استخراج شاخص ویژگی ضمنی برای نظر کاوی مبتنی بر ویژگی(تجزیه تحلیل مبتنی بر سطح ویژگی احساسات)

چکیده

هدف نظر کاوی مبتنی بر ویژگی(تجزیه تحلیل مبتنی بر سطح ویژگی احساسات) مدل سازی روابط بین قطبیت یک سند و اهداف یا ویژگی های کاوش آن می باشد. اگرچه استخراج ویژگی صریح به طور گسترده ای مورد مطالعه قرار گرفته است، با این حال مطالعات محدودی در زمینه استخراج ویژگی های ضمنی و غیر صریح انجام شده اند. یک ویژگی ضمنی، یک هدف کاوش است که به طور صریح در متن مشخص نمی شود. برای مثال، جمله این دور بین شفاف و بسیار ارزان است، یک ایده و نظری را در خصوص ویژگی های ظاهر و قیمت ارایه می کند که با کلمات شفاف و ارزان نشان داده می شود. ما این کلمات را شاخص های ویژگی ضمنی می نامیم(IAI). در این مقاله، ما یک روش جدید را برای استخراج این IAI بااستفاده از میدان تصادفی شرطی پیشنهاد کرده ونشان می دهیم که روش ما عملکرد معنی داری نسبت به رویکرد های موجود دارد. هم چنین در بخشی از این مقاله، ما یک مجموعه ای از متون را برای استخراج IAI از طریق نام گذاری و برچسب گذاری دستی IAI و ویژگی های متناظر آن هادر یک مجموعه متون نظر کاوی شناخته شده توسعه دادیم. تا آن جا که می دانیم، متن ما اولین منبع قابل دسترس عمومی است که ویژگی های ضمنی را همراه با شاخصهای آن ها مشخص می کند.

کلمات کلیدی: نظر کاوی مبتنی بر ویژگی(تجزیه تحلیل مبتنی بر سطح ویژگی احساسات)، تحلیل احساسات، میدان تصادفی شرطی

1- مقدمه

نظر کاوی شامل مجموعه ای از فناوری ها برای استخراج و خلاصه سازی نظرات و عقاید بیان شده در محتواهای تولید شده توسط کاربر مبتنی بر وب می باشد. نظر کاوی موجب بهبود کیفیت حیات و زندگی برای افراد معمولی با اجازه دادن به آن ها برای در نظر گرفتن نظرجمعی کاربران دیگر در خصوص یک محصول، یک چهره و شخصیت سیاسی، یک مقصد توریستی و غیره می شود.هم چنین نظر کاوی موجب بهبود درامد شرکت هامی شود زیرا به این شرکت ها امکان می دهد تا بدانند که مشتریان از چه چیز هایی خوششان می اید و ازچه چیز هایی خوششان

نمی اید. هم چنین نظر کاوی موجب بهبود دموکراسی از طریق اجازه دادن به دولت ها و احزاب وگروه های سیاسی برای ارزیابی دقیق پذیرش اجتماعی برنامه ها و اقدامات خود می شود.

نظر کاوی بستگی به تشخیص دقیق نظرات و عقاید بیان شده در تک تک اسناد نظیر پست های وبلاگ، توییت ها و یا کامنت ها و نظرات کاربران دارد. این تشخیص را می توان در سطوح دقت مختلف انجام داد. برای مثال، قطبیت کل سند را می توان تعیین کرد خواه محقق نظر مثبت را بیان کند یا نظر منفی. برای یک نظر در خصوص این محصول،این سطح از دقت می تواند کافی باشد. با این حال، تعیین جمله به جمله یک ویژگی خاص محصول که در خصوص آن نظرات در یک جمله معین بیان می شود، اغلب مطلوب است.

نظر کاوی مبتنی بر ویژگی(تجزیه تحلیل مبتنی بر سطح ویژگی احساسات(1-2) روابط بین ویژگی های شی مورد نظر و قطبیت وتقارن سند(احساس مثبت و منفی بیان شده در نظر) را در نظر می گیرد. هم چنین ویژگی موسوم به اهداف نظرکاوی است. یک ویژگی،مفهومی است که در خصوص آن یک محقق یانویسنده، نظر خود را در سند بیان می کند.برای مثال، جمله "عدسی این دوربین بسیار خوب است و عمر باطری آن عالی است" را در نظر بگیرید. می توان گفت که قطبیت این دیدگاه در مورد یک دوربین عکس برداری مثبت است. با این حال، به طور اخص، آن چه که نویسنده دوست دارد، عدسی(قدرت دید) وعمر باطری این دور بین است. این مفاهیم، ویژگی های این نظر می باشند.

استخراج ویژگی، فرایند شناسایی ویژگی ها یا اهداف نظر کاوی و یا یک سند معین است. این ویژگی ها مشتمل بر دو قسم هستند: ویزگی های صریح و ویژگی های ضمنی.ویژگی های صریح متناظر یا مرتبط با کلمات و عبارات خاص در سند هستند. در مثال ما، اهداف نظر یعنی عدسی و عمر باطری صریحا در سند قید شده اند. بر عکس، یک ویژگی ضمنی به طور صریح در سند قید نمی شود. جمله "این تلفن ارزان و زیبا است" را در نظر بگیرید. این جمله، نظر مثبت را در خصوص قیمت و ظاهر تلفن بیان می کند. این ویژگی ها در جمله معادل" قیمت این تلفن پایین است و ظاهر آن زیبا است" صریح می باشند.

