آنالیز حس اسناد فارسی با طراحی حوزه تبدیل بهینه

آصف پورمعصومی، هادی صدوقی یزدی، هادی قائمی و زهرا دلخسته

چکیده: با توسعه تعاملات مبتنی بر وب نظیر نظرسنجی‌ها، وبلاگ‌های شخصی و شبکه‌های اجتماعی، آنالیز حس و یا کاوش عقیده به یکی از حوزه‌های تحقیقاتی مهم در علوم کامپیوتر تبدیل شده است. تا کنون روش‌های زیادی مبتنی بر یادگیری ماشین و همچنین پردازش زبان طبیعی در ارتباط با آنالیز حس ارائه شده است. در این مقاله از توزیع کلمات در مجموعه اسناد جمع‌آوری شده به عنوان معیاری جدید برای تشخیص حس جمله استفاده شده است. در روش پیشنهادی با طراحی حوزه تبدیل بهینه مناسب روی توزیع کلمات، دو هدف حداکثرکردن انرژی طیفی کلاس ω_1 در فرکانس‌های پایین و حداکثرکردن انرژی طیفی کلاس ω_2 در فرکانس‌های بالا دنبال می‌شود. با طراحی حوزه تبدیل بهینه، داده‌ها از حوزه فراوانی به حوزه فوریه نگاشت می‌شوند. با این تبدیل بهینه، جداسازی الگوهای دوکلاسی از مفاهیم خوش‌بینی و بدبینی در حوزه تبدیل به راحتی امکان‌پذیر خواهد بود. برای محقق‌شدن مدل ریاضی، استراتژی استفاده از پروفایل نمونه‌ها روی همه نمونه‌های سیگنال نماینده کلاس ω_1 ارائه شده و مسأله حل می‌شود. طیف این پروفایل دارای مؤلفه‌های فرکانس پایین می‌باشد که با فرض تضاد طیفی دوکلاسی ω_1 و ω_2، حداکثرکردن انرژی طیفی کلاس ω_2 نیز ارضا می‌گردد. این روش به روی متون با زبان فارسی و انگلیسی اجرا شده است.

کلیدواژه: **آنالیز حس، حوزه تبدیل، حداکثرکردن انرژی طیفی.**

۱- مقدمه

چه شخصی در انتخابات ریاست جمهوری سال آینده برنده خواهد شد؟ چه سهم‌هایی در روز بعد در بورس با استقبال بیشتری روبه‌رو خواهد شد؟ کدام یک از کالاهای تولید کارخانه دارای مشکلات بیشتری می‌باشند؟ تمایل مردم بیشتر به خرید کدام محصولات است؟ این پرسش‌ها تنها بخشی از سؤالاتی است که با دانستن این نکته که "افراد جامعه در ارتباط با آن رویداد چگونه فکر می‌کنند" می‌توان پاسخ آنها را یافت. پیش از توسعه و فراگیرشدن وب، برای پیش‌بینی رویدادهایی از این دست از نظرسنجی‌های عمومی استفاده می‌شد که هزینه‌ها و مشکلات فراوانی را در پی داشت. نیاز به نیروی انسانی گسترده، صرف ساعت‌ها زمان و طولانی‌شدن فرآیند آمارگیری و عدم تمایل بخش زیادی از جامعه به حضور در فرآیند آمارگیری حضوری، بخشی از مشکلاتی است که در آمارگیری مستقیم از جامعه با آن مواجه می‌شویم. با گسترش و توسعه وب و فراگیرشدن آن و ظهور شبکه‌های اجتماعی و بلاگ‌ها و سایت‌های مختلف نظرسنجی و تولید حجم زیادی از داده‌های آماده پردازش، مبحث آنالیز حس[1] و کاوش عقیده[2] مورد توجه گسترده قرار گرفت به طوری که با استقبال فراوانی در سال‌های اخیر مواجه شده است. شکل‌گیری جامعه مجازی[3]، فرصت فوق‌العاده‌ای را فراهم آورد تا با صرف هزینه‌های اندک، بتوان اطلاعات بسیار ارزشمندی را از میان مجموعه عظیمی از داده‌هایی که کاربران با تمایل خود در وب توسعه می‌دهند، استخراج نمود. آنالیز حس، ابزاری را فراهم می‌آورد تا با بهره‌گیری از آن بتوان دیدگاه‌های موجود در یک متن همانند تشخیص نظرات مثبت یا منفی در مورد یک فیلم را استخراج نمود [۱].

آنالیز حس یکی از زمینه‌های مورد بحث در زمینه پردازش زبان طبیعی می‌باشد که می‌توان آن را به عنوان یک عملیات طبقه‌بندی در نظر گرفت [۲]. در بررسی مقالات مختلف، گاهی آنالیز حس با تعاریف مشابهی نظیر کاوش عقیده و آنالیز موضوعی[4] همراه می‌شود. آنالیز موضوعی به این مطلب اشاره دارد که آیا متن مورد آزمایش شامل عقیده و یا حس می‌باشد یا خیر [۳]. طبق تعریف [۴]، کاوش عقیده "فرآیند پردازش نتایج جستجو برای آیتم‌های داده‌شده است که منجر به تولید لیستی از ویژگی‌های محصول (کیفیت، قیمت و ...) و جمع‌آوری عقاید (خوب، بد و متوسط) در مورد هر کدام از آن ویژگی‌ها می‌شود". این تعریف با گسترش مقالات داده‌شده، آنالیزهای بیشتر را نیز در بر می‌گیرد [۵]. در مقالات بسیاری از آنالیز حس به عنوان کاربردهای مشخص از دسته‌بندی دیدگاه‌ها با توجه به قطبیت آنها (مثبت یا منفی) یاد کرده‌اند [۶]. طبق تعریف [۷] آنالیز حس، "فرآیند تشخیص عقاید، احساسات و ارزیابی‌های مثبت و یا منفی" می‌باشد. علی‌رغم تفاوت‌هایی که در تعاریف وجود دارد اما در بسیاری از مقالات، این سه تعریف معادل یکدیگر به کار می‌روند. تمرکز اصلی در این مقاله بر روی تشخیص مثبت یا منفی‌بودن حس محتواها با استفاده از مشخص‌کردن طیف آنها می‌باشد.

با بررسی آمار و نتایج مقالات مختلفی که تا کنون در زمینه آنالیز حس بر روی داده‌های متنی ارائه شده است می‌توان دریافت که ویژگی‌های متنی استخراج‌شده از متون به خوبی توانایی مدل‌کردن حس افراد را ندارند [۱]. از طرف دیگر روش‌های طبقه‌بندی نیز توانایی‌های محدودی برای مدل‌کردن دارند که نتایج آنها در مقالات مختلف آورده شده است [۱]. از این رو ارائه یک بستر ریاضی مناسب جهت استخراج ویژگی‌های مناسبی که توانایی طبقه‌بندی با کارایی بالا در آنالیز حس را داشته باشد، ضروری به نظر می‌رسد. در این مقاله به دنبال استخراج یک مدل ریاضی مناسب برای طبقه‌بندی مناسب داده‌های دوکلاسه هستیم که کاربردهای فراوانی از جمله در آنالیز حس می‌تواند داشته باشد.

روش پیشنهادی در این مقاله شامل ۳ مرحله اصلی می‌باشد (شکل ۱). در مرحله اول یا مرحله استخراج ویژگی‌های اولیه، ابتدا ویژگی‌های اولیه

این مقاله در تاریخ ۲۲ بهمن ماه ۱۳۹۱ دریافت و در تاریخ ۴ شهریور ماه ۱۳۹۴ بازنگری شد.

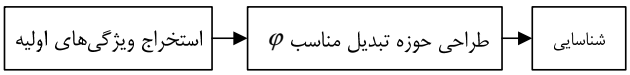
آصف پورمعصومی، دانشکده مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، (email: asef.pourmasoumi@stu-mail.um.ac.ir).

هادی صدوقی یزدی، دانشکده مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، (email: h-sadoghi@um.ac.ir).

هادی قائمی، دانشکده مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، (email: hadi.qaemi@stu-mail.um.ac.ir).

زهرا دلخسته، دانشکده مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، (email: z-delkhasteh@stu-mail.um.ac.ir).

1. Sentiment Analysis
2. Opinion Mining
3. Virtual Community
4. Subjectivity Analysis



شکل ۱: اجزای اصلی روش پیشنهادی.

داده‌ها با استفاده از روش‌های موجود استخراج می‌گردد. در این مقاله از تکنیک‌های معروف در پردازش زبان طبیعی برای استخراج ویژگی‌های اولیه داده‌ها استفاده شده است. در مرحله دوم یا مرحله طراحی حوزه تبدیل مناسب φ، با ارائه یک تابع تبدیل φ مناسب، داده‌ها به حوزه تبدیل بهینه نگاشت می‌یابند. در این فضای جدید، جداپذیری داده‌ها افزایش می‌یابد و خطای طبقه‌بندی به میزان قابل توجهی کاهش می‌یابد. ایده اصلی در این گام، استخراج یک الگوی کلی برای اسناد حس مثبت و اسناد حس منفی و سپس ارائه راهکار برای دسته‌بندی اسناد جدید با توجه به مدل استخراج‌شده می‌باشد. به این منظور با پروفایل‌گیری بر روی همه نمونه‌های کلاس ۱، طیف کلی مربوط به کلاس ۱ تولید می‌شود. سپس با مرتب‌سازی نزولی مؤلفه‌های سیگنال نماینده کلاس اول، انرژی طیفی کلاس اول در فرکانس‌های پایین حداکثر می‌شود. در نتیجه با فرض متمایزبودن دو کلاس مثبت و منفی، انرژی طیفی کلاس دوم در فرکانس‌های بالا حداکثر خواهد شد. این موضوع نهایتاً باعث می‌شود تا تمایز بین دو کلاس برجسته‌تر شده و نمونه‌های جدید به راحتی طبقه‌بندی شوند. در مرحله سوم یا شناسایی، با اعمال یک تبدیل موجک، کامپونت‌های اصلی تشکیل‌دهنده سیگنال هر سند، استخراج می‌شود. سپس عملیات یادگیری با اعمال ماشین بردار پشتیبان انجام می‌شود. نتایج به دست آمده حاکی از آن است که با به دست آمدن پارامترهای یادگیری می‌توان اسناد را با دقت بسیار مناسبی طبقه‌بندی نمود.

ساختار این مقاله به شرح زیر می‌باشد: در بخش ۲ مروری بر کارهای انجام‌شده در زمینه آنالیز حس شده است. در بخش ۳ روش پیشنهادی در این مقاله معرفی شده است. بخش ۴ به نتایج تجربی آزمایشات انجام‌شده پرداخته است. در بخش ۵ در مورد جنبه‌های مختلف روش پیشنهادی بحث شده است. در پایان هم نتیجه‌گیری در بخش ۶ آورده شده است.

