خوشه‌بندی مشخصات کالا در رسانه‌های اجتماعی فارسی زبان با استفاده از الگوریتم‌های یادگیری ماشین

محبوبه منتظری[۱]، زینب‌الهدی حشمتی[۲]، مصطفی صالحی[۳]

[۱] دانشگاه تهران، mbh.montazeri@ut.ac.ir

[۲] دانشگاه تهران، zheshmati@ut.ac.ir

[۳] دانشگاه تهران، mostafa_salehi@ut.ac.ir

چکیده – با گسترش اینترنت و پیدایش وب۲، رسانه‌های اجتماعی به عنوان ابزاری برای تبادل پیام، ایجاد ارتباطات دوسویه را امکان‌پذیر ساخته‌اند. فروشگاه‌های اینترنتی یکی از رسانه‌های اجتماعی هستند که کاربران می‌توانند کالاهای موردنیاز خود را در آن یافته و با دیگر کاربران تبادل نظر انجام دهند. بنابراین با توجه به حجم گسترده این نظرات، تحلیل و جمع‌بندی آنها برای خریدار و فروشنده محصول می‌تواند بسیار مفید باشد. از این‌رو پژوهش‌های گسترده‌ای در حوزه نظرکاوی و تحلیل احساسات کاربران انجام گرفته است. با توجه به اینکه هر کاربر می‌تواند بر روی مشخصات مختلف کالایی مانند دوربین، باتری یا حافظه در یک گوشی تلفن همراه نظر متفاوتی داشته باشد و هریک از این مشخصات با عبارات متفاوتی در نظرات مختلف بیان شوند، خوشه‌بندی هریک از مشخصات کالا می‌تواند دقت بالاتری را در تحلیل احساسات و نظرکاوی حاصل نماید. در حالیکه پژوهش‌های متعددی به منظور خوشه‌بندی مشخصات کالا در زبان‌های مختلف انجام شده است، ولی در زبان فارسی پژوهشی یافت نشد. در این پژوهش با استفاده از روش‌های یادگیری ماشین، مشخصات مربوط به یک خودرو در نظرات کاربران خوشه‌بندی شده‌اند و پس از مقایسه نتایج حاصل از هر روش، الگوریتم امیدبیشینه از نظر معیارهای انتروپی و خلوص منجر به نتایج بهتری شد.

کلید واژه– نظرکاوی، یادگیری ماشین، خوشه‌بندی مشخصات کالا، رسانه‌ی اجتماعی

۱- مقدمه

با گسترش فناوری و آشنایی افراد جامعه در استفاده از وسایل الکترونیکی مانند تلفن همراه، کامپیوتر و تبلت، به اشتراک‌گذاری محتوا در رسانه‌های اجتماعی توسعه پیدا کرده است. نمونه‌ای از مشارکت افراد، فروشگاه‌های اینترنتی است و نظرات کاربران درباره کالاها و خدمات ارائه شده در آن به مشتریان این فرصت را می‌دهد که مسائل خود را به‌طور مستقیم با فروشندگان و یا تولیدکنندگان در میان بگذارند و فروشندگان نیز بتوانند به سرعت و به طور مناسب نسبت به مسائل مطرح شده از سوی مشتریان عکس‌العمل نشان دهند. اما با توجه به حجم انبوه این نظرات، به منظور بهره‌گیری از آن، نظرکاوی و تحلیل احساسات امری مهم و اساسی به شمار می‌رود.

اولین پژوهش‌ها در زمینه تحلیل احساسات، درباره تعیین مستقیم قطبیت (مثبت یا منفی بودن) یک جمله یا سند مشخص انجام می گرفت[۱]. مشکل در سطح تحلیل سند این است که نمی‌توان اطلاعات جزئی‌تر مانند احساسات مثبت یا منفی را با توجه به مشخصات کالا در نظرات به دست آورد. یک کالا مشخصات زیادی دارد و کاربر می‌تواند نظر خود را روی تعداد زیادی از مشخصات بیان نماید. مثلا در جمله " این گوشی صفحه نمایش خوبی دارد، ولی عمر باتری آن کوتاه است"، نظر کاربر بر روی مشخصه صفحه نمایش، مثبت و روی مشخصه باتری، منفی است. بعلاوه در توصیف یک مشخصه، اغلب افراد از کلمات و عبارات متفاوتی استفاده می‌کنند، مانند "عمر باتری" یا "قدرت باتری". بنابراین در این سطح نیاز به مطالعه مشخصات کالا است و می توان گفت سیستم نظرکاوی در سطح تحلیل مشخصات شامل سه سطح استخراج مشخصات[۱] [۲] خوشه‌بندی[۲] [۳،۴] و نظرکاوی[۳] [۱] است است و تمرکز این پژوهش فقط بر روی خوشه‌بندی این مشخصات است.

