگی در دسته‌بندی متن هنوز هم یک زمینه پژوهشی مهم و در حال پیشرفت است. به‌طوری‌که در مطالعه [12] با استفاده از روش‌های فیلتر، ترکیبی از یک روش سراسری و یک روش محلی یک‌طرفه به نام IGFSS[5] موردبررسی قرار گرفت. هدف این مطالعه ایجاد مجموعه‌ای از ویژگی‌ها بود که تاحدودی به‌طور یکسان تمام رده‌ها را نشان دهد. برای این امر، نسبت شانس [12]، [16]، [25] به‌عنوان یک روش انتخاب یک‌طرفه برای استخراج ویژگی‌های منفی استفاده شده بود. نتایج تجربی نشان داد که این روش، عملکرد دسته‌بندی بهتری نسبت به عملکرد فردی روش‌های سراسری دارد؛ اما در برخورد با مجموعه‌داده‌های نامتعادل دارای تعداد زیادی رده، مشکل دارد. در مطالعه [16] برای حل این مشکل یک متغیر جدید انتخاب ویژگی سراسری متغیر (VGFSS[6]) برای دسته‌بندی خودکار اسناد متنی [16] معرفی شد. این روش، ترکیبی از یک روش سراسری و نسبت شانس بود؛ اما نکته قابل‌ذکر این است که دو مطالعه اخیر تنها برخی از روش‌های فیلتر را ترکیب کرده‌اند. شباهت مطالعه ما با پژوهش‌های مطرح‌شده بالا، در استفاده از روش‌های مبتنی بر فیلتر است. تفاوت‌های این مطالعه در اندازه زیاد داده‌ها، جنس داده (داده متنی فارسی تلگرام) و به‌کارگیری نوع و تعداد روش‌های انتخاب ویژگی مبتنی بر فیلتر است.

مطالعات در مورد دسته‌بندی و انتخاب ویژگی در متن فارسی به اندازه کافی انجام نشده است. در این مطالعه به انتخاب ویژگی‌های مناسب جهت دسته‌بندی متن فارسی و شناسایی درخواست پرداخته شده است. مجموعه‌داده مورداستفاده، پیام‌های متنی فارسی تلگرام استخراج‌شده از سامانه ایده‌کاو است. انتخاب ویژگی‌های مناسب جهت افزایش دقت مدل، برای شناسایی درخواست‌های موجود در پیام‌های فارسی تلگرام از مهم‌ترین بخش‌های این پژوهش است. برای محاسبه دقت از ماشین بردار پشتیبان و چند دسته‌بند دیگر نیز استفاده شد. همچنین جهت ارزیابی از معیارهای میانگین میکرو[7] و ماکرو[8] بهره گرفته شد. ورودی‌های مدل شامل تعدادی از زیرمجموعه ویژگی‌های بهینه است. همچنین با هدف افزایش دقت مدل، روش‌های انتخاب ویژگی برای تولید ویژگی‌های مناسب برای هریک از مدل‌ها ارائه شده است؛ سپس از میان همهٔ ویژگی‌های موردبررسی، ویژگی‌های مناسب برای هریک از مدل‌های کاربردی انتخاب ویژگی برگزیده شده‌اند. جهت بیان واضح‌تر، نوآوری‌های اصلی این مطالعه در ادامه بیان می‌شود:

- به‌کارگیری پرکاربردترین روش‌های انتخاب ویژگی مبتنی برفیلتر سراسری و محلی جهت یافتن بهینه‌ترین زیرمجموعه ویژگی در متن فارسی؛
- استفاده از روش‌های ترکیبی برای ایجاد ویژگی‌های مناسب جهت مدل‌های پیش‌بینی دقت در دسته‌بندی متن و کاربرد آن‌ها در شناسایی درخواست‌ها در پیام‌های فارسی تلگرام؛

[1] Embed
[2] Classifier interaction
[3] Local
[4] Global
[5] Local and Global Feature selection
[6] Variable Global Feature Selection Scheme
[7] micro-averaging
[8] macro-averaging

- انتخاب ویژگی‌های مناسب در جهت افزایش دقت و کاهش زمان محاسبات برای هریک از مدل‌های موردبررسی. در این راستا، سعی شده تا علاوه‌بر انتخاب یک الگوریتم کارا، روشی در جهت انتخاب‌های مناسب‌تر ارائه شود؛
- ارزیابی و آزمایش مدل‌های پیشنهادی، برای مجموعه‌ای بزرگ از داده‌های فارسی و تعداد متفاوتی از ویژگی‌ها. مجموعه‌دادگان فارسی تلگرامی مورداستفاده توسط خودمان ساخته شده است.

در ادامه این مطالعه به بخش‌های زیر پرداخته می‌شود: مروری بر کارهای پیشین و همچنین پرکاربردترین روش‌های انتخاب ویژگی مبتنی بر فیلتر در بخش ۲ به‌طور خلاصه توضیح داده شده‌اند. در بخش ۳ روش پیشنهادی توضیح داده شده و در بخش ۴ دادگان مورداستفاده، نتایج مربوط به آزمایش‌ها، دقت‌های به‌دست‌آمده با الگوریتم یادگیری و معیارهای ارزیابی مختلف بر روی داده‌های انتخابی بیان شده است. در آخر به بیان پیشنهادها و کارهای آینده پرداخته‌ایم.

۲- مروری بر کارهای پیشین

در این بخش ابتدا به مرور مطالعاتی که در زمینه شناسایی درخواست و شناسایی سؤال در شبکه‌های اجتماعی پرداخته‌اند می‌پردازیم؛ سپس مطالبی را در مورد روش‌های انتخاب ویژگی در متن بیان می‌کنیم.

۲-۱- مطالعات مربوط به تشخیص درخواست یا سؤال در شبکه‌های اجتماعی

برای شناسایی نوع گفتار در شبکه‌های اجتماعی، یک جمله به انواع مختلفی می‌تواند تقسیم شود. روش‌های مختلفی برای دسته‌بندی جملات وجود دارد. به‌عنوان‌مثال در مطالعه [26] سه نوع جمله فرض شده است: گزاره‌ها[1]، سؤالات[2] و درخواست‌ها[3]. براساس مطالعه *کوین*[4] [26]، در گفتگو، نوع متن گفتاری می‌تواند نوع پاسخ موردنیاز را نشان دهد. سامانهٔ پیشنهادی آن‌ها قادر است، متن گفتاری را با دقت ۸۲ درصد در مجموعه‌داده‌های نیمه‌خودکار شناسایی و دسته‌بندی کند. وظیفه آن‌ها دسته‌بندی جملات با توجه به عملکرد آن‌ها بود. آن‌ها یک روش یادگیری ماشین با استفاده از SVM با ویژگی‌های مختلف NLP ارائه دادند که نتایج خوبی را در یک مجموعه داده متنوع داشت.

مطالعات مختلفی در زمینه شناسایی سؤال در توییتر، رایانامه و صفحات برخط انجام شده است. نمونه‌هایی از آن‌ها در مطالعه [4]، [27-29] آمده است. شناسایی درخواست برای رایانامه نیز در مطالعه [30] بیان شده است. *کوهن*[5] و همکاران [30] در مورد شناسایی درخواست‌ها در جلسات و یا جملاتی که اطلاعات را فراهم می‌کنند، بحث کردند. آن‌ها رده‌های مختلفی مانند درخواست و پیشنهادها ارائه دادند و برای شناسایی این رده‌ها، یک ماشین بردار پشتیبان را با ویژگی‌هایی مانند انگرم[6] و بخشی از عبارات گفتاری پیاده‌سازی کردند. آزمایش‌های آن‌ها نشان داد که بسیاری از دسته‌های پیغام‌ها را می‌توان با دقت بالا و فراخوان متوسط، با استفاده از روش‌های یادگیری دسته‌بندی متن موجود شناسایی کرد.

مقالات مشابهی برای شناسایی درخواست در پیام‌های فارسی تلگرام نیافتیم؛ بااین‌حال در ادامه چند مطالعه مانند بحث شناسایی سؤال را که با شناسایی درخواست تشابه دارند و برای داده‌های غیرفارسی انجام شدند، به‌عنوان مرور کارهای پیشین مطرح می‌کنیم. در شبکه‌های اجتماعی علاوه‌بر درخواست‌ها یا سؤالات درخواستی، سؤالات زیادی پرسیده می‌شود که معانی مختلفی دارند. سؤالات بدیعی[7] نمونه‌ای از این نوع هستند. سؤالات بدیعی، سؤالاتی هستند که به جهت دریافت مستقیم پاسخ پرسیده نمی‌شود و اغلب برای تأکید بر روی نکته‌ای یا درگیر ساختن ذهن مخاطب به‌کار می‌رود.

الگوریتم‌های یادگیری ماشینی در کارهای پیشین برای آموزش دسته‌بندی و شناسایی سؤالات، بیش‌تر برای داده‌های رسانه‌های اجتماعی به کار گرفته شده‌اند؛ مانند مطالعه [29] که از توییتر به‌عنوان مجموعه‌داده برای شناسایی سؤالات بدیعی استفاده شد. همچنین *رانگانات*[8] و همکاران در مطالعه [4] نیز تنها بر شناسایی سؤالات بدیعی و غیر بدیعی از توییت‌ها تمرکز دارند؛ بنابراین، تنها این رده‌های خاص متعلق به اقدامات گفتاری در راه‌حل مورداستفاده قرار می‌گیرند. در این مطالعه، مجموعه‌داده‌ها از سؤالات جمع‌آوری‌شده از رسانه اجتماعی توییتر تشکیل شده است. آن‌ها چارچوبی را برای شناسایی سؤالات بدیعی با الگوبرداری از برخی انگیزه‌های کاربران برای

[1] statements
[2] questions
[3] requests
[4] Quinn

[5] Cohen
[6] ngram
[7] rhetorical
[8] Ranganath

ارسال آن‌ها در رسانه‌های اجتماعی را مدل‌سازی و ایجاد کردند.

همچنین در مطالعه [27] نیز از داده‌های توییتر به‌عنوان داده‌های یک شبکه اجتماعی استفاده شد. نویسندگان این مطالعه شناسایی خودکار سؤالات برای بیان نیاز به اطلاعات را موردبررسی قرار دادند. آن‌ها از خصوصیت‌های زبان‌شناسی، واژگانی و بخشی از ویژگی‌های گفتاری استفاده کردند تا اطلاعات موردنظر را در سؤالات شناسایی کنند. در این مطالعه تجربی، توییت‌های انگلیسی یک‌ساعته را نمونه‌برداری کردند و نتایج تجربی را برای شناسایی سؤال در توییتر گزارش دادند. برای تشخیص توییت، هر دو رویکرد مبتنی بر قانون و رویکرد مبتنی بر یادگیری مورد استفاده قرار گرفت. نتایج آزمایش آن‌ها نشان داد که رویکرد مبتنی بر قانون به‌خوبی عمل کرد ولی رویکرد مبتنی بر یادگیری عملکرد قابل‌توجهی نداشت.

در برخی مطالعات دیگر از داده‌های وب استفاده شد. به‌عنوان‌مثال *اوجوکو*[1] و همکاران در مطالعه [28] مدلی جهت شناسایی و دسته‌بندی سؤالات را در CQA[2] ارائه دادند. در این مطالعه صفحات تارنمای ریسرچ گیت[3] به‌عنوان داده ورودی برای رویکرد نظام‌مند آن‌ها استفاده شد. در این روش ابتدا سؤالات با استفاده از برچسب POS[4] به زبان انگلیسی مشخص شد و پس‌ازآن بر اساس بیشینه مقدار احتمال دسته‌بندی بیز ساده دسته‌بندی شدند.

در زمینه بازاریابی نیز مانند پژوهش ما می‌توان از بحث‌های شناسایی درخواست یا شناسایی سؤال استفاده کرد. با رشد محتوا در شبکه‌های رسانه‌های اجتماعی، شرکت‌ها و ارائه‌دهندگان خدمات به شناسایی سؤالات مشتریان علاقه‌مند شده‌اند. با توجه به رشد متن و افزایش کاربران، شناسایی دستی این سؤالات کار دشواری است؛ لذا با شناسایی خودکار سؤالات و ارجاع دادن آن‌ها به افراد پاسخگو به این سؤالات در بخش خدمات مشتری می‌توان باعث صرفه‌جویی در زمان، افزایش بازخورد مشتری و بهبود مشاغل شد. در مطالعه [31] یک دسته‌بندی دودویی برای دسته‌بندی متن به دو دسته از سؤالاتی که به‌دنبال پاسخ هستند یا به‌دنبال پاسخ نیستند، انجام شده است. در مطالعه آن‌ها از داده‌های توییتر استفاده شد.

