سومین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات

3nd National Conference on Advanced Topics in Computer and Information Technology

بیست وهشتم آذر ماه ۱۳۹۸

ارائه روشی مبتنی بر تحلیل آماری منبع واژگانی وردنت و محتوا به منظور تحلیل عقاید در اسناد فارسی

یاسمن ناصحی

گروه مهندسی کامپیوتر - دانشگاه آزاد اسلامی واحد ماهشهر

Yasaman.n90@gmaill.com

مرجان عبد یزدان

گروه مهندسی کامپیوتر - دانشگاه آزاد اسلامی واحد ماهشهر

Abdeyazdan87@yahoo.com

چکیده

روزانه میلیون‌ها کاربر در سرتاسر دنیا داده‌های خود را از طریق اینترنت به اشتراک می‌گذارند. تحلیل و بررسی این داده‌ها دانش مفیدی را در اختیار ما قرار می‌دهد. در این پژوهش روشی مبتنی بر با تحلیل آماری مجموعه واژگانی وردنت برای دسته‌بندی نظرات در زبان فارسی ارائه شده است و از منبع واژگانی سنتی‌وردنت به منظور گروه‌بندی ویژگی‌ها و انتخاب ویژگی استفاده کردیم.روش پیشنهادی در این مقاله به چندین مرحله تقسیم می‌شود در اولین گام بعد از گرفتن اسناد آن را به جملات تجزیه می‌کردیم بعد از ان عملیات پیش‌پردازش را بر روی جملات انجام شد در ادامه عملیات برچسب زنی انجام شد که جهت زدن برچسب نقش کلمات فارسی بر روی کلمات از نرم‌افزار برچسب نقش کلمات فارسی دانشگاه فردوسی مشهد بهره بردیم در گام بعد ویژگی‌های تشدید کننده و تضعیف کننده را مشخص و نقش آنها در جملات را تحلیل کرده‌ایم و به سراغ اعمال تجزیه‌گر بر روی جملات رفتیم تا با این عملیات سریالی ویژگی‌های خود را جهت عملیات وزن‌گذاری آماده کنیم. در روش پیشنهادی این پژوهش از مجموعه داده استاندارد همشهری جهت ارزیابی بهره گرفته‌ایم و از چهار معیار دقت،صحت،بازخوانی و معیار F1 جهت ارزیابی روش پیشنهادی استفاده کرده‌ایم نتایج نشان دادند که روش پیشنهادی این پژوهش دارای کیفیت بالاتری نسبت به روش‌های پیشین است. نتایج ارزیابی نشان می‌دهد که معیار صحت ۵ درصد و معیار F1 یک درصد بهینه‌تر شده است و نرخ اشتباهات به میزان ۶درصد نسبت به روش‌های پیشین کاهش داشته است.

کلمات کلیدی:دسته بندی، وردنت، ویژگی، پیش پردازش، POS

۱- مقدمه

در سال‌های اخیر با معرفی وب ۲/۰ و ۳۰۰ که بر پایه مشارکت‌ها و تعاملات استوار است، شاهد گسترش فناوری های وب، رونق رسانه‌های اجتماعی و افزایش تعاملات کاربران در اغلب وب سایت‌ها هستیم که حجم زیادی از نظرات غنی و همچنین ارزان را ایجاد کرده‌اند. درصد قابل توجهی از این داده‌ها به صورت متن و صورت‌های دیگر رسانه‌ای نظیر صدا و تصویر نگهداری می‌شوند؛ اما به دلیل نبود یک استاندارد همه جانبه و دقیق در تنظیم متون و ثبت آنها، این داده‌ها طبیعتی غیرساخت‌یافته و یا نیمه‌ساخت یافته دارند [۱۹].

۱-۱-بیان مسئله

با گسترش روز افزون اینترنت و استقبال گسترده که از طرف کاربران مواجه شده است؛ روزانه میلیون‌ها کاربر در سرتاسر دنیا داده‌های خود را از طریق اینترنت به اشتراک می‌گذارند. بخش قابل توجهی از این داده‌ها از نوع متن است. تحلیل و بررسی این داده‌ها دانش مفیدی را در اختیار ما قرار می‌دهد. بنابراین نیاز به ارائه روش‌های خودکار برای تحلیل این داده‌ها به منظور اکتشاف دانش و ارائه آن به عامل انسانی از اهمیت ویژه‌ای برخوردار است. تحلیل احساس،که عقیده‌کاوی نیز نامیده می‌شود، هم اکنون نیز یکی از زمینه‌های تحقیقی فعال در پردازش زبان طبیعی می‌باشد[۲۴]. از دو نوع اطلاعات موجود، اطلاعات مبتنی بر محتوا و مبتنی بر پیکره به منظور مدیریت بردار ویژگی‌ها و همچنین دسته‌بندی ویژگی‌ها استفاده می‌کنیم.

۲-۱- اهمیت پژوهش

در این پژوهش برای اولین بار از دانش آماری پیکره به منظور مهندسی خصیصه‌ها استفاده خواهد شد.این مهم‌ترین اهمیت در انجام این پژوهش می‌باشد، به همین دلیل باید روابط آماری و همچنین فرمول‌های لازم طراحی و ارائه داده شوند. دانش مبتنی بر محتوا نیز به منظور تحلیل عقاید بکار گرفته خواهد شد. بنابراین باید روشی برای ترکیب این دو معیار ارائه داده شود.

۳-۱- ساختار پژوهش

پژوهش جاری در پنج فصل آماده و ارائه شده است که شرح فصول این پژوهش به شرح ذیل می‌باشد:

فصل اول بیان مسئله است فصل دوم این پژوهش به اختصارات مربوط به موضوع پژوهش اختصاص دارد به همین دلیل ضمن بیان تعریفی از متن‌کاوی و تحلیل احساسات به بیان ویژگی‌ها و تحلیل آماری متون پرداخته‌ایم و در اخر اعم سوابق مربوط به روش پیشنهادی اشاره کرده-ایم.

در فصل سوم به بیان روش پیشنهادی پرداخته‌ایم به همین دلیل فلوچارتی علمی جهت آنچه قرار است در این پژوهش اتفاق بیفتد ارائه داده‌ایم سپس به تفکیک و بصورت کامل تمام مراحل فلوچارت را باز کرده و روش کار آن را شرح داده ایم.

فصل چهارم این پژوهش همانگونه که در پاراگرف قبل اشاره شد به شبیه‌سازی و ارزیابی روش پیشنهادی که در فصل سوم بیان شده است اختصاص دارد. در این فصل ضمن معرفی داده‌های استاندارد مورد استفاده به معرفی ابزار مربوط به اجرای روش پیشنهادی و اندازه‌گیری آن می‌پردازیم و در اخر با استفاده از اشکال و نمودارهایی نتیجه روش پیشنهادی را نشان می‌دهیم و در اخر این فصل باز هم با استفاده از نمودارهای روش پیشنهادی را با یکی از کارهای معتبر پیشین مشابه با روش پیشنهادی به مقایسه گذاشته و مزایا و معایب آنها را نسبت به همدیگر اشاره می‌کنیم.

فصل پنجم این پایان‌نامه به نتیجه‌گیری و کارهای آتی اختصاص داده شده است.

۲- پیشینه ی تحقیق

۲-۱- متن‌کاوی

متن‌کاوی که به «کشف دانش در متن»، «داده‌کاوی متنی» و«تحلیل هوشمند متن» مشهور است، به طور کلی به فرآیند استخراج دانش و اطلاعات مورد علاقه و مهم از مجموعه متنی غیرساخت‌یافته اشاره دارد؛ به عبارت دیگر متن‌کاوی فرآیند تحلیل طبیعی متن به منظور کشف و ثبت اطلاعات معنایی برای درونداد و ذخیره‌سازی در یک ساختار سازماندهی شده دانش است [۳۳].

۲-۲- عقیده‌کاوی

عقیده‌کاوی و تحلیل احساسات، فرآیند تحلیل نظرات، احساسات و عواطفی است که در داده‌های متنی بیان شده است. در مکان‌های متنوع و مختلف، درک کردن آن چه دیگران فکر می‌کنند یکی از مهمترین دانش‌ها در فرآیند تصمیم‌گیری می‌باشد. چنین شیوه‌های بدست آوردن اطلاعات را تجزیه و تحلیل احساسات می‌گویند [۴۲].

