



بررسی ابهام زدایی معنایی کلمه ای ذهنی (SWSD) بر روی نظرات کاربران دیجی کالا جهت تجزیه و تحلیل احساسات کاربران

فاطمه زارعی^{۱*}، علیرضا هنرور^۲

Zareifateme38@yahoo.com

Alireza_honarvar@yahoo.co.uk

چکیده

پردازش زبان طبیعی یکی از زیرشاخه‌های مهم در حوزه گسترده هوش مصنوعی و البته در دانش زبان شناسی است. امروزه علاوه بر موضوع ابهام زدایی معنایی کلمات یکی دیگر از موضوعات مهم در زمینه پردازش زبان طبیعی ابهام زدایی معنایی کلمه‌ای ذهنی نیز می باشد که تمرکز آن بر روی بخش احساسی جملات است. حالت و احساسات درونی یک انسان بخشی از ارتباطات روزمره اوست. حالات درونی یک فرد را می توان از برخی از رفتارهای خروجی او مانند اظهارنظرها، حالت چهره و رفتارهای حرکتی اعضای بدنش تشخیص داد. یکی دیگر از رفتارهای خروجی که نشان دهنده حالات درونی فرد می باشد گرامر و نحوه ی نوشتاری او نیز است. ابهام زدایی معنایی کلمه‌ای ذهنی از جمله حوزه‌های تحقیقاتی جدیدی است که اخیرا در برخی از زبان‌ها به ویژه زبان انگلیسی مورد توجه قرار گرفته است، اما از آن جا که تحقیقی بر روی این مسئله در زبان فارسی تاکنون انجام نشده است می توان تحقیق حاضر بر روی این مقوله را یک نوآوری محسوب نمود. به عبارتی در این تحقیق سعی بر این شده است که با استفاده از الگوریتم های یادگیری ماشینی به تجزیه و تحلیل نظرات و احساسات کاربران سایت دیجی کالا در مورد کالای موبایل پردازیم که بتوانیم با تجزیه و تحلیل احساسات موجود، فروش موفق محصولات را پیش بینی نماییم.

کلمات کلیدی: پردازش زبان طبیعی، ابهام زدایی معنایی کلمه ای ذهنی، تجزیه و تحلیل احساسات

۱- مقدمه

پردازش زبان های طبیعی^۱ یکی از زیرشاخه های با اهمیت در حوزه گسترده هوش مصنوعی، و نیز در دانش زبان شناسی است. تلاش عمده در این زمینه ماشینی کردن فرآیند درک و برداشت مفاهیم بیان گردیده با یک زبان طبیعی انسانی می باشد. به تعریف دقیق تر پردازش زبان های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و نوشتاری. با استفاده از آن می توان به ترجمه زبان ها پرداخت، از صفحات وب و بانک های اطلاعاتی نوشتاری جهت پاسخ دادن به پرسش ها استفاده کرد، یا با دستگاه ها مثلا برای مشورت گرفتن به گفتگو پرداخت. در واقع منظور از پردازش زبان طبیعی این است که رایانه ای بتواند زبان انسان را تحلیل و ارزیابی کند و حتی قادر به تولید زبان طبیعی باشد. هدف اصلی در پردازش زبان طبیعی

^۱ Natural Language Processing



این است که با استفاده از الگوریتم‌ها و ساختارهای داده‌ای در علوم رایانه، نظریه‌های محاسباتی زبان را ارائه کنیم. کاربردهای نوشتاری و گفتاری از مهم‌ترین زمینه‌های کاربرد پردازش زبان طبیعی هستند که یافتن اطلاعاتی خاص در یک متن یا ترجمه یک متن به زبانی دیگر را امکان‌پذیر می‌کنند. در سال‌های اخیر، تحقیقات بسیار زیادی در این زمینه انجام شده است که تحقق این اهداف مستلزم داشتن دانشی وسیع از زبان است، بنابراین علاوه بر محققان علوم رایانه، دانش متخصصان زبانشناس نیز در دستیابی به خدمات و سیستم‌های مبتنی بر کاربردهای مختلف پردازش زبان مانند سیستم‌های ارتباط با مشتری از طریق تلفن یا سیستم‌های آموزش به دانش‌آموزان تاثیرگذار خواهد بود.