اگرچه مطالعات بسیاری وجود دارند که به استخراج ویژگی صریح پرداخته اند، استخراج ویژگی ضمنی کم تر مورد مطالعه قرار گرفته است.استخراج ویژگی ضمنی بسیار پیچیده تر از استخراجویژگی صریح است.بااین حال، ویژگی های ضمنی در اسناد فراگیر هستند همان طور که مثال زیر از متن توصیف شده در(1) نشان می دهد:" این

بهترین تلفنی است که می توان داشت. این تلفن همه ویژگی های مورد نیاز در یک موبایل را دارد.این تلفن سبک، براق و جذاب است. این تلفن بسیار کاربر پسند بوده و کار با آن آسان است(دست ورزی آن آسان است).جا به جایی منو ها و حرکت منوی آن نیز بسیار راحت است." در این مثال، عباراتی نظیر سبک وزن، براق و جذاب، کاربرپسند و حرکت منو وسهولت دست ورزی متناظر با ویژگی های وزن، ظاهر، رابط و عملکرد تلفن می باشد. عبارت اخیر یعنی عملکرد اشاره به ویژگی منوی تلفن دارد. اگرچه این مفاهیم به طور صریح در متن ذکر نشده اند، با این حال آن ها به طور ضمنی توسط کلمات موجود معرفی می شوند. این کلمات که نشانه ها و علاماتی برای استنباط ویژگی های ضمنی نظر هستند، موسوم به شاخص های ویژگی ضمنی هستند(IAI).

توجه کنید که در این مقاله، ما هر اسمی را به صورت یک ویژگی در نظر نمی گیریم. در عوض ما فرض می کنیم که یک مجموعه تعریف شده از ویژگی ها(متغیر ها) وجود دارند که شاخص های ویژگی آن ها نشان دهنده ارزش ها می باشند. شاخص های ویژگی از عبارات ویژگی ضمنی تعریف شده توسط لیو(3) به صورت عبارات ویژگی ای که اسم یا عبارات اسمی نیستند از این نظر متفاوت می باشند که شاخص های ویژگی از نظر معنایی اشاره به ارزش های ویژگی های از قبل تعریف شده صرف نظر از بخش سطحی گفتار دارد. در زیر ما مثال هایی را در خصوص شاخص های ویژگی بیان شده با اسم و عبارات اسمی ارایه می کنیم. به جدول 3 مراجعه کنید.

فرایند شناسایی ویژگی های ضمنی یا استخراج ویژگی ضمنی، معمولا در دو مرحله انجام می شود. در ابتدا، شاخص های ویژگی در سند شناسایی می شوند برای مثال، کاربرپسند. سپس، آن ها بر روی ویژگی های متناظر برای مثال "رابط" نگاشته می شوند.در این مقاله، ما به اولین مرحله می پردازیم: شناسایی شاخص های ویژگی،فرایندی که موسوم به استخراج شاخص ویژگی ضمنی یا استخراج IAI می باشد. رویکرد های موجود برای دومین مرحله(نگاشت IAI بر روی ویژگی ها) در بخش 2 ذکر شده اند.

یک IAI می تواند تک بخشی یا تک کلمه ای باشد نظیر براق و یا دو کلمه ای باشد نظیر کاربرپسندو یا حتی یک عبارت کامل باشد نظیر پیمایش و حرکت در منو که در مثال فوق گفته شد.

IAI می تواند بخش های مختلفی از گفتار باشد: در عبارت" این MP3پلیمر واقعا گران است"، شاخص ویژگی گران، نشان می دهد که ویژگی قیمت، یک صفت است . در عبارت " این دور بین عالی به نظر می رسد" عبارت به

نظر می رسد نشان می دهد که ظاهر یک فعل است. در عبارت من از این تلفن بدم می اید، عمر این تلفن کم تر از شش ماه بود، شاخص ویژگی عمر کم، نشان می دهد که دوام تلفن، یک فعل است.

مثال های زیر، شاخص های ویژگی را به صورت اسم یا عبارات اسمی نشان می دهد.در عبارت حتی اگر من قیمت کامل را به این تلفن پرداخت می کردم، این تلفن برای من یک معامله خوب بود. شاخص ویژگی معامله خوب نشان دهنده ویژگی قیمت است.در عبارت شفافیت این تلفن، شاخص ویژگی شفافیت نشان دهنده ویژگی "ظاهر" است. درعبارت(این پلیمر یا پخش کننده موسیقی با خطاهای تصادفی همراه است" شاخص ویژگی خطا های تصادفی نشان دهنده ویژگی کیفیت است.در عبارت این تلفن اشغال است، شاخص ویژگی اشغال نشان دهنده ویژگی کیفیت است.

شاخص های ویژگی ضمنی مختلف می توانند متناظر با یک ویژگی ضمنی باشند. این شاخص های ویژگی ضمنی شاره به ارزش های مختلف این ویژگی دارند برای مثال زیبا یا زشت برای ظاهر و یا اشاره به یک ارزش دارند که در این صورت آن ها مترادف خواهند بود برایمثال زیبا، جذاب یا شفاف و یا در عبارات تقریبا مترادف به کار می روند برای مثال،" نگاه گردن به تلفن جذاب است و یا این که طراح یک سلیقه بسیار خوب را نشان داده است".