در مطالعاتی مانند [31] تنها از روش‌های یادگیری ماشین در شناسایی سؤال استفاده کردند؛ درحالی‌که ما در پژوهش خود جهت شناسایی درخواست، قبل از استفاده از الگوریتم‌های یادگیری از روش‌های فردی و ترکیبی مبتنی بر فیلتر برای انتخاب ویژگی استفاده کردیم. در پژوهش ما از روش‌های انتخاب ویژگی و یادگیری ماشین به‌صورت ترکیبی برای شناسایی درخواست در پیام‌های فارسی تلگرام استفاده شد؛ همچنین علاوه‌بر فارسی‌بودن حجم داده مورداستفاده، از داده‌های مورداستفاده در پژوهش‌های یادشده بیشتر بود. در بخش زیر به شرح روش‌های انتخاب ویژگی می‌پردازیم.

۲-۲- روش‌های پرکاربرد انتخاب‌ویژگی مبتنی برفیلتر

اغلب روش‌های انتخاب ویژگی شناخته‌شده برای دسته‌بندی متن، روش‌های مبتنی بر فیلتر هستند [12]. روش‌های فیلتر، مستقل از الگوریتم یادگیری، ویژگی‌ها را بر اساس یک مرحله پیش‌پردازش انتخاب می‌کنند [11]، [12]، [18]؛ سپس به تک‌تک ویژگی‌ها امتیازی داده می‌شود. ویژگی‌ها بر اساس امتیازات منحصربه‌فرد خود به‌ترتیب نزولی مرتب می‌شوند [7] ,[11] ,[12] ,[18]. در آخرین مرحله تعداد تعیین‌شده‌ای از ویژگی‌هایی که بالاترین امتیاز را دارند، انتخاب می‌شوند [12]، [22]. این تعداد تعیین‌شده یا آستانه ثابت، به‌صورت تجربی به‌دست آمده و برای هر روش می‌تواند متفاوت باشد [12]، [22]. به‌عبارتی این بهترین حد آستانه برای ویژگی‌های انتخاب‌شده، بستگی به مجموعه‌داده مورداستفاده دارد و برای هر مجموعه‌ای متفاوت است [22]؛ درنهایت این مجموعه به‌دست‌آمده به‌عنوان ورودی به الگوریتم دسته‌بندی، داده می‌شود.

روش‌های مبتنی بر فیلتر را می‌توان به دودسته سراسری و محلی تقسیم کرد [12]، [11]. روش‌های انتخاب ویژگی سراسری شناخته‌شده عبارتند از: فراوانی سند[5]، بهره اطلاعات[6]، شاخص‌های جینی[7] و انتخاب‌گر ویژگی متمایز[8]. روش‌های انتخاب ویژگی محلی شامل نسبت شانس[9] و ضریب همبستگی[10] است. در مطالعه *یوسال*[11] [12] به نقل از *اگورا*[12] [19] گفته شده است که روش‌های انتخاب ویژگی مبتنی بر فیلتر بر اساس خصوصیات به دودسته یک‌طرفه[13] و دوطرفه[1] تقسیم

[1] Ojokoh
[2] Community-based Question Answering (CQA)
[3] Research Gate
[4] Part of Speech (POS)

[5] Document frequency (DF)
[6] Information gain (IG)
[7] Gini index (GI)
[8] Distinguishing feature selector (DFS)
[9] Odds ratio (OR)
[10] Correlation coefficient (CC)
[11] Uysal
[12] Ogura
[13] one-sided

می‌شوند. در مطالعه حاضر، جهت یافتن ویژگی‌های بهینه، از پرکاربردترین روش‌های انتخاب ویژگی مبتنی بر فیلتر استفاده شده است؛ لذا در بخش‌های زیر به شرح زمینه‌های ریاضی این روش‌ها پرداخته می‌شود.

۱-۲-۲- نسبت شانس

معیار نسبت شانس، عضویت در کلاس خاصی را با نامزد[2] و عدم عضویت را با مخرج[3] اندازه‌گیری می‌کند. امتیازات عضویت و عدم عضویت، با تقسیم آن‌ها به یکدیگر نرمال شده‌اند؛ بنابراین برای به‌دست‌آوردن بالاترین امتیاز از فرمول، مقدار نامزد باید بیشینه شود و مقدار مخرج باید به حداقل برسد. فرمول نسبت شانس از رابطه (۱) محاسبه می‌شود. نسبت شانس یک روش یک‌طرفه است که لگاریتم عدد محاسبه‌شده را برای رسیدن به مقدار منفی می‌گیرد [32]. تابع لگاریتم نمرات منفی تولید می‌کند درحالی‌که مقدار کسر در این فرمول بین صفر و یک است؛ بنابراین نتیجه می‌گیریم که این روش یک معیار یک‌طرفه است. ویژگی‌هایی که دارای مقادیر منفی هستند به ویژگی‌های منفی اشاره می‌کنند [12]، [32]. به‌طور جداگانه امتیازات نسبت شانس هر ویژگی برای تمام رده‌ها را محاسبه می‌کنیم و بالاترین امتیاز را برای اختصاص یک برچسب رده به ویژگی انتخاب می‌کنیم. برای تعیین برچسب، مقدار امتیاز مطلق را در نظر می‌گیریم تا ویژگی‌های غیر عضو نیز در نظر گرفته شود [32].

$$OR(t|C_i) = log\frac{P(t|C_i)[1-P(t|\overline{C_i})]}{[1-P(t|C_i)]P(t|\overline{C_i})} \qquad (۱)$$

۲-۲-۲- ضریب همبستگی

درواقع ضریب همبستگی (CC) نوعی از مجذور خی است، که در آن $cc^2 = \chi^2$ است. CC را می‌توان به‌عنوان مربع خی "یک‌طرفه" مشاهده کرد. این روش، شایستگی یک زیرمجموعه از ویژگی‌ها را با درنظرگرفتن توانایی پیش‌بینی فردی هر ترم به همراه درجه افزونگی بین آن‌ها، ارزیابی می‌کند. زیرمجموعه ارجح از ویژگی‌ها، همبستگی بالایی در رده و همبستگی پایین بین رده‌های مختلف دارد. در رابطه (۲) ضریب همبستگی CC برای متغیر v یا یک واژهٔ t در یک ردهٔ c_i برای یک مجموعه دارای نمونه‌های N تعریف شده است. در جدول (۱) وابستگی‌ها نشان داده شده است.

$$cc(t,c_i) = \frac{\sqrt{N}[P(t,c_i)P(\bar{t},\overline{c_i}) - P(t,\overline{c_i})P(\bar{t},c_i)]}{\sqrt{P(t)P(\bar{t})P(c_i)P(\overline{c_i})}} \approx \frac{\sqrt{N}(AD-CB)}{\sqrt{(A+C)\times(B+D)\times(A+B)\times(C+D)}} \qquad (۲)$$

مقدارهای مثبت بیان‌گر ویژگی‌هایی هستند که عضویت دارند و مقادیر منفی بیان‌گر عدم عضویت هستند. هرچه مقادیر مثبت بیشتر باشند، ترم‌ها برای نشان‌دادن عضویت قوی‌تر هستند. هرچه مقادیر منفی کوچک‌تر باشند، ترم‌ها برای نشان‌دادن عدم عضویت قوی‌تر هستند. روش انتخاب ویژگی محلی استاندارد CC، ترم‌های با بیشترین مقدار CC را به‌عنوان ویژگی انتخاب می‌کنند. دلیل اصلی این است که ترم‌هایی که از متن‌های نامربوط از یک دسته هستند، بی‌فایده در نظر گرفته می‌شوند. از سویی دیگر، وقتی‌که مقادیر نشان‌دهنده عضویت یا عدم عضویت یک ترم در یک دسته باشند، خی غیر منفی است. بر این اساس ویژگی‌های مبهم در رده پایین‌تر قرار خواهند گرفت. در مقایسه با CC، خی ترم‌های به‌دست‌آمده از هر دو متن‌های مربوط و نامربوط را در نظر می‌گیرد [17]، [25]. دلیل استفاده از روش انتخاب ویژگی ضریب همبستگی این است که زیرمجموعه‌های بهتری نسبت به فرکانس سند تولید می‌کند [17].

(جدول-۱): تاپل‌های وابستگی برای روش‌های انتخاب ویژگی موردبحث

(Table-1): Dependency tuples for the feature selection methods discussed

	Membership in c_i	Nonmembership in c_i
Presence of *t*	(t, c_i)	$(t, \overline{c_i})$
Absence of *t*	(t, c_i)	$(t, \overline{c_i})$

۳-۲-۲- بهره اطلاعات

این روش به‌طور گسترده در آمار و یادگیری ماشین استفاده می‌شود [11]. جهت بیان مفهوم بهره اطلاعات، نسبت سهم وجود یا عدم وجود ترم برای دسته‌بندی صحیح اسناد متنی به‌وسیله نمرات IG نشان داده می‌شود و در دسته‌بندی‌های مبتنی بر درخت تصمیم استفاده می‌شود [11]، [12]، [33]. درصورتی‌که شاخص خوبی برای نسبت دادن سند به هر رده باشد آن‌گاه بیشینه مقدار یک ترم را مشخص می‌کند. بهره اطلاعات، معیار انتخاب ویژگی سراسری است و در رابطه (۳) نشان داده‌شده است. این رابطه برای هر ترم فقط یک نمره تولید می‌کند. M تعداد رده‌ها است. احتمال ردهٔ c_i با $P(c_i)$ نشان داده می‌شود. $P(t)$ احتمال وجود ترم t و $P(\bar{t})$ احتمال عدم وجود ترم t است. $P(c_i|t)$ احتمال شرطی

[1] two-sided
[2] nominator
[3] denominator

ردهٔ c_i برای وجود ترم t است. $P(c_i|\bar{t})$ احتمال شرطی ردهٔ c_i برای عدم وجود ترم t است [12]. محاسبه IG برای هر ترم نشان‌دهنده این هست که آیا این ترم یک نشان‌گر قوی است که در هر ردهٔ خاص گنجانده شود یا خیر [32]. از تعریف آن مشخص است که استراتژی IG یک استراتژی نظارت‌شده است. یعنی، از برچسب‌های اسناد در انتخاب ویژگی‌ها برای حفظ آن ویژگی‌ها استفاده می‌کند [21]، [33].

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

۲-۲-۴- شاخص جینی

یک روش انتخاب ویژگی سراسری برای دسته‌بندی متن است و همین روش به شکل بهبودیافته در الگوریتم درخت تصمیم به‌کار برده می‌شود. رابطه (۴) مربوط به این روش است. در این فرمول $P(t|C_i)$ احتمال وجود ترم t به‌شرط C_i است. همچنین $P(C_i|t)$ احتمال ردهٔ C_i به‌شرط وجود ترم t است [12]، [32].

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 P(C_i|t)^2 \quad (۴)$$

۲-۲-۵- انتخاب‌کننده ویژگی‌های متمایز

روش انتخاب ویژگی مبتنی بر فیلتر، زمانی ایده‌آل است که امتیازهای بالایی را به ویژگی‌های متمایز اختصاص دهد درحالی‌که امتیازات پایین‌تر را به موارد نامربوط اختصاص می‌دهد. در دسته‌بندی متن، هر ترم مشخص با یک ویژگی مطابقت دارد؛ سپس رتبه‌بندی ترم‌ها با توجه به شرایطی انجام می‌شود [34]. یکی از روش‌های انتخاب ویژگی اخیر و مناسب برای دسته‌بندی متن DFS است [34] که یک معیار انتخاب ویژگی سراسری است. این روش ویژگی‌های متمایز را انتخاب می‌کند. انتخاب‌کننده ویژگی‌های متمایز، بهبودی از اطلاعات متقابل با استفاده از کاهش تأثیر احتمالات حاشیه‌ای از ترم‌ها به‌وسیله نرمال‌کردن وزن ترم‌ها است. وزن ترم‌ها را در محدوده [۰،۱] تعریف می‌کند. طریقه محاسبه این روش در رابطه (۵) نشان داده شده است. M تعداد رده‌ها است. $P(c_i|t)$ احتمال شرطی ردهٔ c_i برای وجود ترم t است $P(\bar{t}|c_i)$ احتمال شرطی عدم وجود ترم t برای ردهٔ c_i است $P(t|\overline{c_i})$ احتمال شرطی ترم t برای تمام رده‌ها به‌جز ردهٔ c_i است [12]، [16]، [34].

