

ساخت نیمه‌خودکار یک پیکره از نظرات غیرمستقیم در دامنه دارو و به کارگیری آن برای تعیین قطبیت نظرات

سمیرا نوفrstی^{۱*} و مهرنوش شمس‌فرد^۲

^۱ دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

^۲ آزمایشگاه پردازش زبان طبیعی، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران

چکیده

مسئله نظرکاوی در سال‌های اخیر، بسیار مورد توجه بوده است. تمرکز روش‌های موجود نظرکاوی بر نظرات مستقیم بوده است و اغلب آن‌ها از نظرات غیرمستقیم صرف‌نظر کرده‌اند. در حالی‌که در برخی دامنه‌ها از جمله پژوهشی نظرات غیرمستقیم به دفعات رخ می‌دهند و نادیده گرفتن آن‌ها باعث کاهش دقت سامانه نظرکاوی می‌شود. در این مقاله روشی نیمه‌خودکار برای ساخت پیکره‌ای از نظرات غیرمستقیم به زبان انگلیسی در دامنه دارو ارائه می‌شود. در مرحله نخست روش پیشنهادی، جملاتی که بیان گر یک نظر غیرمستقیم هستند، شناسایی و در مرحله دوم قطبیت آن‌ها با کمک دانش دامنه، الگوهای زبانی و ساختار نظرات تعیین می‌شود؛ سپس از این پیکره در روش‌های یادگیری ماشین جهت تعیین قطبیت نظرات مطرح شده درباره داروها استفاده می‌شود. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی در تعیین قطبیت مجموعه آزمون بدقت ۸۲/۸۱ درصد می‌رسد و بر یک روش برگسته تحلیل نظرات مستقیم به نام الگوهای سنتیک غلبه می‌کند.

وازگان کلیدی: نظرکاوی، نظرات غیرمستقیم، تحلیل احساسات، ساخت پیکره، یادگیری ماشین.

احساس می‌شود. نظرکاوی^۱ که در برخی از پژوهش‌ها تحلیل احساسات^۲ نیز نامیده می‌شود، یک زمینه‌پژوهشی جدید است که به ارائه روش‌های خودکار جهت کاوش، تحلیل، دسته‌بندی و خلاصه‌سازی نظرات می‌پردازد. نظرات به دو دستهٔ کلی تقسیم می‌شوند (لیو، ۲۰۱۲): نظرات مستقیم^۳ و نظرات غیرمستقیم^۴. در یک نظر مستقیم، یک موجودیت یا جنبه‌ای از آن به صورت مستقیم مورد توصیف قرار می‌گیرد. به عنوان مثال جمله "این دارو بسیار مفید است." یک نظر مستقیم با قطبیت مثبت را نشان می‌دهد. در مقابل در یک نظر غیرمستقیم اثر یک موجودیت بر روی موجودیت‌های دیگر بیان می‌شود. به عنوان مثال جمله "پس از مصرف این دارو دچار درد مفاصل شدم." یک نظر غیر مستقیم است که بیان گر اثری منفی از دارو بر روی مفاصل است.

۱- مقدمه

از دیرباز معیار تصمیم‌گیری برای خرید یک محصول، انتخاب یک پژوهش متبحر، داشتن سفری مفرح و یا برگزیدن یک رستوران خوب، مشورت و کسب نظرات افرادی بوده است که تجربه مشابهی در این زمینه داشته‌اند. پیش از پیدایش وب، تعداد این افراد کم و محدود به اقوام، دوستان و آشنایان بود. وب تغییر بزرگی در روش بیان و دسترسی به نظرات ایجاد کرده است. در سال‌های اخیر، گسترش روزافزون سایت‌های نظرسنجی و وبلاگ‌های شخصی منبع عظیمی از نظرات و تجربیات افراد را فراهم آورده است. پژوهش‌ها نشان می‌دهند که این نظرات هم برای کابران در انتخاب یک محصول و هم برای تولیدکنندگان در ارزیابی محصولاتشان مفید بوده است (موراس و همکاران، ۲۰۱۳).

با توجه به حجم فراوان نظرات بر روی وب امکان خواندن و تحلیل همه آن‌ها توسط کابران وجود ندارد؛ بدین منظور نیاز به ایجاد ابزارهای خودکار برای تحلیل نظرات

¹ Opinion mining

² Sentiment analysis

³ Direct opinions

⁴ Indirect opinions



- نواوری‌های این مقاله را می‌توان به صورت زیر خلاصه کرد:
 - ارائه روشی نوین برای ساخت نیمه‌خودکار یک مجموعه برچسب خورده از نظرات غیرمستقیم
 - معرفی، به کارگیری و ارزیابی ویژگی‌های لغوی، نحوی و معنایی مختلف در روش‌های یادگیری ماشین به منظور تعیین قطبیت نظرات غیرمستقیم
- ادامه این مقاله به صورت زیر سازماندهی شده است.
- در ابتدا در بخش ۲ کارهای مرتبط انجام گرفته در زمینه تعیین قطبیت نظرات به اختصار مرور و مشکلات آنها در تعیین قطبیت نظرات غیرمستقیم بررسی می‌شود؛ سپس در بخش‌های ۳ و ۴ جزئیات روش پیشنهادی برای ساخت نیمه‌خودکار پیکرهای برچسب خورده از نظرات غیرمستقیم تشریح و همچنین روش به کارگیری پیکره به منظور تعیین قطبیت نظرات جدید شرح داده می‌شود. در بخش ۵ نتایج آزمایش‌های انجام گرفته به منظور ارزیابی روش پیشنهادی را ارائه خواهیم کرد. در پایان بخش ۶ به نتیجه‌گیری می‌پردازیم.

۲- مرور کارهای پیشین

تلاش‌های متعددی در رابطه با وظایف مختلف مسئله نظرکاوی در دامنه‌های مختلف انجام گرفته است. در (پنگ و همکاران، ۲۰۰۲؛ کمربیا و حسین، ۲۰۱۲؛ لیسو، ۲۰۱۲) روش‌های موجود به خوبی معرفی و دسته‌بندی شده‌اند. در این بخش تنها به معرفی برخی از کارهای مرتبط در زمینه تعیین قطبیت نظرات می‌پردازیم. کارهای انجام گرفته در زمینه تعیین قطبیت نظرات را می‌توان به دو رده کلی تقسیم کرد: روش‌های مبتنی بر واژگان و روش‌های یادگیری ماشین. در ادامه این بخش ابتدا این دو دسته را مرور می‌کنیم و سپس به معرفی کارهای انجام گرفته در زمینه تحلیل نظرات غیرمستقیم می‌پردازیم.

۱-۱- روش‌های مبتنی بر واژگان

این روش‌ها برای تعیین قطبیت نظرات از یک واژگان سنتجمانی^۳ استفاده می‌کنند. یک واژگان سنتجمانی حاوی مجموعه‌ای از لغات و عبارات با بار سنتجمانی مشخص است. تلاش‌های بسیاری جهت ساخت واژگان‌های سنتجمانی انجام گرفته است. تمرکز تلاش‌های اولیه بر روی ساخت واژگان‌های عمومی بوده است (اسلوی و سباستیانی، ۲۰۰۶؛ نویاروسکایا و همکاران، ۲۰۱۱). با این وجود، قطبیت واژه‌ها

یکی از وظایف اصلی در نظرکاوی تعیین قطبیت^۱ یا بار سنتجمانی یک نظر است. هدف این وظیفه دسته‌بندی نظرات در گروه‌های از پیش تعیین شده (غلب در دو گروه مثبت و منفی) است. در سال‌های اخیر مسئله نظرکاوی بسیار مورد توجه قرار گرفته و تلاش‌های متعددی جهت تعیین قطبیت نظرات در دامنه‌های مختلف انجام یافته است. با این وجود تمرکز این تلاش‌ها بر روی نظرات مستقیم بوده است. این در حالی است که در برخی از دامنه‌ها نظری اقتصاد و پژوهشی نظرات غیرمستقیم به دفعات رخ می‌دهند. به طور خاص در دامنه دارو، غالب کاربران به جای بیان یک نظر مستقیم درباره دارو، به توصیف اثرات مثبت و عوارضی که دارو برای آن‌ها داشته است، می‌پردازند.

بررسی‌های ما برروی تارنمای www.druglib.com که یک تارنمای نظرسنجی محبوب در رابطه با داروهاست، نشان می‌دهد که بیش از نیمی از نظرات، غیرمستقیم هستند. بنابراین روش‌های موجود تحلیل نظرات که تنها نظرات مستقیم را مورد توجه قرار می‌دهند، بخش عمده‌ای از اطلاعات مفید در نظرات کاربران را نادیده می‌گیرند که این امر خود سبب کاهش دقت الگوریتم نظرکاوی می‌شود. بنا به دلایل مذکور در این مقاله مسئله تعیین قطبیت نظرات غیرمستقیم مورد توجه قرار می‌گیرد.