۲-۳- پردازش زبان طبیعی

هدف اصلی در پردازش زبان طبیعی، ایجاد تئوری‌های محاسباتی از زبان، با استفاده از الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم رایانه‌ای است. بدیهی است که در راستای تحقق این هدف، نیاز به دانشی وسیع از زبان بوده و علاوه بر پژوهشگران دانش رایانه‌ای، نیاز به دانش زبان شناسان نیز در این حوزه خواهد بود. کاربردهای پردازش زبان طبیعی به دو دسته کلی قابل تقسیم است[۱]:

- کاربردهای نوشتاری
- کاربردهای گفتاری

۲-۴- پردازش زبان فارسی

جایگاه زبان فارسی در میان زبان‌های دیگر را از سه جنبی ورائتی(تاریخی) ، ناحیه‌ای و رده شناختی می‌توان بررسی کرد[۴۹]: از دیدگاه زبان شناسی تاریخی، فارسی همراه با زبان‌های هند آریایی، زیر گروهِ هند - ایرانی را در گروه شرقی زبان‌های هند و اروپایی تشکیل می‌دهند. این زیر گروه شامل زبان‌هایی مانند فارسی، پشتو و کردی می‌باشد. از نظر ناحیه‌ای، به دلیل همسایگی با کشورهای عربی زبان، دارای بسیاری کلمات قرضی و حتی برخی قواعد مشابه با آنها است، فارسی از دیدگاه ویژگی‌های زبانی (رده شناختی)، یک زبان پیوندی و ضمیرانداز است، فارسی از راست به چپ نوشته می‌شود و اگرچه در اصل دارای ترتیب فاعل - مفعول - فعل است ولی مملو از استثنائات مجاز در این ترتیب می‌باشد که حاصل فرآیندهایی چون نامکانی بهم ریختگی، حرکت جهت برجسته سازی، تاخیر، شکافت و شبه شکافت و غیره هستند، و به دلیل استفاده فراوان، عملا فارسی را به یک زبان بدون ترتیب تبدیل می‌کنند[۵۰].

۲-۵- آمار و تحلیل آماری

امروزه به منظور کشفِ « نوع و شدت رابطه ی معنادار » میان متغیرهای مختلف از علم آمار استفاده می‌شود و حتی می‌توان « جهت رابطه » را نیز از طریق آن پیدا نمود. به عبارت دیگر علاوه بر اینکه می‌توانیم نسبت به وجود یا عدم وجود همبستگی معنادار بین دو متغیر ابراز نظر کنیم ، قادریم تا نسبت به اینکه کدام متغیر بر متغیر دیگر اثر می‌گذارد نیز توضیح بدهیم[۳].

بنابراین در تعریف می توان گفت که آمار عبارت است از:

-علمِ طبقه بندی دیتاها و اطلاعات

-علم تصمیم گیری های مبتنی بر منطق

-علم برنامه ریزی های دقیق

-علم توصیف و تبیین آنچه که از مشاهدات می توان فهمید

۲-۶- وردنت

وردنت به انگلیسی یک پایگاه دادگانِ واژگانی برای زبان انگلیسی است. وردنت واژه‌های انگلیسی را به مجموعه‌های مترادفی که synsets نامیده می‌شود، گروه‌بندی می‌کند، تعاریف کوتاه عمومی را به‌دست می‌دهد و روابط مترادف‌های گوناگون را با این مجموعه‌های مترادف ضبط می‌کند.

۲-۷- منبع واژگانی سنتی‌وردنت

سنتی‌وردنت یک منبع واژگانی مبتنی بر وردنت است که به صورت خودکار تولید شده است. نسخه‌های ۱، ۲ و ۳ آن موجود و در دسترس هستند. سنتی‌وردنت ۳ که مبنای این پژوهش قرار گرفته، مبتنی بر وردنت۳ است. سنتی‌وردنت ۳ برای هر مجموعه معنایی موجود از وردنت یک رتبه سه‌تایی ارائه داده است که این اعداد میزان مثبت، منفی و عینی بودن هر مجموعه معنایی را مشخص می‌کنند.

رابطه ۲-۱ برای همه مجموعه‌های معنایی برقرار است[۵۴].

(۲-۱)

$$ObjScore + PosScore + NegScore = 1$$

همانطور که در رابطه ۲-۱ دیده می‌شود، مجموعه رتبه‌های مثبت، منفی و عینی هر مجموعه معنایی باید یک شود. هر معنی از کلمه w با توجه به مقادیر عددی سه‌تایی به مکانی در شکل ۲-۲ نگاشت داده می‌شود. ObjScore با استفاده از رابطه ۲-۲ محاسبه می‌شود.

(۲-۲)

$$ObjScore = 1 - (PosScore + NegScore)$$

۲-۸- ویژگی

این مرحله به انتخاب زیر مجموعه‌ای از ویژگی‌های متن یا کلمات اشاره دارد. با توجه به اینکه تعداد ویژگی‌ها در متن بسیار زیاد می‌باشد و این امر در کاهش کارایی دسته‌بندی تاثیر زیادی خواهد داشت، در مرحله انتخاب ویژگی سعی بر این است که از بین ویژگی‌های موجود مهم‌ترین و اساسی‌ترین ویژگی‌هایی که به افزایش کارایی دسته‌بندی کمک می کنند، انتخاب شوند. با حذف ویژگی‌های غیر مرتبط و غیرقابل تمایز، کارایی دسته‌بندی می‌تواند افزایش یابد[۵].

۲-۹- پیش‌پردازش

تمامی حوزه‌های مرتبط با پردازش زبان طبیعی به نحوی از انحنا با متون واقعی سروکار دارند. صورت‌های غیراستاندارد نویسه‌ها وکلمات به وفور در این نوع متون دیده می‌شوند . قبل از اینکه بتوان از این متون به منظور استفاده در سیستم‌های تبدیل متن به گفتار، ترجمه ماشینی، بازشناسی حروف فارسی، خلاصه‌ساز فارسی، جستجو در متون فارسی استفاده کرد و یا در پایگاه داده ذخیره‌شان کرد، باید ابتدا پیش پردازشی روی آنها انجام گیرد، تا صورت‌های غیر استاندارد به شکل استاندارد تبدیل گردند. طی فرآیند پیش‌پردازش علایم نگارشی، حروف، فاصله‌های بین کلمات، اختصارات وغیره بدون ایجاد تغییرات معنایی در متن به شکل استانداردشان تبدیل می‌گردند[۷].

پس از آماده‌سازی اولیه متون که در آن دادگان در طبقه‌های مختلف قرار می‌گیرند فاز پیش پردازش انجام می‌شود. در واقع پیش پردازش، اولین گام در جهت تطابق مستندات متنی با نمایش آنها در یک قالب مناسب می‌باشد.

۲-۱۰- دسته‌بندی

دسته‌بندی و پیش‌بینی دو نوع عملیات برای تحلیل داده‌ها و استخراج مدل به منظور توصیف دسته‌های مهم داده‌ها، فهم و پیش‌بینی رفتار آینده آنها می‌باشد. مدل‌های دسته‌بندی در پیش‌بینی متغیرهای گسسته و طبقه‌ای بکار رفته و مدل‌های پیش‌بینی یا رگرسیون بیشتر بر روی داده‌های پیوسته بکارمی‌رود.

۲-۱۱- پیشینه ی تحقیق و کارهای مرتبط

سال ۲۰۱۳ عباسی و همکاران روشی به منظور تحلیل عقاید و همچنین انتخاب خصیصه‌ها ارائه داده‌اند.

در سال ۲۰۱۵ پیکیری و همکارانش به تحلیل احساسات در شبکه اجتماعی توییتر با تکنیک متن کاوی پرداخته‌اند.

جلالی و همکارانش در سال ۲۰۱۵ در مقاله ای مورد به تحلیل رفتار مشتریان با استفاده از کاوش کاربری وب پرداخته‌اند.

در سال ۲۰۱۶ رافع و همکارش به ارائه روشی برای آنالیز احساسات در متن نظرات پرداخته‌اند در دنیای امروز حجم عظیمی از اطلاعات به صورت متن می‌باشد.

رضایی و همکارانش در سال ۲۰۱۶ در مطالعه‌ای موردی به بررسی روشهای متن کاوی در تحلیل نظرات و مطلوبیت‌های مشتریان در شبکه‌های اجتماعی پرداخته‌اندآنها از متن‌کاوی به عنوان روشی برای تحلیل نظرات و مطلوبیت‌های مشتریان درشبکه‌های اجتماعی استفاده کرده‌اند.