۲- مروری بر کارهای انجام‌شده

تا کنون روش‌های بسیاری برای آنالیز حس ارائه شده است. روش‌های آنالیز حس از یک دیدگاه به دو دسته کلی تقسیم‌بندی می‌شوند [۱]. دسته‌ای از این روش‌ها مبتنی بر طبقه‌بندی داده‌ها با استفاده از برچسب‌های از پیش مشخص می‌باشد. روش‌های طبقه‌بندی گوناگونی استفاده شده است اما پرکاربردترین آنها در این حوزه، روش SVM[1] می‌باشد که در مقالات مختلف به کار رفته است [۸] تا [۱۰]. دسته دوم هم مبتنی بر لغت‌نامه می‌باشند [۱۱] و [۱۲]. در روش‌های مبتنی بر لغت‌نامه ممکن است دیکشنری به صورت دستی تولید شود [۱۳] و یا این که به صورت خودکار تولید گردد [۱۴]. تعدادی از روش‌ها نظیر [۱۵] تا [۱۷] هم از ترکیب تکنیک‌های مختلف استفاده کرده‌اند. در [۱] مرور کاملی بر روش‌های آنالیز حس و کاوش عقاید انجام شده و خواندن آن برای افرادی که تمایل دارند در این زمینه فعالیت کنند، توصیه می‌شود.

در [۱۸] از ویژگی‌های زبانی برای استخراج حس متون در سایت توییتر استفاده شده است. در این مقاله از ویژگی‌های گوناگونی نظیر ویژگی چندتایی[2]، ویژگی برچسب‌زنی اجزای واژگانی کلام[3]، فرهنگ واژگان و سایر ویژگی‌های سایت‌های میکروبلاگ استفاده شده است. در [۱۲] مرور کاملی بر روش‌های مبتنی بر فرهنگ واژگان شده است. در این مقاله روشی به نام SO-CAL ارائه شده که از یک دیکشنری شامل کلمات نشانه‌گذاری شده به همراه قطبیت و میزان آن استفاده می‌کند. در این روش در ابتدا یک فرهنگ لغت شامل کلمات با حس مثبت و کلمات با حس منفی جمع‌آوری می‌گردد. سپس متن اصلی پردازش شده و ویژگی‌های اصلی آن (کلمات مهم) استخراج می‌گردد. در نهایت با استفاده از هم‌رخدادی ویژگی‌های استخراج‌شده و کلمات موجود در فرهنگ واژگان، وزن نهایی متن با قطبیت مثبت یا منفی تعیین می‌گردد.

در [۱۹] از مفاهیم شبکه‌های اجتماعی برای تشخیص حس افراد استفاده شده است. ایده اصلی در این روش بر این اصل استوار است که افرادی که در شبکه‌های اجتماعی نظیر فیسبوک روابط و تعاملات نزدیکی با یکدیگر دارند، به احتمال زیاد دارای عقاید مشترک بسیاری می‌باشند. بنابراین با آنالیز رفتار دوستان یک فرد می‌توان پی به احساسات و عقاید آن فرد پیرامون مسایل مختلف برد. در این روش از داده‌های سایت اجتماعی پربیننده توییتر برای آزمایش روش پیشنهادی استفاده شده است.

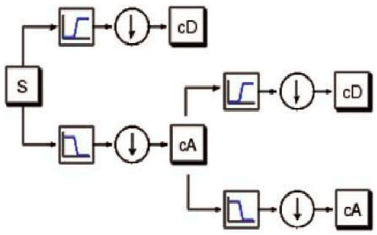
چندین روش شبه‌ناظر[4] نیز تا کنون ارائه شده است [۲۰] و [۲۱]. در [۲۱] از یادگیری شبه‌ناظر مبتنی بر گراف برای استنتاج قطبیت (مثبت یا منفی‌بودن حس) کلمات استفاده شده است. در این روش هر کدام از کلمات یا عبارات به عنوان یک گره یا رأس گراف در نظر گرفته می‌شود و یال‌های وزن‌دار بین گره‌ها هم به نوعی بیانگر شباهت بین آنها می‌باشند. در این مدل استنتاجی، قطبیت بعضی از گره‌ها به صورت دستی مشخص می‌گردد (نمونه‌های مورد استفاده در یادگیری شبه‌ناظر) و سپس حس سایر گره‌ها بر اساس قطبیت گره‌های شبه‌ناظر و همچنین ارتباطات با آنها استنتاج می‌شود. برای ارتباطات بین گره‌ها در این روش از فرهنگ واژگان استفاده شده است. ارتباطات مختلفی نظیر هم‌خانواده‌بودن، متضادبودن، شمول‌بودن و ... بین کلمات در شبکه واژگان تعریف شده است.

در آنالیز حس کار مشابهی در راستای تبدیل از حوزه فراوانی به حوزه فرکانس مشاهده نکردیم اما در حوزه بازیابی اطلاعات چندین مقاله در این زمینه منتشر شده است [۲۲] و [۲۳]. در [۲۳] برای اولین بار موقعیت کلمات در متن نیز به عنوان یک فاکتور جدید برای بازیابی اسناد مرتبط با یک موضوع مورد استفاده قرار گرفته است. در این روش، موقعیت هر کلمه در سند محاسبه می‌شود و سپس با تبدیل این مقادیر به حوزه فرکانس، همانند این است که موقعیت در واحد زمان محاسبه گردد. با این تکنیک، سیگنال موقعیت کلمات در طول سند محاسبه می‌شود و سپس طیف سیگنال با طیف کلمات مربوط به عبارت پرس و جو مقایسه شده و شبیه‌ترین سند انتخاب می‌گردد. در [۲۲] روش مشابهی برای بررسی شباهت بین اسناد ارائه شده است. در این روش شباهت بین اسناد با آنالیز توزیع کلمات دو سند توسط بسط سری فوریه محاسبه می‌گردد.

۲-۱ تبدیل فوریه گسسته

تبدیل فوریه در حوزه ریاضیات، یک مدل خاص از تبدیلات گسسته می‌باشد که در آنالیز فوریه استفاده می‌شود. این تبدیل در بسیاری از حوزه‌های مهندسی برق و کامپیوتر استفاده می‌شود [۲۴]. این تبدیل برای تغییر مبنا و پایه سیگنال به امواج سینوسی مستقل خطی استفاده می‌شود.

1. Support Vector Machine
2. N-Gram
3. Part of Speech Tagging
4. Semi-Supervised Approaches



شکل ۲: شمای کلی یک تبدیل موجک [۲۷] (s سیگنال ورودی، cD ضریب جزئیات و cA ضریب تخمین هستند).

ورودی تبدیل فوریه گسسته [1](DFT)، مجموعه‌ای از اعداد متناهی است که همین موضوع باعث می‌شود تا این تبدیل برای پردازش اطلاعات ذخیره‌شده در رایانه‌ها مناسب گردد. تبدیل فوریه گسسته کاربرد بسیار گسترده‌ای در حوزه پردازش سیگنال دارد. ویژگی مهمی که باعث شده تا این تبدیل در بسیاری از کاربردها استفاده شود این است که در عمل می‌توان این تبدیل را با استفاده از تبدیل سریع فوریه [2](FFT) محاسبه نمود. تبدیل سریع فوریه، الگوریتمی برای محاسبه سریع تبدیل فوریه و معکوس آن می‌باشد. فرم کلی تبدیل گسسته فوریه به صورت زیر می‌باشد [۲۵]

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-i\pi nk}{N}} \qquad (۱)$$

در این رابطه N تعداد ویژگی‌ها یا تعداد نقاط نمونه تابع گسسته ورودی می‌باشد. همچنین x_k نگاشت‌یافته تابع ورودی x_n بر روی موج سینوسی با فرکانس k می‌باشد. به عبارت دیگر تبدیل فوریه، یک سیگنال را از حوزه زمان به حوزه فرکانس نگاشت می‌کند. با مشاهده طیف یک سیگنال است که می‌توان کامپونت‌های فرکانسی عمده که شکل و ساختار اصلی سیگنال را تشکیل می‌دهد، استخراج نمود.

۲-۲ تبدیل موجک گسسته

یکی از روش‌هایی که برای کاهش نویز می‌توان از آن استفاده کرد، روش تبدیل موجک [3] می‌باشد. تبدیل موجک یک سیگنال را به مؤلفه‌های فرکانسی‌اش تجزیه می‌کند. از این تبدیل در حوزه‌های مختلفی نظیر فشرده‌سازی تصاویر، پیش‌گویی زلزله و شاخه‌های مختلف فیزیک و ریاضی استفاده می‌شود. تبدیل موجک باعث افزایش قدرت تفکیک زمانی داده‌ها می‌گردد. مزیت اصلی تبدیل موجک نسبت به تبدیل فوریه در این است که در تبدیل موجک، هم رزولوشن فرکانسی و هم رزولوشن زمانی در نمودار زمان- فرکانس مشخص می‌گردد. تبدیل موجک گسسته [4](DWT) یک روش تبدیل برگشت‌پذیر است که بر روی سیگنال ورودی اعمال می‌شود. تبدیل موجک گسسته به صورت مجموعه‌ای از بانک‌های فیلترهای بالاگذر و پایین‌گذر بر روی داده‌ها اعمال می‌گردد [۲۶]. با اعمال این فیلترها بر روی سیگنال، تعداد نمونه‌ها کاهش یافته و دو سیگنال تقریب کلی و سیگنال جزئی تولید می‌گردد. در هر مرحله می‌توان این عملیات را بر روی سیگنال تقریب کلی اعمال کرده و عملیات را به میزان دلخواه ادامه داد (شکل ۲).

تبدیل موجک پیوسته به صورت زیر تعریف می‌شود

$$CWT_f(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} \Psi(\frac{\overline{x}-\overline{b}}{a}) f(x)\mathrm{d}x \qquad (۲)$$

پارامتر a نشان‌دهنده میزان پهنای موجک (معرف مقیاس) و پارامتر b نشان‌دهنده موقعیت (معرف انتقال) آن می‌باشد و $\psi(x)$ هم تابع موجک مادر نامیده می‌شود. این رابطه نشان می‌دهد که چگونه یک سیگنال $f(x)$ به یک سری توابع موجک $\psi(x)$ تجزیه می‌شود. تبدیل موجک گسسته نیز به صورت زیر تعریف شده است

$$DWT_f(a,b) = \sum\sum 2^{\frac{-j}{2}} \psi(2^{-j}x-k) f(x) \qquad (۳)$$

به طوری که $a = 2^{-j}$ و $b = a.k, k \in \mathbb{Z}$ می‌باشد.