به منظور تعیین قطبیت نظرات غیرمستقیم از روش‌های یادگیری ماشین استفاده شده است. پیش از این روش‌های یادگیری ماشین در تحلیل نظرات مستقیم به کار گرفته شده‌اند و به نتایج خوبی دست یافته‌اند. مشکل اصلی این روش‌ها ساخت یک مجموعه آموزش از نظرات است که قطبیت مثبت و منفی آن‌ها برچسب خورده باشد. ساخت دستی مجموعه آموزش امری زمان بر و هزینه‌بر است. به منظور رفع این مشکل در این مقاله روشی برای ساخت نیمه‌خودکار مجموعه آموزش پیشنهاد می‌شود. برای ساخت مجموعه آموزش از روش‌های نظارت از راه دور^۲ (مینتز و همکاران، ۲۰۰۹) الهام گرفته شده است. در این روش‌ها به جای به کارگیری یک مجموعه آموزش دقيق، از یک مجموعه بزرگ از داده‌های نویه‌ای که به صورت خودکار بدست آمده‌اند، استفاده می‌شود. روش‌های نظارت از راه دور در حوزه‌های مختلف از جمله استخراج روابط (مینتز و همکاران، ۲۰۰۹) و تحلیل احساسات (لین و همکاران، ۲۰۱۲؛ گو و همکاران، ۲۰۰۹) به کار گرفته‌اند و نتایج حاصل از آن‌ها موفقیت‌آمیز بوده است.

³ Sentiment lexicon

¹ Polarity detection

² Distant supervision



در سنتی وردنت واژه "بینایی" بدون قطبیت است. این امر باعث می‌شود، مثال بالا بدون قطبیت در نظر گرفته شود. درحالی که واضح است که این جمله منفی است. به بیانی دیگر روش‌های مبتنی بر واژگان از تحلیل نظرات ضمنی (نظرات فاقد واژه سنجمانی) قاصرند. درواقع مدیریت این نوع نظرات نیاز به تحلیل معنایی عمیق‌تر جملات و عبارات دارد (بالا هور و همکاران، ۲۰۱۱؛ بالا هور و همکاران، ۲۰۱۲).

۲-۲ - روش‌های یادگیری ماشین

روش‌های یادگیری ماشین به آموزش یک دسته‌بند بر روی یک پیکره برچسب‌خورده از نظرات بهمنظور دسته‌بندی نظرات بدون برچسب می‌پردازند. این روش‌ها به نتایج قابل قبولی در تعیین قطبیت نظرات مستقیم دست یافته‌اند (پنگ و همکاران، ۲۰۰۲؛ بویی و موئنز، ۲۰۰۹). ویژگی‌های مختلفی در روش‌های یادگیری ماشین به کار گرفته شده است. از رایج‌ترین ویژگی‌های مورد استفاده می‌توان به کیسه‌لغات^۴ (حضور یک لغت خاص و/یا تعداد رخداد آن) (پنگ و همکاران، ۲۰۰۲)، ترکیب دوتایی لغات^۵ (پنگ و همکاران، ۲۰۰۹)، برچسب‌های نحوی (گو و همکاران، ۲۰۰۲)، شکلک‌ها^۶ (هابرناال و همکاران، ۲۰۱۴) و مسیر بین واژه‌ها در درخت وابستگی (ناکاگاوا و همکاران، ۲۰۱۰) اشاره کرد.

برتری اصلی روش‌های یادگیری ماشین در مقایسه با روش‌های مبتنی بر واژگان این است که این روش‌ها با درنظر گرفتن اطلاعات بافتار و روابط بین واژه‌ها تا حدودی قادر به تعیین قطبیت نظرات ضمنی هستند. با این وجود روش‌های یادگیری ماشین از یک مشکل عمده رنج می‌برند و آن ساخت پیکره آموزش است.

اغلب کارهای پیشین، پیکره آموزش را به صورت دستی ساخته‌اند (هو و لیو، ۲۰۰۴؛ بوسکو و همکاران، ۲۰۱۳). ساخت دستی پیکره آموزش امری زمان بر و هزینه‌بر است که باعث می‌شود، پیکره از لحاظ کمی محدود باشد. بدین دلیل تلاش‌هایی جهت ساخت نیمه‌خودکار و خودکار پیکره آموزش انجام شده است.

در (کاجی و کیتسوری‌گاوا، ۲۰۰۶) روشی معرفی شده است که با به کارگیری ساختار صفحات HTML و تعریف تعدادی الگوی زبانی یک پیکره برچسب‌خورده از نظرات

به طور کامل وابسته به دامنه است. به عنوان مثال واژه "غیرقابل پیش‌بینی" در دامنه "فیلم" می‌تواند دارای قطبیت مثبت باشد؛ درحالی که در رابطه با "دزدگیر ماشین" یک صفت منفی است. به علاوه واژگان‌های عمومی سیاری از واژه‌های خاص دامنه را نمی‌پوشانند که این امر خود باعث کاهش دقت الگوریتم‌های نظرکاوی می‌شود.

به دلیل مشکلات واژگان‌های سنجمانی عمومی، تلاش‌هایی نیز جهت ساخت واژگان‌های خاص دامنه انجام گرفته است (جرجیوت و همکاران، ۲۰۱۲؛ هوآنگ و همکاران، ۲۰۱۴).

روش‌های مبتنی بر واژگان در تحلیل نظرات غیرمستقیم از دو مشکل اصلی رنج می‌برند. نخست این که اغلب واژگان‌های موجود در سطح واژه هستند. بدین معنا که در این واژگان‌ها به هر واژه یا به هر معنای^۷ آن قطبیت نسبت داده می‌شود. این درحالی است که قطبیت و معنای یک جمله تنها از طریق واژه‌ها منتقل نمی‌شود؛ بلکه تعامل واژه‌ها نیز در تعیین قطبیت جمله مؤثر است. یک نظر غیرمستقیم اثر یک موجودیت بر روی موجودیت دیگر را نشان می‌دهد که این اثر اغلب در قالب فعل جمله بیان می‌شود؛ بنابراین فعل جمله تأثیر بهسازی در تعیین قطبیت جمله دارد. به مثال‌های زیر توجه کنید:

"This drug eliminated my acne completely."

"This drug reduced my pain."

روش‌های مبتنی بر واژگان از تعیین قطبیت درست این گونه مثال‌ها قاصرند زیرا واژه‌هایی نظیر "آکنه" و "درد" در واژگان‌های سنجمانی موجود مانند سنتی وردنت^۸ (اسولی و سباسیانی، ۲۰۰۶) منفی هستند. این امر باعث می‌شود دو جمله بالا منفی در نظر گرفته شوند. درحالی که قطبیت کلی هر دو جمله مثبت است. در حالت کلی در دامنه پژوهشی اغلب واژه‌ها مانند "بیماری"، "آلزایمر"، "تب" و "درد" منفی هستند. با این وجود، این واژه‌ها به مراتب در جملات مثبت رخ می‌دهند. بنابراین روش‌های مبتنی بر واژگان در این دامنه‌ها دقت کافی ندارند.

مشکل دوم در رابطه با جملاتی است که حاوی واژه سنجمانی صریح نیستند. به مثال زیر توجه کنید:

"This drug decreased my vision."

¹ sense

² مثال‌های این مقاله از تارنمای www.druglib.com انتخاب شده‌اند. از آنجا که کاربران این تارنما از کشورهای مختلف هستند که گاهی با زبان انگلیسی آشنایی کافی ندارند، در نظرات نوشته شده گاهی غلط املایی و یا گرامری دیده می‌شود.

³ SentiWordNet

⁴ bag-of-words

⁵ bigram

⁶ emoticons

سال ۱۳۹۵ شماره ۲ پیاپی ۲۸

ویژگی‌ها را در تعیین قطبیت نظرات غیرمستقیم نشان می‌دهد.

۳- روش پیشنهادی برای ساخت یک پیکره از نظرات غیرمستقیم

در این بخش در ابتداء مجموعه داده‌ای مورد استفاده برای ساخت پیکره‌ای از نظرات غیرمستقیم درباره داروها توصیف می‌شود؛ سپس جزئیات روش پیشنهادی برای ساخت نیمه‌خودکار این پیکره شرح داده می‌شود.

۱-۱-۳ مجموعه داده‌ای

مجموعه داده‌ای مورد استفاده جهت ساخت پیکره از تارنمای www.druglib.com جمع‌آوری شده است. بدین منظور نظرات مطرح شده درباره ۸۳ دارو گردآوری شده است. این مجموعه دارای ۱۳۱۳ نظر و ۷۰۱۷ جمله است.

هر نظر دارای دو بخش است شکل (۱)؛ متأذدها و متن نظر. متأذدها شامل اطلاعاتی از قبیل سن، جنسیت^۲ بیمار و نیز امتیازی که او به اثربخشی^۳ و عوارض جانبی^۳ دارو می‌دهد، است. متن نظر نیز شامل سه بخش است: فواید، عوارض دارویی و توضیحات. پیش‌فرض این است که کاربر در بخش فواید درباره اثرات مثبت دارو می‌نویسد. به عبارت دیگر انتظار می‌رود حملات این بخش دارای قطبیت مثبت باشند. به طور مشابه فرض می‌شود جملات نوشته شده در بخش عوارض دارویی دارای قطبیت منفی هستند؛ اگرچه این فرضیات به‌الازام همیشه درست نیستند. حملات مطرح شده در بخش توضیحات نیز محدودیتی ندارند.



(شکل ۱)؛ مثالی از یک نظر مطرح شده در تارنمای www.druglib.com

² effectiveness

³ Side effects

می‌سازد. در (گو و همکاران، ۲۰۰۹) از شکلک‌های مورد استفاده در تؤییت‌های کاربران در توئیتر^۱ برای ساخت مجموعه آموزش استفاده شده است.