با توجه به ساختار خاص متن در جملات محاوره‌ای که توسط کاربران تولید می‌شود، مثل استفاده از اصطلاحات عامیانه و مبهم و چندشکلی نوشتن کلمات، مشکلاتی را جهت خوشه‌بندی مشخصات کالا ایجاد می‌کند. این مشکلات در زبان فارسی با توجه به وجود پیشوند و پسوندها، نوشتار فارسی لغات انگلیسی و گاهی نوشتار انگلیسی، حذف

برخی حروف در کلمات، بیشتر است. به علاوه در زمینه خوشه‌بندی مشخصات کالا در زبان فارسی پژوهشی یافت نشده است.

در این مقاله به منظور خوشه‌بندی مشخصات کالا، به ترتیب مراحل پیش پردازش داده‌ها، ایجاد داده‌های برچسب‌خورده و سپس بکارگیری الگوریتم های یادگیری ماشین طی شده است. در مرحله پیش پردازش داده‌ها در سطح محاوره، کتابخانه‌های موجود در بسته هضم[۵] توسعه داده شده و سپس در مرحله ایجاد داده‌های برچسب خورده، تعدادی از مشخصات با استفاده از شباهت واژگانی به عنوان داده های آزمایشی برچسب زده شدند. سپس به منظور خوشه‌بندی از این داده‌ها در الگوریتم امیدبیشینه(EM)[۴] و k_میانگین[۵] استفاده گردید. همچنین از این داده‌ها در الگوریتم k_میانگین نیز بهره گرفته شده و نتایج حاصل با روش‌های تصادفی انتخاب داده‌های آزمایشی و نیز خوشه‌بندی از طریق الگوریتم سلسله‌مراتبی[۶] مقایسه گردیده است. نتایج آزمایشات انجام گرفته در این پژوهش بیانگر بهبود نتایج از نظر معیار های انتروپی[۷] و خلوص[۸] از طریق روش امید بیشینه نیمه‌نظارتی است. به طور خلاصه این مقاله شامل نوآوری‌های زیر است:

- ارائه چارچوبی جهت خوشه‌بندی مشخصات کالا در رسانه های اجتماعی فارسی‌زبان (با توجه به اطلاعات پژوهشی انجام شده، اولین کار در این زمینه)
- پیش پردازش داده‌ها با توجه به محاوره‌ای بودن آنها و کمبود ابزارهای موردنیاز در این زمینه
- جمع‌آوری مجموعه لغات توقف در زبان محاوره
- ایجاد داده‌های برچسب‌خورده با در نظر گرفتن اشتراک کلمات و با توجه به ترتیب لغات تشکیل‌دهنده مشخصه

در ادامه این مقاله، در فصل دوم پژوهش مرتبط با این حوزه بیان شده است، در فصل سوم چارچوب پیشنهادی و در فصل چهارم نتایج حاصل ارائه شده‌اند. فصل پنجم نتیجه‌گیری و کارهای آتی را بیان می کند.

۲- پژوهش‌های مرتبط

به طور کلی می‌توان گفت در موضوع خوشه‌بندی مهم‌ترین مسئله انتخاب معیار شباهت بین خوشه‌ها می‌باشد و اصلی‌ترین این معیارها شامل شباهت معنایی[۹] و نیز شباهت توزیعی[۱۰] می‌باشند[۴]. در روش‌های مبتنی بر شباهت معنایی از لغت‌نامه و یا شبکه معنایی استفاده می‌شود[۶،۷]، ولی این روش‌ها در برخی موارد کارایی لازم را ندارند؛ زیرا بسیاری از عبارات مرتبط با یک مشخصه، عبارات چندکلمه‌ای هستند که به راحتی نمی‌توان در این موارد از لغت‌نامه‌ها استفاده کرد مانند "سرعت خودرو". همچنین بسیاری از عبارات مرتبط با مشخصه، مترادف‌های مختص دامنه هستند؛ مثلا "گران" و "ارزان" می‌توانند مربوط به "قیمت" باشند که هیچ یک از آنها مترادف با هم یا مترادف با "قیمت" نیستند[۴]. برای حل این مشکل کارنینی و همکاران در سال ۲۰۰۵[۸]، روشی را ارائه کرده‌اند که از معیارهای شباهت رشته ای[۱۱]، مترادف‌ها[۱۲] و فاصله واژگانی[۱۳] با استفاده از وردنت استفاده می کند.

