علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۳، شماره ۲، ۱۳۹۴
صفحات ۴۷-۵۹
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

نظرکاوی بین‌زبانی با استفاده از ویژگی‌های معنایی

شیما اسمعیلی تفت آزاده شاکری

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران

چکیده

نظرکاوی یکی از زیربخش‌های متن‌کاوی است. در این حوزه به بررسی متن‌های نظرمند پرداخته می‌شود و هدف تشخیص مثبت و یا منفی بودن مفهوم این متن‌ها است. روش‌ها و راه‌حل‌های پیشنهادی در این حوزه به دو دسته باناظر و بدون ناظر دسته‌بندی می‌شود. از آن‌جا که روش‌های باناظر کارایی و دقت بالاتری نسبت به روش‌های بدون ناظر دارد، سعی می‌شود تا آن‌جایی که امکان دارد شرایط برای استفاده از روش‌های باناظر فراهم شود. اصلی‌ترین نیازمندی این روش‌ها، داده‌های برچسب‌خورده، به عنوان داده آموزش، در دامنه و زبان داده‌های آزمون است. وجود چنین داده‌هایی در تمام جفت دامنه و زبان‌ها محدودیتی برای استفاده از این گونه روش‌ها محسوب می‌شود. با توجه به زمان‌بر و پرهزینه بودن تولید داده‌های برچسب‌خورده به عنوان داده‌های آموزش، معمولاً ایجاد چنین مجموعه داده‌ای به عنوان بهترین راه‌حل در نظر گرفته نمی‌شود. همچنین به دلیل بیان متفاوتی که در دامنه‌ها و زبان‌های متفاوت وجود دارد، استفاده از داده‌های آموزش موجود در دامنه و یا زبان متفاوت به طور مستقیم موجب کاهش کارایی روش‌ها می‌شود. اما وجود داده‌های آموزش در اکثر دامنه‌ها در زبان‌های با منابع غنی انگیزه‌ای برای استفاده غیرمستقیم از این داده‌ها برای نظرکاوی داده‌های آزمون در زبان‌های دیگر ایجاد می‌کند. از این رو روش‌هایی به عنوان روش‌های بین‌زبانی ارائه شد که در آن‌ها از داده‌های آموزش موجود در زبان متفاوت با داده‌های آزمون، برای استخراج اطلاعات و در نهایت انتقال اطلاعات به زبان مورد نظر، بهره می‌برد. در این مقاله روشی برای نظرکاوی بین‌زبانی ارائه می‌شود که این استخراج و انتقال اطلاعات با کیفیت بالایی انجام می‌شود و همچنین برای اکثر زبان‌ها، حتی زبان‌های منابع‌محدود نیز قابل استفاده می‌باشد و به منابع موجود در زبان مورد نظر وابستگی کمی دارد. این روش با استفاده از داده‌های بدون برچسب در هر دو زبان مبدأ و مقصد، یک گراف دوبخشی بین دو دسته از ویژگی‌های محوری و غیرمحوری می‌سازد و ویژگی‌های معنایی را از آن استخراج می‌کند. تنها منبع مورد نیاز برای این روش، یک لغت‌نامه است که به دلیل استفاده از تعداد محدودی از ترجمه‌های آن، میزان وابستگی بالایی به این منبع ندارد.

کلمات کلیدی: نظرکاوی، تحلیل نظرات، ویژگی محوری، ویژگی معنایی، بین‌زبانی، ویژگی مستقل از دامنه، ویژگی وابسته به دامنه، گراف دوبخشی، رده‌بندی.

۱- مقدمه

امروزه با پیشرفت تکنولوژی و گسترش میزان استفاده از اینترنت و فضای مجازی، تولید محتوا توسط همه کاربران امکان‌پذیر شده است. نوشتن وبلاگ، به‌روزرسانی صفحه شخصی در شبکه‌های اجتماعی و وارد کردن نظرات و تجربیات شخصی در مورد یک مطلب از جمله این محتواها است.

نظرات و تجربیات شخصی می‌تواند درباره یک محصول و یا یک خدمت باشد، کاربری که از یک محصول یا خدمت استفاده کرده، بازخوردی در رابطه با آن بیان می‌کند. این نظرات که می‌توانند برای افراد و کاربران دیگر که قصد استفاده از این محصول و یا خدمت را دارند، مفید واقع شود، لحظه به لحظه در حال افزایش و حجیم شدن است. حوزه تحلیل نظرات[۱] و یا نظرکاوی[۲] سعی در تشخیص خودکار مثبت و یا منفی بودن و یا به عبارت دیگر تشخیص جهت احساسی[۳] این نظرات دارد تا بتوان خلاصه‌ای از نظرات وارد شده در مورد یک محصول و یا خدمت به کاربران جدید ارائه داد و آن‌ها را برای تصمیم‌گیری درست هدایت کرد. علاوه بر کاربرد ذکر شده، جمع‌آوری خلاصه‌ای از نظرات داده شده، می‌تواند برای تولیدکنندگان و صاحبان مشاغل نیز بسیار مفید واقع شود، زیرا می‌توانند به ارزیابی مناسبی از محصولات و خدمات ارائه شده خود دست پیدا کنند. این حوزه که زیربخشی از حوزه متن‌کاوی[۴] است، تنها داده‌های نظرمند[۵] را بررسی می‌کند و از داده‌های بدون نظر[۶] صرف نظر می‌کند.

حوزه تحلیل نظرات بیش از یک دهه است که مورد توجه محققین و پژوهش‌گران قرار گرفته است [۱، ۲]. تا به حال، روش‌های باناظر [۲، ۳] و بدون

ناظر [۱، ۴] متعددی برای بهبود کارایی این پیش‌بینی‌ها، پیشنهاد شده است. دسته اول روش‌های باناظر است که با استفاده از داده‌های برچسب‌خورده و استخراج ویژگی‌ها به تحلیل نظرات می‌پردازد. این روش‌ها عموماً دارای کارایی بالا است و وجود داده‌های برچسب‌خورده به عنوان داده آموزش در دامنه و زبان داده‌های آزمون از پیش‌نیازهای آن‌ها به شمار می‌رود. اما در برخی دامنه‌ها و یا زبان‌ها داده‌های برچسب‌خورده در حجم و کیفیت مناسبی در دسترس نیست. دسته دوم روش‌های بدون ناظر است. مزیتی که روش‌های بدون ناظر دارد، عدم نیاز آن‌ها به داده‌های برچسب‌خورده است. این روش‌ها به استخراج کلمات و عباراتی می‌پردازد که دارای جهت احساسی باشد و سپس با استفاده از آن‌ها برچسب نظرات را پیش‌بینی می‌کند. این روش‌ها معمولاً در مقایسه با روش‌های دسته اول دارای کارایی پایین‌تری است.

استفاده از اینترنت و تولید محتوا محدود به افراد و کاربران خاصی نیست و هر فرد با هر زبانی امکان تولید محتوا به زبان خود را دارد. بدیهی است که زبان‌های مختلف، ویژگی‌ها و خصلت‌های متفاوتی دارد و این اختلاف تنها به استفاده از کلمات متفاوت خلاصه نمی‌شود. همچنین برخی زبان‌ها نسبت به سایر زبان‌ها رایج‌تر بوده و در نتیجه منابع و داده‌های بیش‌تری از آن‌ها در اختیار است که در چنین زبان‌هایی کار تحلیل نظرات ساده‌تر انجام می‌شود. اما در زبان‌هایی که با کمبود داده و یا منابع مواجه است، کار دشوارتر می‌شود. از آن‌جا که تولید مجموعه داده‌ای برچسب‌خورده کاری زمان‌بر و پرهزینه است و ممکن است چنین داده‌هایی در حجم مناسب موجود نباشد، یکی از راه‌حل‌ها، استفاده از داده‌های موجود در زبان‌های دیگر است. اما داده‌های دو زبان متفاوت دارای ویژگی‌های کاملاً متفاوتی است که برای به‌کارگیری این داده‌ها نیاز به روش‌هایی برای برطرف کردن مانع است. این روش‌ها که روش‌های بین‌زبانی خوانده می‌شود، تلاش می‌کند از نظرات برچسب‌خورده موجود در زبان دیگر (زبان مبدأ) برای پیش‌بینی برچسب نظرات در زبان مورد نظر (زبان مقصد) استفاده کند.

بدیهی است که دو زبان مختلف، تفاوت‌هایی با هم دارد و به عبارتی دارای توزیع متفاوتی است. روش‌های بین‌زبانی تلاش می‌کند با به‌کارگیری راه‌حلی امکان استفاده از داده‌های موجود در یک زبان برای پیش‌بینی برچسب نظرات در زبان دیگر را فراهم کند تا به کیفیت روش‌های تک‌زبانه نزدیک‌تر شود. ترجمه یکی از داده‌ها به زبان دیگر و هم‌زبان شدن داده‌های آموزش و آزمون از ساده‌ترین راه‌حل‌های موجود به شمار می‌رود [۵، ۶]. اما برای بسیاری از زوج‌زبان‌ها منابع مناسبی برای ترجمه داده‌ها با کیفیت بالا از یک زبان به زبان دیگر در اختیار نیست و در نتیجه لازم است روش‌های دیگری که وابستگی کمتری به این منابع دارد ارائه شود و این محدودیت‌ها را برای این‌گونه از زبان‌ها کاهش دهد.

این مقاله با استفاده از ایده خوشه‌بندی گراف، به ساخت گرافی دوبخشی و دوزبانه می‌پردازد. این گراف بین ویژگی‌های محوری و غیرمحوری استخراج شده از تعداد زیادی نظرات بدون برچسب در هر دو زبان مبدأ و مقصد ساخته می‌شود. ویژگی‌های محوری جفت‌کلمه‌هایی است که هر کلمه، ترجمه کلمه دیگر در زبان دیگری است. با استفاده از یک منبع ترجمه، ویژگی‌های محوری که مستقل از زبان و یا به عبارت دیگری بدون ابهام است، انتخاب می‌شود. اما در مقابل ویژگی‌های غیرمحوری تک‌کلمه‌هایی است که احتمال وجود ابهام در ترجمه آن‌ها بیش‌تر تخمین زده شده است و در نتیجه این ویژگی‌ها ترجمه نمی‌شود و به صورت تک‌کلمه مورد استفاده قرار می‌گیرد. با استفاده از گراف ساخته شده و یکی از الگوریتم‌های پیشنهاد شده برای خوشه‌بندی گراف [۷]، ویژگی‌ها با احتمالی به هر خوشه تعلق می‌گیرد. با انتخاب تعدادی از بهترین خوشه‌ها و استفاده از داده‌های آموزش، یک رده‌بند ایجاد می‌شود که با توجه به دوزبانه بودن خوشه‌ها، برای داده‌های آزمون نیز قابل استفاده است.

روش پیشنهادی توانسته نسبت به سایر روش‌های بین‌زبانی وابستگی خود را به منابع ترجمه کاهش دهد. در این روش نیازی به ترجمه کل داده‌های یک زبان به زبان دیگر نیست و تنها از ترجمه کلمات استفاده می‌شود. همچنین در این روش تمام کلمات ترجمه نمی‌شود و می‌توان با ترجمه کردن تعداد محدودی از کلمات، به کارایی بالایی دست یافت. علاوه بر این، رده‌بند ایجاد شده برای رده‌بندی نظرات در هر دو زبان قابل استفاده می‌باشد و تنها به یکی از زبان‌ها اختصاص ندارد.

نتایج ارزیابی روش پیشنهادی نشان می‌دهد که در مقایسه با دو روش پایه انتخابی دارای اختلاف معنادار از لحاظ آماری است و این نکته بیان‌گر این است که توانسته نمایش بهتری از داده‌ها در دو زبان ارائه کند و از ارتباط بین دو دسته از ویژگی‌ها اطلاعات مفیدی استخراج کند. این روش همین‌طور که ذکر شد، از وابستگی کمی به منبع ترجمه برخوردار است و همچنین حساسیت کمی به تعداد خوشه‌های انتخابی از خود نشان داده است. علاوه بر این‌ها تعداد ویژگی‌های غیرمحوری به عنوان عامل تأثیرگذاری بر کارایی روش پیشنهادی تشخیص داده نشد.

در ادامه ابتدا در بخش ۲ مروری بر کارهای پیشین صورت گرفته در حوزه نظرکاوی صورت می‌گیرد و سپس در بخش ۳ روش پیشنهادی مطرح می‌شود و در بخش ۴ ارزیابی‌های انجام شده برای این روش ارائه می‌شود.

۲- کارهای پیشین

در این بخش به بررسی و مرور کارهایی که در گذشته در این حوزه صورت گرفته‌اند، پرداخته می‌شود. همان‌طور که قبلا ذکر شد، پژوهش‌های انجام‌گرفته در این حوزه را می‌توان به بخش‌ها و زیربخش‌هایی تقسیم کرد. پژوهش‌های بسیاری به تحلیل نظرات به صورت پایه‌ای می‌پردازند [۴-۱]. در [۱] با استفاده از اختلاف اطلاعات متقابل نظرات و یک کلمه مرجع[۷] مثبت و اطلاعات متقابل نظرات و یک کلمه مرجع منفی، به صورت بدون ناظر، به رده‌بندی نظرات پرداخته می‌شود.