۲-۳-۲ تحلیل نظرات غیرمستقیم

اغلب کارهای پیشین تنها نظرات مستقیم را مورد توجه قرار داده‌اند و روش خاصی برای مدیریت نظرات غیرمستقیم ندارند. تعداد محدودی از کارهای پیشین به نظرات غیرمستقیم توجه کرده‌اند که تمرکز آن‌ها بر برچسب‌زندن نظرات غیرمستقیم به صورت دستی بوده است.

در (یو و همکاران، ۲۰۰۹) یک پیکره از نظرات در دامنة اخبار به صورت دستی برچسب زده شده‌اند. در این پیکره علاوه بر نظرات مستقیم، نظرات غیرمستقیم نظری "افزایش قیمت سهام" با برچسب خاص خود مشخص شده‌اند. در (موسات و تراسان- متوا، ۲۰۱۰) روشنی برای تعیین قطبیت نظرات غیرمستقیم در دامنة اقتصاد معرفی شده است که در ابتداء با استفاده از یک فرهنگ لغت خاص دامنة اصلاحات تخصصی متن را استخراج می‌کند. ساختهای اقتصادی به دو گروه مثبت و منفی تقسیم شده است. همچنین تعدادی اصلاح‌کننده مثبت و منفی تعریف شده است؛ سپس با به‌کارگیری یک فرمول ساده قطبیت شاخص در قطبیت اصلاح‌کننده آن ضرب شده و قطبیت کل نظر به دست می‌آید.

در (نوفrstی و شمس فرد، ۲۰۱۵) ابتداء مدلی برای نمایش یک نظر غیرمستقیم معرفی شده است؛ سپس بر اساس این مدل یک پایگاه دانش از نظرات غیرمستقیم ایجاد شده و از روی تاپل‌های موجود در پایگاه دانش تعدادی الگوی معنایی استخراج شده است. با به‌کارگیری تاپل‌های موجود در پایگاه دانش و نیز الگوهای استخراج شده از آن روشنی برای تعیین قطبیت نظرات در دامنة دارو پیشنهاد شده است.

در این مقاله روشنی نیمه‌خودکار برای ساخت یک پیکره از نظرات غیرمستقیم ارائه می‌شود؛ سپس پیکره ساخته شده به عنوان مجموعه آموزش در روش‌های یادگیری ماشین برای تعیین قطبیت نظرات غیرمستقیم مورد استفاده قرار می‌گیرد. بدین منظور تعدادی ویژگی نحوی و معنایی جدید برای استفاده در دسته‌بندهای یادگیری ماشین معرفی می‌شود. نتایج آزمایش‌های انجام گرفته، مفیدبودن این

¹ Twitter



¹ <http://nlp.stanford.edu/software/corenlp.shtml>

² Compound sentence

³ Dependency tree

⁴ Tokenization

⁵ lemmatization

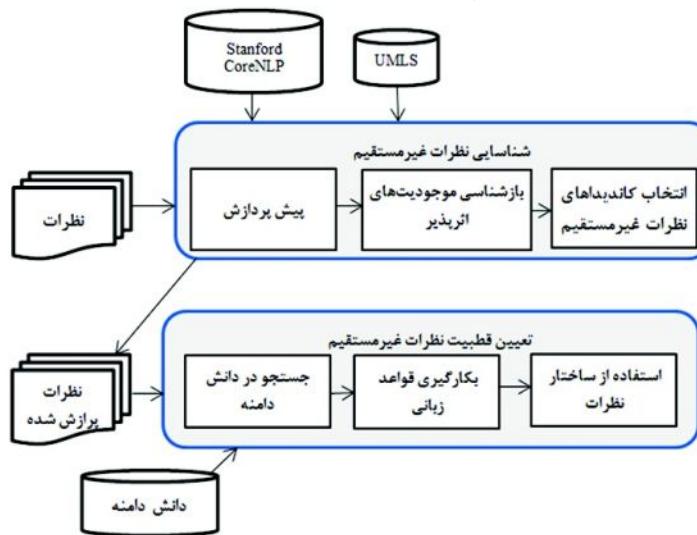
⁶ POS tagging

سال ۱۳۹۵ شماره ۲ پیاپی ۲۸

مرحله دوم قطبیت مثبت و منفی آنها تعیین می‌شود. نمای کلی روش پیشنهادی در شکل (۲) نشان داده شده است. در ادامه، مراحل کار با جزئیات تشریح می‌شود.

۲-۳- ساخت پیکره

بهمنظور ساخت یک پیکره از نظرات غیرمستقیم یک روش دومرحله‌ای معرفی می‌شود. در مرحله نخست جملاتی که بیان گر یک نظر غیرمستقیم هستند، شناسایی و در



(شکل-۲): نمای روش پیشنهادی برای ساخت پیکره

نمای می‌دهد. با این فرض بهمنظور شناسایی نظرات غیرمستقیم گام‌های زیر طی می‌شود (شکل ۲):
پیش‌بردازش: برای هر نظر در مجموعه نظرات، ابتدا با توجه به ساختار نظر، تنها جملاتی که در بخش فواید یا عوارض مطرح شده‌اند، استخراج و از جملاتی که در بخش توضیحات نوشته شده‌اند، صرف‌نظر می‌شود؛ سپس با به‌کارگیری بسته پردازش زبان استنوفورد^۱ جمله‌های یک نظر از یکدیگر جدا می‌شوند؛ سپس هر جمله مركب^۲ به تعدادی واحد کوچک‌تر شکسته می‌شود. این کار با توجه به درخت وابستگی^۳ جمله و ساختار حروف بطریعه کار رفته در آن انجام می‌گیرد (دی مارنف و همکاران، ۲۰۰۶)؛ سپس جملات سؤالی، شرطی و امری حذف می‌شوند. اگرچه برخی از این نوع جملات می‌توانند حاوی نظر باشند؛ اما تکنیک تحلیل آن‌ها متفاوت از جملات خبری است. پس از آن برای هر جمله باقیمانده با کمک ابزارهای موجود در بسته پردازش زبان استنوفورد قطعه‌بندی^۴، لیمایی^۵ و برچسبزنی نحوی^۶ انجام می‌گیرد. در پایان نیز مرجع ضمایر تشخیص داده می‌شوند (لی و

۱-۲-۳- شناسایی نظرات غیرمستقیم

نظر غیرمستقیم اثر یک موجودیت را که اثرگذار نامیده می‌شود، بر موجودیتی دیگر به نام اثرپذیر بیان می‌کند. در ابتدا فرض شد هر جمله‌ای که شامل یک موجودیت اثرپذیر و یک موجودیت اثرگذار است یک نظر غیرمستقیم را نشان می‌دهد. اگر چه ۹۰ درصد جملات استخراج شده با این روش بهدرستی یک نظر غیرمستقیم را نشان می‌دهند؛ اما این روش بسیاری از نظرات غیرمستقیم را استخراج نمی‌کند. علت این امر این است که اغلب نظرات غیرمستقیم فقد موجودیت اثرگذار هستند. از آنجا که کاربران تارنامه‌ای نظرسنجی اغلب درباره یک موجودیت مشخص (موجودیتی که نام آن در عنوان پست ذکر شده است) نظر می‌دهند، بهندرت نام آن موجودیت را در متن نظر خود ذکر می‌کنند؛ بلکه تنها به بیان اثر آن موجودیت می‌پردازند. به مثال‌های زیر توجه کنید:

“reduced amount of cystic acne”

“Lips were chapped and nose was dried.”

“no migraine”

در این نظرات تنها به اثرات دارو اشاره شده است و نامی از دارو نیامده است. بنابراین فرض شده هر جمله‌ای که حاوی یک موجودیت اثرپذیر است، یک نظر غیرمستقیم را

(جدول-۱): مثالی از موجودیت‌های تشخیص داده شده
توسط متامپ

“Over the course of three nights, Avelox caused heart palpitations, confusion, and lasting muscle weakness.”	
Entity	Semantic type
Course	Tmco
three	Qnco
nights	Tmco
Avelox	orch-phsu
caused	ftcn-ence
heart	Bpoc
palpitations	Fndg
confusion	Fndg
muscle weakness	Fndg

همان‌طورکه در جدول (۱) دیده می‌شود، علاوه‌بر مفاهیم مدنظر (یعنی “palpitations” “confusion” و “muscle weakness”) (“muscle weakness” نیز برچسب خورده‌اند. برای حل این مشکل باید مفاهیم مهم خاص دامنه انتخاب و مفاهیم عمومی حذف شوند. مفاهیم عمومی برخلاف مفاهیم خاص دامنه در اغلب دامنه‌ها پرخداد هستند. بنابراین می‌توان مفاهیمی را که در دامنه‌های مختلف پرخداد هستند، به عنوان مفاهیم عمومی در نظر گرفت و آن‌ها را حذف کرد.

به منظور تعیین مفاهیم مهم و مرتبط دامنه معیاری با نام STF_IOF ^۵ (تعداد رخداد نوع معنایی-معکوس تعداد رخداد نظر) تعریف شده است. این معیار اصلاح شده معیار CF_IOF ^۶ است که در (کمپریا و حسین، ۲۰۱۲) به منظور تعیین اهمیت یک مفهوم در یک متن خاص معرفی شده است. STF_IOF که اهمیت یک مفهوم را در یک دامنه خاص اندازه می‌گیرد، به صورت زیر تعریف شده است:

$$STF_IOF_{st,d} = \frac{n_{st,d}}{\sum_k n_{k,d}} \log \sum_k \alpha \cdot \frac{n_k}{n_{st}} \quad (1)$$

که $n_{st,d}$ تعداد رخداد نوع معنایی st در مجموعه نظرات d است. n_k تعداد کل رخدادهای انواع معنایی و α تعداد رخداد نوع معنایی st در کل مجموعه نظرات است. نیز یک پارامتر ثابت است که اهمیت نسبی STF در مقابل IOF را نشان می‌دهد.

