Vol. 3, No. 6, Apr-Jun 2023, (24-37)
Journal of New Achievements in Electrical, Computer and Technology (JNAECT)
مجله دستاوردهای نوین در برق،کامپیوتر و فناوری
دوره ۳، شماره ۶، بهار ۱۴۰۲ (۳۷-۲۴)
journal homepage: https://jnaect.ir

دسته‌بندی متن اخبار فارسی با الگوریتم رگرسیون لجستیک

حمیدرضا لطفی[۱]، محمد علی جوادزاده[۲]

[۱] دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع) h.lotfi@ihu.ac.ir
[۲] استادیار دانشگاه جامع امام حسین (ع) javadzade@ihu.ac.ir

چکیده

با توجه به افزایش روزافزون داده، حجم داده‌های متنی نیز با سرعت بالایی در حال رشد است. استخراج اطلاعات از این داده‌های متنی یکی از ضرورت‌های دنیای مبتنی بر اطلاعات امروزی است. دسته‌بندی متن یکی روش‌های دست‌یابی به اطلاعات این داده‌های حجیم است. در این تحقیق با استفاده از یک مجموعه‌داده استاندارد اخبار فارسی که شامل پنج ویژگی در بیش از ۸۶هزار خبر بود به بررسی عملکرد الگوریتم رگرسیون لجستیک در دسته‌بندی متن فارسی و همچنین مقایسه آن با سایر کارهای مشابه پرداختیم. با توجه مراحل ساخت یک دسته‌بند متن،روش مورد استفاده در بخش بردارسازی را توضیح داده و همچنین اهمیت بخش پیش‌پردازش و مخصوصا روش مورد استفاده در برچسب‌گذاری و تبدیل برچسب‌های فرعی به اصلی را بیان کردیم. در ارزیابی نهایی، با استفاده از تغییر پارامترهای الگوریتم و همچنین اصلاح برچسب‌های اخبار، به نتیجه مطلوب ۹۵٪ در معیار دقت برای دسته‌بندی متن مجموعه‌داده اخبار فارسی رسیدیم.

واژه‌های کلیدی: دسته‌بندی متن، رگرسیون لجستیک، پیش‌پردازش متن، مجموعه‌داده اخبار فارسی

۱. مقدمه

متن یکی از متداول‌ترین انواع داده‌های بدون ساختار است که به شکل گزارش، ایمیل، اخبار و مانند آن در فضای وب موجود است. به دلیل ماهیت نامرتب متن، تجزیه و تحلیل، درک، سازمان‌دهی و مرتب‌سازی داده‌های متنی کاری دشوار و زمان‌بر است. حجم اسناد متنی آنلاین که به دسته‌بندی نیاز دارند، با رشد سریع استفاده از اینترنت در سراسر جهان، به سرعت افزایش می‌یابد. اهمیت دسته‌بندی متن برای زبان فارسی از آن‌جایی مشخص می‌شود که محتوای بیش از ۱.۵٪ داده‌های متنی اینترنت به زبان فارسی است که این زبان را در رتبه دهم این لیست قرار داده است [1].

دسته‌بندی متن برای انسان آسان است، اما برای رایانه که فقط با اعداد سروکار دارد کار دشواری است. زیرا رایانه به صورت ذاتی درکی از زبان طبیعی انسان ندارد و این متون باید به صورت داده‌های قابل فهم برای رایانه تبدیل شوند. هرچه داده‌ها ساختارمندتر باشند، تجزیه و تحلیل آسان‌تر و در نتیجه تصمیم‌گیری بهتر خواهد بود [2].

دسته‌بندی متن با استفاده از یادگیری ماشین می‌تواند به‌صورت خودکار، هر نوع متنی را به روشی سریع و مقرون‌به‌صرفه، مرتب و سازمان‌دهی کند و مجموعه‌ای از کلاس‌های ازپیش‌تعریف‌شده را به متن اختصاص دهد. این دسته‌بندی می‌تواند تحت نظارت یا بدون نظارت باشد. دسته‌بندی با نظارت که دقت بیشتری دارد و در این تحقیق از آن استفاده شده است، به نحوه‌ی آموزش سیستم دسته‌بندی با استفاده از مجموعه‌ای از اسناد متنی برچسب دار بستگی دارد [3].

رگرسیون لجستیک یک تکنیک دسته‌بندی پارامتریک و از گروه دسته‌بندی کننده‌های خطی است و الگوریتمی سریع و نسبتا بدون پیچدگی می‌باشد که تفسیر نتایج آن نیز راحت است. این الگوریتم اگرچه اساساً روشی برای دسته‌بندی دودویی است، اما می‌تواند برای مسائل چند کلاسی نیز اعمال شود [4]. در این تحقیق به بررسی استفاده از رگرسیون لجستیک برای دسته‌بندی متن فارسی می‌پردازیم. در این تحقیق از مجموعه‌داده اخبار فارسی سایت فارس‌نیوز استفاده شده است که شامل ۸۶۷۰۰ خبر در ۱۲ برچسب اصلی می‌باشد. ساخت یک سیستم دسته‌بندی متن معمولاً با یک مرحله پیش‌پردازش برای آماده‌سازی متون برای دسته‌بندی خودکار شروع می‌شود. متون فارسی در برخی از مراحل پیش‌پردازش مانند حذف کلمات توقف، ، وزن‌دهی و انتخاب ویژگی با سایر زبان‌ها مشترک هستند. با این حال با توجه به ویژگی‌های خاص این زبان، انواع خاصی از پیش‌پردازش نیز مورد نیاز است [5]. در ادامه کارهای مرتبط این حوزه را بررسی کرده و همچنین در قسمت ادبیات موضوعی اصطلاحات مورد استفاده را توضیح داده‌ایم. روش پیشنهادی این پژوهش را در قسمت روش پژوهش ارائه کرده و در نهایت در قسمت ارزیابی، عملکرد الگوریتم رگرسیون لجستیک را با معیار دقت نشان داده، و همچنین تاثیر تغییر پارامترهای مختلف در این معیار را بررسی کرده‌ایم.

۲. کارهای مرتبط

در طول مطالعه و تحقیق، مقاله مستقلی در مورد استفاده رگرسیون لجستیک در دسته‌بندی متن فارسی یافت نشد؛ بنابراین از مقالاتی که در دسته‌بندی متن با استفاده از رگرسیون لجستیک، بر روی زبانی مشابه زبان فارسی کار کرده‌اند به عنوان کارهای مرتبط در این زمینه استفاده کرده‌ایم. معیار شباهت در این زبان‌ها، استفاده از شکل نوشتاری مشابه کلمات، الفبای غیر لاتین و ساختار زبانی مشابه با زبان فارسی در نظر گرفته شده است.