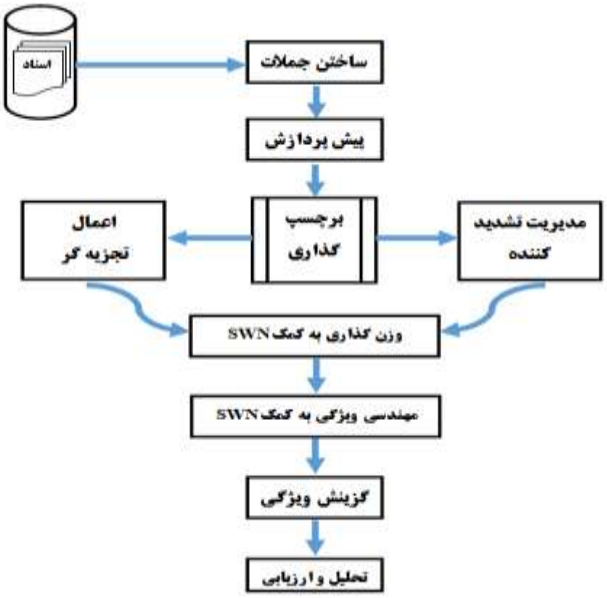
کارگر و همکارانش در سال ۲۰۱۶ به بررسی مدل‌های نظرکاوی و تجزیه و تحلیل احساسات کاربران در محیط وب پرداخته اند بدست آوردن نظرات مردم و مشتریان خود یک تجارت عظیم برای بازاریابی، روابط اجتماعی و حتی رقابت‌های انتخاباتی می باشد.

سهرابی و همکارانش در سال ۲۰۱۶ به تحلیل نظرات کاربران وب سایت‌های تجارت اجتماعی بر اساس روش‌های متن‌کاوی و داده‌کاوی پرداخته‌اند هدف از این پژوهش مطالعه و تحلیل نظرات کاربران وب-سایت‌های تجارت اجتماعی با بهره‌گیری ترکیبی از تکنیک های متن-کاوی و داده‌کاوی بوده است.

۳- روش پیشنهادی

در این مقاله از منبع واژگانی سنتی‌وردنت به منظور گروه‌بندی ویژگی‌ها و انتخاب ویژگی استفاده کرده‌ایم. جزئیات روش پیشنهادی در شکل ۳-۱ نمایش داده شده است. از سنتی‌وردنت برای تبدیل داده‌ها به بردار ویژگی‌ها و همچنین کاهش ابعاد ویژگی‌ها استفاده می‌شود. ضمن اینکه از الگوریتم انتخاب ویژگی مبتنی بر محتوا نیز استفاده می‌کنیم.

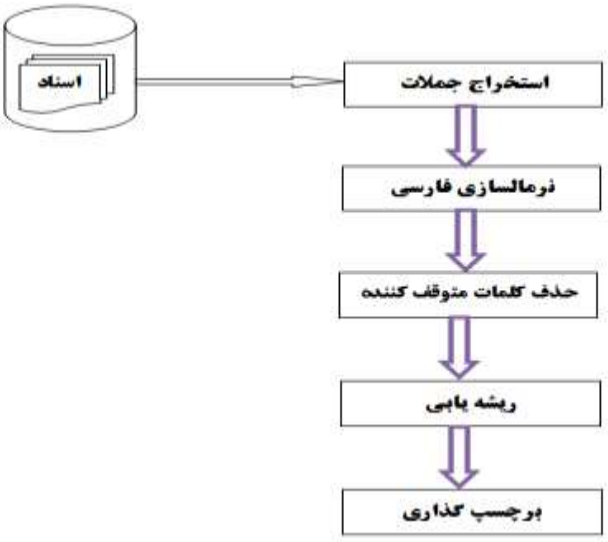
به دلیل اینکه مجموعه واژگانی سنتی وردنت مختص اسناد لاتین است در ادامه به بیان جزئیات روش پیشنهادی برای زبان فارسی خواهیم پرداخت و مشخص می‌کنیم که چه تغییراتی را خواهیم داد تا بتوان این مجموعه واژگانی را برای زبان فارسی هم بکار برد.



شکل ۳-۱ شماتیک کلی روش پیشنهادی

۳-۱- پیش پردازش

در روش پیشنهادی این پژوهش عملیات پیش پردازش در سه مرحله به انجام می‌رسد که مرحله اول استخراج جملات از اسناد و پایگاه داده مورد بررسی است و مرحله دوم در رابطه با نرمالسازی متون اسناد می‌باشد و در مرحله سوم به هر کدام از ویژگی‌ها برچسب گذاری می‌کنیم که این عملیات با POS انجام می‌شود در حقیقت با این برچسب گذاری نقش کلمه از نظر فاعل بودن یا فعل بودن یا مفعول بودن مشخص می‌شود البته در این بین عملیات دیگری هم جهت مدیریت بهتر ویژگی به انجام خواهد رسید. در ادامه به دلیل اینکه روش پیشنهادی این پژوهش برروی متون فارسی کار می‌کند فلوچارتی جهت عملیات پیش‌پردازش متون فارسی ارائه می‌شود همانگونه که در شکل ۳-۲ مشاهده می‌کنید در مرحله اول دیتابیس ما که شامل اسنادی است در قالب جملاتی استخراج می‌شود سپس نرمالسازی متون بر روی آنها انجام شده و عملیات برچسب‌گذاری را جهت استخراج مفیدترین ویژگی‌ها انجام می دهیم و در اخر بردارها را تشکیل می‌دهیم. در ادامه مراحل پیش پردازش را بیشتر توضیح می دهیم.



شکل ۳-۲ مراحل پیش‌پردازش اسناد فارسی

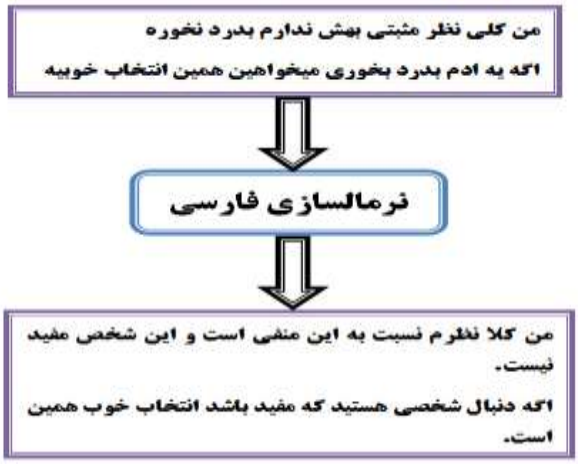
در شکل ۳-۲ پیش‌پردازش‌های که بر روی متون انجام خواهد شد مشخص شده‌اند. در اینجا ورودی عملیات پیش‌پردازش مجموعه‌ای از متون است که اعم ورودی بصورت خام و دارای قالب بدون ساختار هستند که در ادامه شرح اینکه چه اتفاقی برای ساختار می‌افتد بیشتر صحبت می‌کنیم.

همانگونه که در شکل ۳-۳ مشاهده می‌کنید در این مرحله با استفاده از تمیزسازی داده‌ها ، متن را حالت محاوره نرمال می‌کنیم که این امر باعث می‌شود پراکندگی ویژگی‌ها قابل قبولتر شود و همچنین مشکل فاصله و نیم فاصله که یکی از چالش‌های اساسی زبان فارسی در پردازش آن است برطرف شود با نرمالسازی متون باعث می‌شویم که هردو چالش در مکان درست خود قرار گیرند.

۳-۱-۲- نرمال سازی متن

این مرحله قبل از هر مرحله‌ای باید به انجام برسد اسناد دارای جملاتی هستند که هر جمله خود از یکسری ویژگی یا کلماتی تشکیل شده که هر یک از نظر معنایی دارای مفهوم و معنای خاصی هستند بنابراین در روش پیشنهادی جاری قبل از هر نوع عملیات پیش پردازشی به نرمالسازی جملات متون می‌پردازیم تا بدین وسیله هم اینکه عملیات پیش‌پردازش را به روش بهتر انجام دهیم و هم اینکه از حذف ویژگی‌های مهم خوداری کنیم و دیگری اینکه کلمات اضافه را به شکل بهتری جهت بالا بردن سرعت الگوریتم روش پیشنهادی حذف کنیم.

در شکل ۳-۳ نمونه‌ای از نرمالسازی که ما بدنبال آن هستیم را مشاهده می‌کنید در این سند از تعدادی از هواداران فوتبال خواسته شده است نظرات خود را درباره مربی جدید تیم بیان کنند که بعد از بیان نظرات و جمع‌آوری آن در قالبی سندی ورزشی مشاهده می‌کنیم که اکثر نظرات بصورت محاوره‌ای بیان شده‌اند و نیاز به نرمالسازی متن کاملا مشهود است.



شکل ۳-۳ نرمال سازی متن فارسی

۳-۲-۲- حذف کلمات متوقف کننده

یکی از مراحل پیش پردازش حذف کلمات متوقف کننده است. کلمات متوقف کننده به کلماتی گفته می‌شود که در همه اسناد موجود به صورت یکسان به کار خواهند رفت و این مهمترین دلیلی است که نشان می‌دهد اینگونه کلمات به فرایند گروه‌بندی هیچ کمی نخواهند کرد و فقط سرعت اجرای الگوریتم روش پیشنهادی را کند می‌کند. به عنوان مثال اگر به عنوان داده ، سندی داشته باشیم که حاوی صد تکرار از واژه "را" یا "که" یا "و" یا "در" و غیره باشد سوالی که اینجا پیش می‌آید این است که آیا می‌توان با این داده‌ها به دانشی دست پیدا کرد که مشخص کند متن مذکور چه منظوری را به مخاطب می‌رساند یا اینکه چه نوع متنی است؟ بدون شک جواب خیر است .