بسیاری از محققان تنها کلمات قطبیت یا احساس را به صورت شاخص های ویژگی ضمنی در نظر می گیرند. برای مثال، در جمله " این تلفن زیبا است"، عبارت زیبا دارای قطبیت مثبت است و طبیعی است که فرض کنیم نشاندهنده یک نظر و ایده در مورد یک ویژگی می باشد که در این جا این ویژگی " ظاهر "است. توجه داشته باشید که فرض ما در این جا این است که هر دو ویژگی و ارزش به صورت تجمعی با یک عبارت بیان می شوند. اگرچه این رویکرد در بسیاری از موارد موثر است، با این حال در موارد دیگر کارکردی ندارد. برای مثال، در جمله" طراحان این دوربین کارشان عالی بوده است" عبارت طراحان، یک عبارت احساسی نیست بلکه به طور ضمنی نشاندهنده ویژگی" ظاهر" است که به طور ضمنی تنها با کلمه قطبی " خوب" در این جمله نشان داده نشده است. یعنی در این جا ویژگی غیر ضمنی با یک کلمه نشان داده شده و ارزش آن با یک کلمه دیگر نشان داده می شود. شاخص های ویژگی ضمنی و کلمه ای که به آن ارزش می دهد می توانند در جملات متفاوتی ظاهر شوند برای مثال" من این تلفن را دوست دارم". اینتلفن حتی در مناطق با سیگنال پایین کار میکند " که در آن عبارت دوست داشتن، به ویژگی پذیرش، ارزش می دهد.

تصمیم گیری در مورد این که آیا ارزش یک ویژگی به طور ضمنی نشان دهنده نظر مثبت یا منفی است همیشه کم اهمیت نیست. برای مثال، " این تلفن بسیار سنگین است در برابر باطری عمر زیادی دارد" نشان می دهد که وزن بالای یک تلفن بد است و ظرفیت بالای یک باطری خوب است. این موسوم به حقایق مطلوب است حتی اگر متن فاقد یک نظر و ایده صریح در مورد ویژگی خوب یا بد باشد ولی تنها منعکس کننده واقعیت عینی در مورد آن باشد، حقیقت و واقعیت هنوز مطلوب است که به طور ضمنی اشاره به نظر مثبت دارد و یا واقعیت نامطلوب است که به طور ضمنی اشاره به نظر منفی دارد. مثال دیگر" تلفن دارای اخرین نسخه اندروید است" یک واقعیت عینی است و هیچ نظری در این متن وجود ندارد، با این حال برای یک تلفن که اخرین نسخه سیستم عامل است مطلوب است و بنابر این نظری که به طور ضمنی در مورد سیستم عامل نشان داده می شود مثبت است.

در این مقاله، ما یک روش جدید را برای استخراج شاخص ویژگی های ضمنی ارایه می کنیم. ما از یک رویکرد یادگیری نظارت شده بر اساس برچسب زنی متوالی با میدان های تصادفی شرطی استفاده می کنیم. نتایج ما نشان می دهد که رویکرد ما عملکرد بیشتری از رویکرد های موجود دارد.

تا آنجا که می دانیم، هیچ متونی برای فرایند استخراج ویژگی های شاخص ضمنی وجود ندارد. از این روی ما این متن را توسعه دادیم. برای همین منظور، ما به طور دستی شاخص های ویژگی ضمنی و ابعاد متناظر آن ها را در یک متن مشخص برای نظر کاوی(1) نام گذاری کردیم. این متن برای اهداف تحقیقاتی قابل دسترس است.

این مقاله به صورت زیر سازماندهی شده است. بخش دوم در مورد کار ها و مطالعات انجام شده بحث می کند. بخش 3 در مورد طرح ها و ویژگیهای مورد استفاده بحث می کند. بخش 4 به توصیف روش ازمایشی می پردازد. نتایج در بخش 5 ارایه شده است. در نهایت، بخش 6 شامل نتیجه گیری است.

2-مطالعات مربوطه

هو و لیو(1) اولین بار مفهوم استخراج ویژگی را در زمینه نظر کاوی و نیز برای تمایز ابعاد صریح و ضمنی معرفی کردند. با این حال آن ها تنها به ویژگی های صریح (با استفاده از قوانین اماری) پرداخته و ویژگی های ضمنی را در نظر نگرفتند. بعد ها پپسکو و اتیزونی(4) و بلیر گولدنزوت(5) روش آن ها را بهبود بخشیدند.

امروزه، تعدادی از روش ها برای استخراج ویژگی وجود دارند. در این مقاله ما یک روش را بر اساس شیوه یادگیری نظارت شده ارایه می کنیم. از این روی در ادامه این بخش بر روش های یادگیری نظارت شده تاکید خواهیم داشت.

فرایند استخراج ویژگی یک مورد ویژه از فرایند استخراج اطلاعات است. روش های متعددی برای فرایند استخراج اطلاعات(6-7) وجود دارد که رایج ترین آن ها بر اساس برچسب گذاری ترتیبی می باشد. دو روش اصلی برای برچسب گذاری ترتیبی وجود دارد: مدل های مارکوف پنهان(HMM) و میدان های تصادفی شرطی(CRF).

روش های مختلف برای استخراج ویژگی به کار برده شده است. HMM واژگانی برای استخراج نظرات جفت شده با ویژگی های صریح متناظر به کار برده شدند(9).CRF توسط محققان مختلف برای استخراج ویژگی صریح استفاده شده است(10-13).