- الطهراوی [6] در مقاله خود با عنوان دسته‌بندی متن عربی با استفاده رگرسیون لجستیک، و با استفاده از مجموعه‌داده خبرهای عربی پایگاه خبری الجزیره از الگوریتم رگرسیون لجستیک برای دسته‌بندی متن خبرهای عربی استفاده کرده است. از چالش‌های وی در این تحقیق پیش‌پردازش ماهیت خاص زبان عربی می‌باشد. او با استفاده از روش ریشه‌یابی خُجا[1] در یازده مرحله کار پیش‌پردازش متن را انجام داده است. نتایج تجربی به دست آمده از این تحقیق نشان می‌دهد که رگرسیون لجستیک می‌تواند با پیشرفته‌ترین الگوریتم‌های دسته‌بندی متن در زبان عربی رقابت کند. الطهراوی در

[1] Khoja stemming

دسته‌بندی متن عربی با الگوریتم رگرسیون لجستیک، مقدار ۸۶.۵ درصد را برای معیار F-measure به دست آورده است.

- کاندرو و همکاران [7] در مقاله خود تحت عنوان دسته‌بندی متن خبرهای سندی بر اساس روش تجزیه و تحلیل tf-idf، بر روی مجموعه‌داده متن خبرهای زبان سندی که یک زبان از ریشه زبان‌های هندوآریایی است، برای دسته‌بندی متن بر اساس برچسب از پیش تعریف‌شده تمرکز دارد. وی برای جمع‌آوری مجموعه‌های داده، از ابزا استخراج‌گر داده[۲] برای استخراج تیتر اخبار از اکثر روزنامه‌های پرمخاطب سندی مانند عوامی آواز و دیلی جوانگر برای ساخت یک مجموعه‌داده شامل ۲۸۰۰ خبر در پنج دسته استفاده کرده است. برای دسته‌بندی متن در این مجموعه‌داده با استفاده از رگرسیون لجستیک، معیار دقت ۸۳٪ به دست آمده است.
- سازماندهی اخبار و دسته‌بندی مناسب آن‌ها به صورت خودکار از چالش‌های مهم سامانه‌های خبری است. اسد و همکاران [8] در مقاله‌ای با عنوان دسته‌بندی مقالات خبری با استفاده از روش‌های یادگیری ماشینی بانظارت بر روی مجموعه‌داده اخبار اردو و انگلیسی پایگاه خبری پاکستان نیوز، با مقایسه چند الگوریتم و در نهایت با ارائه مدل رگرسیون لجستیک به عنوان بهترین الگوریتم، دقت ۸۹درصدی را به دست آورده است.
- دسته‌بندی خودکار اسناد به دلیل حجم گسترده اسناد متنی که دئما نیز در حال افزایش است، اهمیت فزاینده‌ای دارد. القادی و همکاران [9] بیان می‌دارند که دسته‌بندی متن شامل فرایند برچسب‌گذاری خودکار آن متن با مرتبط‌ترین برچسب است. اما به دلیل عدم وجود مجموعه‌داده مناسب برای زبان عربی، این فرایند با چالش مواجه است. هدف القادی و همکاران در این تحقیق، شناسایی خودکار برچسب یک سند بر اساس ویژگی‌های زبانی آن است. برای رسیدن به این هدف، آن‌ها یک مجموعه‌داده شامل ۹۰ هزار مقاله خبری عربی برچسب‌دار از سایت‌های خبری ایجاد کردند که شامل چهار دسته اصلی تجارت، ورزش، فناوری و خاورمیانه است. در این داده‌ها حروف لاتین، اعداد، علائم نگارشی و کلمات ایستا پاک شده‌اند. القادی و همکاران برای ارزیابی مجموعه‌داده، از چند الگوریتم یادگیری ماشین از جمله رگرسیون لجستیک استفاده کردند. نتایج ارزیابی برای این الگوریتم دقت ۹۸ درصد را نمایش می‌دهد که به همراه الگوریتم ماشین بردار پشتیبان، بالاترین دقت را بین سایر الگوریتم‌ها دارد.

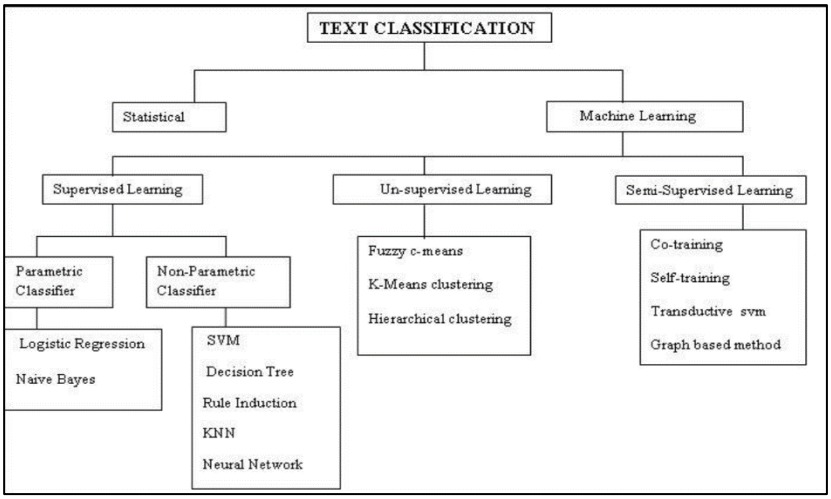
۳. ادبیات موضوعی

الگوریتم استفاده شده در این تحقیق(رگرسیون لجستیک)، یک الگوریتم مبتنی بر یادگیری ماشینی با نظارت و پارامتریک است.در ادامه به توضیح بیشتر ویژگی‌های این الگوریتم می‌پردازیم. الگوریتم‌های دسته‌بندی، پوسته تکنیک‌های متن کاوی را تشکیل می‌دهند. به طور کلی، یک تکنیک دسته‌بندی را می‌توان به رویکردهای آماری و یادگیری ماشین تقسیم کرد. تکنیک‌های آماری صرفاً فرضیه‌های اعلام شده را به صورت دستی برآورده می‌کنند و در کارهای دنیای امروز کمتر مورد استفاده قرار می‌گیرند، اما تکنیک‌های یادگیری ماشین به صورت خاص برای اتوماسیون ایجاد شده‌اند [10]. افزایش حجم، تنوع و رشد داده‌ها باعث به وجود آمدن تکنیک‌های مبتنی بر یادگیری ماشین شده است که از سه زیردسته شامل بدون نظارت، نیمه نظارتی و با نظارت تشکل شده است.