۳-۲-۳- ریشه یابی

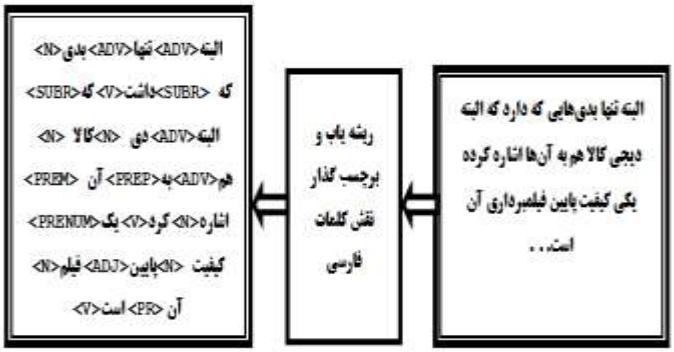
ریشه‌یابی از جمله روش‌های کاهش افزونگی کلمات موجود در اسناد است. برای روشن‌تر شدن مسئله مثالی را بیان می‌کنیم در هر سند احتمال برخورد با کلماتی مشابه ممکن است برخی کلمات مانند "حساب" و "حسابدار" و "حسابرس" و "حسابرسی" معانی بسیار نزدیکی با هم دارند ، به همین دلیل در روش پیشنهادی پیشنهاد ما از یک واژه "حساب" به جای سایر واژه‌ها استفاده خواهیم کرد. با اینکار تعداد ویژگی‌های سند کمتر می‌شود اگر بخواهیم همین مثال را بررسی کنیم تعداد ویژگی‌های مرتبط با حسابدار از ۴ ویژگی به یک ویژگی کاهش پیدا کرد. عملیات ریشه‌یابی در کل دقیقا همین عملکرد را دارد البته اینکه این مرحله در کدام قسمت عملیات پیش‌پردازش قرار بگیرید کاملا به سلیقه و ابتکار محقق وابسته است که در این پژوهش ما ریشه‌یابی را بر روی همه کلمات درون اسناد انجام خواهیم داد و اینجاست که با ریشه‌یابی پراکندگی ویژگی‌ها کمتر شده و تا حدی ابعاد بردار ویژگی هم کاهش داده می‌شود اما بدون شک همچنان مشکل فزونی ویژگی‌ها پابرجاست و نیاز است برای به حداقل رساندن این فزونی عملیات دیگری هم انجام شود.

جهت انجام عملیات ریشه‌یابی الگوریتم‌های مختلفی وجود دارد که در روش پیشنهادی این پایان‌نامه، ما از نرم‌افزار نفشه کلمات دانشگاه فردوسی مشهد بهره گرفته‌ایم.

۳-۳ برچسب‌گذاری کلمات

برچسب نقش کلمات نیز باعث می‌شود کلماتی که قادرند نقش‌های مختلفی در جمله داشته باشند را از یکدیگر تمیز دهیم و به هر یک به عنوان یک ویژگی نگاه کنیم.

به همین دلیل یکی از مهمترین قسمت‌های روش پیشنهادی این پژوهش روش برچسب‌گذاری ویژگی‌ها است. برای ریشه‌یابی نرم‌افزار برچسب‌گذاری نقش کلمات فارسی دانشگاه فردوسی مشهد استفاده شده است همچنین از این نرم‌افزار جهت و برچسب‌گذاری نقش کلمات استفاده خواهد شد این نرم‌افزار توسط تیم پردازش زبان طبیعی دانشگاه فردوسی مشهد طراحی و پیاده‌سازی شد در این نرم‌افزار ابتدا ریشه هر کلمه مشخص خواهد شد و بعد از این به هر کلمه نقش سخن آن اضافه می‌شود . برای روش‌تر شدن مبحث مثالی می‌زنیم جمله" البته تنها بدی‌هایی که دارد که البته دیجی کالا هم به آن‌ها اشاره کرده یکی کیفیت پایین فیلمبرداری آن است" ریشه‌یاب و برچسب گذار نقش کلمات فارسی دانشگاه فردوسی جمله مورد نظر را بعد اعمال ریشه‌یابی و برچسب‌گذاری می کندکه این عملیات و روش آن در شکل ۳-۵ به نمایش گذاشته شده است.



شکل۳-۵ برچسب گذاری نقش کلمات فارسی دانشگاه فردوسی

باید توجه شود که برخی از کلمات با نقشه‌های متفاوتی در جمله حاضر می‌شوند و همین نقش‌های متفاوت است که معانی مختلف را به بار خواهند آورد. این دوگانگی و متفاوت بودن می‌تواند باعث کاهش دقت گروه‌بندی ویژگی‌ها یا هر نوع عملیاتی که قرار است انجام شود، شود.به همین دلیل است که از نقش برچسب استفاده می‌کنیم تا این ابهام را بوسیله آن رفع کنیم.

۳-۴- تشدید کننده

تشدید کننده در نوع وجود دارد که می‌تواند تقویت کننده یا تضعیف کننده باشد به طور کلی تشدید کننده کلمه‌ای یا عبارتی است که ارزش جمله را تغییر می‌دهد. برای مثال کلمه "بسیار" نقش تقویت کننده در جمله را بازی می‌کند و واضح است که این کلمه تشدید

کننده هر چند خود به تنها فاقد معنا و مفهوم خاصی است اما برای جمله مهم است.

در اینجا نقش تجزیه کننده وابستگی به وضوح مشخص است و ما نیاز داریم تا با استفاده یک تجزیه‌گر وابستگی سایر ویژگی‌های که به ویژگی تشدید کننده وابسته هستند را استخراج کنیم تا بدین وسیله نوع اثرگذاری و میزان آن را مشخص کنیم .

۳-۵- تجزیه‌گر معنایی

هدف تحلیل تجزیه‌گر معنایی این است که ابتدا متن را به عبارات تجزیه کند سپس آن متون را به مفاهیمی که می‌توان آنها را در یک ساختار برداری قرار داد، تجزیه می‌کند. بنابراین تجزیه‌گر معنایی، متن را به عبارات تجزیه می‌کند و بعد از آن فعل‌ها و عبارات اسمی و قیدی جمله را از متن استخراج خواهند شد. برای روشن‌تر شدن مسئله بیایید همان جمله نرمال شده قسمت نرمالسازی در مراحل پیش را در نظر بگیریم.

<<اگه یه ادم بدرد بخوری میخواهین همین انتخاب خوبیه>> این جمله حاوی مفاهیم <<اگر مربی‌ای می‌خواهید انتخاب خوب همین است>> می‌باشد که به دو جمله جداگانه <<اگر مربی‌ای می-خواهید>> <<انتخاب خوب همین است>> تجزیه می‌شود.

الگوریتم تجزیه کننده جهت عملیات تجزیه بصورت بالا به پایین عمل می‌کند ولی در الگوریتم پیشنهادی این پژوهش برعکس عمل می‌کند و عناصر که در عمق درخت تجزیه قرار دارند و والدین خود را آنجا تشدید می‌کنند که در عنوان بعدی بیشتر درباره تشدید کننده صحبت می‌کنم.