۳- روش پیشنهادی

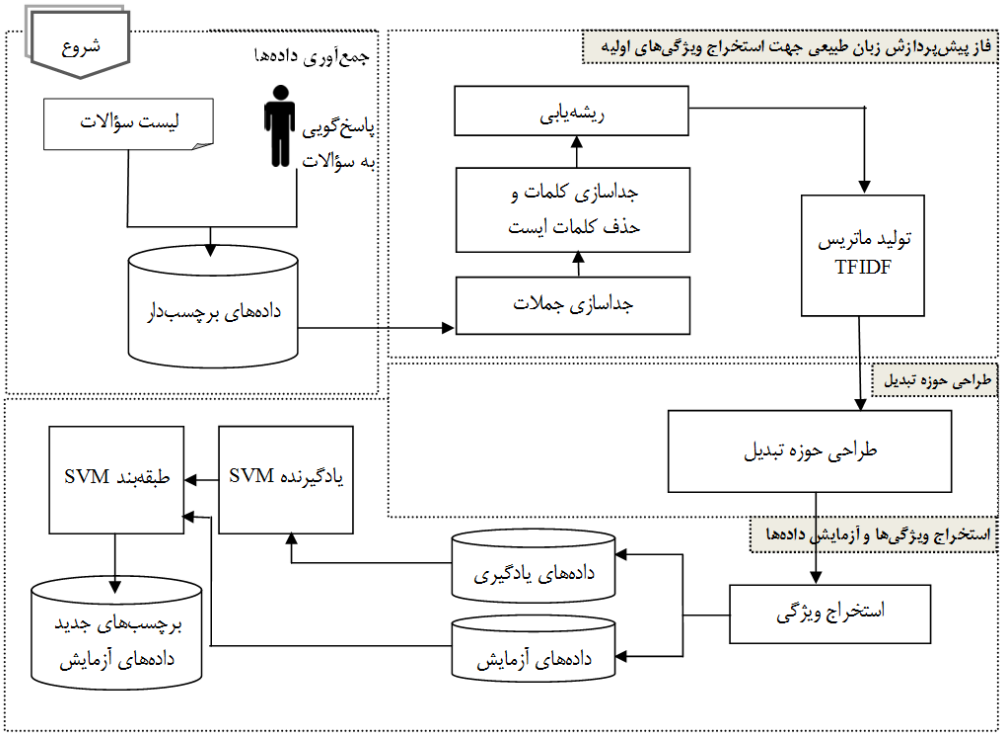
۳-۱ مرحله مقدماتی

نمای کلی سیستم پیشنهادی در شکل ۳ نشان داده شده است. برای پیاده‌سازی و آزمایش این سیستم، سه فاز اصلی پیشنهاد شده است. در فاز اول پیکره مورد نیاز جمع‌آوری می‌گردد. با توجه به این که پیکره متنی با زبان فارسی برای آنالیز حس تا کنون تولید نشده است بنابراین در این فاز با تهیه پرسش‌نامه و جمع‌آوری اطلاعات این پیکره تولید گردیده است. در فاز دوم عملیات پیش‌پردازش زبان طبیعی بر روی داده‌های جمع‌آوری‌شده انجام می‌شود. در این فاز ویژگی‌های مربوط به اسناد استخراج می‌گردد و هر سند با برداری از ویژگی‌های وزنی نشان داده می‌شود. در فاز سوم، داده‌ها از حوزه فراوانی به حوزه بهینه تبدیل می‌شوند و داده‌ها به فضای جدید با ویژگی‌های بسیار کمتر نگاشت می‌یابند. در بخش ۳-۳ نشان داده شده که در فضای جدید، جداپذیری داده‌ها افزایش می‌یابد. سپس در ادامه از روش ماشین بردار پشتیبان (SVM) برای یادگیری و طبقه‌بندی نمونه‌ها استفاده می‌شود. نتایج ارزیابی در بخش ۴ آورده شده است. نتایج نشانگر بهبود قابل توجه در تشخیص حس مثبت یا منفی مربوط به اسناد می‌باشد.

۳-۲ جمع‌آوری داده‌ها

تا کنون کارهای بسیاری در زبان انگلیسی بر روی آنالیز احساسات و یا کاوش عقیده انجام گردیده و در همین راستا مجموعه داده‌های فراوانی تولید شده است. از پرکاربردترین داده‌های مورد استفاده برای آنالیز حس در زبان انگلیسی، داده‌های سایت پرطرفدار توییتر [5] می‌باشد. در تعدادی از مقالات، قطبیت این داده‌ها به صورت دستی برچسب‌گذاری شده است [۱۸] و در بعضی از آنها نیز از شکلک‌ها و یا تگ‌هایی که کاربران برای نشان‌دادن حس خود استفاده کرده‌اند، به عنوان برچسب داده‌ها استفاده شده است [۱۲]. در بسیاری از مقالات انگلیسی هم داده‌ها به صورت پیکره‌هایی کوچک و از طریق مصاحبه و پرسش‌نامه جمع‌آوری شده است. به هر حال تا کنون مجموعه داده مناسبی برای زبان فارسی تولید نشده است. به همین منظور در این فاز، لیستی از سؤالاتی که کاربران در پاسخ به آنها قادر باشند حس خود را منتقل نمایند، تولید گردید. سپس در مصاحبه با افراد مختلف، ۸۰ نفر که خود واقف بر خوش‌بین و یا بدبین بودن خود بودند، انتخاب شدند. پاسخ‌هایی که این افراد به سؤالات دادند، در مرحله بعد برای استخراج ویژگی استفاده می‌شود. جزئیات بیشتر مربوط به این مرحله در بخش ۴ آورده شده است.

1. Discrete Fourier Transform
2. Fast Fourier Transform
3. Wavelet Transform
4. Discrete Wavelet Transform
5. http://demeter.inf.ed.ac.uk



شکل ۳: شمای کلی سیستم.

$$\begin{matrix} & \text{سند} & \\ \text{کلمه} & \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix} & \end{matrix}$$

شکل ۴: ماتریس کلمه- سند.

۳-۳ فاز پیش‌پردازش زبان طبیعی جهت استخراج ویژگی های اولیه

اسناد جمع‌آوری شده در مرحله قبلی، داده‌های خام می‌باشند و نیاز به پردازش دارند. برای این که بتوان ویژگی‌های مناسب را از این اسناد استخراج کرد باید عملیات پیش‌پردازش مختلفی را بر روی متون اعمال نمود. اولین ابزار مورد نیاز، نرمال‌کننده[1] جملات می‌باشد. در پردازش رسم‌الخط زبان فارسی، با توجه به قرابتی که با رسم‌الخط عربی دارد، همواره در تعدادی از حرف‌ها مشکل وجود دارد که از جمله آنها می‌توان به حروف "ک"، "ی"، "همزه" و ... اشاره نمود. چالش‌های دیگری نیز در نگارش زبان فارسی موجود است که با استفاده از یک ابزار نرمال‌کننده متون باید فرمت اسناد را به صورت یکسان نرمال کرده و آنها را به صورت رسم‌الخط فارسی معیار تبدیل نمود. در ادامه با استفاده از ابزارهای استخراج جملات و کلمات، کلمه‌های موجود در سند استخراج خواهند شد. ابزار حذف‌کننده کلمات ایست‌واژه[2]، کلمات پرتکرار بی‌اهمیت را شناسایی و حذف می‌نماید. در ادامه با استفاده از ابزار ریشه‌یاب، کلماتی که از لحاظ ساخت‌واژه‌ای و معنایی، مشابه با یکدیگر می‌باشند در یک دسته قرار خواهند گرفت. مثلاً کلمات "کتاب‌ها"، "کتاب" و "کتابی" در حقیقت از ریشه "کتاب" بوده و و در یک دسته قرار می‌گیرند. در انتها هم ماتریس وزن‌دار TFIDF[3] کلمات ساخته می‌شود. معیار TFIDF یکی از معروف‌ترین روش‌های وزن‌دهی به کلمات می‌باشد که به صورت (۴) محاسبه می‌گردد

$$W_{ij} = (TF - IDF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) \log_2\left(\frac{|D|}{|\{d : t_i \in d\}|}\right) \qquad (۴)$$

که $tf_{i,j}$ فرکانس نسبی کلمه i در سند j و $|D|$ تعداد کل اسناد موجود در پیکره می‌باشد. با تولید ماتریس وزنی کلمات، ویژگی نمونه‌ها و فراوانی آنها استخراج شده و عملیات پیش‌پردازش متن خاتمه می‌یابد. در شکل ۴ شمایی از این ماتریس نشان داده شده است.

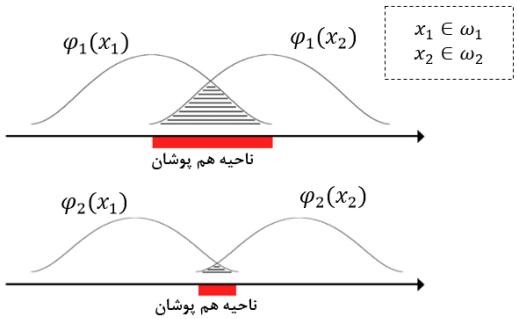
در این مرحله برای پردازش متون فارسی از ابزارهای تولیدشده توسط آزمایشگاه فناوری وب دانشگاه فردوسی[4] استفاده شده است.

۴-۳ طراحی حوزه تبدیل

پس از استخراج فراوانی کلمات، با استفاده از تبدیل فوریه گسسته، این فراوانی‌ها به حوزه فرکانس تبدیل می‌شوند. با محاسبه تعداد رخداد کلمات در اسناد، رخداد کلمات را مانند موقعیت در واحد زمان در حوزه فرکانس می‌توان در نظر گرفت. با استفاده از این تبدیل می‌توان طیف هر نمونه سند را به دست آورد.

حال اگر این تبدیل را به نوعی بتوانیم انجام دهیم که طیف دو کلاس حداکثر تفکیک‌پذیری را داشته باشند، آن گاه می‌توان نمونه را با دقت مناسبی به دو کلاس مثبت و منفی طبقه‌بندی نمود. بنابراین هدف مسأله یافتن تابع تبدیل $\varphi(x)$ می‌باشد به نحوی که دو کلاس مثبت و منفی ω_1 و ω_2 حداکثر تفکیک‌پذیری را داشته باشند. برای این کار می‌توان تابع تبدیل‌های مختلف φ_i را تولید نمود. در این مسأله به دنبال φ_i بهینه‌ای هستیم که منجر به حداکثر تفکیک‌پذیری دو کلاس گردد. در شکل ۵ نحوه عملکرد φ_i برای ارائه دو طیف مناسب از دو سیگنال در دو کلاس مختلف ω_1 و ω_2 نشان داده شده است.

1. Normalizer
2. Stopword Remover
3. Term Frequency-Inverse Document Frequency
4. wtlab.um.ac.ir



شکل ۵: با تغییر φ از φ_1 به φ_2 سعی در کاهش تداخل طیف دو کلاس شده است.

با تغییر φ از φ_1 به φ_2، سطح اشتراک دو کلاس در حوزه طیف تغییر می‌کند. طبیعتاً بهترین φ آن است که حداقل تداخل بین دو طیف را ایجاد نماید (در این تصویر φ_2). هرچه این تداخل کمتر باشد تفکیک‌پذیری داده‌های دو کلاس بیشتر بوده و در نتیجه دقت طبقه‌بندی افزایش می‌یابد. در ادامه با ارائه دو تعریف اولیه، دو قضیه مطرح شده است که در آنها به طرح مشکل تبدیل سیگنال در حوزه فراوانی پرداخته شده که منجر به جداسازی مناسب دو کلاس ω_1 و ω_2 (حس مثبت و حس منفی) می‌گردد.

تعریف ۱: اگر تبدیل $\varphi(x)$ای پیشنهاد شود که حداقل تداخل طیف دو کلاس را منجر نشود، ولی مقدار تداخلی چون k' را ایجاد نماید که میزان آن کم و قابل قبول باشد، آن را $\varphi(x)$ حاشیه‌ای می‌نامیم.

تعریف ۲: جداپذیری در حوزه طیف دو سیگنال از دو کلاس ω_1 و ω_2 یعنی تبدیل فوریه آنها با هم حداقل اشتراک ممکن را داشته باشد.

قضیه: $\varphi(x)$ بهینه یا حاشیه‌ای وجود دارد که قادر است دو کلاس ω_1 و ω_2 را در حوزه طیف جدا نماید.