یادگیری با نظارت پرهزینه‌ترین و سخت‌ترین نوع از این سه دسته است. دلیل این هزینه زیاد و دشواری، نیاز به مداخله انسانی هنگام اختصاص برچسب به کلاس‌ها است که در مجموعه‌داده‌های بزرگ امکان‌پذیر نیست و روش‌های یادگیری ماشین سعی در هوشمند کردن و خودکار کردن این موارد دارند. یکی از پرکاربردترین روش‌های نظارت شده، روش تخمین حداکثر احتمال است [11]. در این روش فرآیند یادگیری را می‌توان با استفاده از مفروضات قبلی ساده کرد. نوع این مفروضات در مورد داده‌ها دو رویکرد پارامتری و ناپارامتریک را معرفی می‌کند [12]. مدلی که می‌تواند داده‌ها را بر اساس پارامترهای اصلی خلاصه کند،

[2] **Scraping**

مدل پارامتریک نامیده می‌شود. در یادگیری با نظارت، انتخاب بهترین برچسب‌ها برای برچسب‌گذاری برای دستیابی به دسته‌بندی خوب، فرصتی برای کاهش هزینه‌های زمانی است [13]. جایگاه رگرسیون لجستیک در علم رده‌بندی[3] الگوریتم‌های دسته‌بند مطابق شکل ۱ می‌باشد.



شکل ۱ –جایگاه رگرسیون لجستیک در رده‌بندی الگوریتم‌های دسته‌بندی متن

رگرسیون لجستیک یک مدل افتراقی[4] است که می‌تواند برای دسته‌بندی احتمالی استفاده شود. این الگوریتم احتمالات پسین را به عنوان خروجی می‌دهد. رگرسیون لجستیک نام خود را از منحنی لجستیک یا سیگموئید[5] (معادله ۱) گرفته است [14].

$$P(y|x)= \frac{1}{1+\exp(-y\alpha^{T} x)} \qquad (1)$$

که در آن α پارامتر مدل است.

از آنجایی که این منحنی با یک انتقال سریع و کنترل‌شده به صفر و یک نزدیک می‌شود، برای دسته‌بندی دودویی مناسب است. رگرسیون لجستیک را می‌توان بر اساس تعداد مقوله‌هایی که در زیر ارائه شده است طبقه بندی کرد [15]:

دوجمله‌ای: فقط دو نوع مقدار در متغیر هدف امکان پذیر است: ۰ یا ۱ که می‌تواند نشان دهنده باخت در مقابل برد، شکست در مقابل برد، زنده در مقابل مرده و ... باشد.

چندجمله‌ای: سه یا چند نوع داده در متغیرهای هدف وجود دارند که مرتب نیستند (یعنی نوع آن‌ها کمی نیست) مانند ویروس آ و ویروس ب و ویروس ث.

ترتیبی: دسته‌های مرتب شده در یک متغیر هدف. به عنوان مثال، نمره ارزیابی را می‌توان به صورت: بسیار خوب، خوب، ضعیف و بسیار ضعیف طبقه بندی کرد. در اینجا، به هر دسته می‌توان نمره‌ای مانند ۰، ۱، ۲، ۳ یا بالعکس داد.

رگرسیون لجستیک را می‌توان به راحتی به چندین کلاس تعمیم داد. یعنی می‌توان یک دسته‌بندی چند کلاسه را به عنوان چندین مسئله دسته‌بندی دودویی در نظر گرفت. تصمیم گیری در مورد برچسب کلاس می‌تواند مبتنی بر مقایسه تخمین احتمال با یک حد آستانه و یا به صورت کلی‌تر، با محاسبه اثربخشی تصمیم مورد انتظار انجام گیرد [16].

[3] **Taxonomy**
[4] **Discriminative**
[5] **Sigmoid**

مجموعه‌داده استفاده شده در این تحقیق، مجموعه‌دادگان اخبار فارسی[6] [17] شامل بیش از ۸۶ هزار خبر با ویژگی‌های عنوان، خلاصه، متن، تاریخ و برچسب، و در ۱۲ موضوع اصلی شامل اقتصاد، ورزش، جامعه، بین‌الملل، فرهنگ، علم، هنر، سیاست، استان، آموزش، زندگی و تفریح است که در سال ۱۴۰۰ جمع‌آوری شده است (شکل ۲). پیش‌پردازش انجام شده بر روی این مجموعه‌داده در قسمت روش پژوهش به صورت کامل توضیح داده شده است.

	news_title	news_text	news_summry	news_tag	news_date
0	...تولید 95 درصد تجهیزات ایمنی و آتش‌نشانی در کشو	خبرگزاری فارس ، گروه شهری : سازمان آتش‌نشانی ...ت	معاون حفاظت و پیشگیری از حریق سازمان ...آتش‌نشانی	جامعه / شهری	8 - 4 - 1401 09 : 23
1	هزار دستگاه خودرو ناقص کماکان در پارکینگ 150 ...خ	به گزارش خبرنگار اقتصادی خبرگزاری فارس ، ...موضوع	دبیر انجمن‌خودروسازان با اذعان براینکه 80 ...درص	اقتصادی / صنعت ، تجارت ، بازرگانی	13 : 15 1400 - 10 - 15
2	مفتدی‌الصدر خواستار احضار سفیر سوئد شد	به گزارش سرویس بین‌الملل خبرگزاری فارس ، « ...مقت	رهبر جریان صدر عراق در پی درگیری‌های دو روز ...اخ	بین‌الملل / غرب آسیا	10 : 48 1401 - 1 - 27
3	کرسنت » یک محور اصلی در تحقیق و تفحص از « ...وزا	مرتضی محمودوند نماینده مردم خرم‌آباد در مجلس ...ش	...نماینده مردم خرم‌آباد در مجلس گفت : مسئله‌کرسن	سیاسی / مجلس	14 : 39 1400 - 10 - 24
4	حجت‌الاسلام خطیب صفحه‌ای در شبکه‌های اجتماعی ...ن	به گزارش خبرگزاری فارس ، روابط عمومی وزارت ...اطل	روابط عمومی وزارت اطلاعات اعلام کرد که ...حجت‌الا	سیاسی / امنیتی و دفاعی	15 : 38 1400 - 6 - 5
...	...	...	...	...	...
86696	...اشتغال و مسکن نخبگان هدف موقوفه البرز است / با	...به گزارش خبرنگار قرآن و فعالیت‌های دینی خبرگزا	...مدیرعامل بنیاد موقوفه البرز با اشاره به لزوم گ	فرهنگ / قرآن و فعالیت‌های دینی	14 : 37 1400 - 11 - 24
86697	هزار مهدکودک و پیش‌دبستانی تحت پوشش سازمان 46 ...	به گزارش گروه تعلیم و تربیت خبرگزاری فارس ، ...حل	رئیس سازمان ملی تعلیم و تربیت کودک گفت : 46 ...هز	تعلیم و تربیت / اموزش و پرورش	19 : 53 1400 - 4 - 13
86698	... آرامی : هزار میلیارد تومان خسارات وارد شده به	به گزارش خبرنگار پارلمانی خبرگزاری فارس ، ...منصو	نماینده بندرعباس در مجلس از معاون اول و وزرای ...	سیاسی / مجلس	12 : 42 1400 - 8 - 26
86699	محصول هاینک‌آزمایشگاهی‌ایران‌ساخت‌رونمایی شد 15	به گزارش گروه علم و پیشرفت خبرگزاری فارس به ...نق	نهمین نمایشگاه ایران‌ساخت با حضور معاون علمی ...و	علم و پیشرفت / علم و فن‌آوری ایران	16 : 01 1400 - 9 - 23
86700	نشست کمیسیون اصل 90 برای رسیدگی به موضوع ...عادی	علی خضریان سخنگوی کمیسیون اصل 90 مجلس ...شورای اس	سخنگوی کمیسیون اصل 90 مجلس گفت : نشست ... کمیسیون	سیاسی / مجلس	19 : 35 1400 - 10 - 8