۳-۶- وزن گذاری کلمات فارسی

یکی از مهمترین عملیات انجام شده در روش پیشنهادی این پژوهش وزن گذاری کلمات است و نیاز است بر اساس وابستگی‌های استخراج شده وزن هر یک از کلمات مشخص و نحوه کارکرد آنها مشخص شود

در روش پیشنهادی این پژوهش اگر بخواهیم وزن کلمات تقویت کننده را مشخص کنیم از رابطه ۳-۱ استفاده خواهیم کرد و اگر بخواهیم وزن کلمات تضعیف کننده را بدست آوریم رابطه ۳-۲ را مورد استفاده قرار خواهیم داد.

$$SS_{w_{i+1}} = SS_{w_{i+1}} + SS_{w_{+1}} * |SS_{w_i}| \qquad ۳-۱$$

$$SS_{w_{i+1}} = SS_{w_{+1}} * |SS_{w_i}| \qquad ۳-۲$$

همانگونه که قبلا ذکر شد مشکل اساسی زبان فارسی این است که منبع واژگانی واحدی برای تعیین وزن گذاری و ارزش کلمات فارسی وجود ندارد و از آنجایی که در این پژوهش قصد داریم از سنتی وردنت استفاده کنیم و از طرفی این مجموعه واژگان فقط برای زبان لاتین کاربرد دارد، باید راهی پیدا کنیم که بتوان وزن کلمات فارسی را جهت عملیات خود بدست آوریم در این پژوهش قصد داریم از یک الگوریتم بین زبانی جهت تعیین وزن کلمات فارسی با کمک سنتی وردنت استفاده کنیم.حال باید هر ویژگی یا حتی کلمه وزن مختص خود را داشته باشد تا با استفاده از وزن مذکور میزان اثرات آن را مشخص کنیم که برای انجام اینکار از الگوریتم شماره ۱ استفاده خواهیم کرد

Algorithm 1: Calculating Persian Feature weight
Input: dataset
Output: Polarities
For each Sen in dataset
    String [] Feature=Tokenize (Sen)
    For each Feature in Features
        TmpFeature=Find Most Similar word to Feature in Persian wordnet
    Polarity=Caculate_Polarity (tmpFeature)
    Polarities [Feature] =polarity
End for
End for
Return Polarities
End

در این پژوهش ویژگی‌ها از نوع Bigram است به همین دلیل باید ورودی الگوریتم شماره دو هم ویژگی‌های Bigram باشد سایر تحقیقاتی که پیش از این انجام شده است، همه ویژگی‌های دوتایی را از متن استخراج می‌کنند بنابراین در اولین گام در این الگوریتم اول باید به دنبال اثبات این باشیم که آیا ویژگی از نوع Bigram است یا خیر، بعد از مشخص شدن این اصل به سراغ مرحله بعد می‌رویم در این مرحله طبق رابطه عملکرد ویژگی را مشخص می‌کنیم اگر ویژگی۱ (F1) یک ویژگی انتشار دهنده باشد از دو حالت خارج نیست یا تقویت کننده است یا تضعیف کننده که اگر تقویت کننده باشد وزن ویژگی ۲ (F2) را افزایش می‌دهد و در صورتیکه ویژگی۱ یک تشدید کننده تضعیفی باشد باعث گاهش وزن ویژگی۲ خواهد شد.

Algorithm 2:
Input Bigram feature F1 F2
If F1 is modifier
    If F1 is amplifier
        SSF2= $SS_{F2} + SS_{F2} * |SS_{F1}|$
    Else if F1 is attenuator
        SSF2= $SS_{F2} * |SS_{F1}|$
End

۳-۷- مهندسی ویژگی‌ها

مهندسی ویژگی‌ها از نظر محقق مبحث مهمی است که در این پژوهش هم به آن پرداخته شده است و با راهکاری های مختلف مانند در نظر گرفتن تشدید کننده ها و استفاده از تجزیه گر سعی بر مهندسی خصیصه ها بود ولی این مجموعه عملیات کافی نیست در این پژوهش مهندسی ویژگی‌ها در دو فاز اصلی انجام می‌شود: فاز اول مجموعه

ویژگی‌ها را با استفاده از سنتی‌وردنت گروه‌بندی می‌کنیم و فاز دوم با الگوریتم انتخاب ویژگی بیان شده در الگوریتم شماره ۴ آن‌ها را کاهش می‌دهیم.

چیزی که مشخص است بسیاری از ویژگی های که در اسناد و پایگاه داده ها وجود دارد برای جهت گروه بندی مهم و کاربردی است در سنتی وردنت وجود ندارد بسیاری از نام های مهم دنیا ویژگی های مهمی هستند که در وردنت وجود ندارد. نام‌ها یا کلماتی که کاربران از ساختار نحوی مختلف مثلا مخفف یا نوشتار عامیانه کلمات استفاده می‌کنند که در سنتی‌وردنت وجود ندارند نباید بدون تحلیل از متن حذف شوند زیرا بعضی مواقع همین ویژگی ها هستند که دقت طبقه بندی را بالا می برد و حذف آنها باعث می شود که دقت عملیات هدف پایین بیایید کاری که در اعم تحقیقات پیشین در نظر گرفته نشده است اما در این پژوهش در نظر گرفته شده است جهت ارضا شدن این هدف دو شرط را در نظرگرفته ایم که جلوتر در باره آن صحبت می کنیم.

ورودی الگوریتم شماره ۳ همه کلمه_ برچسب‌های درون سنتی‌وردنت است و خروجی این الگوریتم نیز بردار گروه‌بندی شده از کلمه_ برچسب‌های است که به عنوان ورودی دریافت شده بودند اما به گونه-ای که هر گروه با شماره شناسایی یکتا مشترک شناخته شوند. در خط ۳ و خط ۴ الگوریتم حلقه‌های تودرتو برای مقایسه همه کلمه_ برچسب درون سنتی‌وردنت به کار گرفته شده‌اند و چیزی که مشخص است اگر دو کلمه_ برچسب شباهت معنایی و آماری داشته باشند، در یک گروه معنایی قرار داده می‌شوند. خط ۵ الگوریتم جهت شناسایی کلمات درون وردنت بکار گرفته شده است این خط بررسی می‌کند که کلمه_ برچسب۱ و کلمه_ برچسب۲ متعلق به یک شماره شناسایی یکتا باشد و این شرطی است که شباهت معنایی آن را معیار اصلی برای گروه‌بندی کلمه_ برچسب‌های مختلف در یک گروه در نظر می-گیرد الیته شباهت معنایی شرط ثانویه هم دارد که اگر شرط اولیه تامین شود باید شرط ثانویه بررسی شود به این صورت که اختلاف میانگین قطبیت کلمه_ برچسب۱ و کلمه_ برچسب۲ از آستانه X1 کم‌تر باشد و تفاوت انحراف استاندارد کلمه_ برچسب۱ و کلمه_ برچسب۲ نیز از آستانه X2 کم‌تر باشد. دو آستانه X1 و X2 استفاده شده به صورت تجربی به ترتیب ۰/۰۱ و ۰/۰۲ تنظیم شده است و اگر دو شرط اولیه و ثانویه برقرار باشند کلمه_ برچسب۱ و کلمه_ برچسب۲ به یک گروه تعلق دارند و یکی از آنها با یک SynID به عنوان کاندید انتخاب خواهد شد این عملیات تا اخرین ویژگی به انجام می رسد.

Algorithm 3: Feature Grouping

```
1.      Input D1 Dictionary of w#POS extracted from SWN
2.      Output D2 reduced Dictionary of w#POS
3.      Foreach w1#pos1 in D1
4.        Foreach w2#pos2 in D1
5.          If (SynID(w#POS) == SynID(w#POS))
6.            If( |μ(1#POS1)-μ(w2#POS2) |<t1 && |σ(w1#POS1)-σ(w2#POS2) |<t2)
7.              D2[SynID(w1)].Add(w2#POS2)
8.            EndIf
9.          EndIf
10.       EndForeach
11.     EndForeach
```

۸-۳- انتخاب ویژگی

در این مرحله همه اسناد مورد بررسی به فرمت تبدیل شده‌اند که توسط ماشین می‌توانند پردازش شوند در مراحل پیش‌پردازش بسیاری از ویژگی‌ها با عملیات‌های متفاوت حذف شده‌اند اما همچنان ویژگی-های در بردار ما وجود دارد که افزونه هستند و باعث پایین رفتن راندمان کار الگوریتم روش پیشنهادی خواهد شد .

این افزونگی همیشگی به دلیل ذاتی بودن اسناد بر می‌گردد و افزونگی در اسناد همیشه وجود دارد اما می‌توان با استفاده از الگوریتم‌های کاربردی این حجم را به حداقل رساند برای اینکار از الگوریتم ۴ استفاده خواهیم کرد.

این الگوریتم بردارها تشکیل شده در مراحل قبل را دریافت خواهد کرد و همه ویژگی‌های بردار بررسی خواهند شد اگر ویژگی جدیدی وجود داشت به فایل اضافه خواهد شد و اگر وجود نداشت کل بردار کنار گذاشته خواهد شد

Algorithm 4 Documents_To_model

```
Input: D the set of document, FeatureVector the set of Feature
Output: OutputFile model file
Foreach (di IN D)
  Foreach (token tj IN di. Sens)
    If (FeatureVector.Contain (tj))
        OutputFile.Add (tj)
```

۴- ارزیابی روش های پیشنهادی

۴-۱- معیار ارزیابی

اگر بخواهیم الگوریتم‌های مربوط به متن‌کاوی را تحلیل و ارزیابی کنیم با معیارهای مختلفی روبه‌رو می‌شویم که هر کدام کاربرد مختص خود را دارند و با استفاده از آنها می‌توان کیفیت الگوریتم را مورد بررسی قرار داد. از جمله معیارهای مهم که در اکثر تحقیقات از آنها استفاده می‌شود Recall، Precision، Accuracy و f-measure است. [21].