اثبات: طبق فرض انجام‌شده تبدیل $\varphi(x)$ای وجود دارد (شکل ۶) که حداقل تداخل بین دو طیف از دو کلاس ω_1 و ω_2 را منجر می‌شود. به عبارتی اگر اندازه تبدیل فوریه یک سیگنال تبدیل یافته $\varphi(x_1) \in \omega_1$ به صورت زیر باشد

$$|F_1(w)| = \left| \int_{-\infty}^{+\infty} \phi(x_1) e^{-jwx} \mathrm{d}x \right| , \; x_1 \in \omega_1 \qquad (۵)$$

و به همین ترتیب اندازه تبدیل فوریه مربوط به سیگنال تبدیل‌یافته دیگر $\varphi(x_2) \in \omega_2$ به صورت زیر باشد

$$|F_2(w)| = \left| \int_{-\infty}^{+\infty} \phi(x_2) e^{-jwx} \mathrm{d}x \right| , \; x_2 \in \omega_2 \qquad (۶)$$

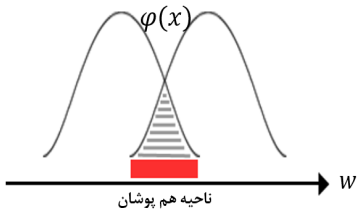
که w فرکانس زاویه‌ای سیگنال و ω_1 و ω_2 به ترتیب کلاس اول و دوم داده‌ها می‌باشند، حال می‌توان میزان هم‌پوشانی فوریه دو سیگنال را با ضرب فوریه‌های آنها و به صورت زیر محاسبه نمود

$$\phi = \min_{\phi} \left| \int_{-\infty}^{+\infty} F_1(w) F_2(w) \mathrm{d}w \right| \qquad (۷)$$

چنانچه φ تغییر کند kهای مختلفی به دست می‌آید. با جستجو روی φ، تبدیل φای مد نظر است که φ^* در (۸) را حاصل کند

$$\phi^* = \arg\min_{\phi} \left| \int_{-\infty}^{+\infty} F_1(w) F_2(w) \mathrm{d}w \right| \qquad (۸)$$

مسئله یافتن بهینه می‌باشد. چنانچه رابطه فوق با روش‌های متداول بهینه‌سازی حل شود که φ بهینه به دست می‌آید و در غیر این صورت با روش‌های ابتکاری می‌توان φ^* حاشیه‌ای را پیشنهاد کرد.



شکل ۶: تبدیل بهینه $\varphi(x)$.

لم: می‌توان نشان داد که مسأله طبقه‌بندی دو کلاس ω_1 و ω_2 به صورت زیر در حوزه تبدیل گسسته فوریه قابل بیان است

$$\min \sum_{x\in\omega_1} \sum_{m\in s_1} \left| \sum_{n=0}^{N-1} \phi(x(n)) e^{\frac{-j2\pi nm}{N}} \right| + \min \sum_{x\in\omega_1} \sum_{m\in s_2} \left| \sum_{n=0}^{N-1} \phi(x(n)) e^{\frac{-j2\pi nm}{N}} \right| \qquad (۹)$$

که S_1 مجموعه فرکانس‌های بالا و S_2 مجموعه فرکانس‌های پایین را اشاره می‌کند و N هم تعداد کلمات یا نقاط نمونه‌گیری می‌باشد.

اثبات: طبق قضیه فوق $\varphi(x)$ای وجود دارد که

$$\min_{\phi} \int_{w=w_T}^{+\infty} \left| \int_{-\infty}^{+\infty} \phi(x) e^{-jwx} \mathrm{d}x \right| \mathrm{d}w \; , \; x \in \omega_1 \qquad (۱۰)$$

یعنی تبدیل فوریه سیگنال مربوط به کلاس ω_1 در فرکانس ω_T تا بی‌نهایت حداقل می‌شود. یعنی $\varphi(x)$ باعث می‌شود که از فرکانس ω_T به بعد حداقل سطح ممکن را داشته باشد و برای تمام نمونه‌های کلاس ω_1 می‌توان نوشت

$$\min_{\phi} \sum_{x\in\omega_1} \left(\int_{w=\omega_T}^{+\infty} \left| \int_{-\infty}^{+\infty} \phi(x) e^{-jwx} \mathrm{d}x \right| \mathrm{d}w \right) , \; x \in \omega_1 \qquad (۱۱)$$

از سویی برای کلاس ω_2 داریم

$$\min_{\phi} \int_{-\infty}^{\omega_T} \left| \int_{-\infty}^{+\infty} \phi(x) e^{-jwx} \mathrm{d}x \right| \mathrm{d}w \; , \; x \in \omega_2 \qquad (۱۲)$$

یعنی اندازه تبدیل فوریه سیگنال مربوط به کلاس ω_2 در فرکانس $-\infty$ تا ω_T حداقل می‌شود. به عبارت دیگر نمونه‌های کلاس ω_2 دارای طیف بالاگذر هستند. این نکته نیز از فرض قضیه قابل استناد است. حال اگر x را به صورت گسسته $x(n)$ و همچنین روی تمام نمونه‌های کلاس ω_2 در نظر بگیریم داریم

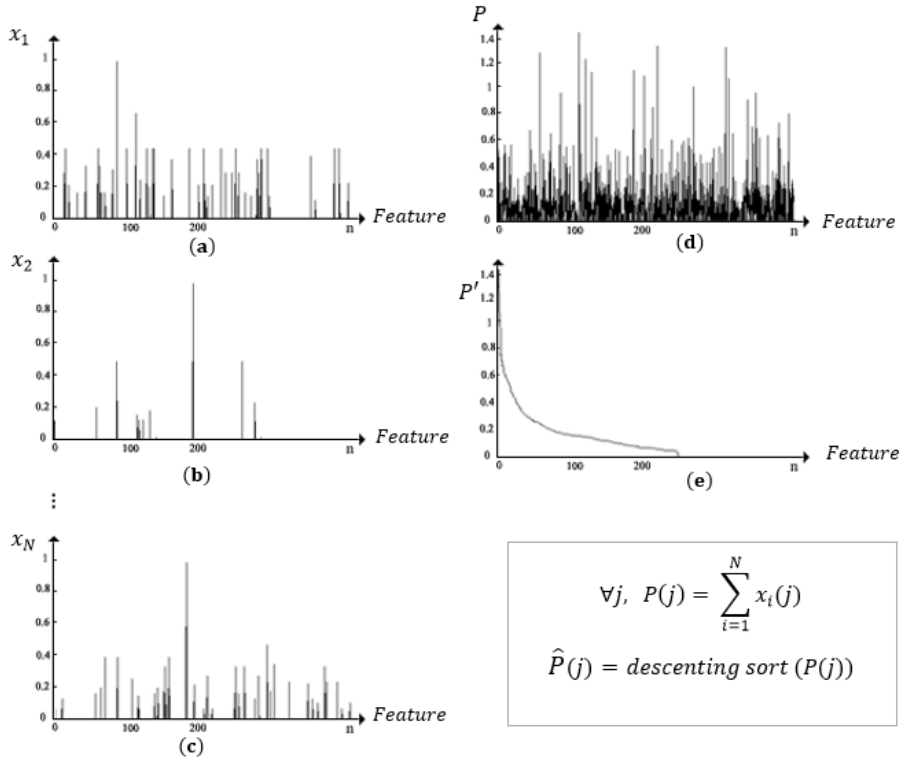
$$\min_{\phi} \sum_{x\in\omega_2} \sum_{m\in s_2} \left| \sum_{n=0}^{N-1} \phi(x(n)) e^{\frac{-j2\pi nm}{N}} \right| \qquad (۱۳)$$

عبارت فوق یعنی روی تمام نمونه‌های کلاس ω_2، مجموع فرکانس‌ها به ازای تبدیل $\varphi(.)$ حداقل می‌شود. به عبارتی تمامی نمونه‌های کلاس ω_2 دارای طیف بالاگذر باشند. لذا به تابع ارائه‌شده در لم می‌رسیم.

در قضیه بعدی تابع $\varphi(x)$ مناسبی که به دنبال آن می‌باشیم معرفی شده است. این تابع باید طوری سیگنال‌ها را تبدیل نماید که تبدیل فوریه سیگنال‌های دو کلاس حداقل مقدار تداخل را داشته باشند.

قضیه ۲: تابع $\varphi(x(n))$ مناسب برای تفکیک دو کلاس ω_1 و ω_2، با محاسبه پروفایل مجموع داده‌های کلاس ω_1 که به صورت نزولی مرتب شده باشد به دست می‌آید.

در توضیح بیشتر صورت قضیه فوق، $\varphi(.)$ می‌تواند $\sum_{x_i\in\omega_1} x_i(j)$ را به صورت نزولی مرتب نماید. یعنی مقدار ویژگی j ام تمامی نمونه‌های



شکل ۷: تابع مناسب برای تبدیل بهینه حاصل از مرتب‌سازی نزولی پروفایل نمونه‌ها روی یک کلاس.

متعلق به کلاس ω_1 با یکدیگر جمع شده و سپس در بردار پروفایل به دست آمده، ویژگی‌ها بر اساس مقدار نزولی مرتب می‌گردند. مرتب‌سازی نزولی ویژگی‌های بردار پروفایل نمونه به دست آمده منجر می‌شود تا فرکانس آن پروفایل نمونه کاهش یابد یا دارای طیف فرکانس پایین‌گذر شود. طبق فرض قضیه ۱ و به دلیل این که ماهیت دو کلاس متفاوت است، می‌توان چنین فرضی کرد که اندیس‌های این مرتب‌سازی نزولی، طیف و فرکانس کلاس دوم را بالا خواهد برد و بنابراین تفکیک‌پذیری دو کلاس آسان‌تر خواهد بود.

اثبات: پس از این توضیح اولیه به اثبات می‌پردازیم. بردار P، پروفایل نمونه‌های کلاس ω_1 می‌باشد. مؤلفه j ام بردار P برابر جمع ویژگی‌های j ام تمام نمونه‌های کلاس ω_1 می‌باشد

$$P(j) = \sum_{i=1}^{N} x_i(j)\ ,\ x_i \in \omega_1 \qquad (۱۴)$$

$$\hat{P}(j) = descentingsort(P(j)) \qquad (۱۵)$$

که N تعداد کل نمونه‌ها می‌باشد. تابع $\hat{P}(j)$ با مرتب‌سازی نزولی مقادیر تابع $P(j)$ به دست می‌آید به طوری که $\hat{P}(1)$ دارای ماکسیمم مقدار جمع $x_i(j)$ها روی همه داده‌های کلاس ω_1 و $\hat{P}(N)$ حداقل مقدار جمع $x_i(j)$ها روی همه داده‌های کلاس ω_1 می‌باشد. در شکل ۷ (a-c) تعدادی از نمونه‌های کلاس فرضی ω_1 نشان داده شده است. در شکل (d) پروفایل نمونه‌ها یا همان بردار P نشان داده شده است. در شکل (e) هم تابع $\hat{P}(j)$ که از مرتب‌سازی نزولی ویژگی‌های بردار P حاصل می‌شود، نشان داده شده است. همان طور که در شکل نیز مشخص است، فرکانس پروفایل مرتب‌شده به صورت نزولی پایین می‌باشد و به عبارت دیگر دارای طیف پایین‌گذر خواهد بود.