برای استفاده از رابطه (۱) نیاز به یک مجموعه از نظرات است که بخشی از نظرات (به نام d) مربوط به دامنه مدنظر و بخش دیگر شامل نظرات در دیگر دامنه‌ها باشد. یک مجموعه از نظرات از سه دامنه دارو، رستوران^۷ و

همکاران، ۲۰۱۳) و هر ضمیر با عبارتی که به آن ارجاع دارد، جایگزین می‌شود.

بازشناسی موجودیت‌های اثربازی: همان‌طورکه پیش از این گفته شد، فرض شده است هر جمله‌ای که شامل یک موجودیت اثربازی باشد، یک نظر غیرمستقیم است. بنابراین به منظور شناسایی نظرات غیرمستقیم، در ابتدا نیاز است انواع ممکن برای موجودیت‌های اثربازی تعیین شود. پیش‌بینی انواع ممکن برای موجودیت اثربازی امری دشوار است. به عنوان مثال اگرچه می‌دانیم “آکوتان” یک داروی پوستی است و به احتمال بسیاری از نظرات، اثر این دارو را بر روی پوست بیان می‌کنند، با این وجود نمی‌دانیم عوارض جانبی این دارو بر روی چه اندام‌هایی اثر می‌گذارد.

در این مقاله به منظور تعیین انواع ممکن برای موجودیت اثربازی یک روش خودکار پیشنهاد شده است. ایده این است که نوع موجودیت‌های پرخداد را به عنوان نامزد در نظر بگیریم. دلیل درست‌بودن این ایده این است که به طور معمول موجودیت‌هایی که اغلب کاربران درباره آن نظر می‌نویسند، موجودیت‌های مهم هستند. اگرچه گاهی کاربران درباره موضوعاتی نامربوط نیز می‌نویسند؛ اما از آنجا که این مطلب برای هر کاربر متفاوت از کاربر دیگر است، موجودیت‌های نامربوط کم‌رخداد هستند. بنابراین می‌توان موجودیت‌های کم‌رخداد را حذف کرد.

برای استخراج موجودیت‌های پرخداد، دویست نظر از تارنمای www.askapatient.com جمع‌آوری و از آن به عنوان مجموعه توسعه^۱ استفاده شده است. برای هر جمله در این مجموعه مفاهیم پژوهشی برچسب زده شده‌اند. برای این کار از نرم‌افزار متامپ^۲ استفاده شده است (آرونسون، ۲۰۰۱). این نرم‌افزار مفاهیم جمله را به مفاهیم تعریف شده در UMLS^۳ نگاشت می‌کند. UMLS یک فرهنگ جامع شامل ۱/۷ میلیون مفهوم است که در ۱۳۰ گروه دسته‌بندی شده‌اند. به هریک از این گروه‌ها یک نوع معنایی^۴ گفته می‌شود.

مشکل در این است که UMLS علاوه‌بر مفاهیم پژوهشی مفاهیم عمومی نظری زمان را نیز دارد. به جدول (۱) توجه کنید. این جدول مثالی را از یک نظر غیرمستقیم که توسط متامپ برچسب زده شده، نشان می‌دهد.

⁵ Semantic Type Frequency – Inverse Opinion Frequency

⁶ Concept Frequency – Inverse Opinion Frequency

⁷ <http://www.cs.cmu.edu/~mehrbod/RR/>

فصل نهم



بیان کند قطبیت آن منفی در نظر گرفته می‌شود. البته برای تعیین قطبیت، نظرات منفی کننده‌ها را نیز مدد نظر قرار داده‌ایم. منفی کننده‌ها واژه‌هایی هستند که قطبیت یک "migraine" جمله را تغییر می‌دهند. به عنوان مثال اگرچه "no" دارد؛ ولی "no migraine" به دلیل وجود منفی کننده "no" دارای قطبیت مثبت است. منفی کننده‌های یک جمله با استفاده از فهرستی از منفی کننده‌های رایج شامل "no"، "never" و "not" شناسایی می‌شوند. اگر این منفی کننده در بازه یک پنجره سه کلمه‌ای از فعل جمله یا موجودیت اثربازیر قرار داشته باشد، قطبیت جمله را معکوس می‌کند؛ در غیر این صورت از آن صرف‌نظر می‌شود. عدد سه با روش سعی و خطاب تعیین شده است.

قواعد زبانی: در این مرحله قطبیت برخی از نظرات که تابه‌حال برچسب قطبیت نخورده‌اند با استفاده از قواعد زبانی تعیین می‌شود. در این مقاله دو قاعدة زبانی مورد استفاده قرار گرفته است (لیو، ۲۰۱۲).

- قاعدة "و": این قاعده می‌گوید جملات و عباراتی که با حروف ربط مشابه "و" به یکدیگر متصل می‌شوند، قطبیت یکسانی دارند. به عنوان مثال در متن "My skin dried up and the acne went away." دو جمله هر دو دارای قطبیت مثبت هستند.

- قاعدة "اما": این قاعده می‌گوید جملات و عباراتی که با حروف ربط مشابه "اما" مرتبط می‌شوند، اغلب دارای قطبیت متصاد هستند. به عنوان مثال دو جمله به کار رفته در متن زیر دارای قطبیت متصاد هستند.

"At first, I noticed an increase of energy, for about 4 days, but then extreme fatigue set it."

با به کارگیری دو قاعدة بالا اگر قطبیت یک جمله مشخص باشد (در گام پیشین قطبیت آن تعیین شده باشد) می‌توان قطبیت جمله دیگر را تعیین کرد.

ساختار نظرات: بسیاری از نظرات کاربران بیان گر اثراتی هستند که در دانش دامنه به‌وضوح و با آن الفاظ دیده نمی‌شوند. برای تعیین قطبیت این گونه نظرات از ساختار نظرات استفاده شده است. در ابتدا فرض شد جملاتی که در بخش فواید یک نظر مطرح می‌شوند، مثبت و جملات مطرح شده در بخش عوارض جانبی، منفی هستند. با این وجود این فرض در بخش فواید درست نیست. به عنوان

مثال متن زیر در بخش فواید یک نظر نوشته شده است:

"Benefits: Imitrex just isn't right for me. I experienced flushing, sensations of tingling/prickling,

محصولات (هو و لیو، ۲۰۰۴) گردآوری شده است. مجموعه نظرات در دامنه داروها همان مجموعه توسعه است که پیش‌تر توصیف شد. با اعمال معیار STF_IOF مفاهیم با اهمیت دامنه دارو استخراج شده است. برای این منظور نوع معنایی موجودیت‌های جملات مجموعه نظرات توسط متمامپ تعیین شده‌اند. اهمیت هر نوع معنایی با استفاده از معیار STF_IOF تعیین شده است. انواع معنایی که مقدار STF_IOF آن از آستانه θ بیشتر باشد به عنوان مفاهیم مهم دامنه انتخاب شده‌اند؛ سپس برای انواع معنایی پرخداد گروه معنایی متناظر با آن‌ها جایگزین شده است.

برای تعیین مقدار θ و α می‌توان از روش سعی و خطاب یا از یک الگوریتم تطبیق پارامتر استفاده کرد. با توجه به موفقیت الگوریتم‌های مبتنی بر اتوماتاهای یادگیر در تطبیق پارامتر (نوفرستی و رجایی، ۲۰۱۱)، برای تعیین مقدار مناسب پارامترها از اتوماتای یادگیر استفاده شده است که بهترتبیب برای پارامترهای θ و α مقادیر ۰.۰۲ و ۰.۱ به دست آمده است.

انتخاب نامزدها: پس از تعیین انواع ممکن برای موجودیت‌های اثربازیر هر جمله‌ای که دارای حداقل یک موجودیت اثربازیر باشد به عنوان یک نامزد برای نظر غیرمستقیم در نظر گرفته شده است.

نتایج آزمایش‌های انجام‌گرفته نشان می‌دهد که روش پیشنهادی برای شناسایی نظرات غیرمستقیم دارای دقت^۱ ۸۵/۰۵ درصد و فراخوانی^۲ ۷۷/۸۱ درصد است.

۲-۲-۳- تعیین قطبیت نظرات غیرمستقیم

به منظور تعیین قطبیت نظرات از دانش دامنه، الگوهای زبانی و ساختار نظرات استفاده شده است. در ادامه به توصیف هر کدام از این موارد می‌پردازیم.