مطالعات کمی به بررسی استخراج ویژگی ضمنی پرداخته اند. اولین سیستم، OPINE(4) است که برای دست یابی به یک طبقه بندی قطبیت بهتر معرفی شد. متاسفانه، این سیستم به خوبی اثبات نشده است وبرای عموم قابل دسترس نیست.

همه روش ها برای استخراج ویژگی ضمنی ، به شاخص های ویژگی ضمنی متکی هستند. در همه کار ها، تنها کلمات احساسی به صورت کاندید هایی برای شاخص های ویژگی ضمنی در نظر گرفته می شوند. خوشه بندی برای تبدیل این شاخص های ویژگی ضمنی به ویژگی های صریح بر اساس آماره های مربوط به ویژگی صریح و کلمات احساسی در جملات استفاده شد(14). قاعده کاوی دو مرحله ای برای ارتباط ویژگی های ضمنی و صریح(15) استفاده شد. در روش مبتنی بر قاعده دیگر، ویژگی های صریح در متن شناسایی شده و سپس ابعاد ضمنی بر روی آن ها با خوشه بندی جفت ابعاد صریح و کلمات احساسی نگاشته شد که جایگزین ویژگی های ضمنی بودند(16). اخیرا چارچوب های مبتنی بر قاعده نتایج بسیار مطلوبی را برای استخراج ویژگی های ضمنی و صریح(17) و باری تحلیل احساس مبتنی بر ویژگی(18-19) به خصوص با استفاده از مفاهیم و نه تک کلمات نشان داده اند(20). شناسایی کلمات احساسی و گرایش احساسی متن به نوبه خود فرایندی است که در آن رویکرد های مبتنی بر قاعده(21)، روش های یادگیری ماشینی(22-24) و منابع واژگانی استفاده می شود(25-26).

3-روش شناسی

در زیر ما به توصیف طرح مورد استفاده برای استخراج شاخص های ویژگی ضمنی و ویژگیهای مورد استفاده در طی ازمایشات می پردازیم.

3-1 استخراج شاخص های ویژگی ضمنی

هدف اصلی برچسب گذاری و نام گذاری کلمات از متن ورودی نظرات به صورت شاخص های ویژگی ضمنی است. خروجی مجموعه ای از اجزای دو گانه است. هر داپل یا جزءدوتایی متشکل از یک جمله و برچسب کلاس تعیین شده با روش استخراج شاخص های ویژگی ضمنی می باشد. برچسب I برای کلاس IAI و برچسب O برای کلاس سایر است. کلمات شفاف و ارزان به صورت IAI طبقه بندی می شوند.

ورودی: این تلفن شفاف و بسیار ارزان است

خروجی(این، O)(تلفن O)،(هست O)،(شفاف،I)(و O)،(بسیار، O)،(ارزان،I)(.،O)

شکل 1 : مثال استخراج شاخص های ویژگی ضمنی

ما فرایند استخراج شاخص های ویژگی ضمنی را به صورت برچسب زنی ترتیبی در نظر می گیریم. فرض کنید که $X = \{x_1, ..., x_m\}$ مجموعه ای از مشاهدات و $Y = \{y_1, ..., y_m\}$ مجموعه ای از برچسب های تعیین شده به این مشاهدات است. هدف اصلی پیش بینی مجموعه ای از برچسب های $Y' = \{y_{m+1}, ..., y_n\}$ با توجه به ورودی های جدید $X' = \{x_{m+1}, ..., x_n\}$ با مدل بدست امده با داده های مشاهده شده X و برچسب های Y است. روش برچسب زنی ترتیبی مورد استفاده میدان های تصادفی شرطی است.

2-3 میدان های تصادفی شرطی

میدان های تصادفی شرطی(CRF) یک چارچو گرافیکی احتمالی برای ایجاد مدل های احتمال گرایانه جهت قطعه بندی و برچسب گذاری توالی های داده ها است. این میدان از یک رویکرد قطعی استفاده می کند. به طور کلی یک CRF یکمدل لگاریتم خطی است که توزیع احتمالی را در توالی ای از داده ها با توجه به یک توالی مشاهده ای خاص تعریف می کند. لافرتی و همکاران(8) یک CR را بر روی مجموعه ای از مشاهدات X و مجموعه ای از توالی های Y تعریف کرد:فرض کنید که $G = (V, E)$ یگ گراف باشد به طوری که$Y = (Y_v)_{v\epsilon V}$ است لذا Y با رئوس G نمایه بندی شده و سپس (X, Y) یک میدان تصادفی شرطی می باشد که در صورتی که مشروط بر متغیر های تصادفی Y_v باشد از ویژگی مارکوف با توجه به گراف پیروی می کند:

$$p(Y_v|X, Y_u, u \neq v) = p(Y_v|X, Y_u, u \sim v), \qquad (1)$$

که $u \sim v$ بدین معنی است که W و V همسایه ها در G هستند. این ویژگی این واقعیت را توصیف می کند که احتمال شرطی یک برچسب Y_v تنها در صورتی بستگی به برچسب Y_u دارد که با Y_v تشابه داشته باشد یعنی $(Y_v, Y_u) \epsilon E$.