86701 rows × 5 columns

شکل ۲ – نمایی از مجموعه‌داده اخبار فارسی سایت فارس‌نیوز

در ادامه این بخش برخی از اصطلاحات استفاده شده در این تحقیق و در بحث پردازش زبان طبیعی را به اختصار توضیح داده‌ایم و در بخش روش پژوهش به توضیح کامل‌تر کارهای انجام شده پرداخته‌ایم.

- **نرمال‌سازی**[7]: در اولین گام، باید متون برای استفاده در گام‌های بعدی به شکلی استاندارد درآیند. از آنجایی که متون مختلف ممکن است بسیار به هم شبیه باشند اما به دلیل تفاوت‌های ساده ظاهری از نظر ماشین متفاوت باشند؛ به همین دلیل سعی می‌شود این تفاوت‌های ساده مانند فاصله‌های اضافی،تنوع نوشتار با حروف بزرگ و کوچک، کلمات چند املایی، اختصارات و... برطرف گردند [18].
- **نشانه‌بندی**[8]: فرایندی است که در آن یک داده متنی داده شده، به واحدهای زبانی کوچک‌تری به نام نشانه[9] تقسیم‌بندی می‌شود. کلمات، اعداد، علائم نقطه‌گذاری و سایر موارد، از جمله واحدهای زبانی هستند که به عنوان نشانه شناخته می‌شوند [19].
- **فرم ریشه**[10]: به فرایند بازگرداندن کلمات به شکل ریشه‌ای[11] آن‌ها، عملیات Stemming گفته می‌شود [20].

[6] **https://github.com/IHU-PersianNewsDataSet-Javadzade-et-al/dataset**
[7] **Normalization**
[8] **Tokenization**
[9] **Token**
[10] **Stemming**
[11] **Root form**

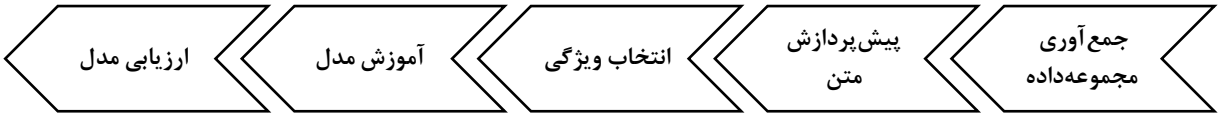
- **واژه‌سازی**[12]: همانند عملیات فرم ریشه ، هدف فرایند واژه‌سازی بازگرداندن کلمات به شکل ریشه‌ای آن‌ها است. با این حال، بر خلاف عملیات فرم ریشه، بخش‌های عطفی و مورفولوژیکی[13] را حذف نمی‌کند، بلکه از پایگاه‌های دانش لغوی جهت پیدا کردن شکل ریشه‌ای صحیح کلمات استفاده می‌کند [21].
- **بردارسازی**[14]: روش‌های مختلفی برای ساختن بردارهای کلمات وجود دارد که همه‌ی آن‌ها به نحوی بدون نیاز به نظارت هستند، بدین معنی که نیاز نیست برای هر کلمه اطلاعات خاصی را به صورت دستی وارد کرد. تنها چیزی که لازم است حجم زیادی متن است. روش‌های تولید بردارهای کلمات شاید از لحاظ محاسبات بعضا پیچیده باشند، اما همگی از یک ایده‌ی کلی برای ساختن بردارها پیروی می‌کنند: کلمات با معنی مشابه در محتوای مشابه ظاهر می‌شوند [22]. کلمه‌ی رئیس‌جمهور را بیشتر در محتوای سیاسی و اطراف کلمات خاصی، مثل قانون، بودجه و مانند این‌ها در متن مشاهده می‌کنید. هدف ساختن بردار برای یک کلمه این است که نشان دهیم یک کلمه از نظر معنایی چقدر و از چه ابعادی به دیگر کلمات شبیه است. روش استفاده شده در این تحقیق برای بردارسازی، روش TF-IDF است. TF-IDF از دو عبارت TF به معنای محدوده فرکانس نرمال و IDF که فرکانس محتوایی معکوس نامیده می‌شود تشکیل شده است [23]. برای به دست آوردن ضریب TF-IDF باید هر کدام از این دو عبارت را مانند معادله ۲ به صورت جداگانه محاسبه نموده و حاصل دو عبارت را در هم ضرب کنیم تا نتیجه، فراوانی وزنی کلمه کلیدی را به ما نشان دهد.

$$w_{i,j}=tf_{i,j}\times\log\left(\frac{N}{df_i}\right) \quad (2)$$

که در آن tf عبارت است از تقسیم تعداد تکرار کلمه بر تعداد کل کلمات محتوا، Idf لگاریتم تقسیم تعداد کل محتوا بر محتواهایی که شامل کلمه مورد نظر هستند. N نیز شامل تعداد کل سندها است.

۴. روش پژوهش یا روش پیشنهادی

مراحل ساخت یک دسته‌بند متن در شکل زیر نشان داده شده است [24]. در ادامه به توضیح هریک از این مراحل می‌پردازیم.



شکل ۳- مراحل ساخت دسته‌بند متن

۴.۱ جمع‌آوری مجموعه‌داده

عدم وجود مجموعه‌داده مناسب در حوزه متن فارسی بسیار پررنگ است به طوری که برای متن‌کاوی و داده‌کاوی در حوزه خبر، چندین پیکره موجود می‌باشد مانند پیکره همشهری، توسعه اندیشه نوین، هزار خبر فارسی و سایرین اما با وجود کارهای تحقیقاتی فراوان که در رابطه با دسته‌بندی متون خبری موجود است، هیچ یک از این تحقیقات با وجود زحمات بسیاری که برای تدوین مجموعه‌داده خود کشیده شده است، اقدام به نشر مجموعه‌داده تولیدی و استفاده شده در تحقیقاتشان نکرده‌اند.