دانش دامنه: بسیاری از فواید و عوارض دارویی شناخته شده هستند. به عنوان مثال "کاهش درد" از فواید شناخته شده "استامینوفن" و "خشکی لب" از عوارض شناخته شده "آکوتان" است. در این مقاله دانش دامنه که شامل فواید و عوارض شناخته شده داروهای مختلف است از تارنمای www.dailymed.com استخراج شده است. اگر در یک نظر غیرمستقیم درباره فایده‌ای شناخته شده از دارو صحبت شده باشد، قطبیت آن نظر را مشیت در نظر می‌گیریم. در مقابل اگر نظر عارضه‌ای شناخته شده از دارو را

^۱ precision

^۲ recall

موجودیت اثربذیر) استخراج می‌شود که البته اصلاح‌کننده می‌تواند مقدار نداشته باشد؛ سپس برای هر دوتایی درجه‌اطمینان محاسبه شده است. تنها دوتایی‌هایی که درجه‌اطمینان آن‌ها از یک آستانهٔ از پیش تعیین شده بیشتر باشد، انتخاب می‌شوند. در این مقاله با روش سعی و خطا مقدار آستانهٔ ۵/۰ تعیین شده است. درجه‌اطمینان را تنها برای دوتایی‌هایی که حداقل سه‌بار در مجموعهٔ نظرات رخداده‌اند محاسبه می‌کنیم و از سایر دوتایی‌ها صرف‌نظر می‌شود. در پایان برای هر دوتایی انتخاب شده اگر تعداد رخداد آن در بخش فواید بیشتر از تعداد رخداد آن در بخش عوارض باشد قطبیت آن را مثبت و در غیر این صورت قطبیت آن را منفی در نظر می‌گیریم. در این مرحله نیز منفی‌کننده‌ها مدنظر قرار گرفته‌اند. به عنوان مثال اگرچه دوتایی (reduce,sosy) دارای قطبیت مثبت است، اما اگر این دوتایی در جمله‌ای که حاوی یک منفی‌کننده است، ظاهر شود قطبیت جمله منفی خواهد بود.

جهت استخراج دوتایی (ریشهٔ فعل، نوع معنایی موجودیت اثربذیر) از یک روش فعل-محور که به‌منظور استخراج رابطه از متون پژوهشی معرفی شده (شاراما و همکاران، ۲۰۱۰) استفاده شده است. در این روش ابتدا فعل اصلی جمله تعیین می‌شود، سپس با استفاده از درخت واستگی جمله اگر موجودیت اثربذیر با فعل جمله رابطه‌ای نظیر مفعول، فاعل، اصلاح‌گر قیدی^۲ و غیره داشته باشد، دوتایی مذکور استخراج می‌شود. مؤلفهٔ نخست این دوتایی ریشهٔ فعل است که از ریشه‌یاب موجود در بستهٔ پردازش زبان استنفورد به‌دست می‌آید و مؤلفهٔ دوم نوع معنایی موجودیت اثربذیر است که توسط متمام‌تعیین می‌شود.

به‌طور مشابه برای استخراج دوتایی (اصلاح‌کننده، نوع معنایی موجودیت اثربذیر) از عبارات فاقد فعل، درابتدا موجودیت اثربذیر مشخص می‌شود؛ سپس به کمک تعدادی قاعدةٔ زبانی که بر روی درخت واستگی عبارت اعمال می‌شود، اصلاح‌کننده موجودیت اثربذیر به‌دست می‌آید.

به عنوان مثال در عبارت "reduction of acne" موجودیت اثربذیر "acne" است. در درخت واستگی این عبارت، لغت "reduction" که با موجودیت اثربذیر "acne" رابطهٔ "prep-of" دارد، به عنوان اصلاح‌کننده آن انتخاب می‌شود؛ سپس ریشهٔ اصلاح‌کننده با استفاده ازتابع getDerivationallyRelatedForms ریشه‌یاب استنفورد به‌دست می‌آید. در مثال بالا برای

² Adverbial modifier

weakness, drowsiness, dizziness, fast/pounding heartbeat. On top of that the packaging is just HORRIBLE. It comes 9 to a package. The packaging is a tri-fold booklet that is very rigid and has individual blisters for each tablet. It's about 6 1/2 by 3 1/2 inches, folded. I like to have migraine meds on me at all times so I ended up cutting the tablets out of the booklet so they would fit in my wallet or purse. Definitely going back to Axert!"

همان‌طور که دیده می‌شود، اغلب جملات متن بالا قطبیت منفی دارند. به‌منظور رفع این مشکل معیاری به نام درجه‌اطمینان تعریف شده است که درواقع بیان‌گر میزان اطمینان ما به مثبت یا منفی بودن یک نظر است. درجه‌اطمینان از رابطهٔ زیر محاسبه می‌شود:

$$\text{confidence}(e) = \frac{|PF-NF|}{PF+NF} \quad (2)$$

رابطهٔ (۲) متغیر e دوتایی (اثر، اثربذیر) که بیان‌گر یک نظر غیرمستقیم است، نشان می‌دهد. به‌منظور عمومی‌سازی به‌جای موجودیت اثربذیر نوع معنایی آن قرار داده می‌شود. همان‌طور که پیش‌تر نیز گفته شد، نوع معنایی توسط نرم‌افزار متمام‌به‌دست می‌آید. به عنوان مثال "کاهش درد" را به‌جای (reduce, pain) (reduce, sosy) نمایش می‌دهیم که sosy طبقهٔ معنایی علائم بیماری‌ها است. دلیل جایگزینی موجودیت اثربذیر با نوع معنایی آن، کاهش مشکل تنکی داده‌ها و افزایش پوشش الگوریتم است. مفاهیم دارای یک نوع معنایی، رفتار مشابهی در الگوریتم تعیین قطبیت نظر دارند. بدین دلیل در مواردی که مشکل تنکی داده‌ها وجود دارد و نمی‌توان یک نظر را به‌طور مستقیم تعیین قطبیت کرد، از نظرات دیگری که موجودیت اثربذیر آن‌ها دارای نوع معنایی یکسان با موجودیت مدنظر است، استفاده می‌شود. NF و PF نیز به‌ترتیب تعداد رخداد e در بخش فواید و عوارض هستند. بر طبق رابطهٔ (۲) اگر e تنها در بخش فواید (یا عوارض) رخداد باشد درجه‌اطمینان آن یک می‌شود. در مقابل اگر e به تعداد یکسان در بخش فواید و عوارض رخداد باشد، درجه‌اطمینان آن صفر می‌شود. درواقع هرچه درجه‌اطمینان یک اثر به یک نزدیک‌تر باشد، میزان اطمینان ما به مثبت (یا منفی) بودن آن اثر بیشتر است.

به‌منظور استفاده از معیار بالا به این صورت عمل شده است که در مجموعهٔ جملاتی که به عنوان نظر غیرمستقیم شناسایی شده‌اند، اگر نظر حاوی یک فعل باشد دوتایی (ریشهٔ فعل، نوع معنایی موجودیت اثربذیر) و اگر نظر یک عبارت بدون فعل باشد، دوتایی (اصلاح‌کننده)، نوع معنایی

¹ modifier





¹ chunk

² Semantic group

سال ۱۳۹۵ شماره ۲ پیاپی ۲۸

۴-۲- ویژگی ها

معرفی و انتخاب ویژگی های مناسب تأثیر بهسزایی در دقت روش های یادگیری ماشین دارد. در زمینه نظر کاوی ویژگی های متعددی معرفی و به کار گرفته شده اند که اغلب لغوی و نحوی هستند (مانند تکوازه ها، زوجوازه ها و برقسب نحوی). در این مقاله برخی از ویژگی های رایج در نظر کاوی در کنار تعدادی ویژگی جدید نظیر عبارت^۱ حاوی موجودیت اثربذیر، فعل جمله، نوع و گروه معنایی موجودیت اثربذیر به کار گرفته شده اند. این ویژگی ها در جدول (۳) توصیف شده اند.

(جدول-۳): ویژگی های مورد استفاده در یادگیری ماشین

اندازه بردار ویژگی	توصیف	ویژگی	نماد
با ریشه یابی	بدون ریشه یابی		
۲۹۹۷	۳۵۹۵	تکوازه ها	unigram F1
۱۵۶۷۶	۱۶۴۳۸	زوجوازه ها	bigram F2
۳۶	۳۶	برچسب نحوی	POS F3
۱۰۵۷	۱۲۰۸	فعل جمله که اثر بیان شده بر روی موجودیت اثربذیر را بیان می کند.	mainVerb F4
۲۰۸۰	۲۱۱۹	عبارت حاوی موجودیت اثربذیر	affected F5
۱۳۰		نوع معنایی موجودیت اثربذیر	type F6
۱۵		گروه معنایی موجودیت اثربذیر	group F7
این ویژگی یک ویژگی دودویی در نظر گرفته شده است.	آیا جمله دارای منفی کننده است؟	negation	F8

همان طور که در جدول (۳) مشاهده می شود، در این مقاله دو ویژگی معنایی به نامهای نوع و گروه معنایی موجودیت اثربذیر به کار گرفته شده است. همان طور که پیشتر گفته شد، مفاهیم UML در ۱۳۰ نوع معنایی گروه بندی شده اند. این انواع خود در پانزده گروه دسته بندی می شوند که هر دسته یک گروه معنایی^۲ نامیده می شود. به عنوان مثال مفهوم "بی خوابی" دارای نوع معنایی "علامت بیماری" و گروه معنایی "اختلال" است. به کار گیری ویژگی های معنایی باعث کاهش مشکل تکی داده ها و

اصلاح کننده "reduction" ریشه "reduce" به دست می آید. بدین ترتیب برای عبارت مثال، دوتایی (reduce, acne) حاصل می شود که در مرحله بعد به جای موجودیت اثربذیر نوع معنایی آن قرار داده می شود. جدول (۲) مثال هایی را از نظرات غیرمستقیم و دوتایی های استخراج شده از آن ها نشان می دهد.