توزیع مشترک در توالی های برچسب Y با توجه به X دارای فرم زیر است

$$p_\theta(y|x) \propto exp\left(\sum_{e\epsilon E,k} \lambda_k f_k(y|_e, x) + \sum_{v\epsilon V,k} \mu_k g_k(v, y|_v, x)\right), \quad (2)$$

که X توالی وترتیب داده ها، y ترتیب برچسب، $y|_S$ مجموعه ای از مولفه های Y مرتبط با رئوس در شبه گرافS، و f_k و g_k توابع ویژگی بوده و θ پارامتر های وزنی است

$$\theta = (\lambda_1, \lambda_2, \lambda_3, ...; \mu_1, \mu_2, \mu_3, ...).$$

توابع ویژگی f_k و g_k مجمو عه ای از توابع می باشد که یک مجموعه ای از مشاهدات X را به عدد حقیقی به خصوص به زیر مجموعه $\{0, 1\}$ نگاشت می کند. این توابع برای مدل سازی مشاهده X_i به صورت یک بردار ساخته می شوند. فرض ما این است که ویژگی ها معین و ثابت هستند. این ویژگی ها معمولا بولین هستندو به صورت دستی ایجاد می شوند. برای مثال، یک ویژگی راس f_k در صورتی حقیقی است که کلمه X_i با حرف بزرگ نوشته می شوند و Y_i اسم است.

برای رویکردپیشنهادی، ما از یک مورد خاص از این چارچوب استفاده کردیم. برچسب توالی میدان های تصادفی شرطی زنجیره خطی(8). این یک روش نظارت شده برای پیش بینی توالی های برچسب با توجه به مجموعه ای از مشاهدات است. روش مورد استفاده در ازمایش ما دسته بند CR-FC در NER استنفورد(27) می باشد. این دسته بند نسخه ای از مدل های توالی CRF زنجیره خطی با رتبه دلخواه است.

3-3 ویژگیها

داده های مورد استفاده برای آموزش برچسب گذار مبتنی بر CRF از مجموعه داده های توصیف شده در بخش 4-1 گرفته شدند. ما داده ها را با حذف نقطه ها،پیش پردازش کردیم. حروف بزرگ به همان صورتی که بودند باقی ماندند.

NER استنفورد شامل یک کلاس جاوا موسوم به NERFeatureFactorمی باشد. این کلاس یا دسته چندین روش استخراج ویژگی را پیاده می کند. به این ترتیب استخراج گننده ویژگی را قادر به استفاده از آن ها با دسته بند CRFC برای ایجاد یک بردار ویژگی می کند. ما از این دسته برای ایجاد این بردار های ویژگی برای ازمایشات خود استفاده کردیم.

با توجه به ترتیب کلمات، ما یک بردار ویژگی را برای هر کلمه نام گذاری شده ایجاد میکنیم. این بردار های ویژگی حاوی ویژگی های زیر هستند:

1- ویژگی های کلمات: این ها ویژگی هایی هستندکه نشان می دهند کدام نوع کلمه ، یک نمونه واقعی ای است که باید برچسب گذاری شود

2- ویژگی های کاراکتر n-grams: این ها ویژگی هایی هستند که نشان می دهند آیا یک زیر رشته در کلمه وجود دارد یا خیر. این نوع از ویژگی ها در فرایند تشخیص نهاد نام گذاری شده(28)مفید بوده اند. زیر رشته ها مربوط به انواع متون می باشند. یک محدودیت در خصوص این n-grams این است که آن ها بزرگ تر از 6 کاراکتر نیستند. دلیل این محدودیت این است که در n-grams های بزرگ تر،اموزش از حیث قدرت محاسباتی با بهره عملکرد طبقه بندی اندک، گران تر می شود. محدودیت دیگر این است که آین ها فاقد قسمت اغاز یا پایان یک کلمه هستند. ما به طور ازمایشی تعیین کردیم که n-grams با این ویژگی ها عملکرد بهتری دارند.

3- ویژگی های شناسه بخش گفتار: شناسه POS کلمه. برای این ویژگی ها، بایستی شناسه POS را برای هر عبارت در جمله به صورت ورودی ارایه کرد. ما از شناسه گذار NLTK برای شناسه گذاری استفاده کردیم.

4- ویژگیهای زمینه و سیاق: این ویژگی ها شامل کلمات، شناسه و ترکیبی از شناسه POS از کلمات پیشین و پسین مورد فعلی می باشند.

5- ویژگی های توالی کلاس: این ها ترکیبی از یک کلمه خاصمی باشند که برچسب آن ها به کلمات پیشین داده شده است. ما از یک پنجره برچسب 2 استفاده کردیم یعنی برچسب های 2 کلمه پیشین به علاوه کلمه فعلی و و ویژگی ها.

6- ویژگی های کلمه Bi-gram

4-شرایط ازمایشی

توصیف کلی شرایط ازمایشی به شرح زیر است: در ابتدا یک متن برای استخراج شاخص های ویژگی ضمنی توسعه یافته و ما سپس متریک های مختلف و روش های اعتبارسنجی را برای ازمایشات خود تعریف کردیم. در نهایت، ما معیار هایی را برای مقایسه عملکرد رویکرد خود تعریف کردیم.

4-1 مجموعه داده ها

ما مشاهده کردیم که هیچ گونه مجموعه داده مناسبی برای ازمایشات ما وجود ندارد. همان طور که در بخش 1 گفته شد، کار های محدودی در استخراج شاخص های ویژگی های ضمنی انجام شده است. به علاوه، فرایند استخراج شاخص های ویژگی های ضمنی تعریف نشد زیرا رویکرد رایج برای استنباط ابعاد ضمنی، استفاده از جملات احساسی می باشد. از این روی، طبیعی است که هیچ گونه منابعی برای استخراج شاخص های ویژگی های ضمنی وجود ندارد. در نتیجه، ما اولین متن را برای استخراج شاخص های ویژگی های ضمنی توسعه دادیم.