۴-۲- منابع مورد نیاز

ما در این پژوهش تمرکز خود را بر روی مجموعه داده‌ای می‌گذاریم که در حوزه متن‌کاوی قابل استفاده باشد و یکی از دلایل استفاده از این مجموعه داده این است که در تحقیقات پیشین نزدیک به موضوع پژوهش مورد استفاده قرار گرفته‌اند.

در اکثر تحقیقات به زبان فارسی در حیطه متن‌کاوی از همین مجموعه داده استفاده می‌کنند مجموعه داده همشهری مربوط به متن اخبار و گروه اخبارهای منتشر شده در روزنامه همشهری مربوط به سال های ۱۳۸۷-۱۳۷۵ می‌باشد.

جدول ۴-۲ ویژگی‌های آماری مجموعه داده همشهری

کلاس	تعداد اسناد
اجتماعی	۲۵۲
اقتصادی	۱۷۶
سیاسی	۲۲۲
علمی	۷
فرهنگی	۱۱۴
طبیعت	۸۴
ورزشی	۱۰۵

جهت انجام عملیات شبیه‌سازی و پیاده‌سازی روش پیشنهادی خود نیاز به ابزار دیگری هم وجود دارد یکی از این ابزار استفاده از نرم‌افزار برچسب زنی نقش کلمات دانشگاه فردوسی مشهد است که جهت عملیات ریشه‌یابی و برچسب‌گذاری در این پژوهش مورد استفاده قرار می‌گیرد.

هر مدلی نیاز به پیاده‌سازی خواهد داشت مدل پیشنهادی در این پژوهش با استفاده از نرم افزار ویژال استادیو ۲۰۱۷ پیاده‌سازی می‌شود

۴-۳ آزمایش و نتیجه‌گیری

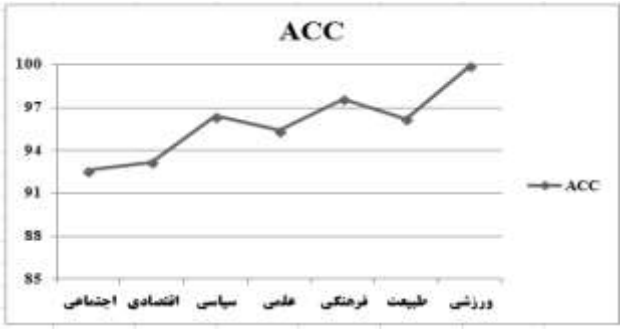
جهت آزمایش کیفیت روش پیشنهادی ما از معیارهای دقت(ACC) و معیار F1 و صحت(PER) بازخوانی (REC) استفاده خواهیم کرد .

جدول ۴-۴ نتایج کلی روش پیشنهادی بر روی هر چهار معیار

برچسب سند	ACC	F1	PER	REC
اجتماعی	92.6	91.1	91.2	91
اقتصادی	93.2	90.978	89.6	92.4
سیاسی	96.4	94.798	94.4	95.2
علمی	95.4	94.609	91.1	98.4
فرهنگی	97.6	97.239	98.3	96.2
طبیعت	96.2	95.747	95.2	96.3
ورزشی	99.9	99.55	99.6	99.5

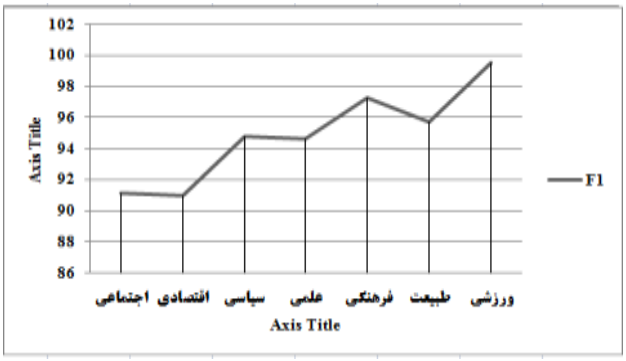
همانگونه که در جدول ۴-۴ مشاهده می‌کنید بعد از انجام عملیات شبیه‌سازی اسناد ورزشی در هر چهار معیار دارای نسبت به طبقه‌های دیگر دارای برتر محسوسی هستند

شکل ۴-۱ نمودار معیار دقت را برای اسناد نشان می‌دهد همانگونه که در شکل مشخص است در این معیار اسناد ورزشی با ۹۹,۹ دارای بالاترین دقت هستند و اسناد اجتماعی با ۹۲,۶ دارای کمترین دقت هستند.



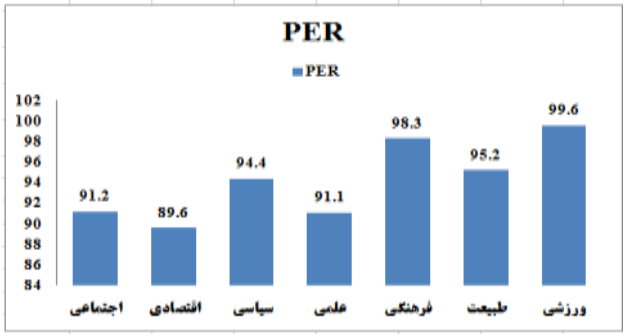
شکل ۴-۱ نتایج معیار دقت روش پیشنهادی بر روی هر کدام از کلاس‌ها

شکل ۴-۲ معیار F1 را نشان می‌دهد در این معیار هم مانند دقت کلاس ورزشی با با مقدار ۹۹,۵۵ دارای بالاترین رتبه است و کلاس اقتصادی هم با مقدار ۹۰,۹۷۸ دارای پایین‌ترین رتبه است.



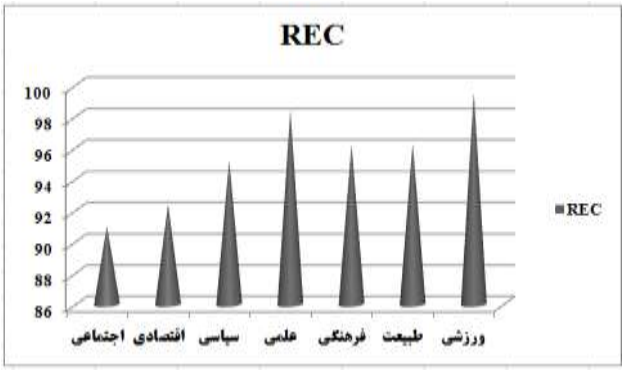
شکل ۴-۲ نتایج معیار F1 روش پیشنهادی بر روی هر کدام از کلاس‌ها

در معیار صحت همانگونه که شکل ۴-۳ نشان می‌دهد بهترین صحت دسته‌بندی باز هم در کلاس ورزشی با مقدار ۹۹,۶ اتفاق افتاده است و بدترین آن به کلاس اقتصادی با مقدار ۹۱,۱ اختصاص دارد.



شکل ۴-۳ نتایج معیار صحت روش پیشنهادی بر روی هر کدام از کلاس ها

اخرین معیار مورد بررسی معیار بـازخوانی اسـت بـا انجـام محاسـبات مربوط به این معیار و همچنین با توجه بـه جـدول ۴-۴ و شـکل ۴-۴ مشخص است که در این معیار بـاز ذهـم بـالاترین نـرخ بـرای کـلاس ورزشی می‌باشد و پایین‌ترین نرخ هم بـه کـلاس اجتمـاعی اختصـاص دارد.



شکل ۴-۴ نتایج معیار بازخوانی روش پیشنهادی بر روی هر کدام از کلاس ها

۴-۴- مقایسه روش پیشنهادی با کارهای پیشین

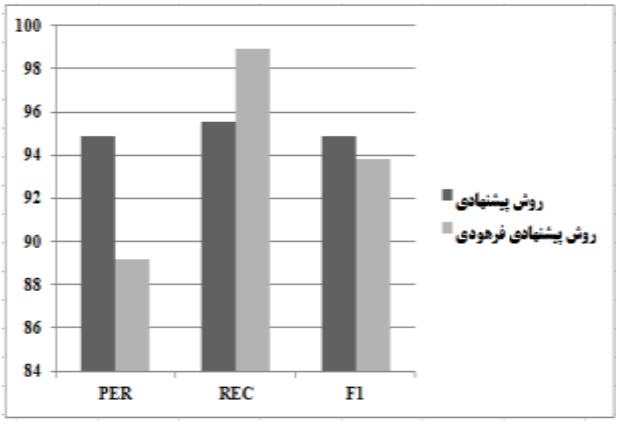
در این قسمت روش پیشنهادی این پژوهش را با روش پیشنهادی که توسط فرهودی و همکارانش در سال ۲۰۱۰ ارائه شد به مقایسه می-گذاریم در سال ۲۰۱۰ فرهودی و همکارانش در مقاله‌ای با استفاده از الگوریتم‌های یادگیری ماشین عملیات خودکار طبقه‌بندی متن فارسی را انجام داده‌اند در این مقاله از روش‌های یادگیری ماشین برای طبقه-بندی خودکار اخبار فارسی استفاده شده است. در همین راستا ، ابتدا سعی کرده‌اند مقدمه زبان و پیش‌پردازش‌ها را بر مجموعه داده‌های همشهری اعمال کنند سپس با استفاده از الگوریتم‌های وزن‌گذاری ویژگی و انتخاب ویژگی‌های سودمند ، یک بردار ویژگی را برای هر سند خبری استخراج کرده‌اند. بعد از ان دو الگوریتم ماشین بردار پشتیبان و k نزدیکترین را آموزش داده‌اند و در آخر نتایج روش پیشنهادی خود را با این دو الگوریتم سنجیده‌اند.