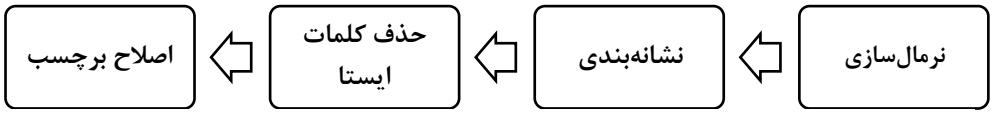
همانطور که در قسمت قبلی اشاره شد مجموعه‌داده استفاده شده در این تحقیق، مجموعه‌دادگان اخبار فارسی [17] می‌باشد. این مجموعه‌داده با پنج ویژگی عنوان خبر، خلاصه، متن اصلی، برچسب موضوعی و تاریخ انتشار جمع‌آوری شده است که در

[12] **Lemmatization**
[13] **Morphological**
[14] **Vectorization**

ویژگی برچسب موضوعی دو نوع برچسب اضافه شده است. این دو نوع برچسب شامل برچسب‌های اصلی و برچسب‌های فرعی هستند که برچسب‌های فرعی به عنوان زیر برچسب‌های هرکدام از برچسب‌های اصلی قرار گرفته‌اند.

۴.۲ پیش‌پردازش متن

از آنجایی که بیشتر اسناد متنی موجود به شکلی بدون ساختار هستند، متنی که فرآیند آموزش بر آن استوار است باید از پیش‌پردازش شود. پیش‌پردازش بخشی ضروری از ساخت یک مدل طبقه بندی کننده است که می‌تواند بر دقت مدل تأثیر مثبت بگذارد [25]. در این تحقیق مرحله پیش‌پردازش شامل چهار مرحله است که در شکل ۴ نشان داده شده است.

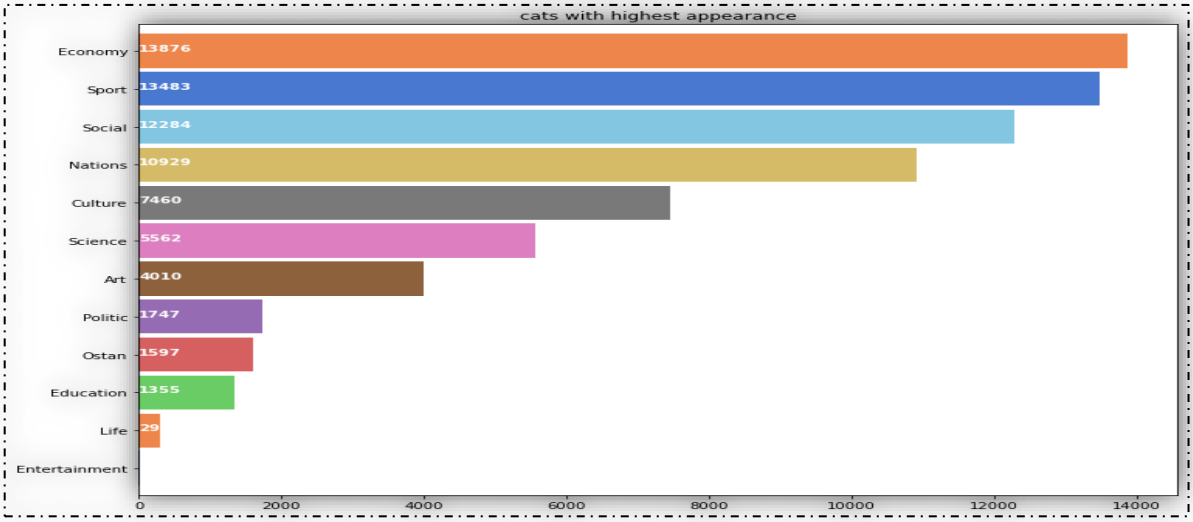


شکل ۴- مراحل پیش‌پردازش

در فرآیند نشانه‌بندی، یک سند متنی به نشانه‌های آن تقسیم می‌شود. در فرآیند نرمال‌سازی، نشانه‌ها و ساختارهای غیر استاندارد در یک سند متنی حذف و یا استاندارد می‌شوند [26]. حذف نیم‌فاصله هم که جزو ویژگی‌های زبان فارسی است در این قسمت انجام می‌شود. کلمات ایستا کلمات متداول در یک سند متنی هستند که حاوی اطلاعات مهمی نمی‌باشند. حذف کلمات ایستا، پیچیدگی مدل را کاهش داده و عملکرد آن را بهبود می‌بخشد.

در مرحله اصلاح برچسب، داده‌های داخل ستون برچسب خبر که شامل برچسب اصلی و زیر برچسب‌های یک خبر خاص هستند، به یک برچسب اصلی تبدیل می‌شوند. یعنی تمام داده‌های یک برچسب در یک گروه جای می‌گیرند. این کار باعث افزایش کارایی مدل هنگام یادگیری و افزایش سرعت و حذف داده‌های پرت می‌شود. این برچسب‌ها، توسط نویسنده از سایت فارس‌نیوز[15] استخراج شده‌اند (شکل ۵).

شکل ۵- لیست برچسب‌های اصلی سایت فارس‌نیوز



شکل ۶- فراوانی برچسب‌های اصلی مجموعه‌داده

[15] **https://www.farsnews.ir/**

۴.۳ انتخاب ویژگی

دسته‌بندی متن معمولاً با مشکل ابعاد بالای فضای ویژگی مواجه است که به معنی وجود تعداد زیادی عبارت در یک سند متنی است. الگوریتم‌های انتخاب ویژگی شامل روش‌های فیلتر و روش wrapper هستند [27]. روش‌های فیلتر، مستقل از الگوریتم یادگیری کار می‌کنند و به زمان کمتری نیاز دارند. این روش‌ها اهمیت هر ویژگی را با استفاده از برخی توابع، اندازه‌گیری و سپس ضروری‌ترین ویژگی‌ها را انتخاب می‌کنند پس برای دسته‌بندی متن کاربردی‌تر می‌باشند.

انتخاب ویژگی الگوریتم‌های مختلفی دارد. ما در این مقاله از روش وزن دهی tf-idf (که به صورت مستقیم روشی برای بردارسازی است و نه انتخاب ویژگی) به عنوان روش انتخاب ویژگی استفاده کرده‌ایم. این روش می‌تواند به ترتیب اهمیت، ویژگی‌های سند متنی را مرتب کرده و در ساخت ماتریس ویژگی به کار بگیرد. بردارساز tf-idf شامل چهار پارامتر اصلی است. این پارامترها برای تنظیم بازه n-گرام‌ها[۱۶]، حد پایین تکرار کلمات[۱۷]، حد بالای تکرار کلمات[۱۸] و حداکثر تعداد ویژگی‌های انتخابی[۱۹] برای ساخت ماتریس ویژگی استفاده می‌شوند [22].