در روش‌های مبتنی بر شباهت واژگانی فرض می‌شود که لغات با معانی مشابه در متون مشابهی ظاهر می‌شوند و برای محاسبه شباهت از معیارهایی مانند فاصله کسینوسی، اقلیدسی، جاکاردین و یا اطلاعات مشترک نقطه‌ای[۱۴] استفاده می شود[۲،۹،۱۰]. برخی پژوهش‌ها نیز از مدل‌سازی موضوع[۱۵] در حل مسائل خوشه‌بندی استفاده می‌کنند[۱۱] و عباراتی با موضوع یکسان را در یک گروه، خوشه‌بندی کرده و از تخصیص پنهان دیریکله[۱۶] استفاده می‌نمایند.

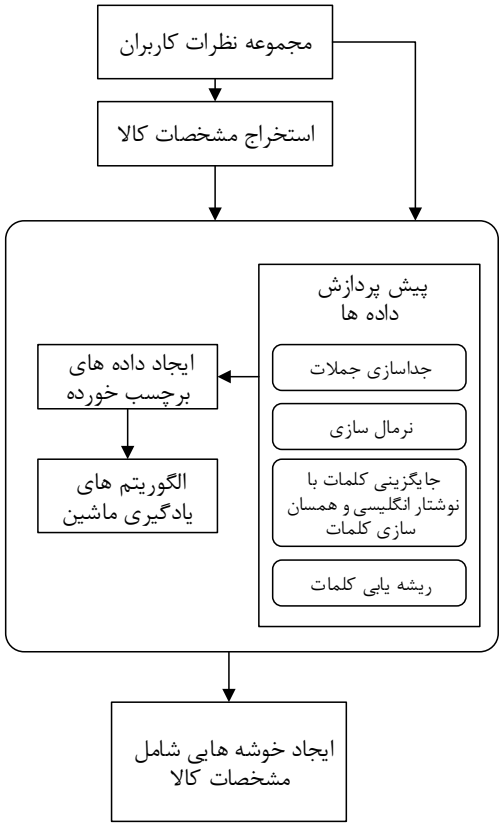
ژای و همکاران در سال ۲۰۱۱ [۴]، یک روش یادگیری نیمه‌نظارتی[۱۷] ارائه کرده‌اند که این روش از الگوریتم امید بیشینه استفاده می‌کند، با این فرض که عبارات مرتبط با مشخصه که واژه‌های مشترک دارند، مثل"توان باتری" و "عمر باتری"، و نیز مشخصاتی که در لغت‌نامه وردنت[۱۸] با هم مترادف هستند، احتمالا مربوط به یک خوشه هستند. نتایج بدست آمده بهبود خوشه‌بندی حاصل را نسبت به استفاده از الگوریتم‌های بدون نظارت[۱۹] نشان می‌دهد. چن و همکاران در سال ۲۰۱۶[۳]، به ازای هر مشخصه مجموعه مشخصات مرتبط و غیر مرتبط را شناسایی نموده و سپس با استفاده از لغت‌نامه و ایجاد بردار حروف برای هر مشخصه، با تعریف سه معیار جدید به منظور محاسبه شباهت، الگوریتم سلسله‌مراتبی را در خوشه‌بندی مشخصات در زبان چینی به کار گرفتند. یتگین و گزوکارا[۱۲] نیز در سال ۲۰۱۵ به منظور تعیین مشخصات و خوشه‌بندی آنها در زبان ترکی، معیارهای جدیدی به منظور محاسبه شباهت و نیز ارزیابی میزان کارایی الگوریتم معرفی نمودند.

با توجه به اهمیت خوشه‌بندی مشخصات کالا در زبان فارسی و نیز تعداد زیاد فارسی‌زبانان، پژوهشی از آن در دست نیست. با توجه به این موضوع، در این پژوهش بر روی خوشه‌بندی مشخصات کالا در نظرات کاربران به زبان فارسی پرداخته شده است.

۳- چارچوب پیشنهادی

به طور کلی خوشه‌بندی مشخصات کالا شامل پیش‌پردازش داده‌ها، ایجاد داده‌های برچسب‌خورده و استفاده از الگوریتم‌های یادگیری ماشین

است. شکل ۱ چارچوب کلی موردنظر را نشان می‌دهد. در ادامه جزئیات هریک از این مراحل بیان می‌شود.