روش دیگری در [۲] ارائه شده است که با استفاده از ویژگی‌های ۱-گرام، ۲-گرام و برچسب مقوله نحوی[۸]، سه روش یادگیری ماشین را با هم مقایسه می‌کند. در [۸] این ایده استفاده از روش‌های یادگیری ماشین بهبود می‌یابد. در این روش جدید ابتدا جملات بدون نظر از متن نظرات حذف می‌شود تا متن نظرات خلاصه‌تر و شفاف‌تر شود. سپس الگوریتم تحلیل نظرات پیشنهاد شده بر روی متن‌های جدید اجرا می‌شوند.

در [۹] تنها با استفاده از صفت‌ها، قیدها و فعل‌های موجود در نظرات، به پیش‌بینی جهت احساسی نظرات پرداخته می‌شود و در [۱۰] با استفاده از محاسبه اختلاف مدل زبانی[۹] نظر با مدل زبانی نظرات مثبت و نظرات منفی، شبیه‌ترین مدل زبانی به نظر شناسایی می‌شود. با توجه به بیان متفاوت احساسات در متن‌های رسمی و غیررسمی، در [۱۱] به بررسی متن‌های غیررسمی پرداخته می‌شود. در سال‌های اخیر کاربرد نمایش طیفی کلمات[۱۰] [۱۲] در تحلیل نظرات با در نظر گرفتن جهت احساسی متون در [۱۵-۱۳] بررسی می‌شود.

یک دسته از پژوهش‌ها در حوزه نظرکاوی، به بررسی نظرات بیان شده به شکل توئیت، در شبکه اجتماعی توئیتر[۱۱] می‌پردازد [۱۶، ۱۷]. توئیت‌ها کوتاه، غیررسمی و حاوی نمادهای خاص و اصطلاحات عامیانه هستند. این نظرات به دلیل تفاوت‌هایی که با نظرات بیان‌شده در تارنماهای مبتنی بر نظر دارند، لازم است روش‌هایی متناسب با خصوصیت‌هایشان ارائه و پیشنهاد شوند.

در ادامه، ابتدا در بخش ۱-۲ کارهای گذشته در حوزه بین‌زبانی مطرح می‌شوند، رویکردهای متفاوت روش‌های گوناگون بیان می‌شود و بعد از آن در بخش ۲-۲ حوزه بین‌دامنه‌ای بررسی می‌شود و خلاصه‌ای از روش‌های پیشنهاد شده در این حوزه معرفی می‌شود.

۲-۱- حوزه بین‌زبانی

کارهای انجام شده در حوزه بین‌زبانی نیز رویکردهای متفاوتی به مسأله دارد. برای مثال، در [۱۸] از یک لغت‌نامه احتمالاتی در سطح کلمه برای ترجمه نظرات در زبان چک به زبان انگلیسی استفاده می‌شود و مسأله بین‌زبانی به مسأله‌ای در حوزه تک‌زبانه تبدیل می‌شود. در [۱۹] با استفاده از سه نوع لغت‌نامه متفاوت و میزان گرایش معنایی موجود برای کلمات در زبان انگلیسی، گرایش معنایی برای کلمات اسپانیایی نیز به دست آورده می‌شود. در [۵] از ایده استفاده کردن از نظرات بدون برچسب استفاده می‌شود و به جای تنها ترجمه نظرات از یک زبان به زبان دیگر، نظرات بدون برچسب چینی به انگلیسی و همچنین نظرات برچسب‌دار انگلیسی به چینی توسط ماشین ترجمه، ترجمه می‌شود. از این رو نظرات بدون برچسب و با برچسب در هر دو زبان تولید می‌شود. درنتیجه رده‌بندی در هر دو زبان به دست می‌آید که با استفاده از ترکیب این دو رده‌بند، نظرات چینی و ترجمه شده آن‌ها به انگلیسی، برچسبی به نظرات زده می‌شود. در [۶] نیز این ایده گسترش می‌یابد که به بهبود کارایی روش منجر می‌شود.

همان‌طور که گفته شد، مشکل اساسی موجود در حوزه بین‌زبانی، اختلاف بین توزیع ویژگی‌ها در زبان‌های مختلف است و با ترجمه کردن داده‌های موجود در یک زبان دیگر به زبان مورد نظر، همچنان این اختلاف وجود دارد. روشی که در [۲۰] ارائه می‌شود، این اختلاف توزیع محاسبه می‌شود و تا حد امکان کاهش می‌یابد. برای این کار ویژگی‌هایی که بین دو توزیع احتمالاتی در دو زبان بیشترین اختلاف را دارد برای استنباط بهتر اختلاف دو توزیع، مفیدتر شناخته می‌شود. ایده این روش وزن‌دهی ویژگی‌ها و نمونه‌های موجود است. که سعی می‌شود به وسیله آن‌ها توزیع دو زبان هر چه بیش‌تر به هم شبیه‌تر شود. استفاده از نیروی انسانی راه دیگری برای افزایش کارایی روش‌های بین‌زبانی است که در [۲۱] مطرح می‌شود. در این روش نظرات بدون برچسب در زبان مورد نظر با استفاده از ماشین ترجمه به زبان مبدأ ترجمه می‌شود. این نظرات ترجمه شده با استفاده از رده‌بند حاصل از داده‌های برچسب‌خورده در زبان مبدأ رده‌بندی و سپس بهترین نمونه‌ها انتخاب می‌شود. این نمونه‌ها توسط یک خبره برچسب زده می‌شود و به داده‌های برچسب‌خورده در زبان مبدأ برای ایجاد یک رده‌بند بهتر اضافه می‌شود. این روند تا رسیدن به کارایی مطلوب تکرار می‌شود.

یکی از روش‌های تحلیل نظرات استفاده از واژه‌نامه[۱۲] است. در حوزه بین‌زبانی سعی می‌شود از یک واژه‌نامه در یک زبان برای برچسب‌گذاری نظرات در زبان دیگر استفاده شود. در [۲۲] یک نمونه از این روش‌ها ارائه می‌شود که از ارتباط درون‌زبانی و میان‌زبانی ویژگی‌ها برای ایجاد واژه‌نامه در زبان مقصد استفاده می‌شود.

در [۲۳] از پیکره موازی برای انتقال اطلاعات از یک زبان به زبان دیگر استفاده می‌شود. در این روش با استفاده از هم‌ترازی در سطح کلمات، حاشیه‌نویسی‌ها از زبان مبدأ به زبان مقصد منتقل می‌شود. رویکرد دیگری که در این حوزه مورد استفاده قرار می‌گیرد، استخراج اطلاعاتی است که به طور واضح و آشکار در داده‌ها قابل دریافت نمی‌باشد. تخصیص دیریشله نهفته[۱۳] از جمله روش‌هایی است که سعی در استخراج این‌گونه اطلاعات دارد. در [۲۴] با استفاده از این روش به همراه ماشین ترجمه، ارتباط میان ویژگی‌های دو زبان استخراج می‌شود و تحلیل نظرات بر روی جنبه‌های نظرات انجام می‌گیرد. در [۲۵] با استفاده از پیکره موازی هم‌ترازشده در سطح جمله، نمایش طیفی عبارت‌ها استخراج می‌شود و از این اطلاعات برای پیش‌بینی جهت معنایی نظرات بهره برده می‌شود.

روش ارائه شده در [۲۶، ۲۷]، حالت بین‌زبانی روشی است که ابتدا در حوزه بین‌دامنه‌ای مطرح شده بود [۲۸]. با ایجاد تغییرات، تطبیق‌ها و پیشنهاداتی که ارائه شد چالش‌های استفاده در حوزه بین‌زبانی برطرف شدند. مهم‌ترین تغییری که انجام گرفت، به استخراج معادل ویژگی‌های زبان مبدأ در زبان مقصد می‌توان اشاره کرد.

۲-۲- حوزه بین‌دامنه‌ای

در روش‌های ارائه شده در حوزه بین‌دامنه‌ای برای تشخیص و کاهش اختلاف موجود بین دو دامنه متفاوت تلاش می‌شود. روش پیشنهاد شده در [۲۸] علاوه بر نظرات برچسب‌خورده در زبان مبدأ از نظرات بدون برچسب در هر دو زبان نیز بهره می‌برد. این‌گونه نظرات را که می‌توان در مقیاس بزرگ‌تری نسبت به نظرات برچسب‌خورده جمع‌آوری کرد، علی‌رغم نداشتن برچسب، حاوی اطلاعات مفیدی است. در این مقاله از این نظرات برای استخراج کوواریانس دو دسته از ویژگی‌ها استفاده می‌شود. یک دسته از ویژگی‌ها که ویژگی‌های محوری نامیده می‌شود، ویژگی‌هایی جهت‌داری است که به مقدار خوبی نشان‌دهنده برچسب نظرات است. این ویژگی‌ها با استفاده از اطلاعات متقابل بین ویژگی و برچسب، استخراج می‌شود و سایر ویژگی‌ها ویژگی‌های غیرمحوری نامیده می‌شود. با محاسبه کوواریانس بین این دو دسته از ویژگی‌ها و کاهش ابعاد آن رده‌بندی نظرات انجام می‌شود.

در [۲۹] نیز از ایده به‌کارگیری نظرات بدون برچسب استفاده می‌شود. مشابه روش پیشنهاد شده در [۲۸] ویژگی‌ها به دو دسته تقسیم می‌شود، اما از تعریف متفاوتی استفاده شده و سعی شده ویژگی‌های وابسته به دامنه را از ویژگی‌های مستقل از دامنه جدا شود. ایده دیگری که مطرح می‌شود ایجاد یک گراف بین این دو دسته از ویژگی است. سپس با استفاده از یکی از روش‌های خوشه‌بندی گراف، ویژگی‌ها خوشه‌بندی می‌شود و با استفاده از نتیجه این خوشه‌بندی و نظرات برچسب‌خورده موجود در دامنه دیگر، برچسب‌گذاری نظرات در دامنه مورد نظر انجام می‌شود.

در [۳۰] نیز رویکرد مشابهی در پیش گرفته می‌شود. با استفاده از الگوریتم بهینه‌سازی پیشنهادی به استخراج کلمات مشترک و کلمات خاص هر دامنه پرداخته می‌شود.

روش پیشنهاد شده در این مقاله در حوزه کارهای بین‌زبانی قرار می‌گیرد و با استفاده از نظرات بدون برچسب در دو زبان مبدأ و مقصد، هم‌زمان به استخراج اطلاعات از زبان مبدأ و انتقال آن‌ها به زبان مقصد می‌پردازد. این روش با استفاده از یک لغت‌نامه و ترجمه تعداد کمی از کلمات می‌تواند این استخراج و انتقال را با کیفیت خوبی انجام دهد. مشابه ایده پیشنهاد شده در [۲۹] گرافی بین دو دسته از ویژگی‌ها ایجاد می‌شود و ویژگی‌های مشترکی برای کلمات کاملاً متفاوت دو زبان استخراج می‌شود. با ارائه ایده‌هایی که در ادامه به آن‌ها پرداخته می‌شود، برای چالش‌ها و محدودیت‌های موجود در این پژوهش راه‌حلی پیشنهاد شده است.

۳- نظرکاوی بین‌زبانی با استفاده از ویژگی‌های معنایی

نظرات موجود در سایت‌ها و صفحات شخصی، هم در دامنه‌های متفاوتی قرار می‌گیرد و هم دارای زبان‌های متفاوتی است. برای تحلیل نظرات به صورت باناظر، بهترین حالت استفاده از نظرات برچسب‌خورده در دامنه و زبان مشابه همان نظرات است. اما به دلائل گوناگونی که قبلاً نیز ذکر شد، امکان دارد این نظرات برچسب‌خورده در دامنه و زبان مورد نظر موجود نباشد، درحالی‌که در دامنه متفاوت و یا زبان متفاوتی، چنین نظراتی یافت می‌شود. از طرفی تولید نظرات برچسب‌خورده پرهزینه و زمان‌گیر است و از طرف دیگر، استفاده مستقیم از نظرات در دامنه و یا زبان دیگر، کارایی مطلوبی ندارد. بنابراین روش‌هایی برای قابل

استفاده شدن نظرات در دامنه و یا زبان دیگر، ارائه می‌شود تا چنین محدودیتی برای نظرات وجود نداشته باشد. روش‌های بین‌دامنه‌ای از نظرات برچسب‌خورده در دامنه متفاوت و زبان یکسان و روش‌های بین‌زبانی از نظرات برچسب‌خورده در دامنه یکسان و زبان متفاوت استفاده می‌کند.