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

۲-۲-۶- فراوانی سند

فراوانی سند مربوط به یک ترم، تعداد اسناد آموزشی است که ترم موردنظر در آن اسناد وجود دارد. از این روش بیشتر در دسته‌بندی متن استفاده می‌شود. به‌این‌علت که هزینه محاسبه الگوریتم کم است [11]، [33]. در بعضی از روش‌های مبتنی بر فیلتر، خوب بودن یک ترم را بر اساس میزان دفعات تکرار در متن مانند فرکانس اسناد ارزیابی می‌کنند [14]. فرکانس سند یک ترم به‌عنوان تعداد کل اسناد موجود در مجموعه اسناد که حاوی این ترم است تعریف می‌شود. اساس DF این است که در دسته‌بندی متن، ترم‌های نادر غیر آموزنده هستند و در هنگام انتخاب ویژگی باید حذف شوند. روش رتبه‌بندی، جهت ارزیابی خوبی یا اهمیت هر ترم در مجموعه‌ای از اسناد استفاده می‌شود. بنابراین DF معیار خوبی برای سنجش اهمیت ترم‌ها است. در این روش N تا از مهم‌ترین ترم‌ها در دسته‌بندی به‌عنوان ویژگی انتخاب می‌شوند و بقیه فیلتر می‌شود؛ با انجام این کار به‌سادگی می‌توان قبل از مرحله انتخاب ویژگی، تعداد زیادی از ویژگی‌های بی‌اهمیت را حذف کرد [35]. به دلیل خطی‌بودن پیچیدگی زمانی DF در اسناد آموزشی، روشی ساده و کارآمد به‌حساب می‌آید. مشکل این روش این است که گاهی ممکن است ترم‌های با فرکانس کم که حذف‌شده‌اند مفید و آموزنده باشند و برخی ترم‌های با فرکانس بالا که به‌عنوان ویژگی انتخاب شده‌اند مفید نباشند [14]، [35]، [36]. به عبارتی DF، ساده‌ترین روش انتخاب ویژگی سراسری است که برچسب سند را بر مبنای بیشترین فرکانس‌ترم مشخص می‌کند. فرمول ریاضی این روش در رابطه (۶) نشان داده شده است. خروجی محصول نقطه‌دارکردن از رده‌ها و احتمال هر ویژگی خاص در آن رده است [21]، [32].

Dfreq = number of true positives + number of false positives

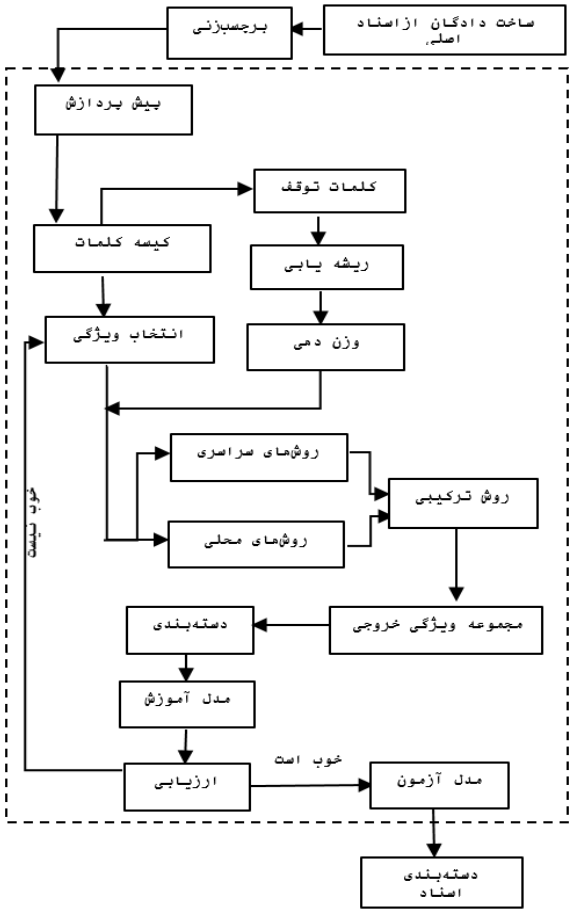
$$DF(a_j) = N.p(a_j) \quad (۶)$$

۳- روش پژوهش

یکی از مشکلات اصلی در دسته‌بندی متن، ابعاد بالای داده‌ها است. ابعاد بالای داده‌ها باعث افزایش زمان محاسبات، هزینه‌ها و کاهش دقت در الگوریتم‌های یادگیری می‌شود [13]، [14]. کاهش ابعاد سبب افزایش

دقت دسته‌بندی متن می‌شود؛ بنابراین باید با روش‌های انتخاب ویژگی این ابعاد را کاهش داد. جهت انتخاب ویژگی روش‌های زیادی وجود دارد. برخی از پژوهش‌ها جهت کاهش بعد، از روش‌های انتخاب ویژگی و ساخت زیرمجموعه بهینه [19]، [22]، [37] استفاده می‌کنند. آن‌ها یک الگوریتم انتخاب ویژگی عملکردی را برای انتخاب یک زیرمجموعه ویژگی کوچک از تعداد زیادی از ویژگی‌ها ارائه داده‌اند. عملکرد دسته‌بندی با این زیرمجموعه، مشابه یا حتی بهتر از استفاده از تمام ویژگی‌ها است [11]، [37-40]. در این پژوهش به‌منظور رفع مشکل ابعاد بالای فضای ویژگی، از روشی ترکیبی جهت ساخت زیرمجموعه بهینه استفاده کردیم. در این راستا ابتدا به بررسی پرکاربردترین روش‌های انتخاب ویژگی مبتنی بر فیلترسراسری و محلی پرداختیم که هرکدام به‌تنهایی نتایج خوبی برای انتخاب ویژگی داشته‌اند. در ادامه نشان دادیم که ترکیب این روش‌ها نتیجه بهتری دارد و سبب ساخت یک زیرمجموعه بهینه‌تر می‌شود. در این بخش مراحل انجام پژوهش به‌صورت گام‌به‌گام تشریح می‌شود. در شکل (۳) مراحل کلی پژوهش نشان داده شده است. عملکرد دسته‌بندی‌ها با استفاده از روش‌های یادگیری می‌توانند سنجیده شوند. با توجه به شکل (۳) درصورتی‌که نتایج ارزیابی به‌اندازه کافی خوب نبود، می‌توانیم دوباره به مرحله انتخاب ویژگی بازگردیم و معیار انتخاب را تغییر دهیم تا به نتیجه موردنظر برسیم [22].

ما برای آزمودن و ارزیابی، از مجموعه‌داده‌های حقیقی متنی استخراج‌شده از سامانه ایده‌کاو استفاده کردیم. از خصوصیات مهم این مجموعه‌داده، وجود جملات فارسی به شکل محاوره‌ای و غیر محاوره‌ای است. هر سند متنی یک بردار است و هر بردار از مجموعه‌ای از واژگان تشکیل شده است [22]. درواقع متن ما دارای ۸۵۷۴۱ پیام فارسی تلگرامی است، که هر پیام به‌عنوان یک جمله در نظر گرفته شد؛ سپس جملات به واژگان تبدیل می‌شود و هر واژه به‌عنوان یک ویژگی در نظر گرفته می‌شود [15]، [22]. واژه یا ترم کوچک‌ترین اجزای تشکیل‌دهنده متن است و نقش مهمی در فرایند دسته‌بندی سندهای متنی دارد [11]، [16].



(شکل-۳): نمودار کلی روش پژوهش
(Figure-3): Outline of Research Method

اندازهٔ دیکشنری ساخته‌شده از مجموعه‌داده ۹۰۴۲۶ است. برای کاهش اندازهٔ دیکشنری می‌توان از روش‌های فیلترینگ و انتخاب‌ترم استفاده کرد. ازجمله این روش‌های فیلترینگ می‌توان به فیلترینگ واژگان توقف اشاره کرد که جزئی از پیش‌پردازش است؛ لذا، ابتدا پیش‌پردازش بر اسناد متنی انجام شد. برای استفاده از این حجم بالای داده، انجام پیش‌پردازش ضروری است تا اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی قرار گیرند. پیش‌پردازش نقش مهمی در کارایی مدل ایجادشده دارد. درواقع داده‌ها را باید برای ورود به الگوریتم‌های یادگیری ماشین و قابل‌تشخیص برای ماشین تبدیل و آماده‌سازی کنیم. در این پژوهش نیاز داریم که متن خود را به اعداد تبدیل کنیم. به فرایند تبدیل متن به اعداد، بردارسازی[1] یا مدل بردار واژگان[2] می‌گویند. ما از یکی از رویکردهای کیسه واژگان[3] که بردارسازی شمارشی[4] بود، جهت بردارسازی متن خود استفاده کردیم. از ابزارها و کتابخانه‌های محیط پایتون برای مراحل پیش‌پردازش و

[1] vectorization
[2] Vector Space Model (VSM)
[3] Bag of Word (BOW)
[4] Frequency/Count Vectorizer

ساخت بردارهای ویژگی استفاده کردیم. پیش‌پردازش و انتخاب ویژگی به‌عنوان مراحل بسیار مهم در کنار استخراج ویژگی و دسته‌بندی شناخته می‌شوند [12]، [16]. پیش‌پردازش‌ها برای تبدیل محتویات متنی به شکل عددی مورد استفاده قرار می‌گیرند و به موارد زیر تقسیم می‌شوند: توکنایز[1]، حذف واژگان توقف[2]، ریشه‌یابی[3] و وزن‌دهی[4] [15]، [16]، [41]. تقسیم‌های پیش‌پردازش، بر کاهش ویژگی‌های زائد با استفاده از کمترین و بیشترین افزونگی تمرکز دارند [24]، [42]، [43]. کتابخانه‌های زیادی مانند Spacy و Nltk جهت پیش‌پردازش متن ارائه شده است. اما تعداد کمی از متن فارسی پشتیبانی کرده و دقت بالایی ندارند. کتابخانه‌های مخصوص زبان فارسی زیاد نیستند و معروف‌ترین آن‌ها می‌توان hazm و Parsivar را نام برد. ما در این پژوهش برای توکنایز و ریشه‌یابی از Parsivar استفاده کردیم. ما از فهرست واژگان توقف خود که برای زبان فارسی (در تلگرام از واژگان عامیانه زیادی استفاده می‌شد) ساخته بودیم استفاده کردیم و آن‌ها را از نمونه‌ها حذف کردیم. در این پژوهش از وزن‌دهی دودویی برای ویژگی‌ها استفاده شد. به این صورت که با استفاده از مجموعه ویژگی‌های انتخابی، برای هر جمله یا نمونه یک بردار ویژگی ساختیم. در هر بردار اگر این ویژگی یا واژه در جمله حضور داشته باشد درایه متناظر با آن ویژگی را یک و در غیر این صورت برابر با صفر قرار دادیم. درنهایت بردار ویژگی برای همه نمونه‌ها به‌صورت ماتریسی از صفر و یک به دست آمد. بعد از مراحل پیش‌پردازش اندازهٔ ویژگی‌ها به ۶۷۸۵ رسید.

ابعاد بالای ویژگی موجب زمان‌برشدن اجرای روش‌های یادگیری می‌شود. به همین دلیل باید با استفاده از روش‌هایی ابعاد ویژگی‌ها را کاهش داد. همان‌طور که در بخش مقدمه بیان‌شده است، یکی از راه‌حل‌های کاهش ابعاد ویژگی استفاده از روش‌های انتخاب ویژگی موجود است. این روش‌های انتخاب ویژگی با استفاده از معیارهای تعریف‌شده خود یک زیرمجموعه مناسب از ویژگی‌ها را پیدا می‌کنند. الگوریتم‌های مبتنی بر فیلتر که به دو دسته محلی و سراسری تبدیل شده‌اند و در بخش ۲ بیان‌شده، نمونه‌هایی از این الگوریتم‌ها برای کاهش ابعاد هستند. در ادامه با استفاده از روش‌های انتخاب ویژگی پیشنهادی، مجموعه ویژگی بهینه که دارای ویژگی‌های بااهمیت بود را یافتیم.

در روش‌های انتخاب ویژگی برای دسته‌بندی متن، امتیازات انتخاب ویژگی که نشان‌دهنده تفاوت میان ترم‌ها در یک مجموعه‌داده است باید محاسبه شود. در مرحله بعد همهٔ ترم‌ها با توجه به امتیازاتی که در مرحله انتخاب ویژگی کسب کرده‌اند، به‌ترتیب نزولی مرتب می‌شوند. بعد از مرتب‌سازی، N ویژگی با بالاترین امتیاز در مجموعه‌ویژگی به‌عنوان ویژگی‌های نهایی انتخاب می‌شوند. مقدار N، یک تعداد تعیین‌شده است که به‌طورمعمول به‌صورت تجربی به‌دست می‌آید و در بعضی موارد به مجموعه‌داده بستگی دارد. هدف مطالعه ما این است که با اصلاح فرایند انتخاب ویژگی که در بالا اشاره شد، عملکرد دسته‌بندی را بهبود بخشد. جهت رسیدن به این عملکرد مطلوب، پرکاربردترین روش‌های انتخاب ویژگی مبتنی برفیلتر که شامل روش‌های محلی و سراسری است باهم ترکیب شده‌اند. مراحل روش انتخاب ویژگی ما به‌صورت زیر است:

۱- ساخت دادگان: داده‌های مورداستفاده را از پیام‌های فارسی تلگرام انتخاب می‌کنیم و برای هر پیام یک برچسب در نظر می‌گیریم.