شکل ۱. مراحل چارچوب پیشنهادی

۳-۱- پیش پردازش نظرات

ابتدا جملات با استفاده از علائم نگارشی مانند (.،:؟ \n) جدا می‌شوند. مرحله نرمال‌سازی (حذف برخی فواصل و اضافه کردن نیم‌فاصله در کلماتی مانند "ال‌سی‌دی" و "خوش‌دست" و حذف علائم نامربوط) انجام می‌شود. همچنین نیاز به همسان‌سازی کلمات است؛ برخی از کلمات مانند "داشتبورد" و "LCD" به ترتیب به "داشبورد" و "ال سی دی" تبدیل می‌شوند. همچنین می‌توان کلمات را ریشه‌یابی نمود؛ مثلا پسوندهای (تون، شون، مون، تان، شان، ها، هاتون) را از انتهای کلمات حذف کرده و برخی از کلمات که این پسوندها جزئی از آنهاست از این قواعد مستثنی می‌شوند. همچنین لیست کلمات زایدی که در دسترس است به منظور استفاده در متون رسمی است و برای استفاده در متون محاوره‌ای این کلمات گردآوری شدند.

۳-۲- ایجاد داده‌های برچسب‌گذاری‌شده

ابتدا باید تعدادی از مشخصات مشابه را در خوشه‌های مرتبط قرار داد. بدین منظور از این فرض استفاده می‌کنیم؛ مشخصاتی دارای کلمات مشترک با احتمال بالایی مربوط به یک خوشه هستند[۴] مانند "کیفیت باتری" و "شارژ باتری"، اما مشخصاتی مانند "کیفیت دوربین" و "کیفیت باتری" دارای کلمه مشترک "کیفیت" می‌باشند. بنابراین به منظور یافتن مشخصات مرتبط با مشخصه‌ای متشکل از بیش از یک کلمه، از اشتراک کلمه اول مشخصه موردنظر با سایر مشخصات صرفنظر می‌کنیم. مثلا برای یافتن مشخصات مرتبط با "کیفیت باتری" فقط به دنبال مشخصاتی دارای کلمه "باتری" هستیم و کلمه "کیفیت" را در اشتراک در نظر نمی‌گیریم. بدین ترتیب تعدادی خوشه بدست می آید که k تعداد از خوشه‌ها که دارای بیشترین عضو هستند را به عنوان داده برچسب‌خورده درنظر می‌گیریم. یکی از خوشه‌ها در نظرات مربوط به گوشی تلفن همراه می‌تواند (باتری، کارکرد باتری، شارژ باتری، مصرف باتری، نگهداری شارژ باتری، عملکرد باتری، شارژ) باشد. پس از این مرحله، باید به ازای هریک از مشخصات، یک سند ایجاد نمود. از آنجا که در جملات محاوره‌ای مرز جملات به طور روشنی مشخص نمی‌باشد، نیاز به درنظر گرفتن یک پنجره محتوایی است؛ بدین ترتیب که a کلمه قبل و b کلمه بعد از مشخصه موردنظر در سند مربوط به این ویژگی قرار می‌گیرند. همچنین اگر مشخصات دیگری در این پنجره قرار داشته باشند و نیز کلمات توقف در سند مربوطه اضافه نمی‌شوند. برای مثال در عبارت "بازی‌هایی با حجم بالا رو هم خیلی خوب میاره شارژش هم خیلی خیلی دیر تموم میشه و فلاش دوربین پشتش هم بسیار عالیه"، سند مربوط به مشخصه "شارژ" پس از حذف کلمات توقف و سایر مشخصات "بازی" و "فلاش دوربین پشت" و پس از ریشه یابی، با در نظر گرفتن پنجره محتوایی a=b=۸ به صورت <حجم، بالا، شارژ، تموم، میشه، فلاش> می‌شود.

۳-۳- استفاده از الگویتم‌های یادگیری ماشین

پس از انجام مراحل قبل با استفاده از اسناد حاضر و k خوشه آماده شده در مرحله قبل به عنوان داده‌های برچسب‌خورده با استفاده از الگوریتم هایی مانند امیدبیشینه، k_میانگین و سلسله‌مراتبی، مشخصات خوشه بندی می‌شوند. الگوریتم‌های یادگیری ماشین بر روی تمام لغات یکتای موجود در اسناد اجرا می‌شود. در ادامه به توضیح این الگوریتم‌ها پرداخته می‌شود.