روش‌های بین‌دامنه‌ای و بین‌زبانی علاوه بر تفاوت‌ها و رویکرد متفاوتی که با هم دارند، دارای اشتراکاتی نیز هستند. هر دو دسته از این روش‌ها از داده‌هایی استفاده می‌کنند که نسبت به داده‌های مورد نظرشان از توزیع متفاوتی برخوردار است و سعی می‌کنند مسأله تفاوت در توزیع دو دسته از داده را به گونه‌ای برطرف سازند. این نکته، این امکان را به ما می‌دهد که از روش‌های بین‌دامنه‌ای بتوانیم در حوزه بین‌زبانی نیز استفاده کنیم. اما روش‌های بین‌دامنه‌ای مزیتی نسبت به روش‌های بین‌زبانی دارد و آن یکسان بودن کلمات استفاده شده در دو توزیع است. درحالی‌که در روش‌های بین‌زبانی دو توزیع متفاوت از کلمات کاملا متفاوت استفاده می‌کند که برای به کار بردن روش‌های بین‌دامنه‌ای در حوزه بین‌زبانی، باید این مسأله در نظر گرفته و راه‌حلی برای رفع آن ارائه شود.

با توجه به نکات بیان‌شده، روش‌های بین‌دامنه‌ای به صورت مستقیم در حوزه بین‌زبانی قابل استفاده نیست. در این مقاله سعی شده روش‌هایی برای برطرف ساختن چالش‌های پیش‌رو برای کمک گرفتن از روش‌های بین‌دامنه‌ای پیشنهاد شود. روش پیشنهادی این مقاله، از یک روش بین‌دامنه‌ای موجود [۲۹] ایده گرفته است. این روش بین‌دامنه‌ای که هم‌ترازی ویژگی‌های طیفی[۱۴] نام دارد، راه‌حلی برای استفاده از نظرات برچسب‌خورده در یک دامنه مبدأ برای برچسب زدن نظرات در دامنه مقصد، به کمک نظرات بدون برچسب در هر دو دامنه، ارائه داده است. چگونگی تشخیص تمایز بین ویژگی‌ها در هر دو زبان و ایجاد گراف دوبخشی که از بخش‌های اصلی این روش به شمار می‌روند، چالش‌هایی بود که در این پژوهش بررسی و روشی برای برطرف ساختن آن‌ها پیشنهاد شد.

در ادامه ابتدا ایده روش هم‌ترازی ویژگی‌های طیفی به صورت مختصر توضیح داده می‌شود و بعد از آن ایده مطرح شده در این مقاله به صورت دقیق‌تر بررسی می‌شود.

۳-۱- روش هم‌ترازی ویژگی‌های طیفی

روش هم‌ترازی ویژگی‌های طیفی برای مسأله تحلیل نظرات بین‌دامنه‌ای مطرح شد. حوزه تحلیل نظرات بین‌دامنه‌ای از نظرات دارای برچسب در دامنه متفاوت و زبان یکسان نسبت به نظرات مورد نظر استفاده می‌کند و هدف کاهش فاصله بین دو دامنه مبدأ و مقصد است که این فاصله، توزیع متفاوت داده‌ها در دو دامنه مختلف است. برای مثال، در دامنه کتاب، کلماتی مثل طولانی، هیجان‌انگیز، یکنواخت، ... برای توصیف یک کتاب استفاده می‌شود، درحالی‌که در دامنه دیگری مانند موبایل این کلمات استفاده چندانی ندارد و در عوض کلماتی مثل طراحی، کیفیت، دوربین، ... برای بیان نظرات بیش‌تر به کار می‌رود که این کلمات هم در دامنه کتاب بسیار کم‌تر مشاهده می‌شود. این روش علاوه بر بهره‌بردن از نظرات دارای برچسب، از نظرات بدون برچسب نیز استفاده می‌کند. این نظرات را به دلیل عدم نیاز به برچسب‌گذاری می‌توان در مقیاس بزرگ‌تری جمع‌آوری کرد. در این روش سعی شده با استفاده از این نظرات، اطلاعات بیش‌تری از دامنه‌ها استخراج شود و از این اطلاعات برای شناسایی برچسب نظرات بهره برده شود.

در این روش ابتدا از بین ویژگی‌های موجود، ویژگی‌های مستقل از دامنه[۱۵] استخراج می‌شود (بخش ۳-۱-۱) و سپس بین این ویژگی‌ها و سایر ویژگی‌ها که ویژگی‌های وابسته به دامنه[۱۶] نامیده می‌شود، گراف دوبخشی تشکیل داده می‌شود (بخش ۳-۱-۲). سپس این گراف دوبخشی با استفاده از روشی که در بخش ۳-۱-۳ به طور کامل توضیح داده می‌شود، به صورت نرم خوشه‌بندی می‌شود که ویژگی‌ها با وزنی به هر خوشه اختصاص داده می‌شود. وزن هر ویژگی در هر خوشه، ویژگی طیفی نامیده می‌شود. با استخراج ویژگی‌های طیفی برای هر نظر و استفاده از یک رده‌بند، با توجه به دوزبانه بودن خوشه‌ها، برچسب نظرات در هر دو زبان را می‌توان پیش‌بینی کرد. در ادامه، هر بخش از این روش به طور دقیق‌تری توضیح داده می‌شود.

۳-۱-۱- انتخاب ویژگی‌های مستقل از دامنه

انتخاب ویژگی‌های مستقل از دامنه اصلی‌ترین و مهم‌ترین قسمت این روش است. این روش وابستگی بالایی به ویژگی‌های مستقل از دامنه انتخاب شده دارد، هر چه روش انتخابی بهتر باشد و ویژگی‌های مناسب‌تری انتخاب شود، فاصله میان دو دامنه به میزان بیش‌تری کاهش می‌یابد. این ویژگی‌ها، ویژگی‌هایی است که در هر دو دامنه مورد نظر پرتکرار و دارای رفتار مشابه است.

با توجه به خصوصیات ذکر شده برای ویژگی‌های مستقل از دامنه، ایده‌ای که در این روش مطرح شد، استفاده از اطلاعات متقابل[۱۷] بین ویژگی‌ها و دامنه‌ها برای استخراج ویژگی‌هایی با رفتار یکسان در دو دامنه است. هر چه این مقدار کم‌تر باشد، نشان‌دهنده استقلال ویژگی از دامنه است. در نتیجه پس از محاسبه اطلاعات متقابل برای تمام ویژگی‌ها، ویژگی‌هایی که کم‌ترین مقدار را دارد، به عنوان ویژگی‌های کاندیدا انتخاب می‌شود.

$$I(X;D)=\sum_{\substack{d\in D\\ D=\{s,t\}}}\sum_{\substack{x\in X\\ X=\{0,1\}}} p(x,d)\log_2\frac{p(x,d)}{p(x)p(d)} \qquad (۱)$$

در این‌جا، X دربردارنده وجود و عدم وجود ویژگی مورد نظر و D دربردارنده دامنه مبدأ و دامنه مقصد است. در نتیجه عبارت بالا برای چهار حالت مختلف محاسبه می‌شود.

پس از مشخص شدن ویژگی‌های کاندیدا، ویژگی‌هایی به عنوان ویژگی‌های مستقل از دامنه انتخاب می‌شود که تعداد رخدادشان از مقدار کمینه تعیین شده بیش‌تر باشد. در این صورت ویژگی‌های مستقل از دامنه، کلمات پراستفاده‌ای است که در هر دو دامنه رفتار مشابهی دارد. در نتیجه این ویژگی‌ها هم‌رخدادی زیادی با سایر ویژگی‌ها دارد و استخراج ویژگی‌های طیفی با دقت خوبی انجام می‌گیرد.

۳-۱-۲- ساخت گراف دوبخشی

در این مرحله، گراف دوبخشی ساخته می‌شود. پس از انتخاب ویژگی‌های مستقل از دامنه، سایر ویژگی‌ها به عنوان ویژگی‌های وابسته به دامنه انتخاب می‌شود. برای تحلیل نظرات بین‌دامنه‌ای و استفاده از برچسب نظرات در یک دامنه برای برچسب‌گذاری نظرات در دامنه دیگر نیاز به وجود نظرات در هر دو دامنه و خوشه‌بندی تمام ویژگی‌ها است. در نتیجه با استفاده از نظرات بدون برچسب در هر دو دامنه، یک گراف دوبخشی ایجاد می‌شود. یال‌های این گراف بین ویژگی‌های مستقل از دامنه و ویژگی‌های وابسته به دامنه است و دارای وزنی به نسبت تعداد هم‌رخدادی دو ویژگی مذکور در نظرات بدون برچسب است.

به عبارت دیگر، این گراف دوبخشی با استفاده از نظرات موجود در هر دو دامنه مبدأ و مقصد ساخته می‌شود. در یک بخش از این گراف ویژگی‌های مستقل از دامنه f_{DI} و در بخش دیگر آن ویژگی‌های وابسته به دامنه f_{DS} قرار دارد. با توجه به تعریف مسأله بین‌دامنه‌ای، تمام این ویژگی‌ها در یک زبان مشترک است و ویژگی‌های وابسته به دامنه در دو دامنه با توزیع متفاوتی دیده شده است. وزن یال بین این دو دسته از ویژگی‌ها به صورت زیر محاسبه می‌شود:

$$w_{ij} = c(S_s, f_{DI_i} \cup f_{DS_j}) + c(S_t, f_{DI_i} \cup f_{DS_j}) \quad (۲)$$

که در این‌جا، $c(S_s, f_{DI_i} \cup f_{DS_j})$ تعداد نظراتی در دامنه مبدأ S_s است که هم iامین ویژگی مستقل از دامنه f_{DI_i} و هم jامین ویژگی وابسته به دامنه f_{DS_j} در آن مشاهده شده باشد. و $c(S_t, f_{DI_i} \cup f_{DS_j})$ تعداد نظرات در دامنه مقصد S_t است که هر دو ویژگی f_{DI_i} و f_{DS_j} در آن رخ داده باشد.

۳-۱-۳- استخراج ویژگی‌های طیفی

همان‌طور که در بخش‌های قبلی توضیح داده شد، با تعریف دو محدودیت برای ویژگی‌های مستقل از دامنه، کلماتی که محدودیت‌ها را برآورده می‌کند کاندیدا برای این دسته از ویژگی‌ها است. اگر فرض کنیم تعداد ویژگی‌های مستقل از دامنه p باشد، p ویژگی پرتکرار با کمترین مقدار اطلاعات متقابل بین ویژگی و دامنه، به عنوان ویژگی‌های مستقل از دامنه انتخاب می‌شود و پس از آن گراف دوبخشی بین ویژگی‌های مستقل از دامنه و ویژگی‌های وابسته به دامنه ساخته می‌شود. از آن‌جایی که در این مراحل از برچسب نظرات استفاده‌ای نمی‌شود، انتخاب ویژگی‌های مستقل از دامنه و ساخت گراف دوبخشی با استفاده از نظرات بدون برچسب صورت می‌گیرد.

اکنون به بررسی روند استخراج ویژگی‌های طیفی پرداخته می‌شود. پس از ساخت گراف دوبخشی بین ویژگی‌های مستقل از دامنه و ویژگی‌های وابسته به دامنه، به سادگی می‌توان ماتریس یال‌های گراف حاصل را تشکیل داد. اگر تعداد کل ویژگی‌ها n و تعداد ویژگی‌های مستقل از دامنه p باشد، ماتریس یال‌های گراف، که M خوانده می‌شود، دارای ابعاد $p \times (n-p)$ خواهد بود. هم‌چنین ماتریس مجاورت این گراف که یک ماتریس مربعی است، ماتریسی به ابعاد $n \times n$ خواهد بود که:

$$A = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \quad (۳)$$

این ماتریس مجاورت A، حاوی وزن خام یال‌ها است. ایده مطرح شده در این الگوریتم، استخراج ویژگی‌های طیفی از ماتریس لاپلاس گراف، به جای ماتریس مجاورت گراف است. فرمول (۴) روش محاسبه ماتریس لاپلاس L را نشان می‌دهد. با محاسبه لاپلاس ماتریس مجاورت، وزن‌های خام به وزن‌های نرمال شده تبدیل می‌شود که این وزن‌های نرمال شده به صورت بهتری نشان‌دهنده میزان اهمیت یال‌ها است.

$$L = D^{-\frac{1}{2}} \times A \times D^{-\frac{1}{2}} \quad (۴)$$

که D در فرمول بالا، ماتریس درجه گراف و یک ماتریس قطری است. درایه‌های قطری ماتریس، درجه رأس متناظر آن درایه است و درایه‌های غیرقطری آن برابر با صفر است.

$$d_{ii} = \sum_j a_{ij} \quad (۵)$$

بعد از محاسبه ماتریس لاپلاس گراف دوبخشی، ویژگی‌های طیفی استخراج می‌شوند که برای استخراج آن‌ها از تجزیه مقدار منفرد[۱۸] استفاده شده است که یکی از روش‌های خوشه‌بندی نرم[۱۹] محسوب می‌شود.

$$U \Sigma V^T = SVD(L) \quad (۶)$$

تجزیه مقدار منفرد با ایجاد تعداد زیادی خوشه، وزن هر ویژگی در هر خوشه را در ماتریس U ذخیره می‌کند و همچنین ماتریس Σ یک ماتریس قطری است که دربردارنده میزان قوت هر خوشه است.