روش پیشنهادی توسط فرهودی و همکارانش از مدل‌های بسیار مختلف برای مدلسازی و تحلیل متن استفاده شده است که همین امر پیچیدگی زمانی روش پیشنهادی آنها را بالا برده است، مثلا برای طبقه بندی متن از الگوریتم SVM با توابع هسته‌ای خطی، چندجمله‌ای rbf درجه‌دو و Mlp استفاده کرده‌اند. اگرچه در روش پیشنهادی آنها هر دو الگوریتم برای طبقه‌بندی متن فارسی نتایج قابل قبولی را نشان می‌دهد اما عملکرد KNN در مقایسه با SVM بهتر است.

جدول و شکل ۴-۵ نتایج هر دو روش را با هم مقایسه می‌کند.

جدول ۴-۵ مقایسه روش پیشنهادی با روش ارائه شده توسط فرهودی

معیار	روش پیشنهادی	روش پیشنهادی فرهودی
PER	94.86	89.18
REC	95.57	98.95
F1	94.86	93.81



شکل ۴-۵ مقایسه روش پیشنهادی با روش ارائه شده توسط فرهودی

همانطور که شکل و جدول ۴-۵ نشان می‌دهند روش پیشنهادی این پژوهش دارای کیفیت بالاتری نسبت به روش پیشنهادی فرهودی و همکارانش است و دلیل آن به این بر می‌گردد که در این روش به مهندسی ویژگی‌ها اهمیت بالاتری داده شد و همین دلیلی شد که روش پیشنهادی در این پژوهش بتواند تعادل بهتری در نتایج داشته باشد و دلیل این امر نیز انسجام روش پیشنهادی بود در روش فرهودی و همکارانش به مهندسی ویژگی‌ها توجه شده بود اما تمرکز اصلی آنها در این روش بر روی ترکیب روش‌های متعدد برای رسیدن به بهترین جواب بود همچنین روش پیشنهادی این پژوهش از لحاظ زمانی و مکانی بسیار بهینه‌تر از روش فرهودی و همکارانش است . قابل توجه است که نسبت به روش فرهودی و همکارانش معیار صحت ۵ درصد و معیار F1 یک درصد، توسط روش پیشنهادی بهینه‌تر شده است و نرخ اشتباهات به میزان ۶درصد نسبت به روش‌های پیشین کاهش داشته است.

۵- نتیجه گیری

جهت صحت کیفیت هر مدلی که طراح می‌شود نیاز به آزمایش و تحلیل نتایج اجرای آن مدل است همانگونه که در فصول قبل ثابت کردیم نتایج روش پیشنهادی نشان از برتری روش پیشنهادی نسبت به روش‌های پیشین می‌دهد.

در روش پیشنهادی این پژوهش از مجموعه داده اسـتاندارد همشـهری جهت ارزیابی بهره گرفته‌ایم و از چهـار معیـار دقت،صـحت،بازخوانی و معیار F1 جهت ارزیابی روش پیشنهادی استفاده کـرده‌ایـم .اشـکال و نمودار نتایج نشان دادند که روش پیشنهادی این پژوهش دارای کیفیت

بالاتری نسبت به روش پیشنهادی فرهودی و همکارانش بود و دلیل آن هم این بود که مهندسی ویژگی‌ها در روش پیشنهادی این پژوهش دقیق تر انجام شده بود، البته در روش فرهودی و همکارانش هم به مهندسی ویژگی‌ها توجه شده بود اما تمرکز اصلی آنها استفاده از ترکیب‌های متفاوت ویژگی بود که همین امر باعث شده بود که پیچیدگی روش پیشنهادی آن بالا برود .
فرهودی و همکارانش در همه معیارها بهتر است نتایج ارزیابی نشان می‌دهد که معیار صحت ۵ درصد و معیار F1 یک درصد، توسط روش پیشنهادی بهینه تر شده است و نرخ اشتباهات به میزان ۶درصد نسبت به روش‌های پیشین نتایج ارزیابی نشان می‌دهد که روش پیشنهادی این امه نسبت به روش فرهودی کاهش داشته است.
از جمله مزایای روش پیشنهای افزایش دقت عملکرد الگوریتم دسته-بندی کننده است و از جمله مشکلاتی هم که روش پیشنهادی ممکن است با آن در آزمایش‌های دیگر مواجه شود عملکرد تجزیه کننده وابستگی است.

منابع

[۱] ویدا شقاقی، ویدا. مبانی صرف، سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها(سمت)، مرکز تحقیق و توسعهِ علوم انسانی، تهران. ۱۳۸۹

[۲] خیامیم مینا و باقری ایوب "ارائه مدلی جدید مبتنی بر داده کاوی برای تجزیه و تحلیل احساسات در سطح عبارت "موسسه آموزش عالی علوم و فن آوری سپاهان اصفهان،سومین کنفرانس بین المللی شهریورماه ۱۳۹۶ فناوری اطلاعات، مهندسی کامپیوتر و مخابرات

[۳] تحلیلگرها : تحلیل آماری انواع داده ها آموزش تحلیل آماری با استفاده از نرم افزار spss و سایر به ادرس http://tahlilgarha.blogfa.com ۱۳۹۶

[۴] کرانی سمیرا "ابزار wordnet درsentiment analysis "مرجع متخصصین علوم داده ایران دانشگاه پیام نور تهران ۱۳۹۷

[۵] برون, غزاله؛ فرهاد راد و حمید پروین، ۱۳۹۵، ارزیابی و مقایسه مناسب ترین الگوریتم های متن کاوی :مطالعه موردی مجموعه داده روزنامه همشهری،چهارمین کنفرانس بین المللی علوم و مهندسی، ایتالیا-رم، موسسه مدیران ایده پرداز پایتخت ویرا

[۶] برومندزاده مصطفی، چراغی فر سعید "بررسی روش های متن کاوی" مشهد: مینوفر، ۹۱ ص ۹۷۸-۶۰۰-۸۰۰-۷۹-۹ سال ۱۳۹۵

[۷] عبدی قویدل هادی، وزیرنژاد بهرام و بحرانی محمد "برچس بزنی موضوعی متون فارسی"

[۸] غضنفری_م, علیزاده _س و تیمورپور_ب. (۱۳۸۷), داده کاوی و کشف دانش انتشارات دانشگاه علم و صنعت ایران.

[۹] تیمورپور بابک، نجفی حیدر، "داده‌کاوی با R به همراه متن‌کاوی و تحلیل شبکه‌های اجتماعی" تهران مرکز تحقیقات و توسعه سازمان اتکا، ۱۳۹۴ : ۵-۶۸-۷۶۲۴-۹۷۸-۶۰۰ مرکز تحقیقات و توسعه ۱۳۹۴

[۱۰] میشلین کمبر،ژان پی، ژیاوی هان مترجم دکتر مهدی اسماعیلی "داده کاوی مفاهیم و تکنیک ها" ویراست سوم-چاپ سوم، انتشارات نیاز دانش، کد محصول: ۹۷۸۶۰۰۶۴۸۱۸۷۶ ،شابک: ۹۷۸۶۰۰۶۴۸۱۸۷۶

[۱۱] پیکری, ناصر؛ سیدعلی اصغر یعقوبی و حمیدرضا طاهری، ۱۳۹۴، تحلیل احساسات در شبکه اجتماعی توییتر با تکنیک متن کاوی، اولین کنفرانس بین المللی وب پژوهی، تهران، دانشگاه علم و فرهنگ

[۱۲] جلالی, شهرزاد و ندا عبدالوند، ۱۳۹۴، تحلیل رفتار مشتریان با استفاده از کاوش کاربری وب (مطالعه موردی: خرده فروشی آنلاین(، دومین کنفرانس ملی تحقیقات بازاریابی، تهران، موسسه اطلاع رسانی نارکیش