(جدول-۲): مثال هایی از دوتایی های استخراج شده از نظرات غیرمستقیم

نظر غیرمستقیم	دوتایی استخراج شده
(experience, sosy)	I experienced severe headache.
(clear, dsyn)	Accutane cleared my acne.
(null, fndg)	Dry lips
(reduce, dsyn)	Reduction of acne

در پایان از نظراتی که قطبیت آن ها نامشخص باقی مانده است صرف نظر می کنیم.

۴- به کار گیری پیکره

به منظور ارزیابی کارآیی پیکره ساخته شده، از آن به عنوان مجموعه آموزش در روش های یادگیری ماشین استفاده شده است. بدین منظور برای هر جمله دارای برقسب ثبت و منفی در پیکره ویژگی های مدنظر استخراج شده است؛ سپس مجموعه جملات به عنوان مجموعه آموزش در یک دسته بند یادگیری ماشین استفاده شده است.

به منظور آموزش دسته بند های یادگیری ماشین بر روی پیکره ساخته شده، پیش پردازش های مختلف بر روی آن انجام شده است. همچنین ویژگی های مختلفی به منظور دسته بندی به کار گرفته شده است که در ادامه به توصیف آنها می پردازیم.

۴-۱- پیش پردازش

پیش پردازش متن یکی از گام های رایج در بسیاری از کاربردهای پردازش زبان طبیعی است. در این مقاله نیز سه پیش پردازش رایج به نامهای حذف هرزوازه ها، تبدیل متن به حروف کوچک و ریشه یابی استفاده شده است. اثر هر یک از این پیش پردازش ها در دقت الگوریتم نظر کاوی در بخش ۵ مورد ارزیابی قرار گرفته است.

۲-۵- ارزیابی روش پیشنهادی برای ساخت پیکره

با به کارگیری روش پیشنهادی برای ساخت پیکره (بخش ۳)، پیکره‌ای از نظرات غیرمستقیم ساخته شد. مشخصات این پیکره در جدول (۴) ارائه شده است.

(جدول -۴): مشخصات پیکره ساخته شده

۳۸۵۶	تعداد جملات
۳۵۰۶۴	تعداد کلمات
۱۳۳۲	تعداد جملات مثبت
۲۵۲۴	تعداد جملات منفی

به منظور ارزیابی کیفیت پیکره ساخته شده، از مجموعه داده‌ای مورد استفاده برای ساخت پیکره (بخش ۱-۳) زیر مجموعه‌ای شامل ۸۲ نظر انتخاب شده است؛ سپس از دو شخص خواسته شده است که نظرات غیرمستقیم این مجموعه را استخراج کرده و به آن‌ها برچسب قطبیت (مثبت یا منفی) اختصاص دهند. در مواردی که این دو شخص اختلاف نظر داشتند شخص سوم قطبیت نهایی نظر را تعیین کرده است. بدین ترتیب مجموعه‌ای شامل ۳۰۷ نظر غیرمستقیم ایجاد شده است. در پایان پیکره ساخته شده با روش پیشنهادی با این پیکره دست‌ساز مقایسه شده است. جدول (۵) دقت روش پیشنهادی برای شناسایی نظرات غیرمستقیم و نیز روش پیشنهادی برای تعیین قطبیت نظرات را نشان می‌دهد. در ستون سوم جدول (۵) نیز دقت پیکره ساخته شده گزارش داده شده است.

(جدول -۵): دقت روش پیشنهادی برای ساخت پیکره

دقت پیکره	دقت روش شناسایی تعیین قطبیت	دقت روش غیرمستقیم نظرات
۷۴/۴۲	۹۰/۵۶	۸۵/۰۵

همان‌طور که در جدول (۵) مشاهده می‌شود، دقت الگوریتم تعیین قطبیت ۹۰/۵۶ درصد است. در حالی که دقت یک روش پایه که تنها با توجه به ساختار نظرات به آن‌ها برچسب قطبیت اختصاص می‌دهد، ۸۳/۴۵ درصد است. درواقع، روش پایه به ترتیب با دقت ۷۹/۸۷ و ۸۷/۳۲ درصد، نظرات مثبت و منفی را تعیین قطبیت می‌کند. بنابراین روش پیشنهادی برای تعیین قطبیت بر روی این روشی که به نظرات بخش

افزایش پوشش الگوریتم یادگیری می‌شود. همچنین از آنجا که در یک نظر غیرمستقیم اغلب اثر موجودیت اثرگذار بر روی موجودیت اثربردار توسط فعل جمله بیان می‌شود، فعل نیز به عنوان یک ویژگی مؤثر در نظر گرفته شده است. در بخش ۵، کارآیی هر یک از این ویژگی‌ها مورد بررسی قرار گرفته است.

۵- نتایج

در این بخش درابتدا مجموعه آزمون مورد استفاده، توصیف می‌شود؛ سپس آزمایش‌های انجام گرفته به منظور ارزیابی روش‌های پیشنهادی ارائه می‌شود. آزمایش‌های انجام گرفته به دو دسته کلی تقسیم می‌شوند. دسته نخست به ارزیابی روش پیشنهادی برای ساخت پیکره می‌پردازد و دسته دوم به سنجش کارآیی روش پیشنهادی برای به کارگیری پیکره جهت تعیین قطبیت نظرات اختصاص دارد.

۱-۵- مجموعه آزمون

از آنجا که یک مجموعه آزمون عمومی که خاص مسئله تعیین قطبیت نظرات غیرمستقیم در دامنه دارو طراحی شده باشد وجود ندارد؛ اقدام به ساخت یک مجموعه دستی از نظرات غیرمستقیم شده است. بدین منظور بخشی از نظرات مطرح شده در باره نوزده دارو در دو تارنمای نظرسنجی www.druglib.com و www.askapatient.com گردآوری شده‌اند. این نوزده دارو به صورت تصادفی از فهرست داروهایی که بیشترین تعداد نظرات را داشته‌اند، انتخاب شده‌اند. همچنین دقت شده است که نظرات انتخاب شده از تارنمای www.druglib.com با مجموعه نظرات انتخاب شده برای ساخت پیکره، هم‌پوشانی نداشته باشند؛ سپس از دو شخص خواسته شده است که نظرات غیرمستقیم این مجموعه را استخراج کرده و به آنها برچسب قطبیت (مثبت یا منفی) اختصاص دهند. تنها جملاتی که توسط هر دو شخص به عنوان نظر غیرمستقیم انتخاب شده‌اند و قطبیت یکسانی نیز برای آن‌ها در نظر گرفته شده است، در مجموعه آزمون نهایی قرار گرفته‌اند. بدین ترتیب مجموعه‌ای از ۲۸۵ نظر غیرمستقیم ساخته شده است که حدود شصت درصد آن‌ها مثبت و بقیه نظرات منفی هستند.

فصل نهم



دیده می‌شود، حساس‌بودن به حروف کوچک و بزرگ باعث افزایش دقت دسته‌بند در حدود ۰/۴ درصد می‌شود.



الف: بررسی اثر حذف هرزواژه‌ها



ب: بررسی حساس‌بودن به حروف کوچک و بزرگ



ج: بررسی اثر ریشه‌یابی

(شکل-۳): اثر پیش‌پردازش‌های انجام گرفته بر روی دقت دسته‌بندی

در آزمایش سوم اثر ریشه‌یابی توکن‌های مجموعه آموزش بررسی شده است. شکل (۳-ج) نشان می‌دهد که ریشه‌یابی دقت دسته‌بند را ۴/۵ درصد افزایش می‌دهد. در آزمایش‌های بعدی پیش‌پردازش‌های حذف هرزواژه‌ها، تبدیل توکن‌ها به حروف کوچک و ریشه‌یابی انجام شده‌اند.

۳-۲-۵- ارزیابی اثربخشی ویژگی‌های معرفی شده برای تحلیل نظرات غیرمستقیم

در سری بعدی آزمایش‌ها تأثیر ویژگی‌های معرفی شده بررسی شده است. شکل (۴) دقت دسته‌بندی‌های بیزین و ماشین بردار پشتیبان (SVM)^۴ را با استفاده از ویژگی‌های مختلف نشان می‌دهد. برای پیاده‌سازی ماشین بردار

فواید قطبیت مثبت و به نظرات بخش عوارض قطبیت منفی تخصیص می‌دهد، برتری دارد.

۳-۵- ارزیابی روش پیش‌نهادی برای به کارگیری پیکره

در این بخش کارآیی روش پیش‌نهادی برای به کارگیری پیکره ساخته شده جهت تعیین قطبیت نظرات، مورد ارزیابی قرار می‌گیرد. آزمایش‌های انجام گرفته بدین منظور به سه دسته تقسیم می‌شوند: ارزیابی پیش‌پردازش‌های انجام‌شده بر روی پیکره، ارزیابی اثربخشی ویژگی‌های معرفی شده برای تحلیل نظرات غیرمستقیم و ارزیابی الگوریتم تعیین قطبیت. در ادامه، به معرفی و شرح نتایج این آزمایش‌ها می‌پردازیم.

۳-۵-۱- ارزیابی پیش‌پردازش‌های انجام‌شده بر روی پیکره

همان‌طور که در بخش (۱-۴) گفته شد، پیش‌پردازش‌های متعددی بر روی مجموعه آموزش انجام گرفته‌اند. در ابتدا اثر هریک از این پیش‌پردازش‌ها بررسی می‌شود. در این آزمایش‌ها تنها از ویژگی‌تکواژه‌ها استفاده شده است.