۲- جداسازی داده‌ها: مجموعه‌دادهٔ به‌دست‌آمده را به دو قسمت تقسیم کرده. هشتاد درصد اولیه را به‌عنوان آموزش و بیست درصد بعدی را به‌عنوان آزمون در نظر می‌گیریم.

۳- پیش‌پردازش: در این مرحله ساخت کیسه واژگان و تشکیل بردار ویژگی انجام می‌شود. واژگان موجود در کیسه واژگان را می‌توانیم به‌عنوان ویژگی در نظر بگیریم. در این مرحله تمامی واژگان را در هر پیام جدا کرده و به هر واژه یک ویژگی می‌گوییم. برای حذف ویژگی‌های نامربوط از روش‌هایی مثل حذف واژگان توقف استفاده می‌کنیم. با این ویژگی‌های نهایی به‌دست‌آمده، بردارهای ویژگی را می‌سازیم. در این مرحله یک هماهنگ‌سازی انجام‌شده و کل متن تبدیل به یک سری واژگان جدا از هم شد؛ سپس واژگان زائد یا ویژگی‌های کم‌اهمیت را [15] طبق روش‌هایی که توضیح داده شد، حذف کردیم تا واژگان مفید و کاربردی که ویژگی‌های مدنظر ما را داشت حاصل شد.

۴- انتخاب مجموعه‌ویژگی‌ها: F را به‌عنوان مجموعه‌ای از تمام ویژگی‌های یک مجموعه تعریف می‌کنیم، برای این بردارهای ویژگی هر دو روش انتخاب ویژگی محلی و سراسری را اجرا می‌کنیم.

[1] Tokenization
[2] Stop words
[3] Stemming
[4] weighting

۵- انتخاب ویژگی‌ها با روش‌های محلی: در این مرحله با استفاده از روش‌های انتخاب ویژگی محلی، برای ویژگی‌های هر رده امتیازات محاسبه می‌شود. یک مجموعه برچسب به نام L_Label برای ویژگی‌ها شامل $2 * c$ برچسب‌های رده ایجاد کنید. در این مجموعه، c تعداد رده‌ها را نشان می‌دهد. برچسب ردهٔ نخست c عضویت را نشان می‌دهد و برچسب c دوم عدم عضویت در این رده را نشان می‌دهد. در انتها برای هر ویژگی بالاترین امتیاز کسب‌شده با روش انتخاب ویژگی محلی را پیدا می‌کنیم. برچسب ردهٔ مربوطه را از مجموعه برچسب L_Label به ویژگی اختصاص می‌دهیم. تمام ویژگی‌هایی را که با روش‌های انتخاب ویژگی محلی به‌دست‌آمده است، به‌صورت نزولی مرتب می‌کنیم. $F\_local \subseteq F$ را به‌عنوان مجموعه‌ای از ویژگی‌های انتخاب‌شده محلی تعریف می‌کنیم. F_local شامل تعداد l ویژگی است.

۶- انتخاب ویژگی‌ها با روش‌های سراسری: با استفاده از روش‌های انتخاب ویژگی سراسری، امتیازات را برای ویژگی‌ها محاسبه می‌کنیم؛ مانند مرحلهٔ قبل برچسب ویژگی‌ها را مشخص، سپس ویژگی‌ها را بر اساس امتیازات به‌صورت نزولی مرتب و این مجموعه مرتب‌شده را G_Label نام‌گذاری می‌کنیم. $F\_global \subseteq F$ را به‌عنوان مجموعه‌ای از ویژگی‌های انتخاب‌شده سراسری تعریف می‌کنیم. F_global شامل تعداد g ویژگی است.

۷- مجموعه‌ویژگی جدید: مجموعه‌ویژگی جدید را F_final نام‌گذاری، اندازه این مجموعه را خودمان مشخص و جهت این کار از خروجی دو مجموعه مرتب‌شدهٔ مراحل قبل استفاده می‌کنیم. از این دو مجموعه ویژگی‌هایی انتخاب می‌شوند که امتیازات بالاتری داشته باشند.

۸- ترکیب: دو مجموعه از مرحله قبل به‌دست می‌آید که مقدار تعیین‌شده‌ای دارد. وزن‌های ویژگی‌های به‌دست‌آمده از هر روش را باهم جمع کردیم تا وزن کلی برای هر ویژگی به دست آید. در آخر وزن‌های جدید را مرتب کردیم. بهترین ویژگی‌ها با وزن بالاتر را انتخاب می‌کنیم.

۹- بخش تکمیلی: اگر تعداد ویژگی‌ها در مجموعه نهایی کمتر از تعداد تعیین‌شده باشد، از ویژگی‌های نادیده گرفته‌شده با امتیازات بالا در مجموعه مرحله هفت به مجموعه نهایی اضافه می‌کنیم.

درنهایت مجموعه به‌دست‌آمده را به الگوریتم‌های دسته‌بندی می‌دهیم. بعد از گذر از مدل آموزشی اگر ارزیابی قابل‌قبول بود به دسته‌بندی می‌پردازیم. در غیر این صورت به مرحله انتخاب ویژگی بازگشته و مجموعه جدید را پیدا می‌کنیم. به‌این‌ترتیب یک مجموعهٔ جدیدی از ویژگی‌ها را به دست آورده‌ایم که شامل ویژگی‌های با بالاترین امتیاز از ترکیب روش‌های محلی و سراسری است. ویژگی‌های منفی گاهی سودمند هستند؛ لذا در این مجموعه‌ویژگی به‌دلیل استفاده از نسبت شانس ویژگی‌های منفی هم مورد ارزیابی قرارگرفته‌اند. حال این مجموعه جدید را با استفاده از روش‌های یادگیری رایج در زمینهٔ دسته‌بندی متن و انتخاب ویژگی مورد آزمایش قرار داده و نتایج را مقایسه می‌کنیم. جزئیات تجربی این مراحل در بخش ۴ آمده است.

۴- نتایج تجربی

در این بخش، عملکرد روش‌های ترکیبی در برابر عملکرد فردی از روش‌های انتخاب ویژگی محلی و سراسری اندازه‌گیری شده است. روش‌های انتخاب ویژگی محلی در این مطالعه، نسبت شانس و ضریب همبستگی بود. روش‌های انتخاب ویژگی سراسری مورداستفاده شامل فراوانی سند، بهره اطلاعات، شاخص‌های جینی و انتخاب‌گر ویژگی متمایز بود. همچنین مراحل پیش‌پردازش از قبیل توکنایز و حذف واژگان انجام شد. برای تأیید عملکرد این روش‌ها، مجموعه‌داده به‌دست‌آمده در بخش ۴-۱ با ویژگی‌های مختلف و معیارهای ارزیابی معرفی‌شده در بخش ۴-۳ در شرایط مختلف مورداستفاده قرار گرفت. در بخش‌های زیر، مجموعه‌داده مورداستفاده و معیارهای ارزیابی و الگوریتم دسته‌بندی به‌طور خلاصه شرح داده می‌شود؛ سپس نتایج مربوط به روش‌های انتخاب ویژگی محلی و سراسری به‌صورت فردی و ترکیبی با نمودار نشان داده شده است. درنهایت عملکرد ترکیب این روش‌ها باهم مقایسه شد و نمودار نتایج مهم‌تر با کرنل‌های مختلف از روش یادگیری رسم شده است.

۴-۱- مجموعه‌داده

در این مطالعه از پیام‌های گروه‌های تلگرام که توسط خودمان استخراج و برچسب‌گذاری شده، استفاده کردیم.

دادگان شناخته‌شده و آماده‌ای در حوزه شناسایی درخواست پیدا نکردیم؛ لذا روش پیشنهادی ما نیازمند مجموعه‌دادگانی از پیام‌رسان تلگرام به زبان فارسی است. جهت رفع این نیازمندی، در ابتدا سامانه ایده کاو را که منبعی جهت جمع‌آوری پیام‌های فارسی تلگرام است، انتخاب کردیم. این سامانه دارای حجم زیادی از پیام‌های فارسی از گروه‌های تلگرامی بود که به‌روز می‌شدند.

جهت انجام این فرایند هفت نفر از دانشجویان مقطع ارشد و دکترای دانشگاه یزد با دریافت نام کاربری و پسورد، در چند جلسه توضیحی در مورد چگونگی برچسب‌زنی، کار خود را شروع کردند. پس از آموزش و دریافت کد کاربری، هر دانشجو با ورود به سایت ایده‌کاو، در قسمت جستجو با به‌کاربردن واژگان کلیدی شروع به جستجوی پرسش‌های کاربران کردند. با استفاده از قوانینی که در جلسات به دانشجویان شرح داده شده بود به برچسب‌زدن جملات پرداختند. در سایت ایده‌کاو پیام‌های تلگرامی جمع‌آوری‌شده و به‌روز شده‌اند. دانشجویان تنها با پیداکردن واژگان پرسشی کلیدی به جستجوی درخواست‌های موردنظر پرداختند. به این صورت که داده‌ها را طبق قوانینی برچسب‌گذاری کرده و پس‌ازآن ارزیابی می‌کنیم. در اینجا برای برچسب‌گذاری باید تشخیص داد که عبارت موردنظر یک درخواست است یا خیر. اگر یک درخواست بود برچسب مثبت و در غیر این صورت برچسب منفی داده می‌شود. این توضیحات بیان‌گر بحث شناسایی درخواست[1] یا شناسایی سؤال[2] است.

فایل دادگان به‌دست‌آمده شامل ۸۵۷۴۱ رکورد است که هر رکورد نشان‌دهنده پیام کاربر است. این فایل چهارده ستون دارد که مشخصات پیام را نشان می‌دهد. ستون‌ها بیان‌گر مشخصاتی ازجمله متن پیام، طول پیام، برچسب مثبت و منفی و یا خنثی توسط فرد نخست و دوم، شناسه گروه، نام گروه، تعداد اعضای گروه، شناسه کاربری که پیام را فرستاده، نوع پیام و زمان ارسال پیام است. جمع‌آوری این فایل از تاریخ ۱۶ بهمن ۱۳۹۶ تا تاریخ هشتم اسفند ۱۳۹۶ زمان برده است. برای اطمینان از صحت برچسب‌های پیام‌ها، هر پیام توسط دو نفر برچسب‌گذاری شده است. آمار برچسب‌زنی طی چند مرحله محاسبه شد. در هر مرحله تعداد داده‌های برچسب‌خورده توسط هر دانشجو متفاوت بوده است. قبل از مرحله آخر آمار به‌صورت زیر گزارش شد:
U1: 15185، U2: 14289،U3: 20462،U4: 9942 ،U5: 14569،U6: 448 ،U7: 439

[1] request identification
[2] question identification

درنهایت آمار پایانی گزارش‌شده یک فایل اکسل دارای ۸۵۷۴۱ رکورد است [51]. برای جستجو از واژگان کلیدی مانند "چگونه"، "چه‌کسی"، "خریدارم"، "نیازمندم" استفاده کردیم. در این پروژه ابتدا داده‌های به‌دست‌آمده را به دو قسمت تقسیم می‌کنیم. مراحل آماده‌سازی داده‌ها را بر روی این تقسیم‌بندی انجام می‌دهیم. هشتاد درصد نخست داده‌ها را به‌عنوان مجموعه آموزشی و بیست درصد بعدی داده‌ها را به‌عنوان مجموعه آزمون در نظر گرفته و مراحل انجام کار مدل‌سازی و ارزیابی را انجام دادیم. با خواندن هر پیام توسط هر فرد برچسب مربوط به آن تعیین می‌شود. جهت اطمینان از صحت برچسب‌زنی هر پیام توسط دو نفر برچسب زده شد. هر پیام سه برچسب دارد. درنهایت برچسبی که بیشتر از بقیه انتخاب‌شده به‌عنوان برچسب اصلی در نظر گرفته شد.

۲-۴- الگوریتم‌های دسته‌بندی

امروزه مؤلفه‌های سخت‌افزاری مانند حافظه و پردازنده به‌مرور پیشرفته‌تر و ارزان‌تر شده‌اند. این امر موجب رایج‌شدن استفاده از الگوریتم‌های یادگیری ماشین برای حل مشکلات دسته‌بندی متن شده است. برای دسته‌بندی به‌طورمعمول از شبکه عصبی مصنوعی[3]، بیزین ساده[4]، نزدیک‌ترین همسایه[5] و بسیاری از الگوریتم‌های دیگر [11]، [13]، [44] استفاده می‌شود؛ اما مهم‌ترین و رایج‌ترین این الگوریتم‌ها ماشین بردار پشتیبان[6] است که توسط *کورتس* و *واپنیک*[7] [45] توسعه یافته است. جهت ارزیابی، اثبات میزان موفقیت و قابل‌قبول بودن روش، از روش‌های یادگیری مختلفی که در مطالعات مربوط به دسته‌بندی متن و انتخاب ویژگی به‌کاررفته‌اند استفاده می‌کنیم. رایج‌ترین روش یادگیری، ماشین بردار پشتیبان بود. در مورد این روش به‌اختصار در بخش زیر توضیح داده می‌شود.