حال اگر فرض کنیم که میزان تابع $\hat{P}(j)$ در نمودار شکل ۷ (d) به صورت خطی کاهش یابد، با داشتن دو نقطه $(1, P_{\max})$ و $(n, 0)$ می‌توان معادله آن را به صورت تابع زیر تخمین زد

$$\hat{P}(j) = \frac{P_{\max}}{1-n}(j-n) \qquad (۱۶)$$

که در آن $P_{\max}$ مقدار $\hat{P}(j)$ است و n ماکسیمم مقدار j یا همان تعداد ویژگی‌ها می‌باشد. همچنین می‌توان تابع $\hat{P}(j)$ را به فرم $\hat{P}(n) = \alpha^n U(n)$ به طوری که $|\alpha| < 1$ تخمین زد. در این صورت تبدیل فوریه زمان گسسته (DTFT) آن به صورت معادله زیر می‌باشد

$$X(\omega) = \sum_{n=0}^{+\infty} \alpha^{-n} e^{-jwt} = \frac{1}{1-\alpha e^{-jw}} \qquad (۱۷)$$

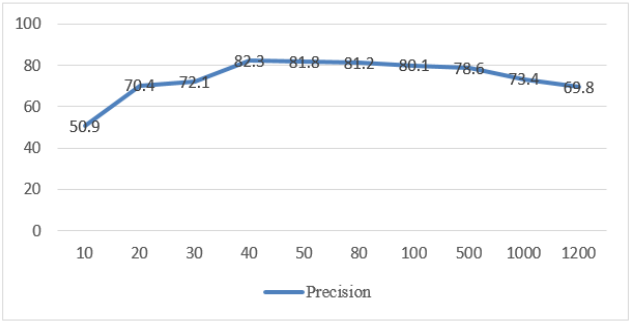
که طیف آن پایین‌گذر است. لذا تابع مرتب‌سازی نزولی یک طیف پایین‌گذر را می‌تواند سبب شود. در حالت کلی تابع مرتب‌سازی نزولی به صورت $\alpha(n)^n U(n)$ قابل بیان است که در یک حالت خاص $(\alpha(n) = \alpha)$، تابع (۱۷) حاصل می‌شود که پایین‌گذر است. با شرط آن که $|\alpha| < 1$، با توجه به نوع تابع $\alpha(n)$، طیف پایین‌گذر سیگنال متفاوت خواهد بود ولی در هر حال فرم کلی (۱۷) را خواهد داشت.

۳-۵ استخراج ویژگی و آزمایش داده‌ها

پس از تبدیل داده‌ها از فضای فراوانی به فضای بهینه و پروفایل‌گیری بر روی نمونه‌ها، با استفاده از تبدیل موجک کامپونت‌های اصلی داده‌ها استخراج می‌شوند. خصوصیاتی چون پردازش مؤلفه‌های فرکانسی سیگنال، بررسی سیگنال در فضای مقیاس، استفاده از موجک گسسته را توجیه می‌نماید. پس از استخراج ویژگی‌های جدید، مرحله یادگیری اعمال می‌گردد. در این مرحله بخشی از داده‌ها به عنوان داده‌های یادگیری استفاده می‌شوند. در سیستم پیشنهادی از یادگیرنده SVM استفاده شده است. پس از به دست آمدن مؤلفه‌های یادگیری، طبقه‌بندی بر روی داده‌های آزمایش انجام می‌شود. طبقه‌بند باید داده‌ها را به دو کلاس مثبت و منفی دسته‌بندی نماید. کلاس مثبت بیان‌کننده حس مثبت نویسنده در پاسخ به سؤالات بوده و حس منفی هم بیان‌کننده حس منفی در پاسخ به سؤالات می‌باشد.

جدول ۱: لیست سؤالات جمع‌آوری شده برای تولید پیکره مورد نیاز.

لیست سؤالات	
شماره	سؤال
۱	آینده شغلی و وضعیت مالی خود را چگونه ارزیابی می‌کنید؟
۲	میزان موفقیت به چه پارامترهایی بستگی دارد؟
۳	دیدگاه خود را در مورد ضرورت ادامه تحصیل بنویسید.
۴	میزان خوشبختی یک فرد را با چه معیارهایی می‌سنجید؟
۵	امکان صعود تیم ملی فوتبال ایران به جام جهانی را چگونه ارزیابی می‌کنید؟



شکل ۸: Precision حاصل پس از جمع ویژگی‌ها با مقادیر مختلف بر روی MR.

جدول ۲: لیست پاسخ به سؤالات یک فرد با حس منفی.

لیست پاسخ سؤالات	
شماره	پاسخ سؤال
۱	با توجه به این که رشته من فقط در مدارس و مراکز تربیت معلم کارایی دارد، به نظر من وضعیت مالی خوبی ندارد.
۲	تلاش، تحصیلات، فرهنگ، خانواده، وضعیت اقتصادی، وضعیت اجتماعی و جایگاه افراد در جامعه پارامترهای مهمی می‌باشند.
۳	افزایش آگاهی، بالارفتن میزان تحصیلات و در صورت اتمام تحصیل رسیدن به هدف مورد نظر، آرامش خاطر، احساس موفقیت و حمایت دیگران برای ادامه تحصیل.
۴	خانواده، اولین رتبه و بعد از آن داشتن وضعیت اقتصادی و مالی مناسب، همسر مورد علاقه و داشتن فرزندان خوب تأثیر به سزایی در خوشبختی یک فرد دارد.
۵	هیچ علاقه‌ای به فوتبال ندارم و مطمئن هستم که مثل همیشه صعود پیدا نمی‌کند و هر هزینه‌ای برای فوتبال را نابه‌جا می‌دانم. این پول باید صرف دیگر ورزش‌ها شود.

جدول ۳: لیست پاسخ به سؤالات یک فرد با حس مثبت.

لیست پاسخ سؤالات	
شماره	پاسخ سؤال
۱	در صورت تلاش واقعی در زمینه تحصیل، آینده خوبی خواهم داشت و در صورت ارائه مطالب جدید در رشته‌ام وضعیت مالی خوبی خواهم داشت.
۲	تلاش، توکل به خدا، امید به زندگی و آینده، انگیزه درونی و بیرونی.
۳	به نظرم امروزه ادامه تحصیل برای اشخاص از مهم‌ترین ضروریات است به طوری که فردی که حتی لیسانس داشته باشد آن چنان از منزلت بالایی برخوردار نیست و فرد باسوادی تلقی نمی‌شود چون توقع و انتظار افراد از حد لیسانس بالاتر رفته و مقاطعی چون فوق لیسانس و دکترا در زندگی امروزه بیشتر مطرح است و منزلت دارند.
۴	دوست‌داشتن چیزهای شخصی و سعی و تلاش برای به دست آوردن چیزهای بهتر و بیشتر. به نظر من فرد نباید در زندگی هیچ وقت قانع باشد و لحظه‌ای از تلاش و تحرک دست برندارد تا به نهایت خوشبختی برسد.
۵	به احتمال خیلی زیاد این تیم با تلاش و پشتکار به جام جهانی خواهد رسید.

٤- نتایج تجربی

٤-١ جمع آوری داده

با توجه به این که سابقه پردازش متن در زبان فارسی جدید می‌باشد، به همین دلیل تا کنون کار مشابهی بر روی آنالیز حس روی متون زبان فارسی مشاهده نشده است. با این توصیف، پیکره‌ای برای آزمایش و بررسی روش پیشنهادی تولید گردید.

به همین منظور ۸۰ نفر از دانشجویان و فارغ‌التحصیلان مقطع کارشناسی و کارشناسی ارشد دانشگاه فردوسی (دختر و پسر) از رشته‌های مختلف تحصیلی در طی مراحلی انتخاب شدند. این افراد طوری انتخاب شدند که ۴۰ نفر آنها خوش‌بین و ۴۰ نفر آنها بدبین باشند. سپس تعدادی سؤال جمع‌آوری شده و پس از بررسی، ۵ سؤال برای تولید پیکره انتخاب گردید. در انتخاب سؤالات سعی شد تا آنهایی انتخاب شوند که خواننده در پاسخ به آنها قادر باشد حس خوش‌بینی و یا بدبینی خود را در قالب کلمات منتقل نماید. لیست این سؤالات در جدول ۱ داده شده است.

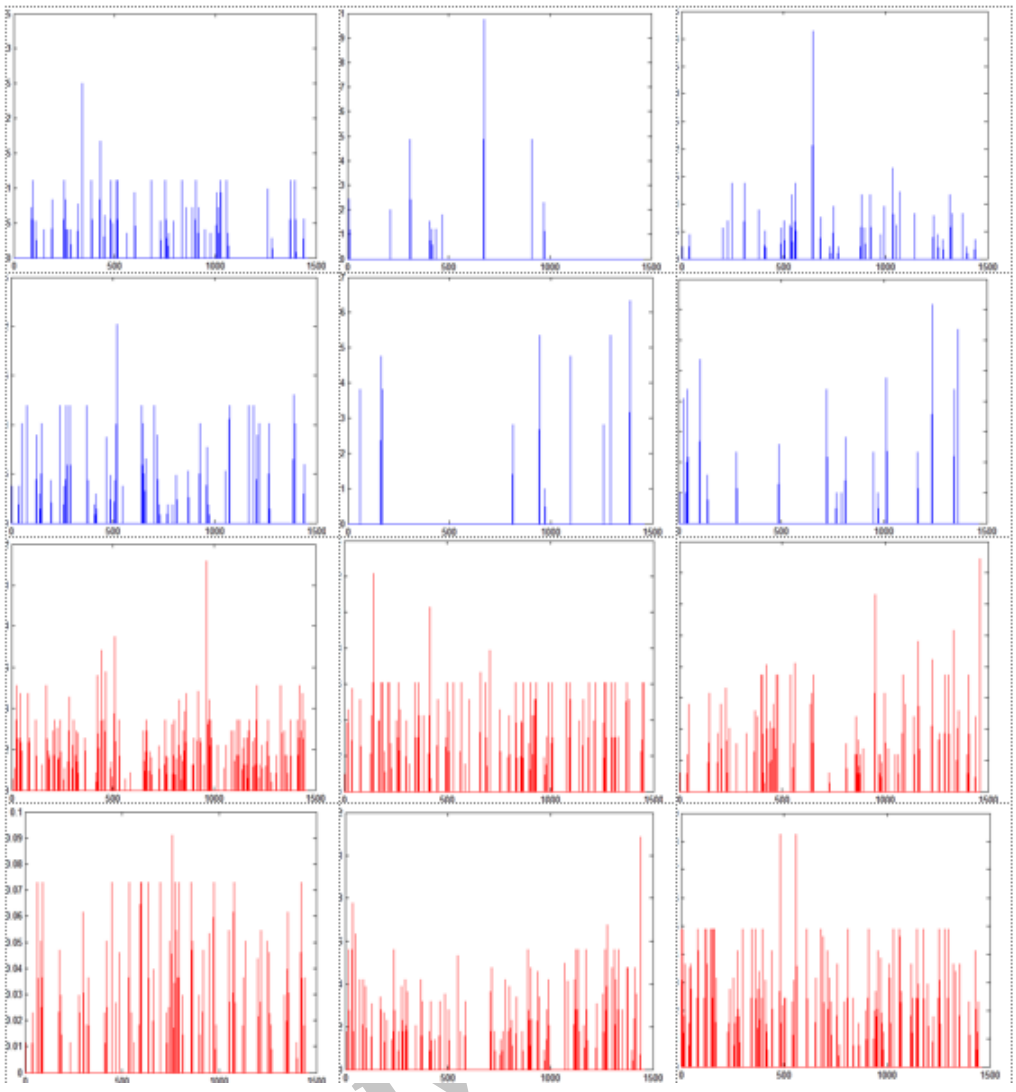
٤-٢ فاز پیش‌پردازش زبان طبیعی جهت استخراج ویژگی‌های اولیه