اگر ویژگی‌های مستقل از دامنه به درستی انتخاب شده باشد و خصلت استقلال از دامنه را داشته باشد، ویژگی‌های وابسته به دامنه نیز به درستی توسط تجزیه مقدار منفرد، به صورت وزن‌دار، خوشه‌بندی می‌شود. در نتیجه برای رده‌بندی می‌توان تنها از ویژگی‌های طیفی استخراج شده ویژگی‌های وابسته به دامنه استفاده کرد. به عبارتی دیگر ویژگی‌های طیفی همان وزن ویژگی‌ها در خوشه‌های ایجاد شده است. با توجه به میزان قوت خوشه‌ها، تنها k تا از بهترین خوشه‌ها برای تشخیص وزن هر نظر در هر خوشه انتخاب می‌شود. این خوشه‌ها که میزان درستی ویژگی‌های طیفی آن‌ها بالاتر است، برای تحلیل نظرات کافی است.

$$\theta = U_{[p:n,1:k]} \in \mathbb{R}^{(n-p)\times k} \quad (۷)$$

با استفاده از θ که حاوی ویژگی‌های طیفی است و نسبت به ویژگی‌های مستقل از دامنه و وابسته به دامنه تعداد بسیار کم‌تری دارد، می‌توان فضای نظرات را به فضای طیفی منتقل کرد و نظرات را در فضای جدید رده‌بندی کرد.

اگر s یک نظر در فضای ویژگی‌های وابسته به دامنه باشد که دارای ابعاد $1 \times (n-p)$ است، s' همان نظر در فضای ویژگی‌های طیفی است و ابعاد آن $1 \times k$ خواهد بود که با توجه به ویژگی‌های استفاده‌شده وزنی در هر خوشه دارد.

$$s' = s \times \theta \quad (۸)$$

در نتیجه، به جای مدل کردن کلمات نظرات، ویژگی‌های طیفی نظرات را استخراج کرده و با استفاده از آن‌ها رده‌بندی صورت می‌گیرد. همان‌طور که قبلاً بیان شد، تعداد ویژگی‌های طیفی نسبت به ویژگی‌ها مستقل از دامنه و وابسته به دامنه، بسیار کم‌تر است و در نتیجه از میزان تنک[۲۰] بودن فضای مسأله بسیار کاسته می‌شود و می‌توان انتظار داشت که رده‌بندی با دقت بالاتری حاصل شود.

۳-۲- روش پیشنهادی

مسأله مورد بررسی در این مقاله مسأله تحلیل نظرات بین‌زبانی است. این حوزه به بررسی روش‌هایی می‌پردازد که می‌خواهد از داده‌های موجود در زبان دیگری برای رده‌بندی داده‌ها در زبان مورد نظرشان استفاده کند. این داده‌ها که معمولاً در یک دامنه است، ویژگی‌های کاملاً متفاوتی با یکدیگر دارد. به عبارت دیگر دو زبان مختلف به علت استفاده از کلمات متفاوت، دارای ویژگی‌های مختلفی است. اما با استفاده از منابع ترجمه، مانند لغت‌نامه، ماشین ترجمه و یا پیکره‌های دوزبانه می‌توان ویژگی‌های دو زبان را به هم نگاشت کرد. در عین حال باید به این نکته توجه داشت که منابع ترجمه حاوی ترجمه‌های مبهم و گاهی نادرست است.

در روش پیشنهادی از روش هم‌ترازی ویژگی‌های طیفی [۲۹] برای حل مسأله تحلیل نظرات بین‌زبانی کمک گرفته شده است. اما به دلیل عدم اشتراک بین ویژگی‌های دو زبان، تعریف و روش محاسبه‌ای که برای ویژگی‌های مستقل از دامنه ارائه شده در حوزه تحلیل نظرات بین‌زبانی قابل استفاده نیست. بنابراین برای محاسبه و شناسایی یک دسته از ویژگی‌ها که نقشی مشابه نقش ویژگی‌های مستقل از دامنه داشته باشد، به معرفی تعریف و روش جدیدی نیاز است. در این مقاله این ویژگی‌ها، ویژگی‌های محوری[۲۱] نامیده شد. ویژگی‌های محوری در این‌جا

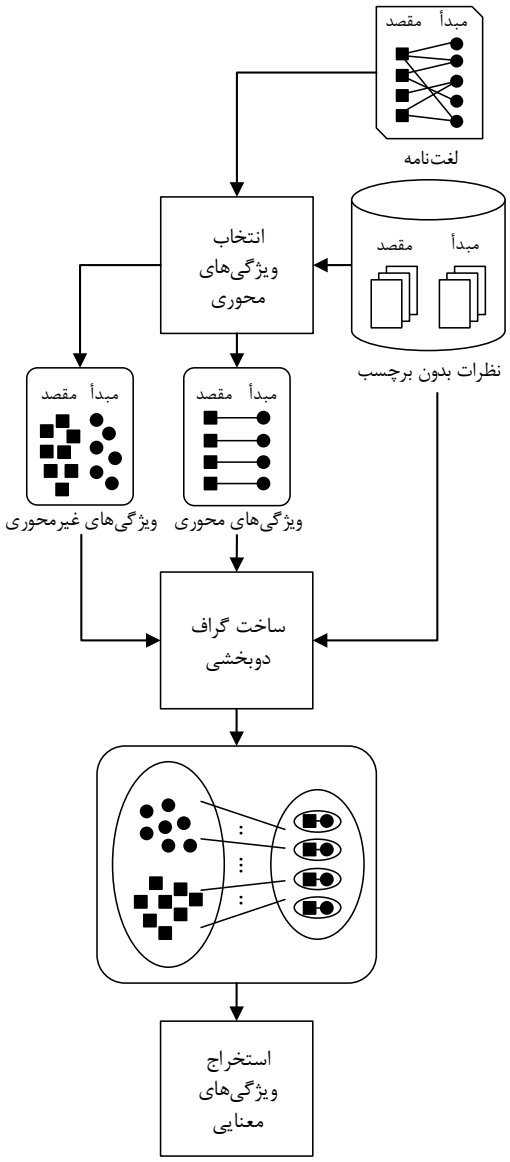
جفت‌کلمه‌هایی است که به عنوان ویژگی‌های مستقل از زبان تعریف شده است و استقلال از زبان، عدم وجود ابهام در ترجمه ویژگی در نظر گرفته شده است. برای شناسایی چنین ویژگی‌هایی احتمال استقلال از زبان برای تمام جفت‌کلمه‌های موجود در منبع ترجمه محاسبه می‌شود. جفت‌کلمه‌هایی که ترجمه‌شان مبهم در نظر گرفته می‌شود جزء دسته دوم یعنی ویژگی‌های غیرمحوری قرار می‌گیرد. از آن‌جایی که وجود ابهام در ترجمه نشان‌دهنده نامناسب بودن این هم‌ترازی‌ها است، راهکاری برای این مسأله نیز ارائه شد. با حذف هم‌ترازی‌های ویژگی‌های غیرمحوری و به عبارتی ترجمه نکردن این ویژگی‌ها، ویژگی‌های غیرمحوری تبدیل به تک‌کلمه‌هایی شد که شامل ویژگی‌هایی از زبان مبدأ و مقصد می‌باشد که در ویژگی‌های محوری جای نگرفته است.

به طور خلاصه، در روش پیشنهادی تعدادی از جفت‌کلمه‌های موجود در منبع ترجمه به عنوان ویژگی‌های محوری انتخاب می‌شود و سایر ویژگی‌های دو زبان به صورت تک‌کلمه در دسته ویژگی‌های غیرمحوری قرار می‌گیرد. ارتباط میان این دو دسته از ویژگی‌ها به کمک ایجاد یک گراف دوبخشی جمع‌آوری می‌شود. سپس از این ارتباطات به دست آمده، ویژگی‌های معنایی استخراج می‌شود و در نهایت نظرات رده‌بندی می‌شود. در این روش علاوه بر نظرات برچسب‌خورده در زبان مبدأ، از نظرات بدون برچسب در هر دو زبان نیز استفاده شده است. این نظرات که به صرف هزینه و زمان برای تعیین برچسب نیاز ندارد، برای جمع‌آوری به وقت و هزینه کم‌تری نسبت به داده‌های آموزش و آزمون احتیاج دارد و در عین حال به دلیل موجود بودن در هر دو زبان به کسب اطلاعات لازم برای تحلیل نظرات بین‌زبانی کمک می‌کند. در شکل ۱ فرآیند استخراج ویژگی‌های معنایی در روش پیشنهادی قابل مشاهده است.

۳-۲-۱- انتخاب ویژگی‌های محوری

ویژگی‌های محوری نقش مشابهی با ویژگی‌های مستقل از دامنه در روش هم‌ترازی ویژگی‌های طیفی دارد. همان‌طور که ویژگی‌های مستقل از دامنه برای انتقال اطلاعات از یک دامنه به دامنه دیگر مورد استفاده قرار می‌گرفت، ویژگی‌های محوری نیز وظیفه انتقال اطلاعات از یک زبان به زبان دیگری را دارد. ویژگی‌های محوری تنها ویژگی‌هایی است که با استفاده از منبع ترجمه، به زبان دیگر ترجمه و به عبارتی با ویژگی‌ای در آن زبان هم‌تراز می‌شود. بنابراین می‌توان ادعا کرد که از این ویژگی‌ها به عنوان پل ارتباطی بین دو زبان استفاده می‌شود. اگر کلمات مناسبی به عنوان ویژگی‌های محوری انتخاب نشود، این ارتباط میان دو زبان به درستی ساخته نمی‌شود و در نتیجه اطلاعات موجود در یک زبان به زبان دیگر به خوبی منتقل نمی‌شود.

برای شناسایی ویژگی‌های محوری، لازم به تعریف محدودیت‌هایی است که نشان‌دهنده خصلت استقلال از زبان باشد و کلماتی که این محدودیت‌ها را برآورده می‌کند، پل ارتباطی خوبی بین دو زبان برقرار کند. به همین منظور پرتکرار و رایج بودن ویژگی به عنوان یک محدودیت برای ویژگی‌های محوری تعریف شد. هر چه یک ویژگی در نظرات بیش‌تری استفاده شده باشد، می‌توان فرض کرد که با تعداد ویژگی‌های متمایز بیش‌تری هم‌رخدادی و ارتباط دارد. دومین محدودیت، مستقل بودن ویژگی از زبان است که از تعاریف اصلی ویژگی‌های محوری به شمار می‌رود. ایده‌ای که در این‌جا برای شناسایی چنین ویژگی‌هایی مطرح می‌شود، ایجاد ارتباط میان ویژگی‌های دو زبان است. وجود چنین ارتباطی میان دو ویژگی از دو زبان متفاوت، نشان‌گر هم‌تراز بودن این ویژگی‌ها در دو زبان است. از منابعی که حاوی ارتباطاتی میان ویژگی‌های دو زبان متفاوت باشد، می‌توان به منابع ترجمه اشاره کرد. در منابع ترجمه، این ارتباطات به صورت ترجمه یک ویژگی در زبان مبدأ به یک یا چند ویژگی در زبان مقصد تعریف می‌شود.



شکل ۱- فرآیند استخراج ویژگی‌های معنایی

منابع ترجمه متفاوتی در زبان‌های مختلف وجود دارد. لغت‌نامه ساده‌ترین و ماشین ترجمه جزء پیچیده‌ترین منابع ترجمه به شمار می‌رود. روش پیشنهادی که مستقل از نوع منبع ترجمه مورد استفاده است، تنها به یک نگاشت کلمات در زبان مبدأ به کلمات در زبان مقصد احتیاج دارد. در نتیجه لغت‌نامه که در اکثر جفت‌زبان‌ها، حتی زبان‌های با منابع محدود موجود است، مورد استفاده قرار گرفته است.

با استفاده از منابع ترجمه، جفت‌کلمه‌هایی مانند (x_s, x_t) به دست می‌آید. اولی کلمه‌ای در زبان مبدأ و دومی کلمه‌ای در زبان مقصد است که یک هم‌ترازی بین این دو کلمه ایجاد شده است. ویژگی‌های محوری مجموعه‌ای از همین جفت‌کلمه‌ها است. این نکته حائز اهمیت است که در منابع ترجمه برای یک کلمه ممکن است چند ترجمه وجود داشته باشد و یا یک کلمه ترجمه چند کلمه باشد. همچنین این ترجمه‌ها ممکن است دارای خطا باشند و یا برخی از هم‌ترازی‌ها تنها در موارد خاصی برقرار باشد. در روش پیشنهاد شده برای انتخاب ویژگی‌های محوری، سعی شده این مسائل در نظر گرفته شود و بهترین و بدون‌ابهام‌ترین ترجمه‌ها انتخاب شود.