۴.۴ آموزش مدل

برای آموزش مدل باید داده‌ها را به دو دسته داده‌های آموزش و داده‌های آزمایش تقسیم کنیم. متداول‌ترین نوع تقسیم، تقسیم ۷۰ به ۳۰ است که ۷۰٪ داده را به کلاس آموزش و ۳۰٪ داده‌ها را به کلاس آزمایش اختصاص می‌دهد [28]. همانطور که در جدول ۲ نیز قابل مشاهده است، در این تحقیق تقسیم ۳۰-۷۰ داده‌ها بهتر از تقسیم ۲۰-۸۰ و یا تقسیم ۴۰-۶۰ داده‌ها عمل کرد. علاوه بر این، استفاده از این نوع از تقسیم باعث می‌شود نتایج این تحقیق با کارهای معتبر، قابل مقایسه باشد.

جدول۱- مقایسه نوع تقسیم داده‌ها بر دقت الگوریتم

ردیف	نوع تقسیم داده‌های آموزش و تست	دقت
۱	۷۰-۳۰	**۹۵.۱۶**
۲	۸۰-۲۰	**۹۱.۲**
۳	۶۰-۴۰	**۸۴.۸**

۴.۵ معیار ارزیابی

برای ارزیابی مدل دسته‌بند کارایی آن را اندازه‌گیری می‌کنند، که این کارایی به معنی توانایی مدل در انجام صحیح دسته‌بندی است. معیارهای مختلفی برای اندازه‌گیری کارایی وجود دارد. در این مقاله از سه معیار ارزیابی درستی[۲۰]، یادآوری[۲۱]، و دقت[۲۲] استفاده شده است [3].

[16] **N-gram_range**
[17] **Min-df**
[18] **Max-df**
[19] **Max-feature**
[20] **precision**
[21] **recall**
[22] **accuracy**

۵. ارزیابی

نتایج دسته‌بندی متن مجموعه‌داده اخبار فارسی با استفاده از مدل رگرسیون لجستیک در جدول ۳ قابل مشاهده است. معیار ارزیابی را درصد دقت قرار داده‌ایم. برای مشاهده تاثیر هر پارامتر در تغییر نتایج، متغیر مورد نظر را تغییر داده و سایر متغیرها را ثابت در نظر می‌گیریم.

جدول ۲- نمایش تاثیر پارامترها بر دقت

ردیف	n-gram_range	Min-df	Max-df	Max-Feature	دقت
۱	۱-۳	۰.۰۱	۰.۷	همه کلمات	۹۵.۱۶
۲	unigram	۰.۰۱	۰.۷	همه کلمات	۹۳.۵
۳	trigram	۰.۰۱	۰.۷	همه کلمات	۴۹.۹
۴	۱-۳	۰.۱	۰.۷	همه کلمات	۸۴.۹۲
۵	۱-۳	۰.۵	۰.۷	همه کلمات	۳۵.۴۶
۶	۱-۳	۰.۰۱	۰.۵	همه کلمات	۹۴.۹۰
۷	۱-۳	۰.۰۱	۰.۹	همه کلمات	۸۵.۹۳
۸	۱-۳	۰.۰۱	۰.۷	۱۰۰	۷۰.۹۳
۹	۱-۳	۰.۰۱	۰.۷	۱۰۰۰	۹۲.۹۳

۵.۱ بررسی تاثیر پارامترها بر روی نتایج

همانطور که مشاهده می‌شود بهترین دقت زمانی حاصل شده است که برای پارامتر n-گرام از بازه ۱تا۳ (یعنی استفاده از یونیگرام و بایگرام و تریگرام)، برای حد پایین تکرار کلمات مقدار ۰.۰۱ (که به معنی نادیده گرفتن کلماتی که در کمتر از ۱٪ سندها تکرار شده‌اند)، برای حد بالای تکرار کلمات مقدار ۰.۷ (یعنی نادیده گرفتن کلماتی که در ۷۰٪ سندها تکرار شده‌اند) و ساخت ماتریس ویژگی با استفاده از کل کلمات، در نظر گرفته‌ایم (ردیف ۱). ما از این نتیجه به عنوان نتیجه معیار برای مقایسه با سایر نتایج استفاده و در هر مرحله تاثیر تغییر هر یک از متغیرها بر دقت نهایی را بررسی کرده‌ایم.

۱) در تغییر بازه n-گرام (ردیف ۲ و ۳) هنگامی که فقط از یونیگرام استفاده کردیم تاثیر زیادی در دقت مشاهده نشد و دقت فقط ۲٪ کم‌تر شده است اما اگر فقط از تریگرام استفاده کنیم دقت به شکل چشم‌گیری کاهش پیدا کرده و به کم‌تر از ۵۰٪ رسیده است.

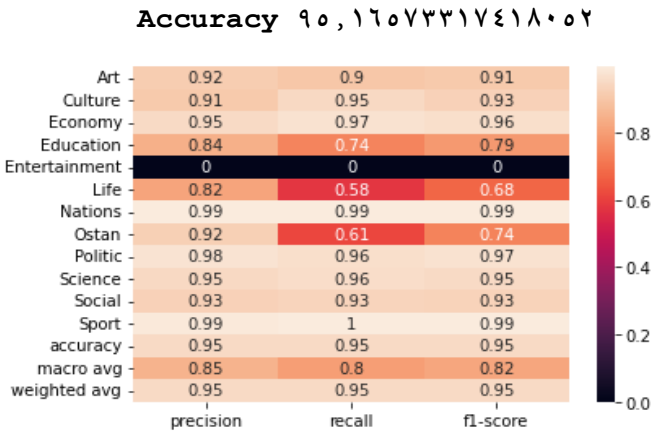
۲) در تعیین حد پایین تکرار کلمات (ردیف ۴ و ۵)، وقتی این مقدار را به ۱۰٪ رساندیم باعث شد دقت‌مان ۱۰٪ کاهش پیدا کند. وقتی این حد را ۵۰٪قرار دادیم و کلماتی که در کمتر از نصف سندها تکرار شده‌اند را در نظر نگرفتیم، معیار دقت به کم‌ترین مقدار خود در این تحقیق یعنی ۳۵٪ رسید. دلیل این افت شدید دقت نادیده گرفتن کلمات کلیدی و پر اهمیت در سندها است که معمولا در کم‌تر از نصف سندها رخ می‌دهند اما در تعیین برچسب سندها نقش بسیار پررنگی دارند.

۳) برای تعیین حد بالای تکرار کلمات (ردیف ۶ و ۷)، این مقدار را به ۵۰٪ کاهش دادیم و کلماتی که در بیشتر از نصف سندها تکرار شده‌اند را نادیده گرفتیم، مقدار دقت فقط ۱٪ کاهش پیدا کرد، ولی وقتی کلماتی که در بیشتر از ۹۰٪ سندها تکرار شده‌اند را نادیده گرفتیم، دقت ۱۰٪ نسبت به حالت معیار کاهش پیدا کرد. این کار باعث می‌شود تعداد بیشتری از کلمات کم اهمیت و عمومی در ساخت مدل دخیل شوند و همین دلیل بر کاهش دقت مدل در تخمین برچسب متن است.