در حال حاضر استخراج خودکار اطلاعات برای صنایع چالش‌بزرگی است. علت اصلی برای این چالش این است که اطلاعات با ارزش، درون زبان طبیعی غیر ساختار یافته‌ای که طبیعتاً مبهم می‌باشد، مخفی شده است (Golkar, J. Ali, D. Shahram, F. Seyed Mohammad Sadegh and S. Mostafa, 2015:1-5).

ابهام زدایی معنایی کلمه‌های^۱ یک مسئله‌ی طولانی مدت در زبان شناسی می‌باشد (G. R. Eneko Agire, 1996:16-22). این روش به طور معمول در چندین حوزه از قبیل یادگیری ماشینی^۲، بازیابی اطلاعات^۳، هدایت ابر متن‌ها، آنالیز محتوا و موضوع، پردازش متن، پردازش گفتاری و تجزیه و تحلیل قواعد دستوری، به کار برده می‌شود (N. Ide and J. Veronic, 1998:1-40). کلماتی در زبان طبیعی وجود دارد که معانی متفاوتی با زمینه‌های متفاوت دارند، اما املاي آنها یکسان می‌باشد. این کلمات، کلمات کثیر المعنی نامیده می‌شوند. ابهام زدایی معنایی کلمه‌ای راه حلی برای این مسئله می‌باشد. ابهام زدایی معنایی کلمه‌ای (Christopher D. Manning, and Hinrich Schutze, 1999:1-704) به طور خودکار معنای صحیح کلماتی را که در یک زمینه خاص دارای معانی متعددی می‌باشند را مشخص می‌کند.

دو روش عمده برای WSD وجود دارد، روش‌های مبتنی بر دانش^۴ و روش‌های مبتنی بر یادگیری ماشینی^۵ (Shallu, 2013:354-360). سیستم‌های ابهام زدایی معنایی کلمه‌ای مبتنی بر دانش، اطلاعاتی را مبتنی بر دانش لغوی^۶ برای اجرای WSD استخراج می‌کنند (Agirre, D. L. Eneko, S. Oier Lopez and Aitor, 2014: 57-84). روش‌های مبتنی بر یادگیری ماشینی، ابهام زدایی را بر اساس زنجیره‌ای از واژگان و آموزش مدلی بر اساس مجموعه‌ای از برچسب‌ها یا غیر برچسب‌ها یا مدل‌های آماری / احتمالی، انجام می‌دهند (Navigli and Roberto, 2009:1-69). متأسفانه، در حال حاضر ما یک بینش صریح در مورد هر یک از روش‌های ابهام زدایی تحت یک شرایط مناسب را در اختیار نداریم (Navigli, P. Roberto and MohammadTaher, 2014: 837-881). سیستم‌های مبتنی بر یادگیری ماشینی دارای سرعت بالایی هستند و این در حالی است که سیستم‌های مبتنی بر یادگیری ماشینی از دقت بالایی برخوردار هستند. هر یک از این روش‌ها نقاط ضعف و قوت خود را دارند، به طوری که پژوهش‌های متعدد ایده‌ی استفاده از ترکیب این دو روش را مطرح کرده‌اند، به عنوان مثال یک پژوهش نشان داده است که ترکیبی از دو روش بهتر از هر یک از روش‌ها به طور جداگانه می‌باشد، به عبارتی این دو رویکرد می‌توانند نقش مکمل یکدیگر را ایفا نمایند (Montoyo, S. Andres, R. Armando, German, Palumar and Manuel, 2005:299-330).

¹ Word Sense Disambiguation

² Machine Learning

³ Information Retrieval

⁴ Knowledge-Based

⁵ Machine Learning Based

⁶ lexical knowledge-based



تجزیه و تحلیل ذهنی^۱ یکی از حوزه های فعال پژوهشی در پردازش زبان طبیعی به شمار می رود ، به طوری که استخراج خودکار عبارات زبانی حالات خاص از جمله فکری^۲ و هیجانی^۳ مانند : عقاید ، باورها ، احساسات ، اهداف ، ارزیابی ها و افکار در یک متن می باشند (Akkaya and Cem, 2013:1-130).

ابهام زدایی معنایی کلمه ای ذهنی^۴ به طور خودکار تعیین می کند که کدام لغت در یک مجموعه از واژگان به عنوان مفهوم ذهنی^۵ و کدام یک به عنوان مفهوم عینی^۶ استفاده می شود ، به عبارتی SWSD باعث بهبود عملکرد سیستم های ذهنی و تجزیه و تحلیل آن ها خواهد شد (Akkaya Cem, Wiebe Janyce and Mihalcea Rada, 2009:190-199).