برای انتخاب ویژگی‌های محوری، محدودیت‌های تعریف شده بر روی جفت‌کلمه‌های موجود اعمال می‌شود. جفت‌کلمه‌هایی محدودیت اول را برآورده

می‌کند که هر دو کلمه در زبان خود پرتکرار باشد در نتیجه کلماتی که از حد آستانه تعریف شده، کم‌تر ظاهر شده باشد، به همراه زوجشان، از لیست کاندیداهای ویژگی‌های محوری حذف می‌شود. محدودیت دوم استقلال ویژگی از زبان است. استقلال ویژگی از زبان تشابه رفتار جفت‌کلمه یعنی رخ دادن و رخ ندادن دو کلمه در نظرات متناظرشان تعریف شده است. ویژگی‌هایی که دو کلمه آن در زبان متناظرشان، رفتار مشابهی با یکدیگر داشته باشد این محدودیت را برآورده می‌کند. این رفتار مشابه و عدم وابستگی بین ویژگی و زبان را می‌توان با استفاده از فرمول اطلاعات متقابل محاسبه کرد. این فرمول در ذیل نشان داده شده است [۳۱].

$$I(X;L)=\sum_{e_x\in\{0,1\}}\sum_{e_l\in\{s,t\}}p(e_x,e_l)\log\frac{p(e_x,e_l)}{p(e_x)p(e_l)}$$
$$=\frac{N_{11}}{N}\log\frac{NN_{11}}{N_{1.}N_{.1}}+\frac{N_{10}}{N}\log\frac{NN_{10}}{N_{1.}N_{.0}}$$
$$+\frac{N_{01}}{N}\log\frac{NN_{01}}{N_{0.}N_{.1}}+\frac{N_{00}}{N}\log\frac{NN_{00}}{N_{0.}N_{.0}} \quad (۹)$$

در فرمول بالا، X ویژگی و یا همان جفت‌کلمه مورد نظر و L زبان است که اگر e_x برابر با یک باشد، احتمال وجود و اگر برابر با صفر باشد، احتمال عدم وجود ویژگی مد نظر است که برای هر زبان، کلمه متناظرش در جفت‌کلمه، برای محاسبه احتمال در نظر گرفته می‌شود. به صورت دقیق‌تر، N_{11} تعداد نظرات حاوی x_s در زبان مبدأ، N_{10} تعداد نظرات حاوی x_t در زبان مقصد، N_{01} تعداد نظرات بدون x_s در زبان مبدأ و N_{00} تعداد نظرات بدون x_t در زبان مقصد است. $N_{1.}$ مجموع N_{11} و N_{10}، $N_{0.}$ مجموع N_{01} و N_{00}، $N_{.1}$ مجموع N_{11} و N_{01} و $N_{.0}$ مجموع N_{10} و N_{00} است. N نیز تعداد کل نظرات در زبان مبدأ و زبان مقصد است.

این فرمول میزان وابستگی دو متغیر ویژگی و زبان را محاسبه می‌کند، در نتیجه هر چه مقدار این احتمال کم‌تر باشد دو متغیر وابستگی کم‌تر و استقلال بیش‌تری نسبت به هم دارند. بنابراین کلماتی به عنوان ویژگی‌های محوری انتخاب می‌شود که کم‌ترین مقادیر اطلاعات متقابل را داشته باشد. در این صورت می‌توان فرض کرد این کلمات علاوه بر استقلال از زبان، دارای کیفیت بالای ترجمه و بدون ابهام در دامنه مورد نظر نیز است.

در صورتی که یک کلمه در زبان مبدأ دارای چند ترجمه مختلف در زبان مقصد باشد، با هر کدام از این ترجمه‌ها، یک جفت‌کلمه تشکیل می‌دهد که هر کدام از این جفت‌کلمه‌ها مقدار اطلاعات متقابل متفاوتی خواهند داشت. جفت‌کلمه‌ای دارای کم‌ترین مقدار است که رفتار آن ترجمه در زبان مقصد به رفتار کلمه در زبان مبدأ نزدیک‌تر باشد. به عبارت دیگر آن ترجمه در دامنه و زبان مورد بررسی، بهترین معادل برای کلمه مورد نظر می‌باشد. در این صورت جفت‌کلمه‌هایی که به عنوان ویژگی‌های محوری انتخاب می‌شوند را می‌توان یک مفهوم غیرمبهم در هر دو زبان در نظر گرفت.

همان‌طور که مشاهده می‌شود، در فرمول (۹) برای محاسبه اطلاعات متقابل، از برچسب نظرات استفاده‌ای نمی‌شود، در نتیجه می‌توان برای انتخاب ویژگی‌های محوری از نظرات بدون برچسب که تعداد بیش‌تری از آن‌ها در اختیار است، استفاده کرد.

۳-۲-۲- ساخت گراف دوبخشی

اکنون ویژگی‌های محوری انتخاب شده است. این ویژگی‌ها رأس‌های یک بخش از گراف دوبخشی را تشکیل می‌دهد و بخش دیگر این گراف از ویژگی‌های غیرمحوری تشکیل می‌شود. همان‌طور که گفته شد، ویژگی‌های محوری به صورت جفت‌کلمه (x_s,x_t) است و به عبارت دیگر یک ویژگی محوری از یک کلمه در زبان مبدأ و یک کلمه در زبان مقصد (که ترجمه کلمه اول است) تشکیل شده است. اما در مقابل ویژگی‌های غیرمحوری به صورت تک‌تک و جدا است، یعنی یک ویژگی غیرمحوری یا یک کلمه در زبان مبدأ y_s و یا یک کلمه در زبان مقصد y_t است که این ویژگی‌ها به ترجمه و هم‌ترازی با ویژگی دیگر در زبان دیگر نیازی ندارد. این خصوصیت موجب بالا رفتن کیفیت روش نسبت به سایر روش‌هایی است که نیاز به ترجمه تک‌تک کلمات دارد. در نتیجه منبع ترجمه‌ای که بتواند کلمه‌ها را ترجمه کند برای این روش کافی است و نیازی به ترجمه جمله در این‌جا وجود ندارد.

پس از مشخص شدن رأس‌های گراف دوبخشی، یال‌های گراف ایجاد می‌شود. گراف دوبخشی، یک گراف وزن‌دار غیرمنفی است که وزن یال‌های آن متناسب با تعداد هم‌رخدادی دو رأس آن در نظرات زبان مبدأ و زبان مقصد است. هم‌رخدادی در این‌جا، وقوع یک کلمه در همسایگی کلمه دیگر است که این همسایگی می‌تواند به صورت یک پنجره اطراف یک کلمه تعریف شود. اندازه این پنجره می‌تواند از ۱ (دو کلمه‌ای هم‌رخدادی دارند که در مجاورت هم در متن ظاهر شده باشند) تا طول سند (هر دو کلمه‌ای که در یک سند ظاهر شده‌اند، هم‌رخدادی دارند) متغیر باشد. در این‌جا اندازه پنجره، طول سند در نظر گرفته شده است. نظری که حاوی یک ویژگی محوری است، وزن یال متناظرش با ویژگی‌های غیرمحوری موجود در همان نظر را یک واحد افزایش می دهد.

بدیهی است که میان ویژگی‌های محوری و همچنین میان ویژگی‌های غیرمحوری یالی وجود ندارد و یال‌ها تنها میان این دو دسته از ویژگی‌ها برقرار می‌شود. طبق توضیحات داده شد، iامین ویژگی محوری را می‌توان مطابق فرمول (۱۰) نشان داد. وزن یال میان iامین ویژگی محوری و jامین ویژگی غیرمحوری، بسته به این‌که ویژگی غیرمحوری مورد نظر متعلق به کدام زبان باشد، به کمک یکی از دو فرمول (۱۱) و (۱۲) محاسبه می‌شود. اگر ویژگی غیرمحوری متعلق به زبان مبدأ باشد، از فرمول (۱۱) و در غیر این صورت از فرمول (۱۲) استفاده می‌شود.

$$f_{p_i}=(x_{s_i},x_{t_i}) \quad (۱۰)$$

$$w_{ij}=c(x_{s_i}\cup f_{n_j},S_s) \quad (۱۱)$$

$$w_{ij}=c(x_{t_i}\cup f_{n_j},S_t) \quad (۱۲)$$

که $c(x_{s_i}\cup f_{n_j},S_s)$ تعداد نظراتی در زبان مبدأ است که دارای هر دو کلمه x_{s_i} و f_{n_j} باشد. x_{s_i} کلمه متعلق به زبان مبدأ در ویژگی محوری iام و f_{n_j} ویژگی محوری jام است و $c(x_{t_i}\cup f_{n_j},S_t)$ تعداد نظرات در زبان مقصد، دارای دو کلمه x_{t_i} و f_{n_j} است.

در این مرحله نیز به علت عدم به‌کارگیری برچسب نظرات، می‌توان از نظرات بدون برچسب استفاده کرد. نظرات بدون برچسب در زبان مبدأ، یال‌های بین ویژگی‌های محوری (x_s,x_t) و ویژگی‌های غیرمحوری در زبان مبدأ y_s و نظرات بدون برچسب در زبان مقصد، یال‌های بین (x_s,x_t) و y_t را می‌سازد. به طور دقیق‌تر، وزن یال بین (x_s,x_t) و y_s، میزان هم‌رخدادی x_s و y_s است و وزن یال بین (x_s,x_t) و y_t، میزان هم‌رخدادی x_t و y_t. در این صورت یک ویژگی محوری که متشکل از کلمات هر دو زبان است با ویژگی‌های غیرمحوری در هر دو زبان یال مشترک دارد. از این رو، در گراف حاصل نقش ویژگی‌های غیرمحوری زبان مبدأ از ویژگی‌های غیرمحوری زبان مقصد، متمایز نیست.

۳-۲-۳- استخراج ویژگی‌های معنایی

با به‌کارگیری دو محدودیت، ویژگی‌های محوری انتخاب و گراف ارتباطشان با ویژگی‌های غیرمحوری ساخته شد. به دلیل دوزبانه بودن گراف و انجام خوشه‌بندی در فضای معنایی، ویژگی‌های استخراجی از این ماتریس ویژگی‌های معنایی نامیده شد. این ویژگی‌ها از نگاه دیگر حاوی مفاهیم کلی‌تر و سطح بالاتری است که به عبارتی معنای خاصی را دربردارد. این ویژگی‌ها بیان می‌کند هر کلمه چه میزان تعلقی به هر خوشه دارد. در ادامه روند استخراج ویژگی‌های معنایی برای مسأله بین‌زبانی توضیح داده می‌شود.

پس از ساخت ماتریس مجاورت طبق فرمول (۳)، ماتریس لاپلاس گراف با استفاده از فرمول (۴) محاسبه می‌شود. همان‌طور که گفته شد، این مرحله وزن‌های خام گراف را به وزن‌های نرمال شده تبدیل می‌کند و در نتیجه این وزن‌های جدید معنی‌دار است. برای مثال اگر در ماتریس اولیه A وزن یال بین f_{p_m} و f_{n_i} با وزن یال بین f_{p_m} و f_{n_j} مساوی و برابر w باشد و f_{n_i} ویژگی پرکاربردتری نسبت به f_{n_j} باشد، در ماتریس L وزن این دو یال برابر نیست و وزن یال f_{p_m} با f_{n_i} کم‌تر از وزن یالش با f_{n_j} است. زیرا مقدار w برای دو ویژگی f_{n_i} و f_{n_j} مفاهیم متفاوتی دارد و برای f_{n_j} حاوی اطلاعات بیش‌تری است.

اکنون با استفاده از فرمول (۶) تمام ویژگی‌های محوری و غیرمحوری، خوشه‌بندی می‌شود که خوشه‌های حاصل شامل جفت‌کلمه‌ها و تک‌کلمه‌های زبان مبدأ و زبان مقصد است. خصلت مهم این خوشه‌ها دوزبانه بودن آن‌ها است. اگر خوشه‌ها را فضای معنایی مستقل از زبان تعریف کنیم، وزنی که هر ویژگی محوری و یا غیرمحوری در هر خوشه دارد ویژگی معنایی آن محسوب می‌شود. با استفاده از فرمول (۸) می‌توان نظرات را از فضای ویژگی‌های محوری و غیرمحوری به فضای معنایی منتقل کرد که با توجه به ویژگی‌های معنایی، نظرات در دو زبان مبدأ و مقصد، ویژگی‌های مشترکی خواهد داشت.

پس از محاسبه میزان اختصاص هر نظر به هر خوشه، با استفاده از یک رده‌بند، تأثیر هر خوشه بر برچسب نظرات آموخته می‌شود. این مرحله بر روی نظرات برچسب‌خورده در زبان مبدأ که داده‌های آموزش به حساب می‌آید، صورت می‌گیرد. از اطلاعات آموخته شده، برای برچسب‌گذاری نظرات در زبان مقصد که داده‌های آزمون محسوب می‌شود، استفاده می‌شود. لازم به ذکر است که به دلیل انتقال نظرات به فضای معنایی و تعریف ویژگی‌ها مشترک برای دو زبان، رده‌بندی به صورت مستقیم و مشابه حالت تک‌زبانه انجام می‌گیرد.