۱-۲-۴- ماشین بردار پشتیبان

ماشین‌های بردار پشتیبان مجموعه‌ای از روش‌های یادگیری نظارت‌شده برای دسته‌بندی، رگرسیون و ردیابی داده‌های خارج از محدوده است [16]، [46]، [47]. نوعی روش یادگیری ماشین است که بر اساس تئوری یادگیری

[3] artificial neural networks
[4] Naive Bayes(NB)
[5] K-Nearest Neighbors(KNN)
[6] Support Vector Machine(svm)
[7] Cortes and Vapnik

آماری ساخته شده است [44]. هدف دسته‌بندی‌کننده SVM بر اساس مفهوم بیشینه‌سازی حاشیه است [16]، [44]. این الگوریتم به‌دنبال یافتن یک سطح تصمیم‌گیری است که بیشینه فاصله را از نمونه‌های متعلق به دو رده داشته باشد. در این راستا بردارهای پشتیبان، نقاط داده‌ای را که در مرز بین دو رده قرار دارند، شناسایی می‌کنند. یکی از مهم‌ترین الگوریتم‌ها در دسته‌بندی است که دارای نسخه‌های متفاوت خطی و غیرخطی است. نسخه خطی این الگوریتم بهترین مورد برای دسته‌بندی متن است و در بیشتر موارد از آن استفاده می‌شود. هسته‌های مختلف ماشین بردار پشتیبان ایجاد شده است. ازجمله هسته خطی، هسته سیگموید و هسته RBF است که دارای پارامترهای بهینه C و گاما است [13]، [44]، [48]. SVM که توسط واپنیک معرفی‌شده، یکی از موفق‌ترین روش‌های یادگیری ماشین مبتنی بر هسته است [13]، [14]، [45]، [48]. در این مطالعه از ابزارهای کتابخانه‌های موجود و پارامترهای پیش‌فرض این الگوریتم استفاده کردیم [12]، [13]، [49].

۳-۴- معیارهای ارزیابی

الگوریتم‌های یادگیری باید با معیارهایی سنجیده شوند. معیارهای ارزیابی رایج مورداستفاده در متن به دودسته داخلی و خارجی تقسیم می‌شود؛ که می‌توان از اندازه‌گیری شباهت به‌عنوان معیاری داخلی و از دقت[1] [11]، صحت[2]، فراخوانی[3] و اندازه‌گیری اف[4] [12] ,[16] به‌عنوان معیاری خارجی اشاره کرد. صحت و فراخوانی، اندازه‌گیری‌های مشترک استفاده‌شده در زمینهٔ استخراج متن بر اساس معادلات را نشان می‌دهند [15]. در این مطالعه از معیارهای ارزیابی که در بخش‌های زیر به‌طور خلاصه توضیح داده‌ایم استفاده کردیم.

Micro−F1 -۴-۳-۱

برای محاسبه میانگین میکرو همهٔ تصمیمات دسته‌بندی در مجموعه‌داده‌ها بدون تبعیض رده‌ای در نظر گرفته می‌شوند. درصورتی‌که کلاس‌ها در یک مجموعه جانب‌دارانه[5] قرار داشته باشند، آن‌گاه رده‌های بزرگ به محدوده‌های کوچک مسلط می‌شوند. معادله مربوط به محاسبه Micro−F1 به‌صورت رابطه (۷) است و در آن p مقدار صحت و r مقدار فراخوانی مربوط به تمام تصمیمات دسته‌بندی در کل مجموعه‌داده است. این معادله یک اندازه‌گیری کلی است و مربوط به یک کلاس خاص نیست [12]، [16]، [32]، [34]، [49].

$$Micro - F1 = \frac{2 \times p \times r}{p + r} \quad (۷)$$

Macro−F1 -۴-۳-۲

برای محاسبه میانگین ماکرو، F-measure [46] برای هر رده در مجموعه‌داده محاسبه می‌شود و میانگین برای تمام رده‌ها به‌دست می‌آید. درنتیجه، بدون درنظرگرفتن توزیع رده، وزن برابری به هر رده اختصاص داده می‌شود. معادله مربوط به محاسبه Macro−F1 به‌صورت رابطه (۸) است و در آن p مقدار صحت و r مقدار فراخوانی ردهٔ k است. این معادله، میانگین محاسبه برای هر ردهٔ خاص است [12]، [16]، [32]، [34]، [49]. نمره $F1$ میانگین هارمونیک صحت و فراخوانی است و به متعادل‌کردن صحت و عملکرد فراخوانی در بهینه‌سازی دسته‌بندی کننده کمک می‌کند [46].

$$Macro - F1 = \frac{\sum_{k=1}^{C} F_k}{C}, \quad F_k = \frac{2 \times p_k \times r_k}{p_k + r_k} \quad (۸)$$

RMSE -۴-۳-۳

این خطای جذر میانگین مربعات است. RMSE نشان می‌دهد که مقادیر پیش‌بینی‌شده با مقادیر واقعی چقدر نزدیک هستند. ازاین‌رو مقدار پایین‌تر RMSE نشان می‌دهد که عملکرد مدل خوب است. یکی از خصوصیات اصلی RMSE این است که واحد مانند متغیر هدف خواهد بود. رابطه (۹) این معادله را نشان می‌دهد [50].

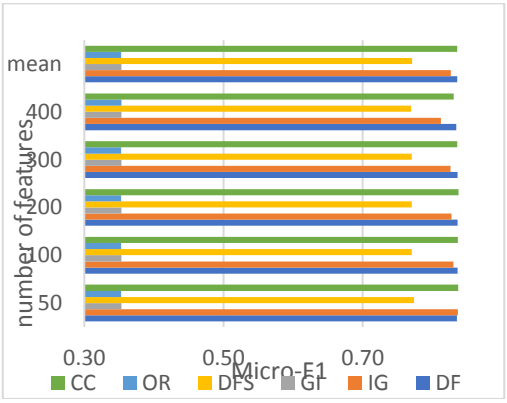
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (۹)$$

۴-۴- ارزیابی روش‌های فردی محلی و سراسری

در این بخش، عملکرد فردی روش‌های انتخاب ویژگی محلی و سراسری مقایسه شد. این مقایسه با توجه به مقادیر Micro-F1 و Macro-F1 به‌دست‌آمده از این روش‌ها انجام شد. تعداد متفاوتی از ویژگی‌ها که توسط هر روش انتخاب می‌شوند، به دسته‌بندی‌کننده‌ها ارسال می‌شود. روش‌های موردبررسی را با تعداد رکوردهای متفاوت از

[1] accuracy
[2] precision
[3] recall
[4] F-measure
[5] biased

داده و همچنین تعداد ویژگی‌های متفاوت انجام داده و نتایج را مقایسه کردیم. نمودار شکل (۴) مقادیر Micro-F1 را که در مجموعه‌داده با ماشین بردار پشتیبان به‌دست‌آمده است، نشان می‌دهد. استفاده از روش‌های انتخاب ویژگی فردی در بیشتر موارد سبب افزایش دقت شد. این افزایش دقت در روش یادگیری SVM و NB بیشتر بوده است. همچنین میزان افزایش با روش انتخاب ویژگی فردی CC نتیجه بهتری داشته است. در نمودار مشاهده می‌کنید که با تعداد ویژگی بالاتر از سیصد به‌طورتقریبی میزان دقت تغییر چندانی نداشته است، لذا ما این زیرمجموعه ویژگی را به‌عنوان زیرمجموعه ویژگی بهینه انتخاب می‌کنیم.



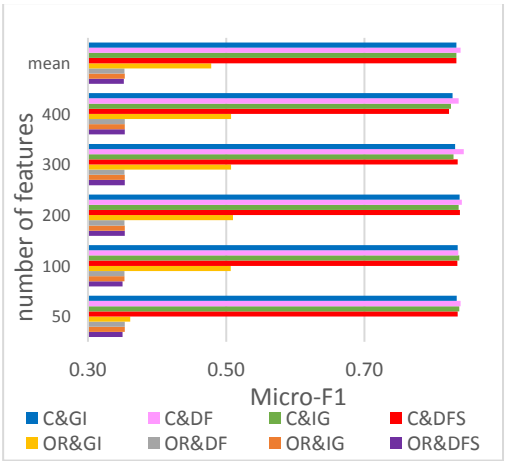
(شکل-۴): نتایج ماشین بردار پشتیبان برای روش‌های فردی
(Figure-4): Support vector machine results for individual methods

در پردازش متن مهم‌ترین گام، ساخت فضای برداری است. واژگان در پرسش‌ها و پیام‌ها، اطلاعات زیادی را به ما می‌دهند. برای این کار از کیسه واژگان استفاده می‌کنیم. در این کیسه واژگان اطلاعات زائد و اضافی مانند واژگان توقف وجود دارد که آن‌ها را حذف می‌کنیم. بعد از ساخت بهترین بردار کیسه واژگان از مناسب‌ترین روش‌های یادگیری استفاده می‌کنیم و با معیارهای ارزیابی نتایج را به‌دست می‌آوریم. در این پژوهش از چهار روش NB، SVM، DT و MLP استفاده شد. روش عملکرد مهم‌ترین این یادگیرنده‌ها در بخش ۳ تشریح شده است. دلیل انتخاب این الگوریتم‌ها این است که مطالعات یادشده، این الگوریتم‌ها را بر روی کیسه واژگان دادگان خود انجام دادند. طبق نتایج به‌دست‌آمده روش یادگیری SVM و NB عملکرد بهتری نسبت به سایر روش‌ها داشت.

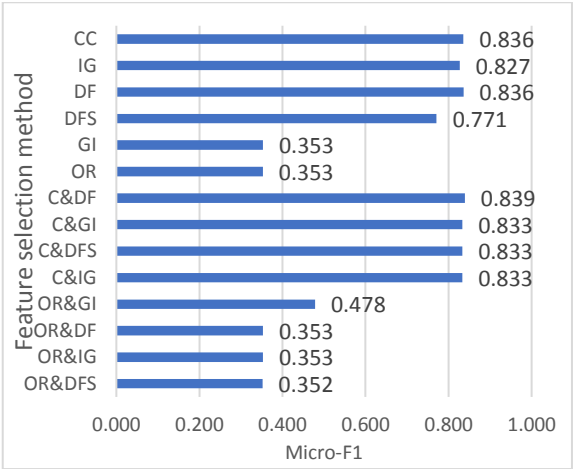
در روش‌های فردی انتخاب ویژگی با تعداد ده‌هزار رکورد، بیشترین مقدار دقت مربوط به الگوریتم NB و روش CC با پنجاه ویژگی بوده است. برای تعداد بیست‌هزار رکورد روش SVM نتیجه بهتری داشته است. در این الگوریتم روش انتخاب ویژگی DFS مقدار دقت بالاتری با یکصد ویژگی داشته است. برای همه رکوردها، روش SVM نتیجه بهتری داشته است. در این الگوریتم روش انتخاب ویژگی CC مقدار دقت بالاتری با دویست ویژگی داشت. برای این روش، میانگین مقادیر معیار Micro-F1 را با تمامی رکوردها و تعداد ویژگی‌های مختلف برای روش‌های فردی در شکل (۴) مشاهده می‌کنید.

۵-۴- ارزیابی روش‌های ترکیبی محلی و سراسری

در این بخش، عملکرد ترکیبی روش‌های انتخاب ویژگی محلی و سراسری مقایسه شد. این مقایسه با توجه به مقادیر Micro-F1 و Macro-F1 به‌دست‌آمده از این روش‌ها انجام شد. تعداد متفاوتی از ویژگی‌ها که توسط هر روش انتخاب می‌شوند، به دسته‌بندی‌کننده‌ها ارسال می‌شود. روش‌های موردبررسی را با تعداد رکوردهای متفاوت از داده و همچنین تعداد ویژگی‌های متفاوت انجام داده و نتایج را مقایسه کردیم. نمودار شکل (۵) مقادیر Micro-F1 را که در مجموعه داده با ماشین بردار پشتیبان به‌دست‌آمده است را نشان می‌دهد. استفاده از روش‌های انتخاب ویژگی ترکیبی در بیشتر موارد سبب افزایش دقت شد. این افزایش دقت درروش یادگیری SVM و NB بیشتر بوده است. همچنین میزان افزایش با روش‌های انتخاب ویژگی ترکیبی شامل CC بهتر بوده است. در نمودار مشاهده می‌کنید که با تعداد ویژگی بالاتر از سیصد به‌طورتقریبی میزان دقت تغییر چندانی نداشته است؛ لذا ما این زیرمجموعه ویژگی را به‌عنوان زیرمجموعه ویژگی بهینه انتخاب می‌کنیم. در شکل (۶) مقایسه‌ای بین روش‌های فردی و ترکیبی دسته‌بند SVM برای میانگین ویژگی‌ها انجام شده است.