در مرحله بعدی، پاسخ سؤالات مربوط به هر فرد در قالب یک سند جمع‌آوری گردید. بدین ترتیب ۸۰ سند تولید گشت که هر سند حاوی حداقل ۵ پاراگراف از پاسخ‌های دانشجویان می‌باشد. با توجه به این که خوش‌بین یا بدبین‌بودن هر فرد مشخص است، بنابراین داده‌ها برچسب‌دار هستند. در جداول ۲ و ۳ یک نمونه سند از یک فرد خوش‌بین و یک نمونه سند از یک فرد بدبین نشان داده شده است. پس از جمع‌آوری اسناد، با استفاده از ابزارهای پیش‌پردازش شامل ابزارهای نرمال‌کننده، جداکننده جملات، استخراج‌گر کلمات، حذف‌کننده ایست‌واژه‌ها، ریشه‌یاب و تولیدکننده ماتریس TFIDF، ویژگی‌های مربوط به اسناد استخراج گردید که در ادامه به صورت کامل بیان می‌شود. ۱۴۶۳ ویژگی (کلمه) در نهایت استخراج شد. بدین ترتیب هر سند دارای یک بردار ویژگی شامل ۱۴۶۳ ویژگی می‌باشد که تعداد زیادی از این ویژگی‌ها صفر می‌باشند.

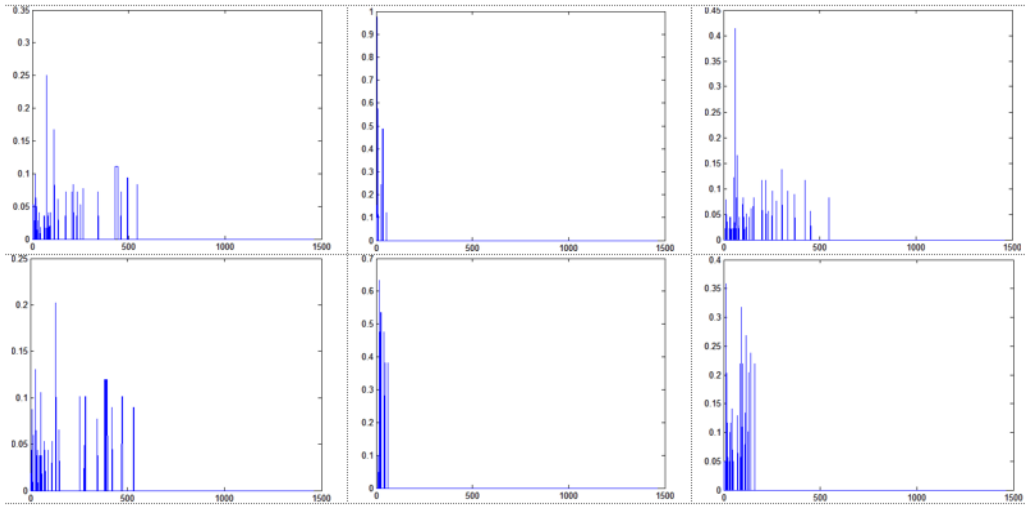
٤-٣ از طراحی حوزه تبدیل

در ادامه، عملیات پروفایل‌گیری بر روی نمونه‌های مثبت و منفی صورت پذیرفت. برای این منظور مقدار ویژگی iام $(1<i<1463)$ نمونه اول (نمونه‌های مثبت) با همدیگر جمع گردیدند.

در شکل ۸ تأثیر انتخاب تعداد ویژگی‌ها بر روی مجموعه داده نمایش داده شده است. همان طور که مشخص است در تعداد بیشتر از ۴۰ نتیجه اختلاف قابل ملاحظه‌ای نداشته است. سپس مقادیر مربوط به پروفایل مجموع نمونه‌های مثبت به صورت نزولی مرتب شدند. با این عمل در



شکل ۹: تصویر مربوط به سیگنال‌های ۶ سند اول حس مثبت (رنگ آبی) و ۶ نمونه اول حس منفی (رنگ قرمز).

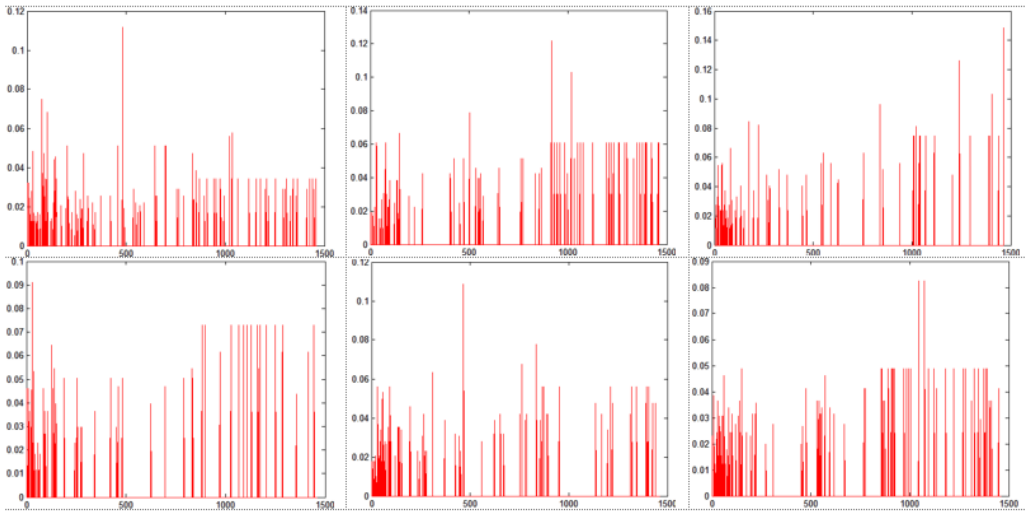


شکل ۱۰: تصویر ۶ نمونه حس مثبت، پس از مرتب‌سازی نزولی مؤلفه‌های پروفایل مجموع حس مثبت.

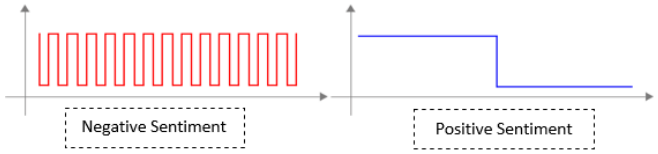
حقیقت ویژگی‌ها جابه‌جا شده و آنهایی که بزرگ‌تر از صفر هستند در کنار همدیگر قرار داده شدند. به عبارت دیگر ویژگی‌هایی که در مثبت‌بودن حس این اسناد تأثیر دارند کنار همدیگر قرار داده می‌شوند. در شکل ۹، تصویر مربوط به سیگنال ۶ نمونه اول حس مثبت و ۶ نمونه اول حس منفی به نشان داده شده است.

همان طور که در شکل ۹ مشخص می‌باشد، مقدار فراوانی‌های ویژگی‌های اسناد مثبت و منفی تقریباً توزیع‌شده است. در شکل ۱۰، تصویر مربوط به شش نمونه اول حس مثبت پس از پروفایل‌گیری و مرتب‌سازی نزولی ویژگی‌ها نشان داده شده است.

همان طور که در شکل ۱۰ نیز مشخص می‌باشد، سیگنال‌های مربوط



شکل ۱۱: تصویر ۶ نمونه اول حس منفی پس از مرتب‌سازی نزولی مؤلفه‌های پروفایل حس مثبت.



شکل ۱۲: مدل کلی داده‌های حس مثبت و منفی.

به نمونه‌های مثبت پس از جابه‌جاکردن ویژگی‌ها، دارای فرکانس پایین می‌باشند. به عبارت دیگر ویژگی‌هایی که تأثیرگذار در مثبت‌بودن حس سند می‌باشند در کنار یکدیگر قرار می‌گیرند. با توجه به فرض جداپذیر بودن داده‌های حس مثبت و منفی، پس از مرتب‌کردن نزولی ویژگی‌های داده‌ها بر اساس نمونه‌های مثبت، باید فرکانس داده‌های حس منفی افزایش یابد. نتایج آزمایشات ثابت‌کننده این ادعا می‌باشد. تصویر مربوط به ۶ نمونه اول داده‌های با حس منفی در شکل ۱۱ آورده شده است. سیگنال‌های منفی در شکل ۱۱ دارای فرکانس بالایی می‌باشند که این به دلیل جابه‌جاکردن ویژگی‌ها می‌باشد. با توجه به این موضوع، تفکیک نمونه‌های مثبت و منفی از یکدیگر بسیار ساده‌تر می‌باشد.

در شکل ۱۲ مدل کلی داده‌های حس مثبت و حس منفی نشان داده شده است. با توجه به فرکانس‌های این دو سیگنال، داده‌ها به خوبی قابل تفکیک می‌باشند. پس از جابه‌جاکردن ویژگی‌ها با اعمال تبدیل موجک و استخراج مؤلفه‌های اصلی، عمل کاهش نویز انجام می‌شود. نتایج اعمال این تابع بر روی سیگنال‌های۶ نمونه اول حس مثبت و منفی (شکل ۹) به ترتیب در شکل ۱۳ و شکل ۱۴ نشان داده شده است.

بعد از استخراج مؤلفه‌های اصلی داده‌ها، از طبقه‌بند SVM برای طبقه‌بندی داده‌ها استفاده شده است. برای بررسی میزان دقت روش پیشنهادی، چندین آزمایش انجام شد و در هر آزمایش بخشی از داده‌ها به عنوان داده‌های یادگیری و بخش دیگر به عنوان داده‌های آزمون به کار رفت. برای تعیین میزان داده‌های آزمایش و یادگیری از روش k-flod استفاده شد و میزان خطای آزمایش در جدول ۴ آورده شده است.

در جدول ۵ خطای حاصل از اجرای SVM بر روی داده‌ها بدون جابه‌جایی ویژگی و بدون اعمال wavelet آورده شده است. هدف از این آزمایش بررسی میزان تأثیر تبدیل از حوزه فراوانی به حوزه بهینه و همچنین بررسی تأثیر جابه‌جاکردن ویژگی‌ها می‌باشد. بدین منظور طبقه‌بند SVM بر روی بخشی از ماتریس TFIDF (حوزه فراوانی) به عنوان داده‌های یادگیری اعمال شد. سپس بخش دیگر داده‌ها به عنوان آزمایش بررسی شد. • خطای حاصل از برچسب‌زنی داده‌ها در جدول۶ آورده شده است. همانطور که قابل مشاهده است میزان خطا نسبت به جدول ۵ در اجراهای مختلف بسیار زیاد می‌باشد.