۴- ارزیابی

در این بخش، آزمایش‌ها انجام شده برای ارزیابی روش پیشنهادی ارائه می‌شود. ابتدا مجموعه داده‌ای مورد استفاده و سپس روش پایه‌ای که برای مقایسه میزان کارایی روش پیشنهادی به کار گرفته شده معرفی می‌شود و در نهایت جزئیات آزمایش‌ها صورت گرفته بیان می‌شود.

۴-۱- مجموعه داده‌ها

مجموعه داده‌ای مورد استفاده، بخشی از داده‌های جمع‌آوری شده توسط پریتنهوفر[۲۲] [۲۷] است. این مجموعه داده‌ای به صورت تقریبی حاوی ۸۰۰ هزار نظر، از نظرات محصولات آمازون برای سه دسته از محصولات کتاب، موسیقی و دی‌وی‌دی است. همچنین این نظرات در چهار زبان انگلیسی، آلمانی، فرانسوی و ژاپنی موجود می‌باشد.

در آزمایش‌ها انجام شده زبان انگلیسی به عنوان زبان مبدأ و زبان آلمانی به عنوان زبان مقصد در نظر گرفته شد. همچنین از نظرات موجود در دامنه کتاب برای آزمایش‌ها استفاده شد. آمار این مجموعه داده‌ای در جدول ۱ قابل مشاهده است.

در داده‌های جمع‌آوری‌شده، هر کاربر علاوه بر ارائه نظر خود برای محصول مورد نظر، یک امتیاز از ۱ تا ۵ نیز به محصول داده است که برچسب نظرات با استفاده از همین امتیاز مشخص می‌شود. به نظرات دارای امتیاز ۴ تا ۵، برچسب مثبت و به نظرات دارای امتیاز ۱ تا ۲، برچسب منفی داده شده است. نظرات با امتیاز ۳ نیز برچسب خنثی تعلق می‌گیرد که از مجموعه داده‌ای حذف شده است. آزمایش‌ها بر روی نظرات مثبت و منفی صورت گرفته است. در نتیجه در این مجموعه داده‌ای اسناد موجود نظرمند هستند.

جدول ۱- آمار مجموعه داده‌ها در دامنه کتاب

	بدون برچسب	نظرات مثبت	نظرات منفی
انگلیسی	۵۰,۰۰۰	۲,۰۰۰	۲,۰۰۰
آلمانی	۱۶۵,۴۵۷	۲,۰۰۰	۲,۰۰۰

از نظرات دارای برچسب انگلیسی برای ساخت مدل رده‌بند استفاده می‌شود و برای تست این مدل، نظرات دارای برچسب آلمانی مورد استفاده قرار گرفته است.

برای منبع ترجمه لغت‌نامه گوگل انگلیسی به آلمانی[۲۳] (بدون استفاده از ماشین ترجمه) انتخاب شده است که احتمالات ترجمه در این آزمایش‌ها در نظر گرفته نشده است. در این لغت‌نامه هر کلمه انگلیسی، یک یا چند ترجمه آلمانی دارد که برای هر کلمه محتمل‌ترین ترجمه از مجموعه ترجمه‌ها، انتخاب شده است. لغت‌نامه منتخب حاوی ۲۶,۱۲۸ کلمه انگلیسی و ۱۷,۳۵۰ کلمه آلمانی است.

۴-۲- روش پایه

برای ارزیابی روش پیشنهادی دو روش پایه برای مقایسه انتخاب شد تا میزان کارایی و بهبود آن بهتر نمایش داده شود. یکی از روش‌های پایه انتخابی، روش یادگیری تناظرات ساختاری بین‌زبانی[۲۴] (CL-SCL) [۲۷] است. این روش که ابتدا برای حوزه بین‌دامنه‌ای مطرح شده بود [۲۸] و بعداً در حوزه بین‌زبانی مورد استفاده قرار گرفت، می‌تواند معیار خوبی برای ارزیابی روش پیشنهادی این مقاله باشد. در این روش نیز تعدادی کلمه به عنوان ویژگی‌های محوری انتخاب می‌شود و سپس تناظرات این ویژگی‌ها با سایر ویژگی‌ها محاسبه می‌شود که از تناظرات محاسبه‌شده برای پیش‌بینی رخداد ویژگی‌های محوری استفاده شده است. به عبارتی دیگر، برای هر جفت ویژگی محوری و غیرمحوری، وزنی استخراج می‌شود که این مقدار با همبستگی دو ویژگی نسبت مستقیم دارد. برای اجرای این روش از کد حاصل از مقاله پریتنهوفر [۲۷] که به صورت آزاد در اختیار قرار گرفته بود، استفاده شد.

روش پایه دیگری که مورد بررسی قرار گرفت، روش مبتنی بر مدل زبانی بین‌زبانی است که کارایی آن برای حوزه تک‌زبانه بررسی شده است [۱۰]. در این روش، با استفاده از مدل زبانی ۱-گرام‌ها، امتیاز مثبت و منفی برای اسناد محاسبه می‌شود. در مدل زبانی استفاده‌شده، با در نظر گرفتن مجموعه‌ای از اسناد، احتمال مشاهده یک سند با استفاده از بیشینه راست‌نمایی[۲۵] و هموارسازی لاپلاس[۲۶] محاسبه می‌شود.

در این روش، مدل زبانی نظرات با برچسب مثبت و مدل زبانی نظرات با برچسب منفی به صورت جداگانه ساخته می‌شود و برای هر نظر امتیاز جداگانه‌ای برای هر برچسب محاسبه می‌شود. این امتیاز، میزان شباهت مدل زبانی نظر مورد بررسی با دو مدل زبانی به‌دست‌آمده، است. برچسب انتخابی برای این نظر، برچسب مدل زبانی با مقدار شباهت بیش‌تر و یا مقدار اختلاف کم‌تر است. این اختلاف دو مدل زبانی با استفاده از روش KL-divergence [۳۲] محاسبه می‌شود (فرمول (۱۳)).

$$D(\theta_d \,||\, \theta^{+/-}) = \sum_{w \in d} p(w \mid \theta_d) \log \frac{p(w \mid \theta_d)}{p(w \mid \theta^{+/-})} \qquad (۱۳)$$

در فرمول بالا θ_d مدل زبانی نظر مورد بررسی و $\theta^{+/-}$ مدل زبانی نظرات مثبت و یا نظرات منفی است. $D(\theta_d \,||\, \theta^{+/-})$ نیز میزان اختلاف مدل زبانی نظر با مدل زبانی نظرات مثبت و یا نظرات منفی است.

برای استفاده از این روش در حوزه بین‌زبانی، با استفاده از لغت‌نامه گوگل، نظرات در زبان مبدأ به زبان مقصد ترجمه می‌شوند و مسأله به یک مسأله تک‌زبانه در زبان مقصد تبدیل می‌شود که در این صورت می‌توان از روش پیشنهاد شده [۱۰] استفاده کرد.

۴-۳- آزمایش‌ها

در این بخش به بررسی آزمایش‌ها انجام شده می‌پردازیم. برای رده‌بندی نظرات از رده‌بند SVM[۲۷] استفاده شده است و برای ارزیابی روش، معیار صحت[۲۸] در نظر گرفته شده است. این مقدار برابر با درصد نسبت تعداد نظرات با برچسب پیش‌بینی‌شده درست، به تعداد کل نظرات است (فرمول (۱۴)).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \qquad (۱۴)$$

در این‌جا، TP و TN به ترتیب تعداد نظرات مثبت درست پیش‌بینی شده و تعداد نظرات منفی درست پیش‌بینی شده است و FP و FN نیز به ترتیب تعداد نظرات منفی که مثبت پیش‌بینی شده و تعداد نظرات مثبت که منفی پیش‌بینی شده است.

انتخاب ویژگی‌های محوری

اولین ارزیابی بر روی روش انتخاب ویژگی‌های محوری انجام شد. در آزمایش اول، ابتدا ویژگی‌های پرکاربرد در هر دو زبان (که پایه‌ای‌ترین روش محسوب می‌شود) به عنوان ویژگی‌های محوری انتخاب شد. این کلمات که تنها از پرکاربردترین کلمات در هر دو زبان محسوب می‌شود، خصلت خاص دیگری ندارد. در آزمایش بعدی ویژگی‌های محوری براساس اطلاعات متقابل بین برچسب نظرات و کلمات به کار رفته در آن‌ها انتخاب شد که لازم بود این روند محاسبه بر روی نظرات برچسب‌خورده صورت گیرد و کلماتی که بیشترین مقدار اطلاعات متقابل را دارد به عنوان ویژگی محوری انتخاب شود. از آن‌جایی که فرض شده این نظرات در زبان مقصد موجود نیست، تنها از نظرات برچسب‌خورده در زبان مبدأ استفاده شد و کلمات انتخاب شده با استفاده از منبع ترجمه، ترجمه و به جفت‌کلمه تبدیل شد. این کلمات به جز پرکاربرد بودن، گرایشی به یکی از برچسب‌های مثبت و یا منفی دارند و دربردارنده یک گرایش احساسی است. در آزمایش بعدی، انتخاب ویژگی‌های محوری طبق روندی که در زیر بخش ۴-۱ توضیح داده شد، صورت گرفت. در این روش که از اطلاعات متقابل بین ویژگی و زبان استفاده می‌کند شهود متفاوتی با دو روش قبلی دارد. در این روش سعی می‌شود ویژگی‌های محوری مستقل از زبان باشد که اطلاعاتی که از گراف دوبخشی حاصل می‌شود برای مسأله بین‌زبانی مفید باشد و در مرحله بعدی ارتباط این اطلاعات با برچسب استخراج شود. نتایج این آزمایش‌ها در جدول ارائه شده است.

ویژگی‌های انتخابی توسط این سه روش از لحاظ خصلت بسیار با هم متفاوت است. ویژگی‌های دسته اول ویژگی‌هایی است که هر دو کلمه آن به دفعات متعددی در نظرات به کار گرفته شده باشد. ویژگی‌های انتخابی دسته دوم، ویژگی‌های پرکاربرد، وابسته به برچسب و جهت‌دار است، اما ویژگی‌های دسته سوم، در هر دو زبان بسیار پرکاربرد و دو کلمه آن دارای رفتار مشابهی است. همان‌طور که مشاهده می‌شود روش سوم نسبت به هر دو روش اول و دوم از لحاظ آماری بهتر عمل کرده است. دلیل آن را می‌توان استفاده از اطلاعات هر دو زبان و خصلت استقلال از زبان این ویژگی‌ها برشمرد.

همچنین روش اول نسبت به روش دوم عملکرد بهتری داشته است که بیان‌گر این نکته است که در روش پیشنهادی ویژگی‌های محوری تنها کافی است نماینده خوبی برای زبان خود باشند و جهت‌دار بودن شرط لازم برای آن‌ها نیست. به عبارتی می‌توان گفت که با استفاده از اطلاعات متقابل بین کلمه و برچسب، امکان انتخاب ویژگی‌های محوری مناسب وجود دارد. اما تضمینی وجود ندارد که با استفاده از این روش، ویژگی‌های مستقل از زبان انتخاب شود. همچنین این ویژگی‌ها فقط با استفاده از داده‌های زبان مبدأ انتخاب می‌شود و ممکن است برای زبان مقصد ویژگی مناسبی نباشد.

جدول ۲- مقایسه روش‌های انتخاب ویژگی‌های محوری

معیار انتخاب	صحت
پرکاربرد بودن	۷۸/۱۴۸
پرکاربرد بودن + اطلاعات متقابل بین کلمه و برچسب	۷۸/۲۷۴
پرکاربرد بودن + اطلاعات متقابل بین ویژگی و زبان	**۸۱/۷۵۲**[۲۱]

هم‌ترازی کلمات

در روش پیشنهادی برای جفت کردن ویژگی‌های محوری ایده استفاده از یک لغت‌نامه مطرح شد. همان‌طور که در بخش ۲ گفته شد، در سال‌های اخیر روش‌های مبتنی بر نمایش طیفی کلمات در حوزه‌های مختلفی مورد استفاده قرار گرفته است. این دسته از روش‌ها در حوزه بین‌زبانی نیز وارد شده است. در [۳۳] با استفاده از ایده‌ای که در [۳۴] مطرح شده بود، با بهره‌گیری از یک پیکره موازی[۲۹] و داده‌های تک‌زبانه برای دو زبان مورد نظر، بردارهایی برای کلمات هر دو زبان استخراج می‌شود و می‌توان با استفاده از توابع محاسبه شباهت دو بردار، بردارهای نزدیک به هم را شناسایی کرد. در نتیجه برای کلمه از زبان مبدأ، کلماتی از زبان مقصد که دارای بردارهای نزدیک به بردار کلمه مورد نظر است، به دست می‌آید.