در آزمایش نخست، تأثیر حذف هرزواژه‌ها بررسی شده است. بدین منظور دسته‌بند بیزین ساده (NB)^۱ در نرم‌افزار وکا^۲ (هال و همکاران، ۲۰۰۹) یکبار بر روی مجموعه آموزش با وجود هرزواژه‌ها و بار دیگر بدون هرزواژه‌ها اجرا شده است. بهمنظور حذف هرزواژه‌ها از دو فهرست هرزواژه استفاده شده است: فهرست موجود در وکا و فهرست از هرزواژه‌های به کارگرفته شده بهمنظور نظرکاوی در دامنه فیلم^۳. نتایج به دست آمده در شکل (۳-الف) از این شده‌اند. همان‌طور که در شکل دیده می‌شود، حذف هرزواژه‌ها با استفاده از فهرست خاص نظرکاوی در دامنه فیلم دقت را به میزان ۹/۶ درصد افزایش می‌دهد. دقت عبارت است از درصدی از مثال‌های مجموعه آزمون که به درستی دسته‌بندی شده‌اند. با توجه به نتایج به دست آمده در آزمایش‌های بعدی هرزواژه‌ها حذف شده‌اند.

در آزمایش بعدی یکبار دسته‌بند به حروف بزرگ و کوچک حساس است و بار دیگر تنها تکواژه‌ها با حروف کوچک در نظر گرفته شده‌اند. همان‌طور که در شکل (۳-ب)

¹ Naïve Bayes

² Weka

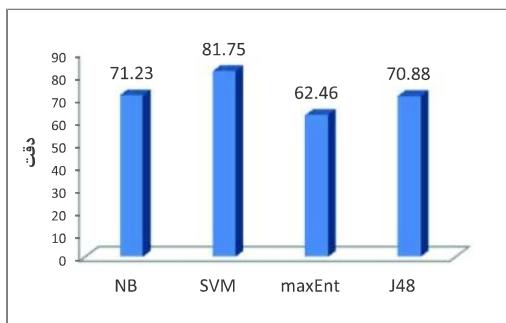
³ <https://github.com/kuberkaul/SentimentAnalysis-MovieReviews/blob/master/sentiment%20analysis%28usin%20naive%20bayes%29/src/stopwords.txt>

⁴ Support Vector Machine

سال ۱۳۹۵ شماره ۲ پیاپی ۲۸

قابل ملاحظه‌ای افزایش می‌دهد. در دسته‌بند ماشین بردار پشتیبان، بیشترین دقต با به کارگیری کلیه ویژگی‌های معروف شده در جدول (۳) حاصل شده است.

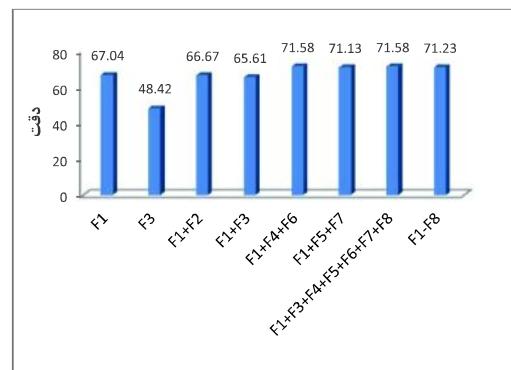
۳-۳-۵- ارزیابی الگوریتم تعیین قطبیت
برای تعیین قطبیت نظرات مجموعه آزمون، یک الگوریتم دسته‌بندی با آموزش بر روی پیکرۀ پردازش شده و با به کارگیری ویژگی‌های معروف شده در جدول (۳) به یادگیری می‌پردازد. بدین منظور می‌توان از الگوریتم‌های دسته‌بندی و روش‌های وزن‌دهی مختلف استفاده کرد. در این راستا، در ابتدا آزمایش‌هایی جهت ارزیابی دقت دسته‌بندی‌های مختلف و نیز بررسی اثر وزن‌دهی به ویژگی‌ها انجام گرفته است؛ سپس دقت الگوریتم پیشنهادی برای تعیین قطبیت با دقت یک روش موجود تحلیل نظرات مستقیم مقایسه شده است.
در آزمایش نخست دقت دسته‌بندی‌های مختلف با یکدیگر مقایسه شده است (شکل ۵). دلیل انتخاب این دسته‌بندی‌ها استفاده مکرر آن‌ها در تعیین قطبیت نظرات مستقیم و رضایت‌بخش بودن نتایج حاصل از آن‌ها است. این دسته‌بندی‌ها با استفاده از ترکیب همه ویژگی‌های نامبرده در جدول (۳) اجرا شده‌اند. شکل (۵) نشان می‌دهد که SVM به نتایج بهتری دست یافته است. همان‌طور که در جدول (۳) مشاهده می‌شود، اندازه بردار ویژگی‌های مورد استفاده بزرگ، ولی فضای برداری متون خلوت است. در پژوهش‌های پیشین نیز نشان داده شده است که الگوریتم‌های افزایشی مانند SVM که بایاس استنتاجی دارند، عملکرد خوبی با ماتریس‌های خلوت نشان می‌دهند (کیوین، ۱۹۹۵).



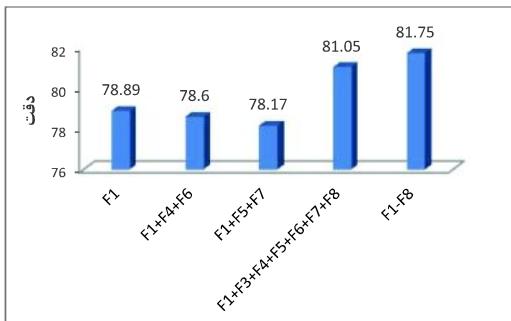
(شکل ۵): مقایسه دقت دسته‌بندی‌های مختلف در تعیین قطبیت نظرات غیرمستقیم

در آزمایش‌های قبلی، حضور/عدم حضور تکوازه‌ها و زوجوازه‌ها به عنوان ویژگی‌های دودویی در نظر گرفته شده است. این دو ویژگی به عنوان مبنای در بسیاری از کارهای

پشتیبان نیز از ابزار وکا با کرنل چندجمله‌ای استفاده شده است.



الف: دسته‌بند بیزین



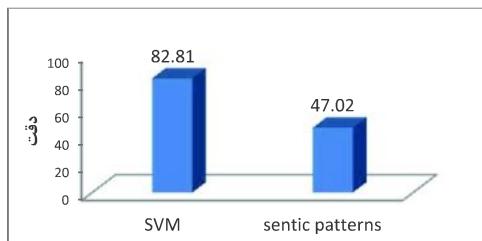
ب: دسته‌بند SVM

(شکل ۴-۴): اثر به کارگیری ترکیب‌های مختلف از ویژگی‌های معروف شده در دقت الگوریتم تعیین قطبیت

در شکل (۴-الف) دو حالت پایه در نظر گرفته شده است: زمانی که تنها از تکوازه‌ها استفاده می‌شود و زمانی که تنها از برچسب نحوی استفاده می‌شود. همچنین ترکیب تکوازه‌ها با برچسب نحوی و با زوجوازه‌ها که در نظر کاوی پرکاربرد می‌باشد، نیز استفاده شده است. علاوه بر آن ترکیب تکوازه‌ها با ویژگی‌های معنایی معرفی شده نیز به کارگرفته شده است. همان‌طور که در شکل مشاهده می‌شود استفاده از کلیه ویژگی‌های معرفی شده در جدول (۳) به جز ویژگی زوجوازه‌ها باعث افزایش $\frac{4}{5}$ درصدی و به کارگیری کلیه ویژگی‌های معرفی شده در جدول (۳) باعث افزایش $\frac{4}{2}$ درصدی دقت دسته‌بند بیزین در مقایسه با حالت پایه استفاده از تکوازه‌ها شده است.

در شکل (۴-ب) نیز اثر ترکیب ویژگی‌های معنایی معرفی شده با حالت پایه تکوازه‌ها در دسته‌بند ماشین بردار پشتیبان بررسی شده است. همان‌طور که مشاهده می‌شود افزودن ویژگی‌های معنایی دقت دسته‌بند را به طور

در شکل (۶) مشاهده می‌شود که روش پیشنهادی در مقایسه با الگوهای سنتیک به نتایج بهتری دست یافته است. با توجه به این که الگوهای سنتیک یک روش برجسته موجود در زمینه تحلیل نظرات مستقیم است، می‌توان نتیجه گرفت روش‌های موجود نظرکاوی در زمینه تعیین قطبیت نظرات غیرمستقیم، ضعیف عمل می‌کنند. به بیانی دیگر مسئله تعیین قطبیت نظرات غیرمستقیم نیازمندی‌های خاص خود را دارد و برای رسیدن به نتایج قبل قبول باید این نیازمندی‌ها مورد توجه قرار گیرند.