۴) همان‌طور که قبلا اشاره شد، در این تحقیق برای تعیین حداکثر تعداد ویژگی در ساخت ماتریس ویژگی برای مدل از تمام کلمات استفاده کرده‌ایم. به صورت عمومی این کار باعث افزایش هزینه و زمان آموزش مدل می‌شود، چرا که در مجموعه‌داده‌های واقعی تعداد کلمات بسیار زیاد است و به سخت افزار و نرم‌افزار بسیار قدرتمندی برای پردازش نیاز است. در این مقاله برای رسیدن به حداکثر میزان دقت، ماتریس ویژگی با تمام کلمات ایجاد شد که نتیجه مطلوب را نیز در پی داشت. هنگامی که مقدار این پارامتر را ۱۰۰ در نظر گرفتیم (یعنی ماتریس ویژگی را با ۱۰۰ کلمه پراهمیت کل مجموعه‌داده ایجاد کردیم) مقدار دقت در مقایسه با نتیجه معیار، ۲۵٪ کاهش یافت (ردیف ۸) و وقتی ماتریس ویژگی را با ۱۰۰۰ کلمه پر اهمیت این مجموعه‌داده در نظر گرفتیم، میزان دقت تنها ۲٪ درصد کمتر از بهترین نتیجه شد (ردیف ۹). بنابراین اگرچه در این مقاله نتیجه معیار، بهترین دقت را به ما می‌دهد اما در مجموعه‌داده‌های بزرگ‌تر و سخت‌افزار محدودتر و به توجه به نمودار هزینه-زمان، این نتیجه بهینه‌تر خواهد بود.

۵.۲ بررسی تفاوت برچسب‌ها در شاخص‌های ارزیابی

با توجه به شکل ۷، برچسب‌های مختلف در مقدار معیارهای ارزیابی نتایج مختلفی داشته‌اند. وقتی معیار ارزیابی را f1-score در نظر گرفتیم، برچسب ورزشی بهترین نتیجه و برچسب زندگی بدترین نتیجه را داشت. دلیل بالا بودن این مقدار برای برچسب ورزشی تعداد بالای نمونه در مجموعه‌داده و همچنین یادگیری بهتر مدل به دلیل مشخص و خاص بودن کلمات موجود در داده‌های آموزش این برچسب برای تخمین برچسب داده‌های آزمایش بود. اما در برچسب زندگی و استان که کمترین مقدار را داشته‌اند، هم وجود داده‌های کم در مجموعه‌داده در مقایسه با سایر برچسب‌ها و هم نامشخص و عمومی بودن محتوای متن این برچسب‌ها باعث شده است که مدل نتواند به خوبی آموزش ببیند و در نتیجه در تخمین برچسب داده‌های آزمایش به مشکل بخورد.

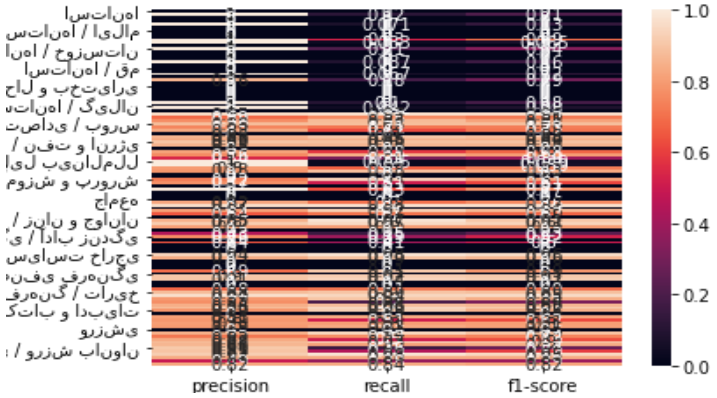


شکل ۷- نتایج ارزیابی در حالت برچسب‌های اصلاح شده

برچسب تفریح نیز به دلیل تعداد بسیار کم در این مجموعه‌داده، مقدار صفر گرفته است. همچنین در شکل ۸ نیز تاثیر استفاده از برچسب‌های اصلاح نشده در یادگیری مدل به جای حالت اصلاح‌شده‌ی آن‌ها به نمایش درآمده است که نسبت به نتیجه معیار، ۱۲٪ درصد کم‌تر است. اگر برچسب‌ها اصلاح نشوند، هر برچسب منحصربه‌فرد در مجموعه‌داده که از برچسب اصلی و زیربرچسب‌ها

تشکیل شده است به عنوان یک برچسب اصلی در نظر گرفته می‌شود. این کار باعث می‌شود که مدل برای بسیاری از این برچسب‌ها تعداد کافی داده برای آموزش نداشته باشد و در نتیجه نتواند در داده‌های آزمایش تخمین مناسبی داشته باشد.

Accuracy 83.58736571481039



شکل۸- نتایج به دست آمده در حالت بدون اصلاح برچسب‌ها

۱۲. نتیجه‌گیری

در این تحقیق ابتدا در مورد مقدمات دسته‌بندی متن و همچنین الگوریتم رگرسیون لجستیک و کاربرد آن در این حوزه توضیح مختصری دادیم. سپس کارهای مرتبط در حوزه دسته‌بندی متن شبه فارسی با استفاده از الگوریتم رگرسیون لجستیک را بررسی کردیم و در ادامه اصطلاحات به کار رفته در دسته‌بندی متن را نیز توضیح مختصری دادیم.در شرح روش پیشنهادی این تحقیق، مراحل ساخت دسته‌بندی متن را شرح داده و هر قسمت را جداگانه بررسی کردیم و نوآوری‌های این تحقیق در قسمت پیش‌پردازش و برچسب‌گذاری را معرفی کردیم. در نهایت نیز نتایج به دست آمده را از نظر تاثیر تغییر پارامترهای مختلف و همچنین از نظر تفاوت برچسب‌ها بررسی و مقایسه کردیم. در نتایج به دست آمده مشاهده شد که روش پیشنهادی ما با در نظر گرفتن معیار دقت برای ارزیابی، عملکرد بهتری نسبت به کارهای مشابه قبلی از خود به نمایش می‌گذارد. در کارهای آتی نیز می‌توان از روش ارائه شده بر روی سایر مجموعه‌داده‌های معتبر فارسی و همچنین برای بررسی میزان دقت الگوریتم در مقایسه با سایر کارهای حوزه دسته‌بندی متن فارسی استفاده کرد.