در این بخش جفت‌کلمه‌های کاندیدا برای ویژگی‌های محوری را با استفاده از این روش به دست می‌آوریم و با روش پیشنهادی که استفاده از یک لغت‌نامه است، مقایسه می‌کنیم.

در آزمایش اول، با استفاده از یک ابزار آماده[۳۰]، پیکره موازی آلمانی- انگلیسی و نظرات بدون برچسب، بردارهای کلمات استخراج شد و با استفاده از میزان شباهت کسینوسی، برای هر کلمه از زبان مبدأ، ۵ کلمه از زبان مقصد با نزدیک‌ترین بردار به بردار کلمه زبان مبدأ، انتخاب شد. کلماتی که از حد آستانه

کم‌تر ظاهر شده‌اند نیز از کاندیداهای ویژگی‌های محوری حذف شد. در نهایت مانند روش پیشنهادی با محاسبه مقدار اطلاعات متقابل میان ویژگی و زبان، جفت‌کلمه‌های مستقل از زبان به عنوان ویژگی‌های محوری انتخاب شد. در آزمایش دوم روند انجام شده طبق روندی است که در بخش ۳-۲ گفته شد. نتایج این دو آزمایش در جدول ۳ قابل مشاهده می‌باشد.

جدول ۳- مقایسه روش‌های متفاوت برای هم‌ترازی کلمات

روش هم‌ترازی	صحت
نمایش طیفی کلمات	۵۵/۰۲۸
لغت‌نامه	**۸۱/۷۵۲**

همان‌طور که مشاهده می‌شود، با استفاده از نمایش طیفی کلمات معادل خوبی برای کلمات به دست نیامده است. در نتیجه با استفاده از این ویژگی‌های محوری چگونگی ارتباط میان دو زبان به خوبی تشخیص داده نشده که منجر به کاهش بسیار زیاد کارایی شده است.

انتخاب ویژگی‌های غیرمحوری

آزمایش بعدی بر روی روش انتخابی ویژگی‌های غیرمحوری انجام شد. در ابتدا سعی شد برای این دسته از ویژگی‌ها از یک واژه‌نامه برای هر زبان استفاده شود. این واژه‌نامه‌ها حاوی کلمات مثبت، منفی و خنثی بود. ابتدا تنها کلمات جهت‌دار (مثبت و منفی) به عنوان ویژگی‌های غیرمحوری انتخاب شد. در آزمایش بعدی از کلمات خنثی نیز استفاده شد تا تاثیر وجودشان بررسی شود. در مرحله بعدی به جای استفاده از واژه‌نامه، از کلمات استفاده شده در نظرات به عنوان ویژگی‌های غیرمحوری استفاده شد. در یک آزمایش کلماتی انتخاب شدند که تعداد رخدادشان از یک مقدار کمینه بیش‌تر باشد و در آزمایش بعدی مقدار کمینه بسیار کم‌تری انتخاب شد تا تعداد کلمات بیش‌تری (بیش از ۴ برابر) برای ویژگی‌های غیرمحوری انتخاب شود و عملاً تنها از کلماتی که تنک بودند، استفاده نشد. نتایج این آزمایش‌ها در جدول ۴ آمده است.

جدول ۴- مقایسه روش‌های انتخابی ویژگی‌های غیرمحوری

ویژگی‌های غیرمحوری	صحت
کلمات جهت‌دار واژه‌نامه	۷۲/۵
واژه‌نامه	۷۲/۴۵
کلمات پرکاربرد	۷۹/۲
کلمات رایج	**۸۱/۴۵**

این آزمایش‌ها نشان می‌دهد وجود کلمات خنثی در ویژگی‌های غیرمحوری تأثیر منفی قابل توجهی بر کارایی روش نمی‌گذارد. همچنین هر چه تعداد ویژگی‌های غیرمحوری بیش‌تر باشد و در نتیجه ویژگی‌های معنایی برای کلمات بیش‌تری استخراج شود، کارایی روش پیشنهادی افزایش می‌یابد.

بررسی تأثیر ایست‌واژه‌ها[۳۱]

در آزمایش بعدی وجود و یا عدم وجود ایست‌واژه‌ها بررسی شده است. یک بار ایست‌واژه‌ها در متن نظرات حفظ شد و بار دیگر در مرحله پیش‌پردازش این کلمات از متن نظرات حذف شد. نتایج این بررسی در جدول ۵ نشان داده شده است.

جدول ۵- ارزیابی تأثیر ایست‌واژه‌ها در کارایی

	صحت
با وجود ایست‌واژه‌ها	۷۲/۵
بدون وجود ایست‌واژه‌ها	**۸۱/۷۱۲**

همان‌طور که مشاهده می‌شود وجود ایست‌واژه‌ها در این روش تأثیر منفی دارد و وجود این کلمات در این روش مؤثر واقع نمی‌شود و بهتر است حذف ایست‌واژه‌ها به عنوان یک مرحله از پیش‌پردازش بر روی نظرات انجام شود. علی‌رغم این‌که می‌توان ایست‌واژه‌ها را ویژگی‌های مستقل از زبان و همچنین غیرمبهم فرض کرد (در عمل نیز با در نظر گرفتن ایست‌واژه‌ها، تعداد زیادی از آن‌ها در مجموعه ویژگی‌های محوری قرار گرفت)، فاقد اطلاعات معنایی بودنشان را می‌توان از دلایل بروز این رفتار برشمرد که این خصوصیت ایست‌واژه‌ها سبب می‌شود گراف دوبخشی، نسبت به حالت حذف ایست‌واژه‌ها، حاوی اطلاعات مفید کم‌تری باشد و در نتیجه خوشه‌بندی با دقت کم‌تری صورت گیرد.

مقایسه با روش‌های پایه

در این قسمت نتایج مقایسه روش پیشنهادی با روش‌های پایه ارائه می‌شود. با توجه به نتایج آزمایش‌ها قبلی، روش پیشنهادی (CLSFD[۳۲]) بر روی مجموعه داده‌ای انجام گرفت. ابتدا ایست‌واژه‌ها از مجموعه داده‌ای حذف شد. سپس ۲۵۰ ویژگی محوری با استفاده از اطلاعات متقابل بین ویژگی و زبان انتخاب شد و کلمات کاربردی نظرات به عنوان ویژگی‌های غیرمحوری انتخاب شد و در نهایت ۱۰۰ ویژگی معنایی از گراف دوبخشی حاصل استخراج شد. در ضمن تأثیر تعداد ویژگی‌های محوری و تعداد ویژگی‌های معنایی در ادامه بررسی خواهد شد. برای روش CL-SCL نیز تعداد ویژگی‌های محوری ۲۵۰ انتخاب شد. به دلیل مفید بودن ایست‌واژه‌ها برای این دو روش پایه، در آزمایش‌ها روش CL-SCL و مدل زبانی، ایست‌واژه‌ها در مرحله پیش‌پردازش حذف نشد.

این آزمایش‌ها به صورت اعتبارسنجی متقابل ۵ بخشی انجام شد. در هر آزمایش ۳۲۰۰ نظر (۱۶۰۰ نظر مثبت و ۱۶۰۰ نظر منفی) در زبان مبدأ به عنوان داده آموزش و ۸۰۰ نظر (۴۰۰ نظر مثبت و ۴۰۰ نظر منفی) در زبان مقصد به عنوان داده آزمون انتخاب شد. میانگین صحت این ۵ آزمون در جدول ۶ گزارش شده است. برای مقایسه آماری این روش‌ها و بررسی معنادار بودن اختلاف کارایی روش‌ها از لحاظ آماری، آزمون زوج‌شده t بر روی نتایج حاصله انجام گرفت. نتایج این آزمایش‌ها نیز در جدول ۶ گزارش شده است. همان‌طور که مشاهده می‌شود، روش پیشنهادی نسبت به دو روش پایه بهتر عمل می‌کند و اختلاف کارایی آن نسبت به دو روش پایه از نظر آماری معنادار است.

جدول ۶- مقایسه روش پیشنهادی با روش‌های پایه

	صحت
مدل زبانی	۷۲/۷۷۵
CL-SCL	۷۹/۳۲۲
CLSFD	**۸۱/۷۱۲**[۲،۱]

تحلیل دیگری که می‌توان برای مقایسه این روش‌ها انجام داد، دلیل تأثیر متفاوت ایست‌واژه‌ها در این سه روش است. همان‌طور که ذکر شد، دو روش پایه در حالت وجود ایست‌واژه‌ها عملکرد بهتری دارد و همچنین از منبع ترجمه برای ترجمه تمامی ویژگی‌ها استفاده می‌کند، اما روش پیشنهادی در حالت عدم وجود این کلمات بهتر عمل می‌کند و از منبع ترجمه تنها برای ترجمه تعداد محدودی

ویژگی بهره می‌برد. از آن‌جا که ترجمه اکثر ایست‌واژه‌ها ترجمه غیرمبهم و با دقت خوبی صحیح است، وجود این کلمات در حالتی که تمام ویژگی‌ها ترجمه می‌شود می‌تواند کیفیت ترجمه را به مقدار زیادی افزایش دهد. در نتیجه این رفتار برای این روش‌ها قابل پیش‌بینی و توجیه‌پذیر است.

روش پیشنهادی بین‌زبانی، علاوه بر کارایی بهتر از لحاظ آماری نسبت به دو روش دیگر، وابستگی کم‌تری به منبع ترجمه دارد که علت آن ترجمه شدن تنها ویژگی‌های محوری در این روش است. اما در دو روش دیگر لازم است تمام کلمات ترجمه شود. اگر کلمه‌ای در منبع ترجمه موجود نباشد و یا به خوبی ترجمه نشود، امکان کاهش کارایی در این روش‌ها وجود دارد. در روش پیشنهادی نبود ترجمه کلمات در منبع ترجمه تنها باعث حذف آن از کاندیداهای ویژگی محوری می‌شود به طوری که حضور این کلمات در ویژگی‌های غیرمحوری باعث استخراج ویژگی‌های معنایی برای چنین کلمات می‌شود. در نتیجه روش پیشنهادی علاوه بر وابسته نبودن به منبع ترجمه، از کلماتی که ترجمه‌ای برای آن‌ها موجود نیست هم بهره می‌برد. در صورتی که منبع ترجمه بهتری در دسترس باشد احتمال افزایش کارایی هر سه روش وجود دارد و با توجه به وابستگی بیش‌تر روش‌های پایه به منبع ترجمه، می‌توان پیش‌بینی کرد این افزایش کارایی برای روش‌های پایه نسبت به روش پیشنهادی بیش‌تر باشد.

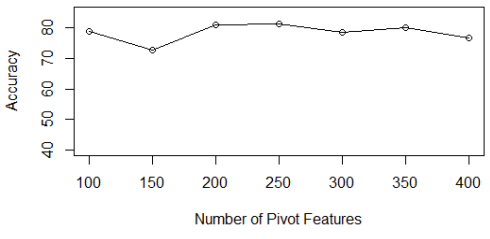
همان‌طور که در بخش قبل توضیح داده شد، نظرات در دو زبان مبدأ و مقصد با انتقال به فضای معنایی، دارای ویژگی‌های مشترکی می‌شود. در نتیجه مدل رده‌بندی که از روی نظرات زبان مبدأ ساخته می‌شود هم برای رده‌بندی نظرات زبان مبدأ و هم برای رده‌بندی نظرات زبان مقصد، قابل استفاده است. به همین دلیل با استفاده از ویژگی‌های معنایی استخراج‌شده، علاوه بر آزمایش‌ها بین‌زبانی، آزمایش‌ها تک‌زبانه را نیز می‌توان انجام داد. نتایج این آزمایش‌ها در هر دو زبان انگلیسی و آلمانی در جدول ۷ قابل مشاهده است. همان‌طور که مشاهده می‌شود کارایی روش در زبان آلمانی به میزان قابل توجهی از کارایی در زبان انگلیسی بهتر است. به نظر می‌رسد این تفاوت در کارایی به دلیل اختلاف تعداد نظرات بدون برچسب در دو زبان انگلیسی و آلمانی باشد. بنابراین می‌توان نتیجه گرفت این نظرات تأثیر بالایی در کارایی روش دارند و تعداد بیش‌تر این نظرات به استخراج بهتر ویژگی‌های معنایی کمک می‌کند.