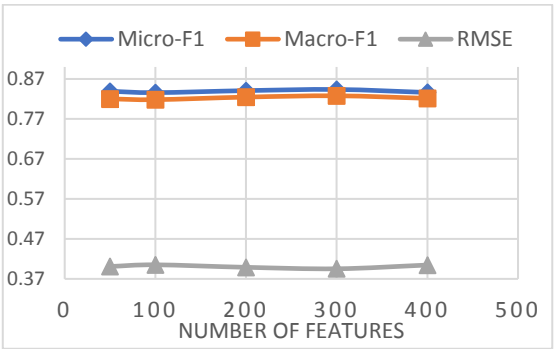
(شکل-۵): نتایج ماشین بردار پشتیبان برای روش‌های ترکیبی
(Figure-5): Support vector machine results for hybrid methods



(شکل-۶): مقایسه دقیق‌تر میانگین روش‌های فردی و ترکیبی در SVM

(Figure-6): A more accurate comparison of the average of individual and hybrid methods in SVM

در روش‌های ترکیبی برای تمامی رکوردها روش یادگیری SVM بیشترین درصد افزایش دقت را نسبت به دقت بدون کاهش ویژگی داشته است [51]. از بین روش‌های ترکیبی برای این الگوریتم روش ترکیبی CC&DF برای تعداد سیصد ویژگی بیشترین عدد دقت یعنی ۰/۸۴۳ را داشته است. مقدارمعیارهای Micro-F1 و Macro-F1 و RMSE در شکل (۷) نشان داده شده است. میزان دقت برای این روش بدون انتخاب ویژگی ۰/۶۹۶ بوده است.

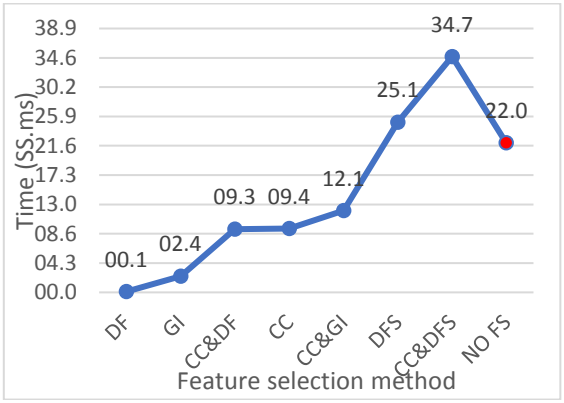


(شکل-۷): نتایج ماشین بردار پشتیبان برای روش CC&DF

(Figure-7): Support vector machine results for CC&DF method

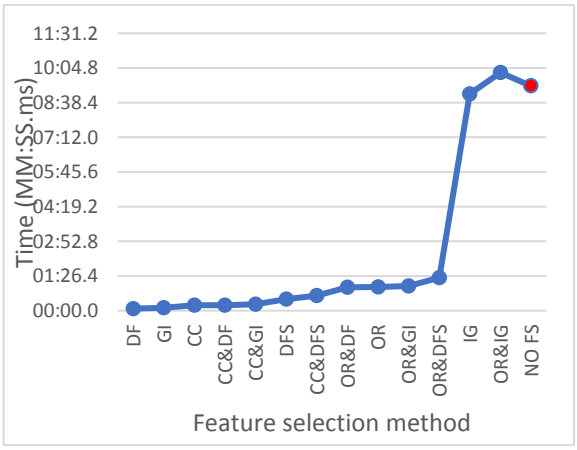
با توجه به جدول (۲) که در پیوست آورده شده است، مشاهده می‌شود که تمامی روش‌های ترکیبی با CC دارای نتایج بالاتری بود. از بین روش‌های ترکیبی با OR هم، ترکیب با GI دارای نتیجه بالاتری است. در بقیه ترکیبات این روش، میزان دقت افزایش پیدا نکرده ولی با توجه به کاهش میزان ویژگی‌ها، این روش با ویژگی‌های کمتر همان دقت را داده است؛ لذا اعمال این روش‌های ترکیبی، در کاهش ابعاد و زمان محاسبه تأثیرگذار است.

بیشتر تمرکز این پژوهش بر روی افزایش دقت بوده است؛ بااین‌حال استفاده از روش‌های انتخاب ویژگی باعث کاهش زمان محاسبات نیز شده‌اند و به همین دلیل نمونه‌ای از نمودارهای زمانی در شکل (۸) برای NB و شکل (۹) برای SVM، جهت مقایسه زمان با و بدون کاهش ویژگی برای روش‌هایی که دقت‌های مطلوبی داشتند برای ده‌هزار نمونه نشان داده شده است. در این نمودارها مشاهده می‌کنید هنگامی‌که از روش‌های انتخاب ویژگی استفاده شده، زمان کاهش یافته است؛ به‌خصوص در SVM علاوه‌بر این‌که در بیشتر روش‌های انتخاب ویژگی مقدار دقت نسبت به بدون انتخاب ویژگی افزایش داشت، در تمامی موارد به‌کارگیری انتخاب ویژگی نیز، زمان به مقدار زیادی نسبت به بدون انتخاب ویژگی کاهش داشته است. در برخی به‌کارگیری‌های روش‌های انتخاب ویژگی، زمان نسبت به بدون انتخاب ویژگی افزایش داشت، ولی این افزایش زمان در برابر افزایش دقت و کاهش ویژگی‌ها ناچیز و قابل‌چشم‌پوشی بود.



(شکل-۸): نتایج مقایسه زمان الگوریتم دسته‌بندی NB با و بدون انتخاب ویژگی

(Figure-8): Results Comparison of NB classification algorithm time with and without feature selection

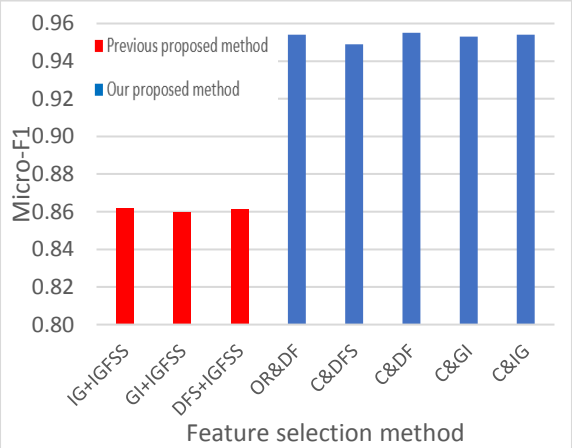


(شکل-۹): نتایج مقایسه زمان الگوریتم دسته‌بندی SVM با و بدون انتخاب ویژگی

(Figure-9): Results Comparison of SVM classification algorithm time with and without feature selection

۱-۵-۴- مقایسه با روش ترکیبی مقالات دیگر

مجموعه‌دادگان فارسی تلگرامی مورداستفاده در این پژوهش، توسط خودمان ساخته شد و مشابه آن را نیافتیم؛ لذا جهت مقایسه و ارزیابی با مقالات دیگر، از یک مجموعه‌دادگان غیرفارسی به نام Reuters-21578 که در پژوهش‌های زیادی در زمینه دسته‌بندی متن و انتخاب ویژگی کاربرد دارد، استفاده کردیم. این مجموعه‌دادگان در مطالعه [12] که روش‌های مبتنی بر فیلتر را باهم ترکیب کرده بود، برای انتخاب ویژگی به کار برده شد. ما نیز روش خود را با این مجموعه‌دادگان به‌کار بردیم و نتایج را مقایسه کردیم. نتایج مقایسه نشان داد که روش ترکیبی ما نتیجه بهتری در مقایسه با روش ترکیبی مقاله پیشین جهت انتخاب ویژگی داشت. در این مقایسه از SVM استفاده شده است. در روش‌های فردی روش CC با مقدار ۰/۹۵۴ دارای نتیجه بالاتری است. در روش‌های ترکیبی روش CC&DF با مقدار ۰/۹۵۵ نتیجه بالاتری داشت. این در حالی است که بالاترین نتیجه برای مطالعه [46] دارای مقدار ۰/۸۶۲ بود. نمودار مربوط به این مقایسه در شکل (۱۰) آورده شده است.



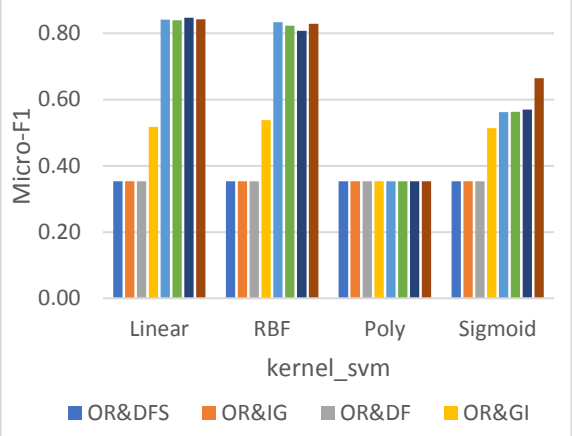
(شکل-۱۰): میانگین نتایج الگوریتم SVM با معیار Micro-F1 جهت مقایسه روش پیشنهادی ما و روش مقالات پیشین با داده Reuters-21578

(Figure-10): Average SVM algorithm results with the Micro-F1 criterion for comparing our proposed method and the method of previous articles with Reuters-21578 data

۶-۴- ارزیابی ماشین بردار پشتیبان با کرنل‌های مختلف

SVM یک یادگیرنده مبتنی بر هسته است. می‌توان این یادگیرنده را با مقدار هسته‌های مختلف مورد ارزیابی قرارداد. نوع هسته حالت‌های خطی، چندجمله‌ای، سیگموید و تابع پایه شعاعی دارد. در این پژوهش به‌این‌علت که SVM نتایج بهتری را ارائه داد، کرنل‌های مختلف این روش یادگیری را هم محاسبه کردیم. در پژوهش‌های مشابه که در بخش مقدمه بیان شدند محاسبه کرنل‌های مختلف انجام نشده بود. با انتخاب کرنل‌های متفاوت، ویژگی‌های متفاوتی انتخاب می‌شوند و این امر سبب می‌شود که مقدارهای دقت به‌دست‌آمده نیز تغییر کند. در ادامه به بیان جزئیات پرداخته‌شده است.

نتایج هسته‌های مختلف را در نمودار شکل (۱۱) مشاهده می‌کنید. هسته خطی عملکرد بهتری داشت و ساده‌تر بود. در Linear SVM با روش ترکیبی CC&DF و برای سیصد ویژگی دارای نتیجه ۰/۸۴۶ است که نسبت به سایر کرنل‌ها مقدار بیشتری بود. در Poly SVM افزایش دقتی نداشتیم. در Sigmoid SVM برای روش ترکیبی CC&GI دارای مقدار۰/۶۶۴ بود. در RBF SVM برای روش ترکیبی CC&DFS نتیجه برابر با ۰/۸۳۳ است. در هرکدام از این انواع هسته با تغییر پارامترهای گاما و C نتایج فرق می‌کند. ما از مقادیر پیش‌فرض برای این پارامترها استفاده کردیم.



(شکل-۱۱): نتایج ماشین بردار پشتیبان با کرنل‌های مختلف

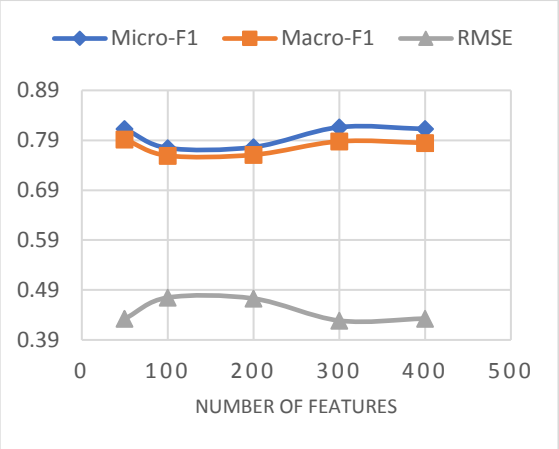
(Figure-11): Support vector machine results with different kernels

نتایج نشان داد که ترکیب روش‌های مبتنی بر فیلتر محلی و سراسری، عملکرد دسته‌بندی بهتری نسبت به روش‌های فردی داشته است. این ترکیب‌ها، با تولید زیرمجموعه بهینه از تمامی ویژگی‌های مهم و کارآمد، ابعاد فضای ویژگی را کاهش دادند. این امر موجب افزایش دقت شد. از طرفی الگوریتم‌هایی که بهترین زیرمجموعه ویژگی را انتخاب می‌کنند از نظر زمان بسیار بهینه بوده و زمان محاسبات را کاهش دادند. درواقع مدل یادگیری، مبتنی بر ویژگی‌های کمتری به‌دست آمده است که چنین مدلی قابلیت تعمیم بیشتری در نمونه‌های جدید را دارد و زمان موردنیاز را نیز کاهش می‌دهد؛ این موضوع نشان‌دهنده

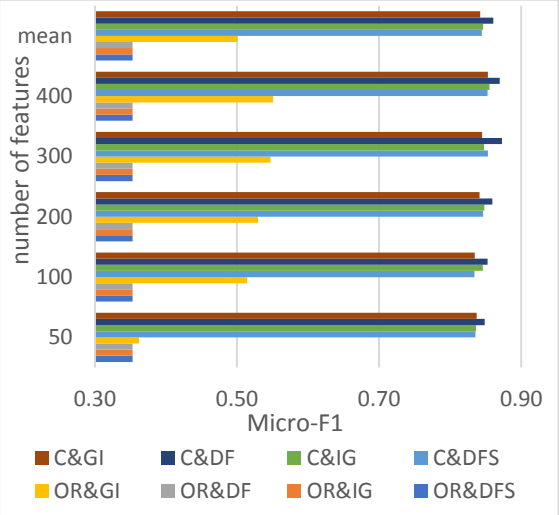
اهمیت انتخاب ویژگی در دسته‌بندی متن است. ارزیابی و مقایسه عملکرد این روش‌ها در نمودارهای بالا نشان داده شده‌اند. با این روش‌های ترکیبی می‌توانیم بهینه‌ترین زیرمجموعه ویژگی را به‌دست آوریم. در هر مرحله عملیات انتخاب ویژگی انجام می‌شود و مدل یادگیری مربوط به آن آموزش داده می‌شود. این مراحل را تا جایی ادامه می‌دهیم که به بهترین نرخ کاهش ویژگی و صحت دسته‌بندی برسیم.