با توجه به نتایج جدول ۵ تبدیل از حوزه فراوانی به حوزه بهینه و همچنین جابه‌جایی ویژگی‌ها باعث کاهش خطای ۳۰ درصدی شده است که بسیار چشم‌گیر می‌باشد. در حقیقت جابه‌جاکردن ویژگی‌ها و استخراج مؤلفه‌های اصلی باعث می‌شود که دو کلاس به راحتی قابل تفکیک باشند.

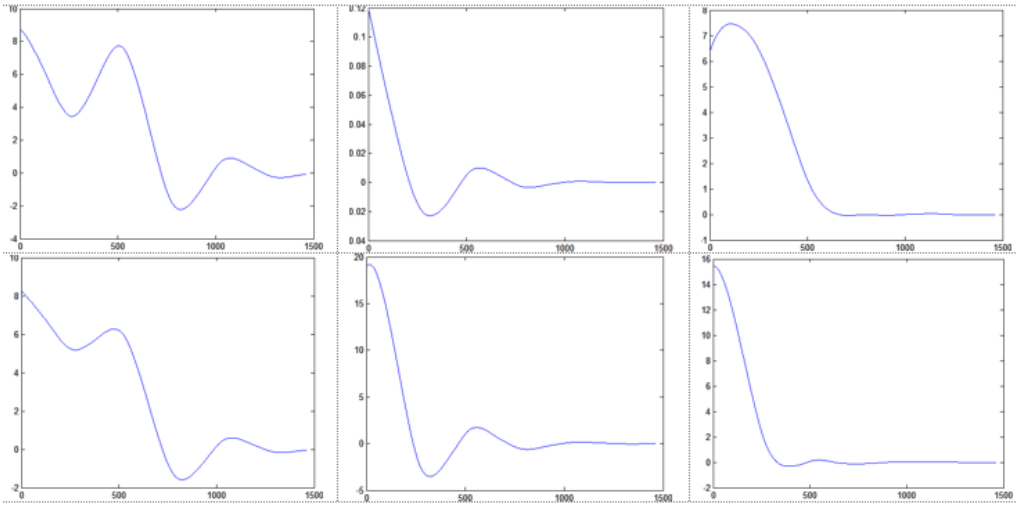
۵- بحث

۵-۱ بررسی تأثیر جابه‌جاکردن ویژگی‌ها

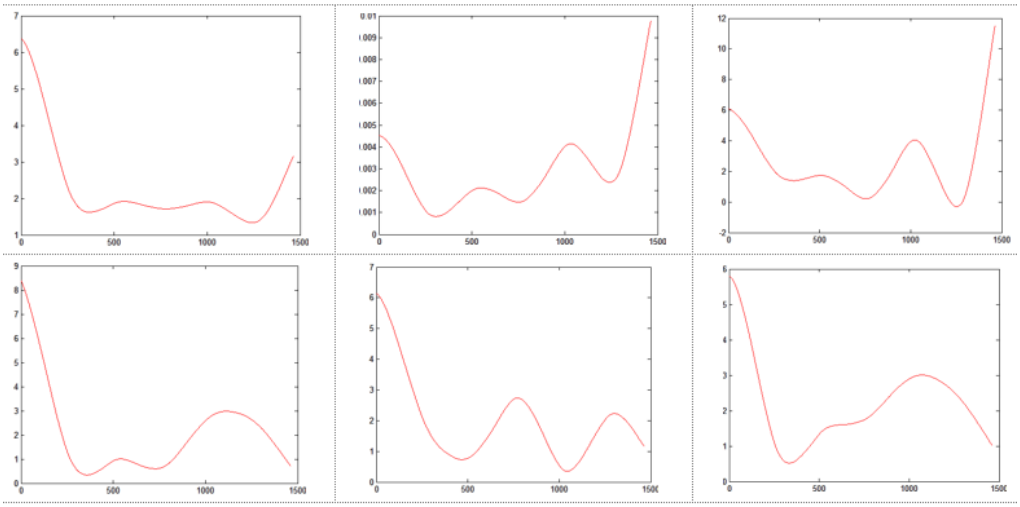
در آزمایش قبلی تأثیر همزمان تبدیل از حوزه فراوانی به حوزه بهینه و همچنین جابه‌جاکردن ویژگی‌ها را مشاهده کردیم. با این تغییرات خطا در حدود ۳۰ درصد کاهش یافت. برای این که اهمیت موضوع بیشتر مشخص شود در این آزمایش جدید، صرفاً داده‌ها را از حوزه فراوانی به حوزه بهینه منتقل نمودیم اما از جابه‌جاکردن ویژگی‌ها خودداری کردیم. خطا به طور متوسط تا میزان ۵۶ درصد افزایش می‌یابد. جابه‌جاکردن ویژگی‌های نمونه‌های مثبت در آزمایشات قبلی باعث شد تفکیک‌پذیری داده‌ها بسیار ساده‌تر گردد. همان طور که پیشتر اشاره شد، تبدیل داده‌ها از حوزه فراوانی به حوزه بهینه منجر به افزایش خطا تا میزان ۵۶ درصد در جدول شده است. دلیل این امر را می‌توان این طور تفسیر نمود که در این حالت، پس از تبدیل از حوزه فراوانی به حوزه بهینه، چون مؤلفه‌های اول انتخاب می‌شوند، در حقیقت به نحوی فیلتر پایین‌گذر اعمال می‌گردد و این موضوع باعث می‌شود که میزان خطا افزایش یابد. ممکن است این طور تصور شود که بدون تبدیل داده‌ها به حوزه جدید و صرفاً با جابه‌جاکردن ویژگی‌ها می‌توان خطا را به میزان قابل توجهی کاهش داد. آزمایش بعدی این فرضیه را رد می‌کند.

۵-۲ بررسی تأثیر تبدیل داده‌ها از حوزه فراوانی به حوزه بهینه

در این آزمایش تأثیر تبدیل داده‌ها از حوزه فراوانی به حوزه بهینه بررسی می‌شود. برای این منظور در این آزمایش بدون آن که داده‌ها به حوزه بهینه منتقل گردند، ویژگی‌های مجموع داده‌های حس مثبت طوری جابه‌جا می‌شوند که ویژگی‌های تأثیرگذار در حس مثبت در کنار یکدیگر قرار گیرند. نتایج جالب می‌باشد، این بار نیز در حدود ۳۷ درصد خطا وجود دارد که به دلیل عدم تبدیل فضا به وجود آمده است.



شکل ۱۳: سیگنال ۶ نمونه اول حس مثبت پس از اعمال تبدیل موجک.



شکل ۱۴: سیگنال ۶ نمونه اول حس مثبت پس از اعمال تبدیل بهینه.

جدول ۴: خطای آزمایش حاصل از ۱۰ اجرای مختلف به ازای مقادیر مختلف آزمایش و یادگیری (آزمایش روی ۸۰ نمونه).

Value of *k*	Number of Training Data	Number of Test Data	Error Rate									
			run۱	run۲	run۳	run۴	run۵	run۶	run۷	run۸	run۹	run۱۰
۲-Fold	۴۰	۴۰	٪۵	٪۵	٪۵	٪۵	٪۵	٪۵	٪۵	٪۵	٪۵	٪۵
۳-Fold	۵۴	۲۶	٪۴	٪۴	٪۴	٪۴	٪۴	٪۴	٪۴	٪۴	٪۴	٪۴
۵-Fold	۶۴	۱۶	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳
۸-Fold	۷۰	۱۰	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳	٪۳

جدول ۵: خطای آزمایش حاصل از ۱۰ اجرای مختلف SVM به روی داده‌ها بدون جابه‌جایی ویژگی و بدون اعمال WAVELET به ازای مقادیر مختلف آزمایش و یادگیری (آزمایش روی ۸۰ نمونه).

Value of *k*	Number of Training Data	Number of Test Data	Error Rate									
			run۱	run۲	run۳	run۴	run۵	run۶	run۷	run۸	run۹	run۱۰
۲-Fold	۴۰	۴۰	**٪۳۶**	**٪۳۷**	**٪۳۶**	**٪۳۶**	**٪۳۷**	**٪۳۷**	**٪۳۷**	**٪۳۷**	**٪۳۷**	**٪۳۷**
۳-Fold	۵۴	۲۶	٪۳۷	٪۳۶	٪۳۶	٪۳۷	٪۳۶	٪۳۷	٪۳۷	٪۳۷	٪۳۶	٪۳۷
۵-Fold	۶۴	۱۶	٪۳۵	٪۳۵	٪۳۵	٪۳۵	٪۳۶	٪۳۵	٪۳۶	٪۳۶	٪۳۵	٪۳۵
۸-Fold	۷۰	۱۰	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۵	٪۳۶	٪۳۵	٪۳۵	٪۳۶

با توجه به آزمایش‌های انجام‌شده در جدول‌های ۶ و ۷، جابه‌جایی ویژگی‌ها و تبدیل از حوزه فراوانی به حوزه بهینه باعث می‌شود که تفکیک‌پذیری دو کلاس به میزان قابل توجهی افزایش یافته و دقت تشخیص را به میزان قابل توجهی افزایش دهد.

۳-۵ بررسی عملکرد روش ارائه‌شده بر MR و CR

مجموعه داده MR که یک مجموعه داده نظرسنجی از فیلم‌ها می‌باشد یکی از معروف‌ترین و پرکاربردترین مجموعه داده‌های موجود در آنالیز

جدول ۶: خطای آزمایش حاصل از ۱۰ اجرای مختلف SVM به روی داده‌ها در حوزه تبدیل، بدون جابه‌جایی ویژگی‌ها و به ازای مقادیر مختلف آزمایش و یادگیری (آزمایش روی ۸۰ نمونه).

Value of k	Number of Training Data	Number of Test Data	Error Rate									
			run۱	run۲	run۳	run۴	run۵	run۶	run۷	run۸	run۹	run۱۰
۲–Fold	۴۰	۴۰	٪۵۶	٪۵۶	٪۵۷	٪۵۶	٪۵۷	٪۵۶	٪۵۶	٪۵۷	٪۵۸	٪۵۷
۳–Fold	۵۴	۲۶	٪۵۸	٪۵۸	٪۶۰	٪۵۸	٪۵۸	٪۵۹	٪۵۹	٪۵۸	٪۵۹	٪۵۹
۵–Fold	۶۴	۱۶	٪۶۰	٪۶۱	٪۶۰	٪۶۱	٪۶۰	٪۶۰	٪۶۰	٪۶۰	٪۵۹	٪۶۰
۸–Fold	۷۰	۱۰	٪۶۲	٪۶۲	٪۶۲	٪۶۲	٪۶۲	٪۶۲	٪۶۲	٪۶۲	٪۶۲	٪۶۲

جدول ۷: خطای آزمایش حاصل از ۱۰ اجرای مختلف SVM به روی داده‌ها در حوزه فراوانی با جابه‌جایی ویژگی‌ها و به ازای مقادیر مختلف آزمایش و یادگیری (آزمایش روی ۸۰ نمونه).