۱۲. منابع و مراجع

1. "w3techs," World Wide Web Technology Surveys, 28 April 2023. [Online]. Available: https://w3techs.com/technologies/overview/content_language. [Accessed 28 April 2023].
2. Hassani, Hossein., Beneki, Christina., Unger, Stephan., Taj Mazinani, Maedeh., Yeganegi, Mohammad Reza., "Text mining in big data analytics," Big Data and Cognitive Computing, vol. 4, no. 1, 2020.
3. Kowsari, Kamran., Jafari Meimandi, Kiana,., Heidarysafa, Mojtaba., Mendu, Sanjana., Barnes, Laura., Brown, Donald., "Text classification algorithms: A survey," Information, vol. 10, no. 4, 2019.
4. Shah, Kanish., Patel, Henil., Sanghvi, Devanshi., Shah, Manan., "A comparative analysis of logistic regression, random forest and KNN models for the text classification.," Augmented Human Research, vol. 5, pp. 1-16, 2020.
5. Rezaeian, Naeim., Novikova, Galina.,, "Persian text classification using naive Bayes algorithms and support vector machine algorithm," Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 8, no. 1, pp. 178-188, 2020.
6. Al-Tahrawi, Mayy M, "Arabic text categorization using logistic regression," International Journal of Intelligent Systems and Applications, vol. 7, no. 6, 2015.
7. Kandhro, Irfan Ali., Jumani, Sahar., Lashari, Ajab Ali., Nangra, Saima., Lakhan, Qurban Ali., Taimoor Baig, Mirza., Guriro, Subhash., "Classification of Sindhi headline news documents based on TF-IDF text analysis scheme," Indian Journal of Science and Technology, vol. 12, no. 33, pp. 1-10, 2019.
8. Asad, Muhammad Imran, Abubakar Siddique, Muhammad., Hussain, Safdar., Naveed Hassan, Hafiz., Munawwar Gul, Jam., "Classification of News Articles using Supervised Machine Learning Approach," Pakistan Journal of Engineering and Technology, vol. 3, no. 03, pp. 26-30, 2020.
9. Qadi, Leen Al., Rifai, Hozayfa El., Obaid, Safa., Elnagar, Ashraf., "Arabic Text Classification of News Articles Using Classical Supervised Classifiers," in 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), Madrid, 2019.
10. Luo, Xiaoyu., "Efficient English text classification using selected machine learning techniques," Alexandria Engineering Journal, vol. 60, no. 3, pp. 3401-3409, 2021.
11. Kadhim, Ammar Ismael., "Survey on supervised machine learning techniques for automatic text classification," Artificial Intelligence Review, vol. 52, no. 1, pp. 273-292, 2019.
12. Park, Junyong, "Simultaneous estimation based on empirical likelihood and general maximum likelihood estimation," Computational Statistics & Data Analysis, vol. 117, pp. 19-31, 2018.
13. Thangaraj, Muthuraman., Sivakami, Muthusamy., "Text classification techniques: A literature review," Interdisciplinary journal of information, knowledge, and management, vol. 13, p. 117, 2018.
14. Gasparetto, Andrea., Marcuzzo, Matteo., Zangari, Alessandro., Albarelli, Andrea., "A survey on text classification algorithms: From text to predictions," Information, vol. 13, no. 2, p. 83, 2022.
15. Hassan, Sayar Ul., Ahamed, Jameel., Ahmad, Khaleel., "Analytics of machine learning-based algorithms for text classification," Sustainable Operations and Computers, vol. 3, pp. 238-248, 2022.

16. Pranckevičius, Tomas., Marcinkevičius, Virginijus., "Application of logistic regression with part-of-the-speech tagging for multi-class text classification," EEE 4th workshop on advances in information, electronic and electrical engineering (AIEEE), pp. 1-5, 2016.
17. Javadzade., Mohammad Ali., Hosseini, Hossein., Ghalenoei, Mohammad., Mokhtari, Mohammad Mahdi., "Design and production of Persian news data set IHU-PersianNewsDataSetJavadzade-et-al Imam Hossein Comprehensive University," Journal of New Achievements in Electrical, Computer and Technology, vol. 2, no. 3, pp. 103-121, 2022.
18. HaCohen-Kerner, Yaakov., Miller, Daniel., Yigal, Yair., "The influence of preprocessing on text classification using a bag-of-words representation," PloS one, vol. 15, no. 5, 2020.
19. Alyafeai, Zaid., Al-shaibani, Maged S., Ghaleb, Mustafa., Ahmad, Irfan., "Evaluating various tokenizers for Arabic text classification," Neural Processing Letters, pp. 1-23, 2022.
20. Almuzaini, Huda Abdulrahman., M. Azmi, Aqil., "Impact of stemming and word embedding on deep learning-based Arabic text categorization," IEEE Access, vol. 8, pp. 127913-127928, 2020.
21. Kurasinski, Lukas., Mihailescu, Radu-Casian., "Towards machine learning explainability in text classification for fake news detection," in 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020.
22. Wendland, André., Zenere, Marco., Niemann, Jörg., "Introduction to text classification: impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique," in Systems, Software and Services Process Improvement: 28th European Conference, EuroSPI 2021, Krems, Austria, 2021.
23. Zhou, Hai, "Research of Text Classification Based on TF-IDF and CNN-LSTM," Journal of Physics: Conference Series, vol. 2171, no. 1, 2022.
24. Sebastiani, Fabrizio., "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1-47, 2002.
25. Mostafavi, Sareh., Pahlevanzadeh, Bahareh., Falahati Qadimi Fumani, Mohammad Reza., "Classification of Persian News Articles using Machine Learning Techniques," Computer and Knowledge Engineering, vol. 4, no. 1, pp. 1-10, 2021.
26. Kobayashi, Vladimer B., T. Mol, Stefan., A. Berkers, Hannah., Kismihok, Gábor., N. Den Hartog, Deanne., "Text classification for organizational researchers: A tutorial," Organizational research methods, vol. 21, no. 3, pp. 766-799, 2018.
27. Sammons, Mark., Chris todoulopoulos, Christos., Kordjamshidi, Parisa., Khashabi, Daniel., Srikumar, Vivek., Roth, Dan., "EDISON: Feature Extraction for NLP, Simplified," in Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 2016.
28. Joseph, V. Roshan., "Optimal ratio for data splitting," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 15, no. 4, pp. 531-538, 2022.

Classification of Persian news text with logistic regression algorithm

Hamid reza Lotfi, Mohammad ali Javadzadeh
Master's student of Imam Hossein University (AS), h.lotfi@ihu.ac.ir
Assistant Professor of Imam Hossein University (AS), javadzade@ihu.ac.ir

Abstract— Due to the ever-increasing amount of data, the amount of textual data is also growing at a high speed. Extracting information from these textual data is one of the necessities of today's information-based world. Text classification is one of the methods of obtaining information from this massive data. In this research, using a standard dataset of Persian news, which included five features in more than 86 thousand news, we investigated the performance of the logistic regression algorithm in the classification of Persian text and also compared it with other similar works. Considering the steps of creating a text category, we have explained the method used in the vectorization section and also stated the importance of the pre-processing section, especially the method used in tagging and converting sub-tags to main ones. In the final evaluation, by changing the algorithm's parameters and modifying the news tags, we reached the desired result of 95% in the accuracy criterion for the text classification of the Persian news dataset.

Keywords: Text classification, Logistic regression, Text preprocessing, Persian news dataset