[۱۳] وکلی گارماسه, عاطفه &وحید رافع, ۱۳۹۵, ارائه روشی برای آنالیز احساسات در متن نظرات ,نخستین کنفرانس ملی تحقیقات بین رشته ای در مهندسی کامپیوتر، برق، مکانیک و مکاترونیک ,قزوین, مرکز آموزش عالی فنی مهندسی بوئین زهرا, پارک علم و فناوری استان قزوین

[۱۴] رضایی, سمیرا؛ حسین دستخوان و محمدصالح اولیاء، ۱۳۹۵، روشهای متن کاوی در تحلیل نظرات و مطلوبیت های مشتریان در شبکه های اجتماعی: مطالعه موردی در بازار محصولات دیجیتال ایران، سیزدهمین کنفرانس بین المللی مهندسی صنایع، بابلسر، دانشگاه علوم و فنون مازندران

[۱۵] شاه طالبی, نجمه؛ محمدجواد کارگر و کمال میرزائی، ۱۳۹۵، بررسی مدل های نظرکاوی و تجزیه و تحلیل احساسات کاربران در محیط وب، دومین کنفرانس بین المللی وب پژوهی، تهران، دانشگاه علم و فرهنگ

[۱۶] سهرابی بابک، رئیسی وانانی ایمان و خداپرست فرشته "تحلیل نظرات کاربران وب سایت های تجارت اجتماعی بر اساس روش های متن کاوی و داده کاوی" مقاله ۴، دوره ۱۱، شماره ۲، پاییز و زمستان ۱۳۹۵، صفحه ۱۷۹-۱۶۳

[۱۷] بنی طالبی, افسانه و فرساد زمانی، ۱۳۹۵، نظرکاوی شبکه های اجتماعی با استفاده از الگوریتم های یادگیری ماشین، کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات، تهران، دبیرخانه دایمی کنفرانس

[۱۸] عبدیزدان, مرجان، ۱۳۹۷، ارایه روشی مبتنی بر تحلیل آماری منبع واژگانی وردنت و محتوا به منظور تحلیل عقاید در اسناد لاتین، دومین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات اطلاعات، بندرماهشهر، سازمان نظام صنفی رایانه ای استان خوزستان، دانشگاه آزاد اسلامی واحد ماهشهر.

[19] T. Pilehvar, H. Faili, M. Soltani, Classification of Persian textual documents using Learning Vector Quantization, 4rd IEEE Conference on Knowledge Engineering and Natural Language Processing, NLP-KE, 2009.

[20] Bijankhan, M. & J. Seikhzadeghan & M. Bahrani & M. Ghayoomi, "Lessons from Creation of a Persian Written Corpus: Peykare". Language Resources and Evaluation Journal.Vol. 45, No. 2.143-164, 2011.

[21] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.
[22] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features in Machine Learning", 10th European Conference on Machine Learning, pp. 137-142, 1998.
[23] M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," Expert Syst. Appl., vol. 40, no. 16, pp. 6266–6282, 2013.
[24] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.
[25] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 246–253.
[26] T. O'Keefe and I. Koprinska, "Feature selection and weighting methods in sentiment analysis," in Proceedings of the 14th Australasian document computing symposium, Sydney, 2009, pp. 67–74.
[27] C. Priyanka and D. Gupta, "Identifying the best feature combination for sentiment analysis of customer reviews," in Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, 2013, pp. 102–108.
[28] J. M. Chenlo and D. E. Losada, "An empirical study of sentence features for subjectivity and polarity classification," Inf. Sci. (Ny)., vol. 280, pp. 275–288, 2014.
[29] S. O. Orimaye, "Learning to classify subjective sentences from multiple domains using extended subjectivity lexicon and subjective predicates," in Asia Information Retrieval Symposium, 2013, pp. 191–202.
[30] I. Pavlopoulos, "Aspect based sentiment analysis," Athens Univ. Econ. Bus., 2014.
[31] A. Hogenboom, F. Frasincar, F. De Jong, and U. Kaymak, "Polarity classification using structure-based vector representations of text," Decis. Support Syst., vol. 74, pp. 46–56, 2015.
[32] K. Sarvabhotla, P. Pingali, and V. Varma, "Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents," Inf. Retr. Boston., vol. 14, no. 3, pp. 337–353, 2011.
[33] Pons-Porrata A, Berlanga-Llavori R, RuizShulcloper J. Topic discovery based on text mining techniques. Information Processing & Management 2007; 43(3): 752-68.
[34] Ramezani H, Alipour Hafezi M, Momeni E. Scientific maps: methods and techniques. Journal of the Popularization of Science 2014; 5(6): 53-84. Persian
[35] Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Research Synthesis Methods 2011; 2(1):1-14.
[36] Sumathi S, Mohanapriya S, Nagasandhiyalakshmi B, Shanmugapriya N. Prediction of outbreak of heart disease using text mining. Discovery 2016; 52(245): 1070-7.
[37] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence 2009; 1(1): 60-76.
[38] Piedra, D., & Ferrer, A. Text Mining and Medicine : Usefulness in Respiratory Diseases. Archivos de Bronconeumología, 2014 ;50(3), 113–119
[39] Ananiadou S, McNaught J. Text Mining for Biology and Biomedicine. Boston, London: Artech House; 2006.
[40] Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 2001;17 Suppl 1:S97-106
[41] Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988;31(4):526-57.
[42] M. Shams, A. Shakery, and H. Faili, "A non-parametric LDA-based induction method for sentiment analysis," in Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on, pp. 216–221, IEEE, 2012
[43] Mohammad Ehsan Basiri, Ahmad Reza Naghsh-Nilchi, Nasser Ghassem-Aghaee."A Framework for Sentiment Analysis in Persian." Open Transactions on Information Processing ,pp. 1-14.2014
[44] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies,pp. 1–167, 2012.
[45] A. Bagheri, M. Saraee, and F. de Jong, "Sentiment classification in Persian: Introducing a mutual information-based method for feature selection," in Electrical Engineering (ICEE), 2013 21st Iranian Conference on, pp. 1–6, IEEE, 2013.
[46] Liao, X., Cao, D., Tan, S., Liu, Y., Ding, G., and Cheng X.Combining Language Model with Sentiment Analysisfor Opinion Retrieval of Blog-Post. Online Proceedings of Text Retrieval Conference (TREC) 2006
[47] Mishne, G. Multiple Ranking Strategies for Opinion Retrieval in Blogs. Online Proceedings of TREC, 2006.
[48] Natural Language Processing Research Group, Department of Computer Science ,University of Sheffield,Regent Court ,211 Portobello, Sheffield, S1 4DP , UK ,+44 (0)114 222 1901 nlp-enquiries@shef.ac.uk,
[49] Mehrnoush bazrafkan. (2014). A review of persian language processing problems using computer systems. national conference on computer engineering and information technology management. [in Persian]
[50] Mehrnoush shamsfard. (2005). Challenges and Open Problems in Persian Text Processing
[51] Mehrnoush shamsfard. (2006). Persian text processing: past achievements, challenges ahead. The second workshop on farsi and computer. [in Persian]
[52] Publication of the university of science and technology. (2009). system analysis automatic find keywords farsi. [in Persian]
[53] A. Esuli and F. Sebastiani, "SentiWordNet : A High-Coverage Lexical Resource," pp. 1–26, 2006.
[54] A. Esuli and F. Sebastiani, "SENTIWORDNET: A high-coverage lexical resource for opinion mining," Evaluation, pp. 1–26, 2007.

[55] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.,” in LREC, 2010, vol. 10, pp. 2200–2204.
[56] Shams, M.; Shakery, A.; Faili, H. A non-parametric LDA-based induction method for sentiment analysis, in Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International. Symposium on, IEEE, 2012, pp. 216–221
[57]A. Hassan, A. Abbasi, and D. Zeng, “Twitter sentiment analysis: A bootstrap ensemble framework,” in Social Computing (SocialCom), 2013 International Conference on, 2013, pp. 357–364.
[58] O. Appel, F. Chiclana, J. Carter, and H. Fujita, “A hybrid approach to the sentiment analysis problem at the sentence level,” Knowledge-Based Syst., vol. 108, pp. 110–124, 2016.[59] Zhou, Xiaofei, Yue Hu, and Li Guo. "Text Categorization Based on Clustering Feature Selection." Procedia Computer Science 31 (2014): 398-405.
[60]Y. Ren, R. Wang, and D. Ji, “A topic-enhanced word embedding for Twitter sentiment classification,” Inf. Sci. (Ny)., vol. 369, pp. 188–198, 2016.
[61] Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Systems. 2018 Dec 1;161:124-33.
[62] Wu C, Wu F, Wu S, Yuan Z, Liu J, Huang Y. Semi-supervised dimensional sentiment analysis with variational autoencoder. Knowledge-Based Systems. 2019 Feb 1;165:30-9.
[63] Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H. Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications. 2019 Mar 1;117:139-47.