در ده‌هزار رکورد بیشترین میزان افزایش دقت با پنجاه ویژگی بود. در بیست‌هزار رکورد بیشترین میزان افزایش دقت با یکصد ویژگی بود. در تمام رکوردها بیشترین میزان افزایش دقت در سیصد ویژگی بود و این مجموعه به‌عنوان زیرمجموعه بهینه در نظر گرفته شد؛ لذا می‌توان نتیجه گرفت که با افزایش حجم دادهٔ مورد پردازش، زیرمجموعه ویژگی بهینه هم افزایش می‌یابد. در برخی از موارد درصد افزایش بسیار کم، در برخی قابل‌توجه و در برخی دیگر درصد افزایشی نداشته‌ایم؛ اما نکته قابل‌ذکر آن است که کاهش زمان محاسبه با توجه به مقدار کاهش ویژگی قابل‌توجه بوده است.

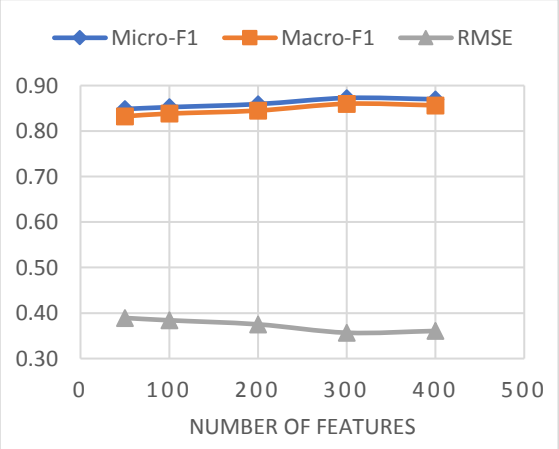
طبق یافته‌ها با توجه به معیارهای ارزیابی مختلف استفاده‌شده در آزمایش‌ها، نتایج حاصل بیان‌گر بیشترین تأثیر روش ترکیبی در میزان افزایش دقت به‌ترتیب مربوط به روش‌های یادگیری SVM، NB و MLP بوده است؛ اما با روش DT افزایشی نسبت به مقدار دقت بدون کاهش ویژگی نداشته است که در مقایسه با کاهش ابعاد ویژگی قابل‌ذکر نیست. تعداد متفاوتی از ویژگی‌ها که توسط هر روش انتخاب می‌شوند، به دسته‌بندی کننده‌ها ارسال شدند. جدول (۲) در پیوست، امتیازات Micro-F1 و Macro-F1 برای مجموعه‌داده با دسته‌بندی‌هایی که بیشترین افزایش را داشته‌اند، نشان می‌دهد. قسمت‌های برجسته در جداول نشان‌دهنده بیشترین امتیاز برای یک روش خاص است. همان‌طور که مشاهده می‌کنید، نتایج تجربی نشان می‌دهد که بیش‌تر روش‌های ترکیبی عملکرد دسته‌بندی را از نظر Micro-F1 و Macro-F1 بهبود بخشیده است. همچنین مقدارهای Micro-F1 از Macro-F1 در تمامی روش‌ها بیشتر است. نتایج و نمودارهای مربوط به ارزیابی دسته‌بندی کننده‌های NB، MLP و DT در نمودار شکل‌های (۱۲) تا (۱۹) آورده شده است. در شکل (۱۹) یک مقایسه کلی از نتایج دو روش دسته‌بندی‌کننده SVM و NB که بالاترین مقدار را داشتند، نشان داده‌ایم.



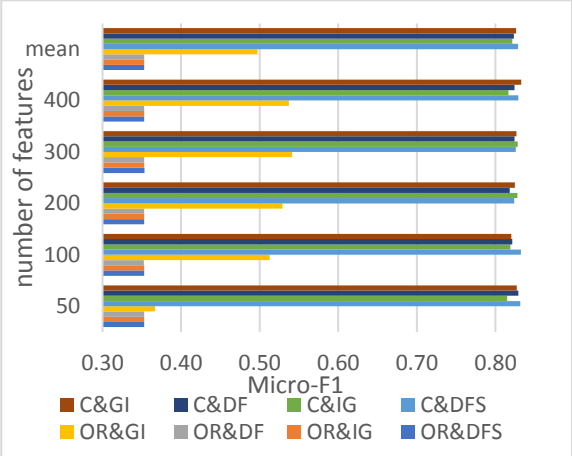
(شکل-۱۴): نتایج بیزین ساده برای روش CC&DF
(Figure-14): Naive Bayes (NB) results for CC&DF method



(شکل-۱۵): نتایج MLP برای روش‌های ترکیبی
(Figure-15): MLP results for hybrid methods

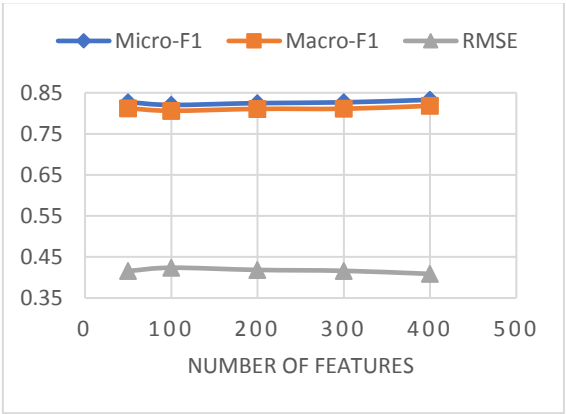


(شکل-۱۶): نتایج MLP برای روش CC&DF
(Figure-16): MLP results for CC&DF method



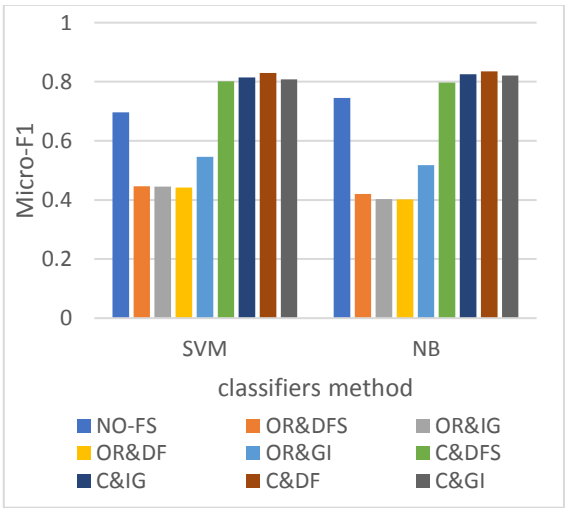
(شکل-۱۷): نتایج درخت تصمیم برای روش‌های ترکیبی

(Figure-17): Decision Tree (DT) results for hybrid methods



(شکل-۱۸): نتایج درخت تصمیم برای روش CC&DF

(Figure-18): Decision Tree (DT) results for CC&DF method



(شکل-۱۹): مقایسه میانگین روش‌های دسته بندی کننده SVM و NB

(Figure-15): Comparison of the mean of SVM and NB classification methods

۵- نتیجه‌گیری و کارهای آینده

در این مطالعه بررسی جامعی بر روی پرکاربردترین روش‌های انتخاب ویژگی مبتنی بر فیلتر انجام شد. با استفاده از مجموعه‌داده‌ها، الگوریتم‌های دسته‌بندی و معیارهای ارزیابی، اثربخشی روش‌های ترکیبی موردبررسی قرار گرفت و در مقایسه با کارایی فردی از روش‌های انتخاب ویژگی مبتنی بر فیلتر مقایسه شد. نتایج نشان داد که روش‌های ترکیبی عملکرد بهتری نسبت به روش‌های فردی داشته‌اند. طبق نتایج به‌دست‌آمده استفاده از روش ترکیبی پیشنهادشده می‌تواند میزان دقت را به مقدار زیادی افزایش و همچنین زمان محاسبه را کاهش دهد. این دو مورد معیارهای تأثیرگذاری در روش‌های یادگیری ماشین هستند. از بین روش‌های یادگیری مورداستفاده، روش‌های ماشین بردار پشتیبان و بیزین ساده درصد بالاتری از افزایش را داشتند. روش ترکیبی CC&DF برای تعداد سیصد ویژگی بیشترین دقت را نسبت به بقیه روش‌ها نشان داد. این مجموعه به‌عنوان زیرمجموعه بهینه شناخته شد. در این پژوهش روش SVM برای دسته‌بندی متن استفاده شد. از میان کرنل‌های مختلف این روش، کرنل خطی بهترین نتیجه را داشت. در پژوهش‌های مشابه علاوه‌بر کم‌بودن حجم داده نسبت به پژوهش ما، استفاده از کرنل‌های متفاوت را نیز مشاهده نکردیم؛ همچنین چنین ترکیبی از این روش‌های انتخاب ویژگی، برای داده فارسی استفاده نشده بود.

یکی از کارهایی که جهت انجام در آینده می‌توان به آن اشاره کرد این است که تمامی این روش‌ها را با روش‌های یادگیرمبنا و یادگیرحاصل ترکیب و نتایج به‌دست‌آمده ازنظر افزایش دقت و زمان را مقایسه کنیم. پیام‌های زیادی در سایر شبکه‌های اجتماعی وجود دارد که علاوه بر داده‌های کسب‌وکاری، دارای سیگنال‌های بورسی هستند و استخراج این داده‌های پنهان می‌تواند سودمند باشد؛ لذا از این روش کاهش ویژگی می‌توان برای به‌دست‌آوردن یک راه‌حل پیشنهادی جهت پیش‌بینی بازار بورس استفاده کرد.

6- References ۶- مراجع

[۱] خبرگزاری تحلیلی ایران، "پیام‌رسان تلگرام در کدام کشورها طرفدار دارد؟"، ۱۱ تیر ۱۳۹۸، برگرفته از لینک: khabaronline.ir/news/1275665، به نقل از سایت: digitalinformationworld.com، تاریخ استخراج: ۱۴ دی ۱۳۹۸.

[1] Iran Analytical News Agency, "In which countries do telegram messengers favor?", *khabaronline.ir*, July. 2, 2019. [Online]. Available: khabaronline.ir/news/1275665. [Accessed:4 January 2020].

[۲] اقتصادنیوز، "آخرین آمار از محبوب‌ترین شبکه‌های اجتماعی در ایران" ، اقتصادنیوز سایت مرجع اقتصاد

ایران، ۲۰ فروردین ۱۳۹۸، برگرفته از لینک: https://b2n.ir/661242، تاریخ استخراج: ۱۴ دی ۱۳۹۸.

[2] Economics News, "Latest statistics from the mostpopular social networks in Iran", *eghtesadnews.com*, April. 9, 2019. [Online]. Available: https://b2n.ir/661242. [Accessed:4 January 2020].

[3] Wikipedia contributors, "Telegram (software)," *Wikipedia, The Free Encyclopedia*, 27 December 2019, 15:24 UTC. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Telegram_(software)&oldid=932678184. [Accessed:4 January 2020].

[4] S. Ranganath, X. Hu, J. Tang, S. Wang, and H. Liu, "Understanding and Identifying Rhetorical Questions in Social Media,", *ACM Transactions on Intelligent Systems and Technology*, vol. 9, pp. 1–22, 2018.

[5] J. Zhang, A. Spirling, and C. Danescu-Niculescu-Mizil, "Asking too much? The rhetorical role of questions in political discourse,", *arXiv preprint arXiv:1708.02254*, 2017.

[6] S. Ranganath, S. Wang, X. Hu, J. Tang, and H. Liu, "Facilitating Time Critical Information Seeking in Social Media,", *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 2017.