الگوریتم امید بیشنیه

نیگام و همکاران در سال ۲۰۰۰ [۱۳] طریقه دسته‌بندی متون با استفاده از الگوریتم امید بیشینه را ارائه نمودند. در کاربردهای خوشه‌بندی که برچسب داده ها در اختیار نیست می‌توان با ایجاد تعدادی داده برچسب خورده این روش را به صورت زیر به کار گرفت:

$$P(w_t|c_j) = \frac{1+\sum_{i=1}^{|D|} N_{ti} P(c_j|d_i)}{|V|+\sum_{m=1}^{|V|}\sum_{i=1}^{|D|} N_{mi} P(c_j|d_i)} \qquad (۱)$$

$$P(c_j) = \frac{1+\sum_{i=1}^{|D|} P(c_j|d_i)}{|C|+|D|} \quad (۲)$$

$$P(c_j|d_i) = \frac{P(c_j)\prod_{k=1}^{|d_i|} P(w_{d_i,k}|c_j)}{\sum_{r=1}^{|C|} P(c_r)\prod_{k=1}^{|d_i|} P(w_{d_i,k}|c_r)} \quad (۳)$$

ابتدا بر روی داده‌های برچسب‌خورده، فرمول‌های ۱ و ۲ اجرا می‌شود. به ازای هر کلمه w_t که در مجموعه کلمات وجود دارد، احتمال متعلق بودن این کلمه در خوشه jام محاسبه می‌شود که j = 1, …, k و نیز $P(c_j)$ به صورت نسبت تعداد اعضای خوشه jام به تعداد کل مشخصات محاسبه می‌شود. $P(w_t|c_j)$ برابر است با نسبت تعداد دفعاتی که w_t در کل اسناد تکرار شده به مجموع تعداد دفعاتی که w_t در هر یک از اسناد مربوط در خوشه jام تکرار شده است.

در این روش با فرض مستقل بودن کلمات از دسته‌بند بیز ساده[۲۰] با بیشترین احتمال توزیع پسین، استفاده می‌کنیم. طبق قضیه نایو بیز با فرض مستقل بودن ویژگی‌ها از یکدیگر $P(c_j|w_t) \propto \mathrm{P}(c_j) * \prod \mathrm{P}(w_t|c_j)$. پس از این مرحله می توان $P(c_j|w_t)$ را بر روی همه داده‌ها محاسبه نمود. در مراحل بعدی هر سه فرمول بر روی همه داده‌ها اجرا می‌شوند تا زمانی که خوشه مربوط به مشخصات تغییر نکند.

الگوریتم k_میانگین

یکی دیگر از الگوریتم‌های پیاده‌سازی شده به منظور خوشه‌بندی مشخصات، روش k_میانگین است. در خوشه‌بندی متون، به طور مرسوم، اسناد به صورت گروهی از کلمات در نظر گرفته می‌شوند و هر کلمه در یک سند به صورت یک ویژگی تعریف می‌شود، بنابراین هر سند به صورت یک بردار ویژگی چند بعدی نمایش داده می‌شود که هر بعد متناظر با مقدار وزنی کلمه در مجموعه اسناد است. این مقدار وزنی می‌تواند با استفاده از شاخص میزان تکرار کلمه در معکوس تکرار سند(tfidf)[۲۱] محاسبه شود و به صورت $tf = tf * Idf$ محاسبه می شود. tf برابر با نسبت تعداد تکرار کلمه w در سند موردنظر به تعداد کل کلمات سند و Idf به صورت لگاریتم نسبت تعداد کل اسناد شامل کلمه w به تعداد اسناد شامل کلمه w محاسبه می‌شود.

در این الگوریتم ابتدا داده‌ها در k خوشه قرار می‌گیرند. در هر خوشه یک مرکز در نظر گرفته می‌شود و سپس فاصله هریک از تمام داده‌ها تا این مرکز محاسبه می‌شود. هر داده در خوشه‌ای قرار می‌گیرد که کمترین فاصله تا مرکز آن خوشه را دارا باشد. سپس مرکز خوشه‌ها تغییر کرده و فاصله‌ها محاسبه می‌شود و داده‌ها به همین ترتیب در خوشه‌ها قرار می‌گیرند تا زمانی که داده‌های موجود در خوشه‌ها تغییر نکنند.