(شکل-۶) : مقایسه روش پیشنهادی با الگوهای سنتیک

۶- نتیجه‌گیری

در این مقاله در ابتدا روشی برای ساخت نیمه‌خودکار یک پنجه از نظرات غیرمستقیم ارائه شد. با به کارگیری این روش یک پنجه شامل ۳۸۵۶ نظر غیرمستقیم دارای برچسب مثبت/منفی ایجاد شد؛ سپس از این پنجه به عنوان مجموعه آموزش در روش‌های یادگیری ماشین برای دسته‌بندی نظرات یک مجموعه آزمون استفاده شد. بدین منظور اثر پیش‌پردازش‌های مختلف بر روی پنجه مورد بررسی قرار گرفت. همچنین مجموعه‌ای از ویژگی‌های لغوی، نحوی و معنایی در چهار دسته‌بند بیزین ساده، ماشین بردار پشتیبان، J48 و بیشینه آنتروپی به کار گرفته شد. نتایج آزمایش‌های انجام گرفته به منظور ارزیابی روش پیشنهادی نشان می‌دهد که به کارگیری ترکیب ویژگی‌های معروف شده، دقت الگوریتم تعیین قطبیت را به میزان قابل توجهی در مقایسه با حالات پایه (استفاده از تکوازه‌ها و استفاده از برچسب نحوی به تنها) افزایش می‌دهد. همچنین روش پیشنهادی در تعیین قطبیت نظرات مجموعه آزمون به نتایج بسیار بهتری در مقایسه با یک روش برجسته موجود در تعیین قطبیت نظرات مستقیم به نام الگوهای سنتیک می‌رسد.

پیشین مورد استفاده قرار گرفته‌اند. در جدول (۶) نتایج دو آزمایش دیگر ارائه شده است. در آزمایش نخست به جای ویژگی‌های دودویی از ویژگی‌های وزن دار استفاده شده است. وزن یک ویژگی با معیار TF-IDF^۱ تعیین شده است. جدول (۶) نشان می‌دهد که وزن دهی به ویژگی‌ها دقت الگوریتم‌های یادگیری ماشین را افزایش داده است.

در آزمایش دوم برای کاهش فضای ویژگی تنها تکوازه‌ها و زوج‌وازه‌های پر تکرار انتخاب شده‌اند. در این آزمایش حداقل تعداد تکرار یک ویژگی در بسیاری از کارهای پیشین (هابرناال و همکاران ۲۰۱۴) مورد استفاده قرار گرفته است. همان‌طور که در جدول (۶) مشاهده می‌شود، کاهش فضای ویژگی دقت دسته‌بندها را کاهش داده است.

(جدول-۶): بررسی تأثیر وزن دهی به ویژگی‌ها و نیز

به کارگیری ویژگی‌های پر تکرار در دقت الگوریتم تعیین قطبیت

ویژگی‌های پر تکرار	TF-IDF	بدون وزن دهی و حذف ویژگی‌های کم رخداد	دسته‌بند
۷۰/۸۸	۷۳/۳۳	۷۱/۲۳	NB
۸۱/۰۵	۸۲/۸۱	۸۱/۷۵	SVM

در آزمایش پایانی نتایج به دست آمده توسط دسته‌بند SVM با وزن دهی به ویژگی‌ها با نتایج حاصل از یکی از روش‌های برجسته موجود در تحلیل نظرات مستقیم به نام الگوهای سنتیک^۲ (پوریا و همکاران ۲۰۱۴) مقایسه شده است. این روش که متنکی بر قواعد زبانی، محاسبات عرفی^۳ و یادگیری ماشین است، از سنتیکنت^۴ که یک پایگاه دانش از مفاهیم عرفی قطبی است، استفاده می‌کند. اگرچه در سنتیکنت بسیاری از مفاهیم رایج در دامنه پژوهشی مانند "آکنه"، "معده درد"، "افزايش وزن" و "کاهش استرس" وجود دارند که می‌توانند به درستی برخی از نظرات غیرمستقیم را تعیین قطبیت کنند، با این وجود این روش در تعیین قطبیت اغلب نظرات غیرمستقیم با شکست مواجه می‌شود. دلیل اصلی این امر نیز این است که الگوهای سنتیک بیشتر متنکی بر مفاهیم به کار گرفته در جمله عمل می‌کند و در اغلب موارد اثر فعل جمله را نادیده می‌گیرد.

¹ Term Frequency - Inverse Document Frequency

² sentic patterns

³ common sense computing

⁴ www.sentic.net

۷- مراجع

- Kaji, N., and Kitsuregawa, M. Automatic construction of polarity-tagged corpus from HTML documents. Proceedings of the COLING/ACL on Main conference poster sessions, 2006.
- Kivinen, J., and Warmuth, M. K. The perceptron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. Proceedings of the eighth annual conference on Computational learning theory. ACM, 1995.
- Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics, 2013, 39(4), pp. 885-916 .
- Liu, B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012, 5(1), pp. 1-167 .
- Liu, K.-L., Li, W.-J., and Guo, M. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. AAAI, 2012.
- Mintz, M., Bills, S., Snow, R., and Jurafsky, D. Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, Vol. 2.
- Moraes, R., Valiati, J. F., and Gavião Neto, W. P. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 2013, 40(2), pp. 621-633 .
- Musat, C., and Trausan-Matu, S. A comparative study of the relevance of indirect and direct opinions in economic texts. Annals of DAAAM & Proceedings, 2010 .
- Musat, C., and Trausan-Matu, S. The impact of valence shifters on mining implicit economic opinions. Artificial Intelligence: Methodology, Systems, and Applications, Springer, 2010, pp. 131-140.
- Nakagawa, T., Inui, K., Kurohashi, S. Dependency tree-based sentiment classification using CRFs with hidden variables. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010.
- Nevarouskaya, A., Prendinger, H., and Ishizuka, M. SentiFul: A lexicon for sentiment analysis. Affective Computing, IEEE Transactions, 2011, 2(1), pp. 22-36 .
- Noferesti, S., and Rajaei, M. A Hybrid algorithm based on ant colony system and learning automata for solving Steiner tree problem. International Journal of Applied Mathematics and Statistics™, 2011, 22 (S11), pp. 79-88 .
- Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium, 2001.
- Balahur, A., Hermida, JM., Montoyo, A., and Muñoz, R. EmotiNet: a knowledge base for emotion detection in text built on the appraisal theories. Natural Language Processing and Information Systems: Springer, 2011, pp. 27-39.
- Balahur, A., Hermida, JM., and Montoyo, A. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. IEEE Transactions on Affective Computing, 2012, Vol. 3, pp. 88-101.
- Boiy, E., and Moens, M. F. A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval, 2009, 12(5), pp. 526-558.
- Bosco, C., Patti, V., and Bolioli, A. Developing corpora for sentiment analysis and opinion mining: the case of irony and senti-tut. IEEE Intelligent Systems, 2013.
- Cambria, E., and Hussain, A. Sentic computing: Springer, 2012.
- De Marneffe, M.-C., MacCartney, B., and Manning, C. D. Generating typed dependency parses from phrase structure parses. Proceedings of LREC, 2006.
- Esuli, A., and Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. Proceedings of LREC, 2006.
- Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009, pp. 1-12
- Goeuriot, L., Na, J.-C., Min Kyaing, W. Y., Khoo, C., Chang, Y.-K., Theng, Y.-L., and Kim, J.-J. Sentiment lexicons for health-related opinion mining. Proceedings of the 2nd ACM SIGKDD International Health Informatics Symposium, 2012.
- Habernal, I., Ptáček, T., and Steinberger, J. Supervised sentiment analysis in Czech social media. Information Processing & Management, 2014, 50(5), pp. 693-707 .
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009, 11(1), pp. 10-18 .
- Hu, M., and Liu, B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.
- Huang, S., Niu, Z., and Shi, C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowledge-Based Systems, 2014, 56, pp. 191-200 .



Noferesti, S., and Shamsfard, M. Resource construction and evaluation for indirect opinion mining of drug reviews. *PLoS One*, 2015, 10(5).

Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002, Vol. 10.

Poria, S., Cambria, E., Winterstein, G., and Huang, G.-B. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems.*, 2014

Sharma, A., Swaminathan, R., and Yang, H. A verb-centric approach for relationship extraction in biomedical text. *Semantic Computing (ICSC)*, 2010.

Yu, B., Diermeier, D., and Kaufmann, S. The Wal-Mart Corpus: A multi-granularity corporate opinion corpus for opinion retrieval, classification and aggregation, working paper.



سمیرا نوفرستی در سال ۱۳۸۲ دوره کارشناسی خود را رشته مهندسی کامپیوتر در دانشگاه صنعتی شریف به پایان رسانده است. در سال ۱۳۸۴ مدرک

کارشناسی ارشد خود را در همان رشته از دانشگاه صنعتی امیرکبیر اخذ کرده است. در حال حاضر دانشجوی دکترای دانشگاه شهید بهشتی و عضو هیئت علمی دانشگاه سیستان و بلوچستان است.

نشانی رایانامه ایشان عبارت است از:

snoferesti@ece.usb.ac.ir



مهرنوش شمس فرد مدرک کارشناسی و کارشناسی ارشد ایشان در رشته مهندسی کامپیوتر از دانشگاه صنعتی شریف گرفته شده است. مدرک دکترای ایشان در

گرایش هوش مصنوعی در سال ۱۳۸۱ از دانشگاه صنعتی امیرکبیر اخذ شده است. ایشان در حال حاضر عضو هیئت علمی دانشگاه شهید بهشتی و سرپرست آزمایشگاه پردازش زبان طبیعی در دانشکده مهندسی و علوم کامپیوتر هستند.

نشانی رایانامه ایشان عبارت است از:

m-shams@sbu.ac.ir