[7] A. D. Walker, P. Alexopoulos, A. Starkey, J. Z. Pan, J. M. Gómez-Pérez, and A. Siddharthan, "Answer Type Identification for Question Answering, ", in *Joint International Semantic Technology Conference*. Springer, Cham, 2015, pp. 235–251.

[8] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry,", *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, Jun. 2013.

[۹] عزیزی وامرزانی, حامد و مریم خادمی. بررسی کاربرد و چالش‌های کلان داده در تحلیل عقاید. هفتمین کنفرانس ملی مهندسی برق و الکترونیک ایران، گناباد، دانشگاه آزاد اسلامی گناباد, ۱۳۹۴.

[9] H. A. Vamerzani and M. Khademi, "Exploring the Uses and Challenges of Big Data in Opinion Analysis," in *Proceedings of the 7th Iranian Conference on Electrical and Electronics Engineering, Gonabad, Islamic Azad University of Gonabad, 2016.*

[۱۰] کیانی نژاد, محمد؛ طاهره هاشمی و محسن رشیدی. متن‌کاوی شبکه‌های اجتماعی برای احساسات و تمایلات مصرف‌کننده برند، ششمین کنفرانس بین‌المللی اقتصاد، مدیریت و علوم مهندسی، بلژیک، مرکز بین‌المللی ارتباطات دانشگاهی، ۱۳۹۴.

[10] M. Kiani nejad, T. hashemi, and M. rashidi, " Text mining social networks for consumer brand feelings and desires," in *Proceedings of the 6th International Conference on Economics, Management and Engineering Sciences, Belgium, International Center for Academic Communication, 2016.*

[11] D. Ö. Şahin and E. Kılıç, "Two new feature selection metrics for text classification," *Automatika*, vol. 60, no. 2, pp. 162–171, 2019.

[12] A. K. Uysal, "An improved global feature selection scheme for text classification," *Expert systems with Applications*, vol. 43, pp. 82–92, 2016.

[13] M. Nekkaa and D. Boughaci, "Hybrid harmony search combined with stochastic local search for feature selection," *Neural Processing Letters*, vol. 44, no. 1, pp. 199–220, 2016.

[14] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review,", *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3797–3816, 2019.

[15] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and O. A. Alomari, "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, 2017.

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

[17] G. BİRİCİK, B. Diri, and A. C. SÖNMEZ, "Abstract feature extraction for text classification," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 20, no. Sup. 1, pp. 1137–1159, 2012.

[18] A. Melo and H. Paulheim, "Local and global feature selection for multilabel classification with binary relevance," *Artificial intelligence review*, vol. 51, no. 1, pp. 33–60, 2019.

[19] H. Ogura, H. Amano, and M. Kondo, "Distinctive characteristics of a metric using deviations from Poisson for feature selection," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2273–2281, 2010.

[20] R. Saidi, W. Bouaguel, and N. Essoussi, "Hybrid Feature Selection Method Based on the Genetic Algorithm and Pearson

Decision: An Empirical Study for Appraisal of Feature Selection Methods,", in *2018 24th International Conference on Automation and Computing (ICAC)*, 2018, pp. 1–6.

[33] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, "Feature selection methods for text classification," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 230–239.

[34] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification,", *Knowledge-Based Systems*, vol. 36, pp. 226–235, 2012.

[35] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive Bayes for text categorization," *IEEE transactions on knowledge and data engineering*, vol. 28, no. 9, pp. 2508–2521, 2016.

[36] S. L. Lam and D. L. Lee, "Feature reduction for neural network based text categorization," in *Proceedings. 6th International Conference on Advanced Systems for Advanced Applications*, 1999, pp. 195–202.

[37] N. G. R. Chawla, "Improved Feature Subset Selection using Hybrid Ant Colony and Perceptron Network," *International Journal of Scientific Research and Management*, vol. 5, no. 8, pp. 6764–6770, 2017.

[38] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Computing*, vol. 22, no. 3, pp. 811–822, 2018.

[39] S. Choi, J. H. Shin, J. Lee, P. Sheridan, and W. D. Lu, "Experimental demonstration of feature extraction and dimensionality reduction using memristor networks," *Nano letters*, vol. 17, no. 5, pp. 3113–3118, 2017.

[40] H. Naji, W. Ashour, and M. Alhanjouri, "A New Model in Arabic Text Classification Using BPSO/REP-Tree," *JOURNAL OF ENGINEERING RESEARCH AND TECHNOLOGY*, vol. 4, pp. 28–42, 2017.

[41] N. Kumar, S. Mitra, M. Bhattacharjee, and L. Mandal, "Comparison of Different Classification Techniques Using Different Datasets," Singapore, 2019, pp. 261–272.

[42] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," *Engineering Applications of Artificial Intelligence*, vol. 70, pp. 25–37, 2018.

[43] M. A. Hall, "Correlation-based feature selection for machine learning," *Doctoral dissertation, University of Waikato, Dept. of Computer Science*, 1999.

Correlation Coefficient," in *Machine Learning Paradigms: Theory and Application*, Springer, 2019, pp. 3–24.

[21] N. Nicolosi, "Feature selection methods for text classification," *Department of Computer Science, Rochester Institute of Technology, Tech. Rep*, 2008.

[22] C. Huang, J. Zhu, Y. Liang, M. Yang, G. P. C. Fung, and J. Luo, "An efficient automatic multiple objectives optimization feature selection strategy for internet text classification," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 5, pp. 1151–1163, 2019.

[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

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

[25] Z. Zheng and R. Srihari, "Optimally combining positive and negative features for text categorization," In Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, 2003.

[26] K. Quinn and O. Zaiane, 'Identifying questions & requests in conversation', in *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering*, 2014, pp. 1–6.

[27] B. Li, X. Si, M. R. Lyu, I. King, and E. Y. Chang, 'Question identification on twitter', in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2477–2480.

[28] B. Ojokoh, T. Igbe, A. Araoye, and F. Ameh, 'Question identification and classification on an academic question answering site', in *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 2016, pp. 223–224.

[29] S. Ranganath, X. Hu, J. Tang, S. Wang, and H. Liu, 'Identifying Rhetorical Questions in Social Media.', in *ICWSM*, 2016, pp. 667–670.

[30] W. Cohen, V. Carvalho, and T. Mitchell, 'Learning to classify email into "speech acts"', in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 309–316.

[31] A. Ramzy and A. Elazab, 'Question Identification in Arabic Language Using Emotional Based Features', *arXiv preprint arXiv:2008.03843*, 2020.

[32] B. Z. Abbasi, S. Hussain, S. Bibi, and M. A. Shah, "Impact of Membership and Non-membership Features on Classification

and Telecommunication (JIST), vol. 32, pp. 249–262, 2021.

[44] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, "A new feature selection method based on a validity index of feature subset,", *Pattern Recognition Letters*, vol. 92, pp. 1–8, 2017.

[45] V. Vapnik and V. Vapnik, "Statistical learning theory Wiley," *New York*, pp. 156–160, 1998.

[46] D. Sarkar, "Text Classification,", in *Text Analytics with Python*, Springer, 2019, pp. 275–342.

[47] M. B. Dastgheib and S. Koleini, "Persian Text Classification Enhancement by Latent Semantic Space," *International Journal of Information Science and Management (IJISM)*, vol. 17, no. 1, p. 33, 2019.

[48] C. Qi, Z. Zhou, Y. Sun, H. Song, L. Hu, and Q. Wang, "Feature selection and multiple kernel boosting framework based on PSO with mutation mechanism for hyperspectral classification," *Neurocomputing*, vol. 220, pp. 181–190, 2017.

[49] A. K. Uysal, "On two-stage feature selection methods for text classification,", *IEEE Access*, vol. 6, pp. 43233–43251, 2018.

[50] M. Swamynathan, *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. Apress, 2019.

[51] Z. Khalifeh-Zadeh, M. A. Z. Chahooki, "An Effective Method of Feature Selection in Persian Text for Improving the Accuracy of Detecting Request in Persian Messages on Telegram," *Journal of Information Systems*

زهرا خلیفه‌زاده مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه شیراز و مدرک کارشناسی ارشد را در همان رشته از دانشگاه یزد دریافت کرد. زمینه علاقه‌مندی وی یادگیری ماشین و داده‌کاوی است.

نشانی رایانامه ایشان عبارت است از:

zahra.kh2005@gmail.com

محمدعلی زارع‌چاهوکی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۷۸ از دانشگاه شهید بهشتی و مدرک کارشناسی ارشد و دکترا را به‌ترتیب در سال ۸۳ و ۹۱ در رشته مهندسی نرم‌افزار از دانشگاه تربیت مدرس دریافت کرد. ایشان هم‌اکنون دانشیار دانشکده مهندسی کامپیوتر دانشگاه یزد بوده و زمینه علاقه‌مندی وی یادگیری ماشین، بینایی ماشین و مهندسی نرم‌افزار است.

نشانی رایانامه ایشان عبارت است از:

chahooki@yazd.ac.ir

پیوست

(جدول-۲): نتایج میکرو و ماکرو برای تعداد ویژگی مختلف در روش‌های ترکیبی

(Table-2): Micro and macro results for a number of different properties in hybrid methods

Micro-F1 & Macro-F1 scores for dataset using SVM &NB										
SVM	0.696					0.694				
	Micro-F1					Macro-F1				
number of features	50	100	200	300	400	50	100	200	300	400
Odds_Ratio&DFS	0.353	0.353	0.353	0.353	0.353	0.261	0.261	0.353	0.261	0.261
Odds_Ratio&Information_Gain	0.353	0.353	0.353	0.353	0.353	0.261	0.261	0.261	0.261	0.261
Odds_Ratio&DF	0.353	0.353	0.353	0.353	0.353	0.261	0.261	0.261	0.261	0.261
Odds_Ratio&Gini_Index	0.361	0.507	0.510	0.507	0.507	0.277	0.502	0.506	0.503	0.503
Correlation&DFS	0.835	0.834	0.838	0.835	0.823	0.812	0.812	0.817	0.815	0.807
Correlation&Information_Gain	0.837	0.837	0.836	0.829	0.825	0.817	0.818	0.820	0.812	0.809
Correlation&DF	0.839	0.836	0.841	**0.844**	0.836	0.820	0.818	0.825	**0.828**	0.822
Correlation&Gini	0.834	0.835	0.838	0.831	0.827	0.817	0.817	0.820	0.816	0.813
Document_Frequency	0.836	0.837	0.837	0.836	0.835	0.815	0.817	0.819	**0.821**	0.820
Information_Gain	0.837	0.831	0.828	0.827	0.812	0.815	0.809	0.808	0.808	0.797
Gini_Index	0.353	0.353	0.353	0.353	0.353	0.261	0.261	0.261	0.261	0.261

DFS	0.774	0.771	0.771	0.771	0.770	0.765	0.762	0.762	0.762	0.761
Odds_Ratio	0.353	0.353	0.353	0.353	0.353	0.261	0.261	0.261	0.261	0.261
Correlation	0.837	0.837	**0.838**	0.836	0.831	0.817	0.819	0.820	0.819	0.815
NB	0.745					0.673				
	Micro-F1					Macro-F1				
number of features	50	100	200	300	400	50	100	200	300	400
Odds_Ratio&DFS	0.665	0.674	0.682	0.682	0.689	0.450	0.480	0.518	0.535	0.559
Odds_Ratio&Information_Gain	0.662	0.672	0.665	0.689	0.692	0.444	0.477	0.487	0.551	0.566
Odds_Ratio&DF	0.649	0.661	0.668	0.673	0.678	0.419	0.458	0.489	0.509	0.539
Odds_Ratio&Gini_Index	0.653	0.659	0.659	0.672	0.674	0.424	0.450	0.475	0.510	0.518
Correlation&DFS	0.780	0.654	0.774	0.715	0.659	0.771	0.417	0.701	0.709	0.658
Correlation&Information_Gain	0.791	0.810	0.765	0.676	0.806	0.779	0.782	0.753	0.675	0.768
Correlation&DF	0.813	0.775	0.776	**0.816**	0.813	**0.792**	0.759	0.761	0.788	0.785
Correlation&Gini	0.772	0.700	0.756	0.761	0.731	0.762	0.696	0.744	0.748	0.723
Document_Frequency	0.813	0.804	0.784	0.793	0.804	0.790	0.780	0.763	0.770	0.775
Information_Gain	**0.826**	0.653	0.761	0.695	0.673	**0.806**	0.415	0.690	0.541	0.492
Gini_Index	0.647	0.653	0.657	0.660	0.660	0.399	0.422	0.448	0.457	0.464
DFS	0.655	0.665	0.678	0.685	0.692	0.419	0.452	0.493	0.512	0.529
Odds_Ratio	0.649	0.657	0.670	0.676	0.680	0.412	0.439	0.497	0.528	0.546
Correlation	0.789	0.757	0.793	0.781	0.804	0.777	0.746	0.774	0.764	0.767