Value of k	Number of Training Data	Number of Test Data	Error Rate									
			run۱	run۲	run۳	run۴	run۵	run۶	run۷	run۸	run۹	run۱۰
۲–Fold	۴۰	۴۰	٪۳۷	٪۳۷	٪۳۷	٪۳۶	٪۳۷	٪۳۶	٪۳۷	٪۳۷	٪۳۷	٪۳۷
۳–Fold	۵۴	۲۶	٪۳۶	٪۳۷	٪۳۶	٪۳۶	٪۳۷	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶
۵–Fold	۶۴	۱۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۶
۸–Fold	۷۰	۱۰	٪۳۵	٪۳۶	٪۳۵	٪۳۶	٪۳۶	٪۳۶	٪۳۶	٪۳۵	٪۳۶	٪۳۶

جدول ۸: ارزیابی روش پیشنهادی با استفاده از داده‌ها CR و MR.

Authors	Precision	
	MR	CR
[۲۸]	۶۶٫۳	۶۵٫۲
[۳۱]	۷۷٫۳	۸۱٫۱
[۳۲]	۷۹٫۴	۸۰٫۸
[۳۳]	۷۹٫۱	۸۱٫۴
[۳۴]	۷۹٫۵	۷۸
[۳۵]	۸۱٫۵	۷۹
$OUR_{without}$	۶۹	۶۵
OUR_{with}	۸۲٫۳	۸۱٫۹

حس است که به صورت رایگان در اینترنت قرار داده شده است[1]. این مجموعه داده در [۲۸] توسعه یافته که حاوی ۵۳۳۱ نظر مثبت و ۵۳۳۱ نظر منفی می‌باشد. مجموعه داده CR[2] یک مجموعه داده نظرسنجی از ۵ محصول می‌باشد که در [۲۹] توسعه یافته است. در این مجموعه داده نظرات مثبت و منفی هر محصول در یک فایل در کنار هم قرار گرفته است.

در ادامه این بخش نتایج حاصل از اعمال روش ارائه‌شده در مقاله بر روی این دو مجموعه داده بیان می‌شود. در مرحله قبل برای پیش‌پردازش از ابزارهای ایجادشده برای زبان فارسی استفاده شد اما مجموعه داده MR و CR به زبان انگلیسی هستند و برای همین منظور برای مرحله پیش‌پردازش و استخراج ویژگی‌ها از ابزارهای GATE [۳۰] استفاده شده است. همان طور که در بخش قبل بیان شد با تبدیل از حوزه فراوانی به حوزه بهینه و همچنین جابه‌جایی ویژگی‌ها بهترین نتیجه حاصل شد. به همین دلیل باید انتظار همین نتیجه را برای داده‌های MR و CR داشت. برای بررسی صحت مطلب بیان‌شده مجموعه داده در طی دو مرحله یک بار بدون استفاده از تبدیل بهینه و جابه‌جایی (OUR_{with}) و یک بار با تبدیل بهینه و جابه‌جایی ($OUR_{without}$) عملیات ارزیابی شده که نتایج آن در جدول ۸ ذکر شده است. برای مقایسه عملکرد روش ارائه‌شده در جدول ۸، دقت نتایج روش پیشنهادی با کارهای دیگر مقایسه شده است.

1. http://www.cs.cornell.edu/People/pabo/movie-review-data/
2. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

٦- نتیجه‌گیری

در این مقاله روش جدیدی برای آنالیز حس متون فارسی ارائه شد. در روش پیشنهادی در ابتدا ویژگی‌های اولیه اسناد با استفاده از ابزارهای پیش‌پردازش متن استخراج می‌گردد. سپس با انتقال ویژگی‌ها از حوزه فراوانی به حوزه تبدیل بهینه، امکان جداسازی کلاس‌های حس مثبت و منفی با دقت بسیار بالاتری فراهم می‌آید. در این مقاله ثابت شده که تبدیل بهینه‌ای وجود دارد که با استفاده از آن با انتقال داده‌ها از حوزه فراوانی به حوزه جدید، می‌توان قابلیت تمایز دو کلاس را به شدت افزایش داد به طوری که دقت طبقه‌بندی به طور قابل ملاحظه‌ای افزایش یابد. این تبدیل با پروفایل‌گیری نمونه‌ها بر روی کلاس مثبت و یا منفی و مرتب‌سازی نزولی آنها حاصل می‌گردد. روش پیشنهادی برای اولین بار بر روی پیکره‌ای از متون فارسی که از طریق مصاحبه با ۸۰ نفر به دست آمده آزمایش گردید که نتایج آن در بخش‌های قبل آورده شده است.

مراجع

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval 2(1-2)*, vol. 2, no. 1-2, pp. 1-135, 07 Jul 2008..

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering J.*, vol. 5, no. 4, pp. 1093-1113, Dec. 2014.

[3] R. F. Bruce and J. M. Wiebe, "Recognizing subjectivity: a case study in manual tagging," *Natural Language Engineering*, vol. 5, no. 2, pp. 187-205, Jun. 1999.

[4] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proc. of 12th Int. Conf. on World Wide Web*, *WWW'03,* pp. 519-528, 2003.

[5] O. Nasraoui, "Book review: web data mining-exploring hyperlinks, contents, and usage data," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 2, pp. 23-25, Dec. 2008.

[6] B. Liu, *Sentiment Analysis and Subjectivity*, Handbook of Natural Language Processing, 2010.

[7] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347-354, 2006.

[8] S. M. Liu and J. H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1083-1093, 15 Feb 2015.

[9] L. Yung-Ming and L. Tsung-Ying, "Deriving market intelligence from microblogs," *Decision Support Systems*, vol. 55, no. 1, pp. 206-217, Apr. 2013.

[10] R. Moraes, J. F. Valiati, and W. P. GaviãO Neto, "Document-level sentiment classification: an empirical comparison," Between SVM and ANN," *Expert Systems with Applications*, vol. 40, no 2, pp.621-633, Feb. 2013.
[11] F. L. Cruz, J. A. Troyano, F. EnríQuez, F. J. Ortega, and C. G. Vallejo, "Long autonomy or long delay? the importance of domain in opinion mining," *Expert Systems with Applications*, vol. 40, no. 8, pp. 3174-3184, Jun. 2013.
[12] M. Taboada, *Lexicon-Based Methods for Sentiment Analysis*, Association for Computational Linguistics, 2011.
[13] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussions," in *Working Notes of the ACM SIGIR Workshop on Operational Text Classification*, 6 pp., Mar. 2001.
[14] P. Turney and M. Littman, "Measuring praise and criticism: inference of semantic orientation from association," *ACM Trans. on Information Systems*, vol. 21, no. 4, pp. 315-346, Sep. 2003.
[15] Y. Dang, Y. Zhang, and H. Chen, "A lexicon enhanced method for sentiment classification: an experiment on online product reviews," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 46-53, Aug. 2010.
[16] P. Rudy and M. Thelwall, "Sentiment analysis: a combined approach," *J. of Informetrics*, vol. 3, no. 2, pp. 143-157, Apr. 2009.
[17] S. Dasgupta and V. Ng, "Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification," in *Proc. of ACL-IJCNLP*, vol. 2, pp. 701-709, 2009.
[18] E. Kouloumpis, "Twitter sentiment analysis: the good the bad and the OMG!," in *Proc. 5th Int. AAAI Conf. on Weblogs and Social Media*, pp. 538-541, Barcelona, Catalonia, Spain, 17-21 Jul. 2011.
[19] C. Tan, *et al*., "User-level sentiment analysis incorporating social networks," in *Proc, of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, PP. 1397-1405, 2011.
[20] D. Rao and D. Ravichandran, "Semi supervised polarity lexicon induction," in *Proc. of the European Chapter of the Association for Computational Linguistics, EACL'09*, pp. 675-682, Apr 2009.
[21] O. Tackstrom and R. McDonald, "Semi-supervised latent variable models for sentence-level sentiment analysis," in *Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies*, *HLT'11,* vol. 2, pp. 569-57, 2011.
[22] P. Galeas, R. Kretschmer, and B. Freisleben, "Document relevance evaluation via term distribution analysis using Fourier series expansion," in *Proc. of the 9th ACM/IEEE-CS Joint Conf. on Digital libraries*, pp. 277-284, Mar. 2009.
[23] A. F. Laurence, K. Ramamohanarao, and M. Palaniswami, "Fourier domain scoring: a novel document ranking method," in *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 5, pp. 529-539, May 2004.
[24] S. Steven, "Chapter 8: the discrete Fourier transform," *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd Ed., San Diego, CA, USA: California Technical Publishing, 1999.
[25] M. R. Spiegel, *Schaum's Outline of Theory and Problems of Fourier Analysis*, New York, NY, USA: McGraw Hill, 1974.
[26] S. Mallat, *A Wavelet Tour of Signal Processing*, New York, NY, USA: Academic Press, 1999.
[27] E. C. Mundim, H. A. Schots, and J. M. Araujo, "WTdecon, a colored deconvolution implemented by wavelet transform," *The Leading Edge*, vol. 25, no. 4, pp. 398-401, Apr. 2006.
[28] B. Pang and L. Lillian, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05*, pp. 115-124, 2005.
[29] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. of the 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, *KDD'04*, pp. 168-177, 2004.
[30] H. Cunningham, K. Humphreys, R. Gaizauskas, and Y. Wilks, *Developing Language Processing Components with GATE Version 8*, University of Sheffield Department of Computer Science, Nov. 2014.
[31] T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency tree-based sentiment classification using CRFs with hidden variables," in Proc. The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie, pp. 786-794, 2010.
[32] R. Socher, B. Huval, C. Manning, and A. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP'12,* pp. 786-794, 2010.
[33] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. of the 2013 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP'13*, pp. 121-135, Aug. 2012.
[34] L. Dong, F. Wei, S. Liu, M. Zhou, and K. Xu, "A Statistical Parsing Framework for Sentiment Classification," *Computational Linguistics*, vol. 14, no. 2, pp. 293-336, Jun 2014.
[35] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. of the 2014 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP'14*, pp. 135-151, Sep. 2014.

آصف پورمعصومی تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار به‌ترتیب در سال‌های ۱۳۸۷ و ۱۳۹۰ در دانشگاه فردوسی مشهد به پایان رسانده است و هم‌اکنون دانشجوی دکتری مهندسی نرم افزار دانشگاه فردوسی مشهد می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش زبان طبیعی، متن‌کاوی، آنالیز حس و نظر، خلاصه‌سازی متن، فرایند کاوی.

هادی صدوقی یزدی کارشناسی خود را در رشته مهندسی برق از دانشگاه فردوسی مشهد در سال ۱۳۷۳ اخذ نمود و تحصیلات خود را در مقطع کارشناسی ارشد و دکتری در رشته مهندسی برق در دانشگاه تربیت مدرس تهران با اخذ مدارک تحصیلی مربوطه به ترتیب در سال‌های ۱۳۷۵ و ۱۳۸۴ ادامه داد. در حال حاضر دکتر هادی صدوقی یزدی استاد دانشکده فنی و مهندسی دانشگاه فردوسی مشهد است. تحقیقات بنیادی ایشان در زمینه شناسی الگو و هوش مصنوعی می‌باشد.

هادی قائمی مدرک کارشناسی ارشد خود را در سال ۱۳۹۳ در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه فردوسی مشهد اخذ نموده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش زبان طبیعی، پردازش الگو، مباحث یادگیری ماشین.

زهرا دلخسته مدرک کارشناسی خود را در سال ۱۳۹۲ در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه فردوسی مشهد اخذ نموده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش زبان طبیعی، پردازش الگو.