سلسله مراتبی (پایین به بالا)

در این الگوریتم ابتدا هر داده به عنوان یک خوشه در نظر گرفته می شود و سپس فاصله هر یک از داده‌ها با یکدیگر محاسبه می‌شود. در این مرحله داده‌هایی با فاصله کمتر به صورت یک جفت در یک خوشه قرار می‌گیرند و فاصله خوشه‌ها با یکدیگر محاسبه می‌شوند. این کار ادامه می‌یابد تا همه داده‌ها در یک خوشه قرار گیرند، به منظور ایجاد k خوشه می‌توان پس از ایجاد k خوشه الگوریتم را پایان داد.

۴- ارزیابی کارایی چارچوب پیشنهادی

۴-۱- مجموعه داده

مجموعه داده مورد استفاده در این پژوهش نظرات کاربران در بازه زمانی پنج ساله از آبان ماه سال ۹۰ تا مهرماه سال ۹۵ درباره خودروی تیبا می‌باشد که شامل ۶۱۹ نظر است. این نظرات از سایت www.pedal.ir با استفاده از یک نرم‌افزار خزنده[۲۲] جمع‌آوری شده‌اند. فرض می‌شود همه این مشخصات توسط یکی از الگوریتم‌های موجود استخراج شده‌اند[۲]. مثلا در نظرات مربوط به خودرو "شتاب"، "سرعت"، "قدرت موتور" به عنوان مشخصات درنظر گرفته می‌شوند.

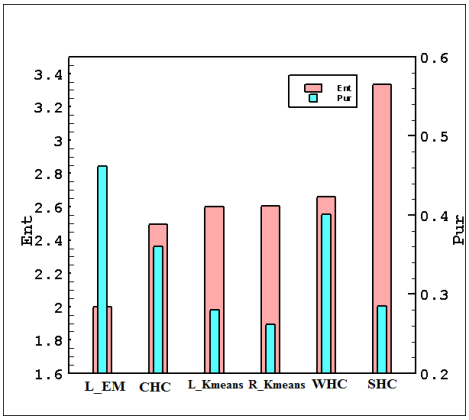
۴-۲- نتایج ارزیابی‌ها

پس از پیش پردازش داده‌ها و ایجاد داده‌های برچسب‌گذاری شده، از الگوریتم‌های امید بیشینه، k_میانگین و سلسله مراتبی جهت خوشه بندی استفاده شد. در الگوریتم امید بیشینه، از داده‌های برچسب‌گذاری شده که در بخش ۳-۲ توضیح داده شد، استفاده گردید. همچنین از مقدار وزنی کلمات با استفاده از میزان تکرار کلمه و معکوس تکرار سند و نیز میانگین داده‌های برچسب گذاری موجود در هر خوشه به عنوان مرکز خوشه‌ها در الگوریتم k_میانگین استفاده شد. همچنین این الگوریتم با استفاده از مراکز تصادفی با ده بار اجرا نیز استفاده گردید و به منظور محاسبه شباهت از فاصله کسینوسی استفاده نمودیم.

از مقدار وزنی کلمات در الگوریتم سلسله مراتبی نیز استفاده گردید. به منظور محاسبه شباهت از فاصله کسینوسی با معیار تک اتصالی[۲۳] و اتصال کامل[۲۴] و نیز از فاصله کسینوسی با معیار فاصله وارد استفاده شده است.

به منظور ارزیابی نتایج از معیارهای انتروپی و خلوص استفاده شده است. در واقع انتروپی میزان توزیع داده‌ها در خوشه‌های مختلف و معیار خلوص درصد درستی خوشه‌ها را نشان می‌دهند[۹]. هرچه میزان انتروپی به صفر نزدیک باشد و مقدار خلوص به یک نزدیکتر باشد، نتایج بهتری حاصل شده‌اند. این معیارها در مقالات متعددی مربوط به خوشه

بندی مشخصات کالا [۲،۴] استفاده شده اند. نتایج حاصل از الگوریتم های مذکور در شکل (۲) آورده شده است.

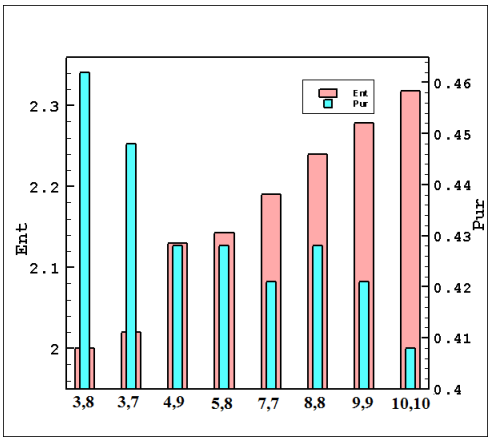


شکل ۲: نتایج تجربی روی مجموعه داده خودرو

محور عمودی سمت چپ، میزان انتروپی و محور عمودی سمت راست، مقدار خلوص را نشان می‌دهند.