هو ولیو(1) متنی را برای استخراج شاخص های ویژگی های ضمنی توسعه دادند. این متن در بسیاری از فرایند های نظر کاوی استفاده می شود. ما از متون فوق باری ایجاد یک متن جدید برای استخراج شاخص های ویژگی های ضمنی استفاده کردیم. ما یک متن را برچسب گذاری کردیم که نشاندهنده استخراج شاخص های ویژگی های ضمنی و ویژگی های ضمنی متناظر آن هاست. ماتنها جملاتی را انتخاب کردیم که تنها دارای یک ویژگی ضمنی برای برچسب گذاری متن ارایه می کند. از این روی ما همه جملات را برچسب گذاری نکردیم.

جدول 1 برخی از ویژگی های متن استخراج شاخص های ویژگی های ضمنی را نشان می دهد. این متشکل از 314 ارزیابی امازون از 5 محصول در ابزار های الکترونیکی است:یک دی وی دی پلیمر(ستون DVDدر جدول)، دوربین کنون، یک MP3 پلیر،یک دوربین نیکون، و یک موبایل نوکیا. این جدول، برخی از نظرات و ارزیابی ها رادر هر سند توصیف می کند. هم چنین این سند توصیف می کند که یک ارزیابی چه تعداد کلمات و جملات را دارد.

ویژگی های اماری در سطوح دقت مختلف در جدول 2 نشان داده شده است. نام هر ستون مشابه با نام ستون های گزارش شده در جدول 1 است. این جدول در 3 بخش برای هر سطح دقت تقسیم شده است.

- سطح جمله
- سطح نشانه
- سطح تیپ

بخش سطح جمله نشان می دهد که چه تعداد جملات در سند وجود دارند، چه تعداد از جملات سند حداقل دارای یک استخراج شاخص های ویژگی های ضمنی می باشد(که در ردیف IAI نام گذاری شده است) و چه درصدی از جملات حداقل دارای یک IAI هستند. بخش سطح تیپ و نشانه، ویژگی های یکسان را برای این سطوح دقت نشان می دهد.

جدول 1: ویژگی های متن

	DVD	Canon	MP3	Nikon	Phone
ارزیابی ها	99	45	95	34	41
کلمات در هر ارزیابی	572.3	1236.4	1575.5	924	1085.3
جملات در هر ارزیابی	7.47	13.26	18.90	3.64	13.31

جدول 2: ویژگی های اماری

	DVD	Canon	MP3	Nikon	Phone
Sentence level					
جملات	740	597	1796	346	546
IAI#	147	63	155	36	44
IAI%	19.86%	10.55%	9.03%	10.40%	8.05%
Token level					
نشانه ها	56661	55638	149676	31416	44497
IAI#	164	79	214	50	66
IAI%	0.289%	0.141%	0.142%	0.159%	0.148%
Type level					
تیپ ها	1767	1881	3143	1285	1619
IAI#	72	63	136	40	42
IAI%	4.07%	3.34%	4.32%	3.11%	2.59%

توزیع POS برای IAI برچسبگذاری شده در متن در جدول 3 نشان داده شده است. هر ردیف بیانگر شناسه کلی Penn Treebank POS است. نخستین ردیف بیانگر همه شناسه هایی است که به صورت صفت هستند (JJ,

(JJR, JJS، و دومین ردیف نشان دهنده شناسه های اسم است(NN,NNS, NNP, NNPS) و ردیف سوم نشاندهنده شناسه فعل (VB, VBD, VBG, VBN, VBP, VBZ)) است. اخرین ردیف باقی مانده شناسه های مشاهده شده با شاخص های ویژگی های ضمنی است. ستون شاخص های ویژگی های ضمنی نشان می دهد که چه تعداد واژگان را می توان با یک شناسه معین برچسب گذاری کرد. سومین ستون توصیف می کند که چه تعداد کلمات با شناسه معین را می توان در جملات با شاخص های ویژگی های ضمنی دید. چهارمین ستون توزیع شناسه مشاهده شده در شاخص های ویژگی های ضمنی را نشان می دهد.

جدول 3: توزیع POS

POS	IAI	POS in IAI Sentence	P(IAI)
JJ	157	527	0.2818
NN	167	1692	0.3000
VB	220	1112	0.3836
other	19	3900	0.0346

2-4 متریک ها وروش های ارزیابی

ما از علامت اختصاری به عنوان استاندارد طلایی استفاده کردیم. شاخص های ویژگی های ضمنی برچسب گذاری شده شامل ترکیبات و عبارات هستند. کلمات برچسب گذاری شده نظیر شاخص های ویژگی های ضمنی مطابق با کلمات برچسب گذاری شده به صورت شاخص های ویژگی های ضمنی در علامت اختصاری هستند که به صورت مثبت حقیقی در نظر گرفته می شود. این کلماات که مطابق با هم نیستند به صورت مثبت کاذب در نظر گرفته می شوند(FP). منفی های کاذب(Fn)، کلمات برچسب گذاری شده به صورت IAI هستند. ما دقت و بازخوانی را اندازه گیری کردیم دقت P و بازخوانی R به صورت زیر تعریف می شوند

$$P = \frac{tp}{tp + fp}, \qquad R = \frac{tp}{tp + fn}.$$

متریک عملکرد مورد استفاده امتیاز F1 است. این امتیاز به صورت زیر تعریف می شود

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}.$$

نتایج در یک شرایط اعتبار سنجی و ارزیابی متقابل ده برابر بدست امدند.