در حال حاضر ابهام زدایی معنایی کلمه ای ذهنی به یکی از زمینه های مهم تحقیقاتی پردازش زبان طبیعی تبدیل شده است که یک فیلد مهم مطالعاتی است که سعی می کند احساسات، رفتار، نظرات و تحلیل افراد مختلف را نسبت به موجودیت ها و ویژگی های آن بیان کند. این موجودیت می تواند محصول ها، سرویس ها، سازمان ها، افراد و رخدادها و موضوعات باشد. این نظر می تواند مثبت و یا منفی و یا هیچکدام از این دو باشد و به صورت نظر خنثی طبقه بندی شود. به عبارتی ابهام زدایی معنایی کلمه ای ذهنی شاخه ای از علوم کامپیوتر و پردازش زبان طبیعی است که سعی دارد ماشین و هوش مصنوعی را با احساس و عواطف انسانی آشنا سازد و تشخیص آن ها از هم را میسر سازد. هدف از این مقاله تجزیه و تحلیل احساسات نظر کاربران با استفاده از داده های جمع آوری شده در مورد کالای موبایل است، که در ادامه این نوشتار و در قسمت دوم پیشینه تحقیق، در قسمت سوم روش اجرای تحقیق و در قسمت آخر جمع بندی و نتیجه گیری از مطالب آورده شده است.

۲- پیشینه تحقیق

تحقیقات در زمینه پردازش زبان طبیعی از سال های بسیار دور در جریان بوده و مربوط به دهه ۱۹۴۰ میلادی است. از پردازش زبان طبیعی، در طول جنگ جهانی دوم و در حدود سالهای ۱۹۴۰ برای رمزگشایی کدهای دشمن استفاده شد. اما زمان صلح فرصت خوبی به محققان داد تا در پی کاربردهای جدیدی در حوزه پردازش زبان طبیعی باشند. نخستین تلاش ها در این زمینه که در پی اهداف ترجمه ماشین بودند، از موفقیت چندانی برخوردار نبود. به طور کلی تاریخچه پردازش زبان طبیعی از دهه ۱۹۵۰ میلادی شروع می شود. در ۱۹۵۰ آلن تورینگ مقاله معروف خود را درباره ی آزمایش تورینگ که امروزه به عنوان ملاک هوشمندی شناخته می شود، منتشر ساخت. نخستین تلاش ها برای ترجمه توسط رایانه ها ناموفق بودند ، به طوری که ناامیدی بنگاه های تأمین بودجه ی پژوهش از این حوزه را نیز در پی داشتند. پس از اولین تلاش ها آشکار شد که پیچیدگی زبان بسیار بیشتر از چیزی است که پژوهشگران در ابتدا پنداشته بودند. بی گمان حوزه ای که پس از آن برای استعانت مورد توجه قرار گرفت زبان شناسی بود. اما در آن دوران نظریه ی زبان شناسی وجود نداشت که بتواند کمک شایانی به پردازش زبان ها نماید. در سال ۱۹۵۷ کتاب ساختارهای نحوی اثر نوام چامسکی زبان شناس جوان آمریکایی که از آن پس به شناخته شده ترین چهره ی زبان شناسی نظری تبدیل شد به چاپ رسید. از آن پس پردازش زبان با حرکت های تازه ای دنبال شد اما هرگز قادر به حل کلی مسئله نشد.

¹ Subjectivity analysis

² mental

³ emotional

⁴ Subjectivity Word Sense Disambiguation

⁵ subjective

⁶ objective

در دهه ۱۹۸۰ و ۱۹۹۰ پژوهش بر روی WSD به نقطه عطف خود رسید، درست زمانی که منابع واژگان مانند لغت نامه ها و اصطلاح نامه ها به طور گسترده در دسترس قرار گرفتند به طوری که کارهای انجام شده در این حوزه از لحاظ تئوری جالب بود اما کاربرد چندانی نداشت- (Montoyo, S. Andres, R. Armando, German, Palumar and Manuel, 2005:299-330). بسیاری از محققان از یک لغت نامه ماشین قابل خواندن^۱ به عنوان منبعی از دانش واژگان برای ابهام زدایی معنایی کلمه استفاده نمودند، استفاده از این روش ها باعث بهره‌گیری از دانش موجود در فرهنگ لغت ها و در نتیجه اجتناب از مقادیر زیادی از اصول آموزشی خواهد شد (Lesk, M, 1986:24-26).