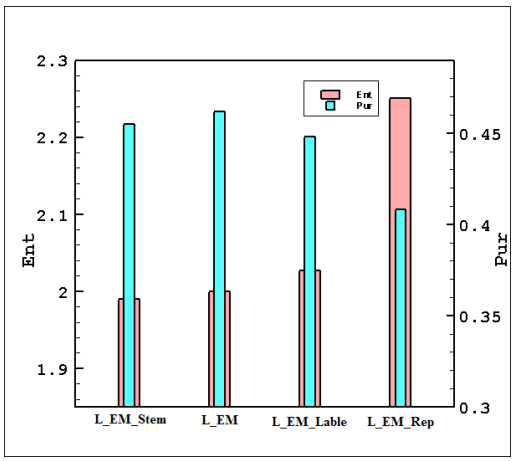
همانطور که در این شکل قابل مشاهده است، نتایج حاصل از الگوریتم امیدبیشینه با استفاده از داده‌های برچسب‌خورده به دلیل استفاده از احتمال وقوع کلمات و نه فقط شباهت واژگانی بهتر از دو الگوریتم دیگر عمل می‌کند. همچنین الگوریتم k میانگین با استفاده از داده‌های برچسب خورده بهتر از مراکز تصادفی عمل می‌کند.

الگوریتم سلسله مراتبی نیز هنگام استفاده از اتصال وارد بهتر از دو نوع اتصال دیگر نتایج بهتری را ارائه کرده است.

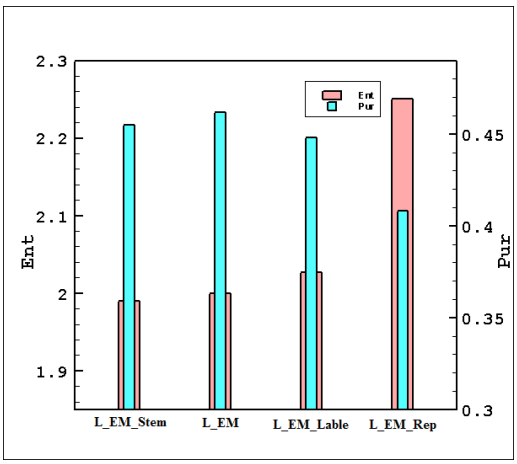


شکل ۳: انتخاب پنجره محتوایی در الگوریتم امید بیشینه

همچنین نتایج حاصل از انتخاب پنجره محتوایی در شکل ۳ آمده است و بهترین نتایج برای سه کلمه قبل و هشت کلمه بعد از هر مشخصه حاصل شده است.



۴. (الف)



۴. (ب)

شکل ۴: تاثیر ریشه‌یابی کلمات، اصلاحیه در برچسب‌گذاری داده‌ها و جایگزینی نوشتار انگلیسی و همسان‌سازی کلمات در نتایج. ۴.(الف) الگوریتم امید بیشنیه، ۴.(ب) الگوریتم k_میانگین

نتایج حاصل از حذف کلمات با نوشتار انگلیسی و نیز همسان‌سازی کلمات، ریشه‌یابی کلمات و نیز انتخاب داده‌های برچسب‌گذاری شده در شکل ۴ آورده شده است. همانطور که در این شکل مشخص است، همسان‌سازی و جایگزینی کلمات با نوشتار انگلیسی بیشترین تاثیر را در نتایج داشته است و بدون آن، معیار انتروپی بسیار زیاد و معیار خلوص بسیار کم شده‌اند.

۵- نتیجه‌گیری

در این پژوهش سعی کردیم به منظور خوشه‌بندی مشخصات کالا، علاوه بر بهبود دادن روش‌های پیش‌پردازش از جمله ریشه‌یابی کلمات محاوره، نرمال‌سازی کلمات و همسان‌سازی آنها، در مرحله ایجاد داده‌های برچسب‌گذاری شده در روش ارائه شده توسط ژای و همکاران[۴] بهبود حاصل نماییم و در پایان با استفاده از روش‌های یادگیری ماشین به خوشه‌بندی مشخصات کالا پرداختیم. نتایج حاصل، تاثیر بهبود کارهای انجام شده را نشان می‌دهند.