جدول ۷- نتایج آزمایش‌ها تک‌زبانه

	صحت
انگلیسی	۷۷/۳۷۸
آلمانی	۸۴/۰۲۶

بررسی حساسیت به تعداد ویژگی‌های محوری

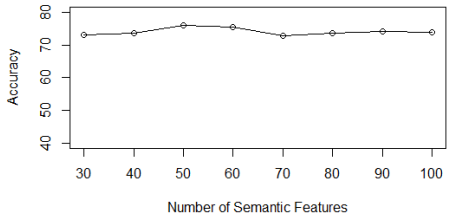
با تغییر تعداد ویژگی‌های محوری انتخابی و ثابت ماندن دیگر تنظیمات، تأثیر این پارامتر در روش پیشنهادی بررسی شد. نتایج این آزمایش‌ها در شکل ۲ نشان داده شده است. همان‌طور که ملاحظه می‌شود روش پیشنهادی در یک بازه بزرگ از ۲۰۰ تا ۳۵۰ به تعداد ویژگی‌های محوری حساسیت زیادی ندارد.



شکل ۲- حساسیت روش به تعداد ویژگی‌های محوری

بررسی حساسیت به تعداد ویژگی‌های معنایی

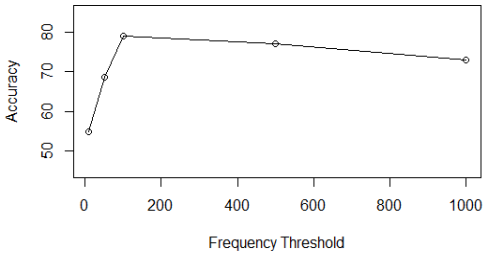
علاوه بر آزمایش‌ها گفته‌شده، حساسیت روش پیشنهادی نسبت به تعداد ویژگی‌های معنایی نیز بررسی شد. در این بررسی تعداد متفاوتی از ویژگی‌های معنایی با سایر تنظیمات مشابه، استخراج شد که روند تغییر کارایی روش در شکل ۳ نشان داده شده است. همان‌طور که ملاحظه می‌شود، تعداد ویژگی‌های معنایی استخراج‌شده تأثیر قابل ملاحظه‌ای بر کارایی روش نداشته و کارایی روش با تغییر مقدار این پارامتر تغییر زیادی نمی‌کند.



شکل ۳- حساسیت روش به تعداد ویژگی‌های معنایی

بررسی تأثیر میزان حدآستانه تکرار ویژگی‌های محوری

همان‌طور که گفته شد، یکی از محدودیت‌های تعریف‌شده برای انتخاب ویژگی‌های محوری، پرتکرار بودن هر دو ویژگی یک جفت ویژگی در زبان‌های متناظرشان است. خصلت پرتکرار بودن با انتخاب یک حد آستانه برای تعداد تکرار ویژگی‌ها در داده‌های بدون برچسب بررسی می‌شود. آزمایش‌های متعددی برای بررسی تأثیر مقدار این حد آستانه انجام گرفت. نتایج این آزمایش‌ها در شکل ۴ نشان داده شده است.



شکل ۴- تأثیر مقدار حد آستانه ویژگی‌های محوری بر روش پیشنهادی

همان‌طور که ملاحظه می‌شود، در صورتی که مقدار حد آستانه انتخابی بسیار کوچک باشد، ویژگی‌های خوبی به عنوان ویژگی‌های محوری انتخاب نمی‌شود. همچنین با انتخاب مقدار بسیار بالا برای حد آستانه، تعداد زیادی از ویژگی‌های مناسب، از کاندیداهای ویژگی‌های محوری حذف می‌شوند. در نتیجه مقداری برای حد آستانه مناسب است که هم از انتخاب ویژگی‌های کم‌تکرار و نامناسب جلوگیری کند و هم باعث حذف ویژگی‌های خوب نشود.

۵- نتیجه‌گیری

در این مقاله مسأله بین‌زبانی در حوزه نظرکاوی بررسی شد. روش پیشنهادی این مقاله، ابتدا با انتخاب دو دسته ویژگی، ویژگی‌های محوری و ویژگی‌های غیرمحوری، در هر دو زبان مبدأ و مقصد، ساخت یک گراف دوبخشی و در نهایت

استخراج ویژگی‌های معنایی برای ویژگی‌های غیرمحوری، توانست به کارایی خوبی دست پیدا کند. در این روش که از یک لغت‌نامه دوزبانه برای ترجمه ویژگی‌های محوری استفاده شد، برای ساخت بهتر گراف و استفاده از داده‌های بیشتر، از نظرات بدون برچسب در هر دو زبان بهره برد. از مزیت‌های این روش نسبت به روش‌های پایه بررسی‌شده، می‌توان به وابستگی کم این روش به حجم و کیفیت منبع ترجمه اشاره کرد. با توجه به تنظیمات و منابع مورد نیاز روش پیشنهادی، این روش می‌تواند برای زبان‌هایی که ماشین ترجمه مناسبی ندارد نیز مفید واقع شود. وجود نظرات بدون برچسب از ضروریات این روش است و هر چه تعداد این نظرات بیش‌تر باشد، دقت روش نیز افزایش پیدا می‌کند.

با توجه به خاصیت دوزبانه بودن گراف دوبخشی، ویژگی‌های معنایی برای ویژگی‌های غیرمحوری هر دو زبان به دست آمد که سبب شد امکان استفاده از مدل به‌دست‌آمده برای رده‌بندی نظرات در هر دو زبان فراهم شود. در آینده می‌توان حالت گسترش‌یافته این گراف را برای بیش از دو زبان بررسی کرد و از مدل حاصل برای رده‌بندی نظرات در چند زبان متفاوت استفاده کرد. همچنین با ایجاد منابع ترجمه با کیفیت‌های متفاوت می‌توان تأثیر میزان کیفیت منبع ترجمه بر کارایی روش را بررسی کرد. افزایش تعداد نظرات بدون برچسب در زبان انگلیسی و بررسی میزان تغییر در کارایی روش نیز از جمله کارهایی است که در آینده می‌توان به آن پرداخت.

با توجه به وابسته نبودن روش پیشنهادی به یک زوج زبان خاص، می‌توان با ایجاد مجموعه داده‌ای مناسب برای زبان فارسی، کارایی این روش را برای زوج زبان فارسی- انگلیسی نیز بررسی کرد.

مراجع

[1] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. Annu. Meet. ACL*, 2002.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," *Proc. Conf. EMNLP*, 2002.

[3] D. Tang, and et. al., "Sentiment Embeddings with Applications to Sentiment Analysis," *IEEE Trans. Knowledge and Data Engineering*, vol. 28, pp. 496-509, 2016.

[4] T. Zagibalov, and J. Carroll, "Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text," *Proc. Int'l Conf. COLING*, 2008.

[5] X. Wan, "Co-Training for Cross-Lingual Sentiment Classification," *Proc. ACL/AFNLP Int'l Joint Conf.*, 2009.

[6] X. Wan, "Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews," *Computational Linguistics*, vol. 37, pp. 587-616, 2011.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Advances in NIPS*, 2001.

[8] B. Pang, and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. Annu. Meet. ACL*, 2004.

[9] 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, pp. 46-53, 2010.

[10] Y. Hu, and et. al., "A Language Modeling Approach to Sentiment Analysis," *Proc. Int'l Conf. ICCS*, 2007.

[11] M. Thelwall, and et. al., "Sentiment in Short Strength Detection Informal Text," *JASIST*, vol. 61, pp. 2544-2558, 2010.

[12] T. Mikolov, and et. al., "Distributed Representations of Words and Phrases and their Compositionality," *Advances in NIPS*, 2013.

[13] A. L. Maas, and et. al., "Learning Word Vectors for Sentiment Analysis," *Proc. Annu. Meet. ACL*, 2011.

[14] I. Labutov, and H. Lipson, "Re-Embedding Words," *Proc. Annu. Meet. ACL*, 2013.

[15] D. Tang, and et. al., "Sentiment Embeddings with Applications to Sentiment Analysis," *IEEE Trans. Knowledge and Data Engineering*, vol. 28, pp. 496-509, 2016.

[16] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," *Proc. Int'l Conf. COLING*, 2010.

[17] C. Tan, and et. al., "User-Level Sentiment Analysis Incorporating Social Networks," *Proc. ACM Int'l Conf. SIGKDD*, 2011.

[18] J. S. Olsson, D. W. Oard, and J. Hajic, "Cross-Language Text Classification," *Proc. Annu. Int'l ACM Conf. SIGIR*, 2005

[19] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-Linguistic Sentiment Analysis: From English to Spanish," *Recent Advances in NIPS*, 2009.

[20] C. Wan, R. Pan, and J. Li, "Bi-Weighting Domain Adaptation for Cross-Language Text Classification," *Proc. Int'l Joint Conf. IJCAI*, 2011.

[21] M. S. Hajmohammadi, R. Ibrahim, A. Selamat, and H. Fujita, "Combination of Active Learning and Self-Training for Cross-Lingual Sentiment Classification with Density Analysis of Unlabelled Samples," *Information Sciences*, vol. 317, pp. 67-77, 2015.

[22] D. Gao, and et. al., "Cross-Lingual Sentiment Lexicon Learning with Bilingual Word Graph Label Propagation," *Computational Linguistics*, vol. 41, pp. 21-40, 2015.

[23] M. S. C. Almeida, and et. al., "Aligning Opinions: Cross-Lingual Opinion Mining with Dependencies," *Proc. ACL/AFNLP Joint Conf.*, 2015.

[24] H. Guo, and et. al., "OpinionIt: A Text Mining System for Cross-Lingual Opinion Analysis," *Proc. ACM CIKM*, 2010.

[25] S. Jain, and S. Batra, "Cross Lingual Sentiment Analysis using Modified BRAE," *In Proc. Conf EMNLP*, 2015.

[26] P. Prettenhofer, and B. Stein, "Cross-Language Text Classification Using Structural Correspondence Learning," *Proc. Annu. Meet. ACL*, 2010.

[27] P. Prettenhofer, and B. Stein, "Cross-Lingual Adaptation Using Structural Correspondence Learning," *ACM TIST*, vol. 3, p. 13, 2011.

[28] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification," *Proc. Annu. Meet. ACL*, 2007.

[29] S. J. Pan, and et. al., "Cross-Domain Sentiment Classification via Spectral Feature Alignment," *Proc. Int'l Conf. WWW*, 2010.

[30] G. Zhou, and et. al., "Cross-Domain Sentiment Classification via Topical Correspondence Transfer," *Neurocomputing*, vol. 159, pp. 298-305, 2015.

[31] K. W. Church, and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol. 16, pp. 22-29, 1990.

[32] T. M. Cover, and J. A. Thomas, "Elements of Information Theory," *Wiley-Interscience*, 1991.

[33] S. Gouws, Y. Bengio, and G. Corrado, "BilBOWA: Fast Bilingual Distributed Representations without Word Alignments," *Proc. Int'l Conf. Machine Learning*, 2015.

[34] Mikolov, Tomas, Quoc V. Le, and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," *CoRR*, abs/1309.4168, 2013.

شیما اسمعیلی تفت مدرک کارشناسی خود را در رشته مهندسی فناوری اطلاعات از دانشگاه تهران در سال ۹۲ دریافت کرد. در همان سال نیز با استفاده از سهمیه استعدادهای درخشان در مقطع کارشناسی‌ارشد مشغول به تحصیل شد. سپس در سال ۹۵ موفق به اخذ مدرک کارشناسی‌ارشد در رشته مهندسی فناوری اطلاعات از دانشگاه تهران گردید. علایق پژوهشی او متن‌کاوی، بازیابی اطلاعات و داده کاوی می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shima.esmaeili@ut.ac.ir

آزاده شاکری استادیار دانشکده مهندسی برق و کامپیوتر پردیس دانشکده‌های فنی دانشگاه تهران است. او مدرک دکترای خود را در سال ۱۳۸۷ از دانشگاه ایلینویز اوربانا- شمپین در آمریکا دریافت کرد. زمینه‌های پژوهشی مورد علاقه وی مدیریت اطلاعات متنی، بازیابی اطلاعات، متن کاوی، و داده کاوی می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shakery@ut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۴/۰۸/۱۶

تاریخ اصلاح: ۱۳۹۴/۱۰/۱۳

تاریخ قبول شدن: ۱۳۹۴/۱۰/۲۳

نویسنده مرتبط: دکتر آزاده شاکری، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران.

[1]Sentiment Analysis
[2]Opinion Mining
[3]Sentiment Orientation
[4]Text Mining
[5]Subjective
[6]Objective
[7]Seed

[8]Part of Speech Tag
[9]Language Model
[10]Word Embedding
[11]Twitter
[12]Lexicon
[13]Latent Dirichlet Allocation
[14]Spectral Feature Alignment
[15]Domain-Independent Features
[16]Domain-Specific Features
[17]Mutual Information
[18]Singular Value Decomposition
[19]Soft Clustering
[20]Sparse
[21]Pivot Features
[22]Prettenhofer
[23]https://translate.google.com/#en/de/
[24]Cross-Lingual Structural Correspondence Learning
[25]Maximum Likelihood
[26]Laplace Smoothing
[27]http://svmlight.joachims.org/
[28]Accuracy
[29]Parallel Corpus
[30]https://github.com/gouwsmeister/bilbowa
[31]Stopwords
[32]Cross-Lingual Semantic Feature Derivation