3-4 رویکرد معیار

اولین معیار برچسب گذاری کلمات احساس به صورت IAI برای هر جمله در یک ارزیابی می باشد(14-16). ما این را BSLN1 معیار می نامیم. ما از واژه احساسی مورد استفاده در(2) برای تعیین قطبیت نظرات استفاده می کنیم. این واژه مطابق با دو فهرست است. اولین فهرست دارای کلمات و عبارات مثبت است که شامل کلماتی هستند که یک نظر مثبت را در یک زمینه خاص نظر دهی شده پیشنهاد می کند. دومین لیست با کلمات منفی است. این کلمات نشان دهنده یک نظر منفی(برای مثال وحشتناک) است.

الگوریتم این معیار به صورت زیر است: برای هر کلمه در جمله، ما تعیین می کنیم که آیا این کلمه در دو فهرست واژگان قرار دارد یا خیر. در این صورت، آن را به صورت IAI برچسب گذاری می کنیم.

ما دومین معیار را بر اساس طبقه بندی و دسته بندی متن پیشنهاد میکنیم که BSLN2 نام دارد. ما دسته بند متنی بیز ساده رااجرا کردیم. این دسته بند با متون متشکل از علامات اختصاری نشان داده می شود. وظیفه این دسته بند تعیین این است که آیا یک جمله دارای حداقلیک IAI است یا خیر. در صورتی که جمله به صورت جمله دارای IAI طبقه بندی شود، عبارت به صورت IAI برچسب گذاری می کنیم.

ویژگی های مورد استفاده در دسته بند NB به شرح زیر هستند

- ریشه واژه اختصاری. ما کلمات توقف را حذف کردیم
- بهترین 500 مجموعه bi-grams بدست امده با یک شاخص ارتباط اطلاعات متقابل نقطه ایپ

در نهایت ما برچسب گذار توالی مدل مارکوف پنهان درجه دوم را پیاده سازی کردیم. این یک روش استاندارد برای برچسب گذاری توالی است. این روش موسوم به BSLN3 است. این برچسب گذار را با علامت اختصاری خود اموزش دادیم. چون برچسب گذار HMM درجه دوم است، ما از بیگرم ها و تری گرم ها به عنوان ویژگی ها استفاده کردیم. داده های اموزشی به صورت زیر پیش پردازش می شوند

- کلماتی که کم تر از 5 برابر در متن(کلمات نادر) ظاهر می شوند، در داده های اموزشی برای برچسب PARE تغییر می یابند.
- کلمات نادر که دارای حداقل یک کاراکتر عددی هستند برای برچسب NUMERIC تغییر می یابند

- کلمات نادر که متشکل از حروف بزرگ هستند برای برچسب ALLCAPS تغییر می یابند

همه معیار ها در نرم افزار فیتون اجرا شدند. ما از دسته بند NB در NLTK برای BSLN2 استفاده کردیم.

	Precision	Recall	F1-Score
Extraction of Sentence with IAI	0.25	0.37	0.30

جدول 4. BSLN2 عملکرد طبقه بندی جملات را بیان می کند

5- نتایج

جدول 4 عملکرد BSLN2 را نشان می دهد که جملات با حداقل یک IAI را دسته بندی می کند

جدول 5عملکرد معیار ها و رویکرد مبتنی بر CRF را با مجموعه ویژگی های مختلف مقایسه می کند.WT ترکیبی از کلمات و شناسه ها است(نکات 1 و 3 از توصیف ویژگی در بخش 3-3). ویژگی های CNG، ویژگی های n-grams(نکته 2) می باشند.CNTX ویژگی های زمینه و ویژگی های بیگرم (نکته 4 و 6) می باشند.CLS ، ویژگی های توالی دسته(نقطه 4) است.

ما مشاهده می کنیم که ویژگی های WT، بیشترین دقت را دارند. با این حال بازخوانی ضعیف است.

ویژگی های CNG منجر به تقویت بازخوانی می شوند. این ویژگی ها، ویژگی های مورفولوژیکی و ریخت شناسی کلمات(ریشهها، پیشوند ها وپسوند ها) را پوشش می دهند. کلمات با ویژگی های مورفولوژیکی مشابه از نظر معنایی مشابه هستند. برای مثال جمله، این تلفن عالی به نظر می رسد می تواند به صورت ظاهر این تلفن عالی است و یا حتی این تلفن با این قاب عالی به نظر می رسد بازنویسی شود. ریشه" به نظر می رسد"در جملات پیشین، بهترین IAI است که به ویژگی ظاهر بر می گردد. از این روی کلمات با اینریشه بایستی دارای احتمال استخراج بیشتری به صورت IAI باشند. مشکل اصلی این است که این ویژگی ها موجب کاهش دقت کلی می شود زیرا کلمات بیشتر که IAI نیستند با این حال کاراکتر N-grams دارای احتمال استخراج بالایی است. ویژگی های Cntx و CLS موجب بهبود دقت و بازخوانی می شود. بهترین عملکرد با ترکیبی از ویژگیهای WT, CNG, CNTX

CLS بدست می اید.