در کارهای آتی به منظور خوشه‌بندی مشخصات کالا در زبان فارسی، می‌توان از روش برداری کلمات [۱۵] که میزان تشابه مفهومی کلمات در بدنه متن را ارائه می‌دهد[۱۶]، و یا با استفاده از لغت‌نامه فارس‌نت[۱۴]، از شباهت معنایی در ایجاد داده‌های برچسب گذاری شده استفاده نمود.

مراجع

[۱] حسین اکبریان، مصطفی صالحی، هادی ویسی، "تعیین جهت‌گیری نظرات در رسانه‌های اجتماعی فارسی زبان" بیست و چهارمین کنفرانس مهندسی برق ایران، شیراز، ۱۳۹۵.

[2] Golpar-Rabooki E, Zarghamifar S, Rezaeenour J, "Feature extraction in opinion mining through Persian reviews" AI and Data Mining, 3(2), pp. 169-79, 2015.

[3] Chen Y, Zhao Y, Qin B, Liu T, "Product Aspect Clustering by Incorporating Background Knowledge for Opinion Mining" PloS one, 11(8), 2016.

[4] Zhai Z, Liu B, Xu H, Jia P, "Clustering product features for opinion mining" In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 347-354. ACM, 2011.

[5] github.com/mojtaba-khallash/JHazm/tree/master/JHazm [Access date: 26/11/2016]

[۶] محمد نجانی، احمد براآنی دستجردی، "روش جدید خوشه‌بندی مستندات متنی الکترونیکی فارسی به کمک واژ–هستان شناسی فارس نت"، اولین کنفرانس ملی دانش پژوهان کامپیوتر و فناوری اطلاعات، تبریز، ۱۳۹۰.

[7] Poria S, Chaturvedi I, Cambria E, Bisio F, "Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis" In Neural Networks (IJCNN), 2016 International Joint Conference, pp. 4465-4473. IEEE, 2016.

[8] Carenini G, Ng RT, Zwart E, "Extracting knowledge from evaluative text" In Proceedings of the 3rd international conference on Knowledge capture, pp. 11-18. ACM, 2005.

[9] Xiong S, Ji D, "Exploiting Capacity-Constrained K-Means Clustering for Aspect-Phrase Grouping" In International Conference on Knowledge Science, Engineering and Management, pp. 370-381. Springer, 2015.

[10] Gupta P, Kumar S, Jaidka K, "Summarizing Customer Reviews through Aspects and Contexts" In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 241-256. Springer International Publishing, 2015.

[11] Zhai Z, Liu B, Xu H, Jia P, "Constrained LDA for grouping product features in opinion mining" In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 448-459. Springer Berlin Heidelberg, 2011.

[12] Yetgin Z, GÖZÜKARA F, "New metrics for clustering of identical products over imperfect data" Turkish Journal of Electrical Engineering & Computer Sciences, 23(4), pp. 195-208, 2015.

[13] Nigam K, McCallum AK, Thrun S, Mitchell T, "Text classification from labeled and unlabeled documents using EM" Machine learning, 39(2-3), pp. 103-134, 2000.

[14] Shamsfard M, Hesabi A, Fadaei H, Mansoory N, Famian A, Bagherbeigi S, Fekri E, Monshizadeh M, Assi SM, "Semi automatic development of farsnet; the persian wordnet" In Proceedings of 5th Global WordNet Conference, Mumbai, India, 29, 2010.

[15] Segura-Bedmar I, Suárez-Paniagua V, Martınez P, "Combining conditional random fields and word embeddings for the CHEMDNER-patents task" In Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, pp. 90-93, 2015.

[16] https://en.wikipedia.org/wiki/Word2vec [Access date: 26/11/2016]

---

۱ Feature extraction

۲ Feature clustering

۳ Opinion mining

۴ Expectation Maximization

۵ kmeans

۶ Hierarchy

۷ Entropy

۸ Purity

۹ Semantic similarity

۱۰ Distributional similarity

۱۱ String similarity

۱۲ Synonyms

۱۳ Lexical distance

۱۴ Pointwise Mutual Information(PMI)

۱۵ Topic modeling

۱۶ Latent Dirichlet Allocation(LDA)

۱۷ Semisupervised learning

۱۸ Wordnet

۱۹ Unsupervised learning

۲۰ Naïve Bayse

۲۱ Term Frequency* Inverse Document Frequency

۲۲ Crawler

۲۳ Single Linkage

۲۴ Complete Linkage