در سال ۱۳۷۲، در آزمایشگاه پردازش زبان طبیعی دانشگاه صنعتی شریف، مهنوش شمس فرد (۱۳۷۳) سامانه ای به نام دنا طراحی و پیاده نمود. در این سامانه جهت پردازش زبان فارسی با به کارگیری نظریه وابستگی مفهومی شنک (۱۹۷۵)، جمله های فارسی به شبکه ای از مفهوم ها و روابط میان آنها تبدیل می شوند. گام های گوناگون پردازش جملات عبارت اند از واکافت واژه ای، واکافت ساخت واژی، واکافت نحوی، واکافت معنایی و استنتاج. پس از آن، سه پروژه دیگر با تمرکز بر بخش های گوناگون سامانه دنا ۱ تعریف و اجرا شدند. هر یک از این پروژه ها سعی در گسترش یکی از بخش های این سامانه داشتند.

بالاخره Cem Akkaya و همکارانش در سال ۲۰۰۹ و به طور مجزا در سال ۲۰۱۳ به معرفی و تحقیق بر روی موضوع جدید ابهام زدایی معنایی کلمه ای ذهنی نیز پرداختند که در آن تجزیه و تحلیل احساسات با بهره گیری از روش مبتنی بر لغت نامه انجام گرفته است.

۳- روش اجرای تحقیق

در یک مجموعه از واژگان، کلمات به دو دسته Objective و Subjective تقسیم می‌شوند. در دسته‌بندی Objective نظرات به دو گروه عینی (خنثی) و ذهنی (مثبت و منفی) نیز تقسیم می‌شوند که هدف آن جداسازی جملات دارای احساس با جملاتی است که خبری (خنثی) هستند. در طبقه‌بندی Subjective هدف ما جداسازی نظرات مثبت و منفی از یکدیگر هستند، به عبارتی در دسته بندی Subjective تمرکز ما بر روی بخش احساسی جملات می‌باشد.

افراد نسبت به موضوعات مختلفی مثل فیلم‌ها، محصولات و یا ... ابراز نظر می‌کنند که در این صورت می‌توان با SWSD همه اینها را استخراج کرد و از این اطلاعات برای اهداف بعدی خود استفاده کرد. به عنوان مثال جمله ی زیر را در نظر بگیرید: دوربین گنوم زوم عالی داره ولی تو شب ها اصلا خوب عکس نمی گیره ایراد دیگه ای که داره خیلی سنگینه. در این جمله نسبت به زوم حس مثبت و نسبت به وزن و عکاسی در شب نظر منفی ابراز شده است. و یا شبکه‌ی اجتماعی را تصور کنید که لحن و احساس شما را از آپدیت روزانه ی که پست کردید تشخیص می دهد و شما را با افرادی با حس مشابه در تماس قرار میدهد تا با هم بر سر مشکل مشترکتان (مثلا حقوق کم) درد و دل کنید! همچنین اگر شبکه ی اجتماعی دارای جامعه‌ی بزرگی باشد اطلاعات جمع آوری شده‌ی آن برای امور آماری - روان شناسی بی نظیر خواهند بود. مثلا روان شناسان می توانند شروع یک اپیدمی افسردگی در یک منطقه ی خاص را تشخیص دهند. از آنجا که میزان جرم وقتی که مردم یک منطقه مشکلات روحی دارند بیشتر است پلیس نیز می‌تواند نیرو های خود را بر این اساس در جاهای مختلف تقسیم کند تشخیص احساس در زمان حال بیشترین کاربرد را در سایت های عرضه انواع کالا دارد. بسیار مهم است که مدیر سایت و شرکت عرضه کننده ی محصول بدانند کدام نقد و بررسی ها از یک محصول مثبت هستند و کدام منفی و اینکه شدت منفی

^۱ machine-readable dictionaries



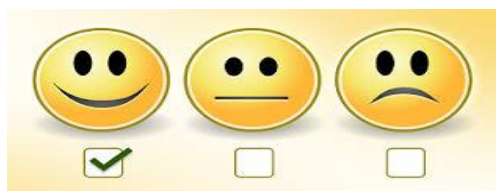
یا مثبت بودن چقدر است . کاربر از محصول کمی ناراضی است یا شدیداً ؟ شرکت های بزرگی همچون Amazon ، Ebay،Google از SWSD برای بررسی نظرات کاربران استفاده می کنند.