جدول 5: عملکرد استخراج IAI با ویژگی های مختلف

	دقت	بازخوانی	امتیاز F1
BSLN1	0.0381	**0.3158**	0.0681
BSLN2	0.1016	0.1379	0.1170
BSLN3	0.5307	0.1439	0.2264
WT	**0.6271**	0.0575	0.1053
CNG	0.4765	0.1925	0.2742
CNTX	0.5030	0.1148	0.1869
WT,CNG	0.4697	0.1992	0.2795
WT,CNG,CNTX	0.5209	0.2031	0.2932
WT,CNG,CNTX,CLS	0.5458	0.2064	**0.2970**

رویکرد ما برای این کار عالی است. این رویکرد عملکرد بهتری از BSLN1 وbsln2 دارد. دقت آن بسیار بالاست. با این حال بازخوانی کم تر از BSLN1 می باشد.

برای ایجاد تعادل بین دقت و بازخوانی در رویکرد مبتنی بر CRF، ما از دسته بندCRF استفاده کردیم(29). این رویکرد امکان ایجاد اریبی در دسته های مختلف را می دهد. این اریبی ها می تواند مقدار واقعی را اختیار کند. چون اریبی کلاس Aبه سمت نامتناهی میل می کند دسته بند قادر به پیش بینی برچسب های A است و به سمت مقدار بینهایت منفی است که برچسب A را پیش بینی می کند. این اریبی ها برای تعدیل تعادل دقت- بازخوانی استفاده می شود.

ما یک ازمایش را با اریبی کلاس IAI انجام دادیم. ما مقدار این اریبی را در یک دامنه 1.5تا 3.5 تغییر دادیم. ما مقدار اریبی کلاس سایر را برابر با 1 قرار دادیم. مجموعه ای از ویژگی های مورد استفاده در ردیف پایانی جدو.ل 5 نشان داده شده اند. جدول 6 دقت، بازخوانی و امتیاز F1 را از چندین ازمایش با مقادیر اریبی متفاوت کلاس IAI نشان می دهد. شکل 2 گرافیک این داده را نشان می دهد.

جدول 6: دقت، بازخوانی و امتیاز F1 با اریبی کلاس IAI متفاوت

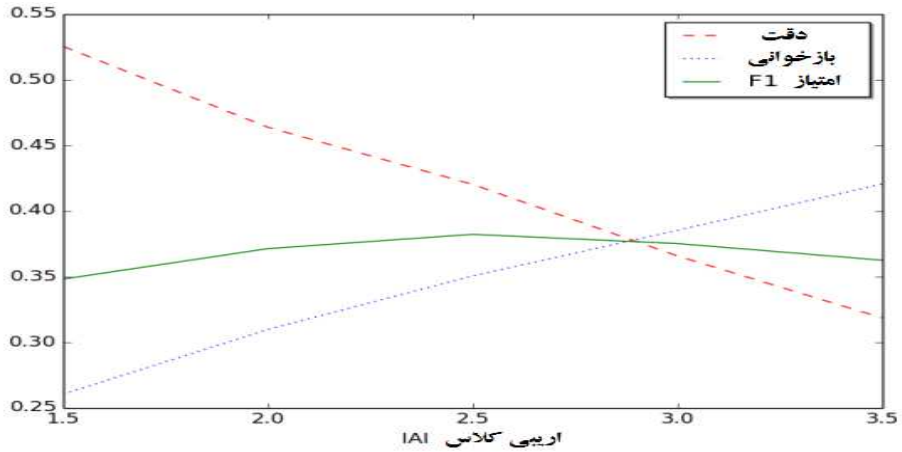
IAI Class Bias	Precision	Recall	F1 Score
1.5	**0.5252**	0.2602	0.3479
2.0	0.4636	0.3095	0.3711
2.5	0.4201	0.3503	**0.3820**
3.0	0.3656	0.3850	0.3750
3.5	0.3184	**0.4203**	0.3623

بهترین امتیاز F1 با مقدار اریبی کلاس IAI 2.5 بدست می اید. این خود مقدار 28.61 را از حیث عملکرد استخراج IAI بدون اریبی کلاس IAI نشان می دهد. به علاوه،هر دو دقت و بازخوانی بالاتر از مقدار معیار هستند.

7- نتیجه گیری

ما مدل استخراج شاخص های ویژگیهای ضمنی را توصیف کرده ایم که شامل کلماتی هستند که به طور ضمنی بیانگر ویژگی های ضمنی یک سند با استفاده از میدان های تصادفی شرطی هستند. ما یک مجموعه داده را برای این کار بر اساس یک متن مشخص برای نظر کاوی توسعه دادیم. همچنین ما ارزیابی عملکرد مقایسه ای را با سه معیار ارایه کردیم. نتایج نشان داد که رویکرد ما بر تر از این معیار هاست. ویژگی های مورد استفاده توصیف شده و در استخراج IAI بسیار ساده وموثر بودند.

برای مطالعات آینده، ما قصد داریم تا در مورد ویژگی های جدید برای این فرایند تحقیق کنیم. ما بر این باوریم که ویژگی های وابستگی نحوی قادر به بهبود عملکرد(30-31) می باشند. در نهایت ما بر روی مدل استخراج ویژگی های ضمنی بر اساس IAI کار میکنیم. ما به بررسی رویکرد های مختلف برای نگاشت IAI با ویژگی های ضمنی با استفاده از تشابه معنایی(32-35) خواهیم پرداخت.



شکل 2: دقت، بازخوانی و امتیاز F1 با CRF اریب