تشخیص احساس، ماشین های فعلی را یک نسل به جلو رانده و باور عمومی بر بی احساس بودن ماشین ها را کم رنگ تر می سازد . در پی آن باعث آسان تر شدن بسیاری از جهات زندگی انسان نیز خواهد شد . علی رغم این که هیچ پیشینه ی پژوهشی خاصی در رابطه با این موضوع در کشورمان موجود نمی باشد ولی با توجه به اهمیت این موضوع سعی بر این شده است که بتوان با استفاده از داده های جمع آوری شده و استخراج و بازیابی آن ها گامی موثر در جهت پیشبرد این موضوع برداشت. به همین جهت ما در این پژوهش سعی بر این داریم که با بهره گیری از داده های استخراج شده از سایت ایرانی دیجی کالا احساس یک گوینده یا نویسنده را که در پیکره ی یک متن موجود می باشد را استخراج نماییم.

در زمینه ابهام زدایی معنایی کلمه ای ذهنی تاکنون روش های مختلفی پیشنهاد شده است که به اختصار به توضیح هر یک از آن ها می پردازیم :

۳-۱- شکلک ها

همان طور که در شکل ۱ مشاهده می شود، ساده ترین روش تشخیص احساس یک نویسنده مشاهده ی شکلک هایی است که او در یک متن به کار برده است. با این شکلک های بسیار ساده می توان مفاهیم احساسی پیچیده ای را انتقال داد. اما با این وجود بسیاری از پروژه های فعلی، شکلک ها را به دلیل استفاده نادرست کاربران از آن ها و کاربرد بسیار کم شکلک ها در متن های جدی از جمله محیط های رسمی نادیده می گیرند (J. Park, V. Barash, C. Fink, and M. Cha, 2013:466-475).



شکل ۱ - نمونه ای از شکلک ها

۳-۲- روش مبنی بر لغت نامه^۱

در این روش با توجه به مثبت یا منفی بودن کلمه در یک مقیاس مشخص ارزشی معادل با آن به کلمه اختصاص داده می شود و متن با استفاده از پردازش زبان طبیعی آنالیز می گردد. به عبارتی در این روش از قواعد گرامری برای شناخت زبان و از دیکشنری های لغوی مانند: Wordnet^۲ و Senti Wordnet (یک نسخه ی ویرایش شده از ورد نت است) برای ارزش دهی به کلمات استفاده می کند. مهم ترین ضعف این روش وابستگی به دیکشنری های لکسیکال مانند Wordnet می باشد. این دیکشنری ها برای زبان انگلیسی تهیه شده اند و برای سایر زبان ها از جمله زبان فارسی کاملاً بلا استفاده اند .

^۱ Lexicon-based Approach

^۲ یک پایگاه داده ی لغوی برای زبان انگلیسی است که اطلاعاتی در مورد چهار بخش گفتار در اختیار می گذارد که عبارتند از اسم، صفت، فعل و قید. واژه نامه مجموعه ای از لغات احساسی با قطبیت آن کلمه و درجه شدت آن است. معمولاً برای ایجاد واژه نامه ابتدا یک لیست اولیه تشکیل می شود و سپس با استفاده از هم معنی و مخالف معنی این کلمات، این لیست لغات گسترش پیدا می کند.



برای رفع این مشکل باید به طور دستی یک دیکشنری مختصر مربوط به دامنه‌ی کاری ایجاد شود که بسیار وقت گیر خواهد بود. این روش از لحاظ سرعت و قدرت پردازش بسیار سریع‌تر از سایر روش‌ها عمل می‌کند. هرچند میزان خطای بالای آن و همچنین الزام نوشتن قواعد بسیار برای زبان باعث شده این روش به ندرت مورد استفاده قرار بگیرد. امروزه از روش‌های یادگیری ماشینی به دلیل سادگی و همچنین عدم نیاز به دسترسی به ساختار جمله بیشتر استفاده می‌شود (Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede, 2011:267-307)

۳-۳- روش مبنی بر یادگیری ماشینی

در این روش بر خلاف روش قبل نیاز به در نظر گرفتن ساختار گرامری جمله ندارد و از رده بند هایی^۱ مانند الگوریتم درخت تصمیم، بیز و ... استفاده می‌کند. در عوض باید به طور مناسبی با داده‌های آموزشی^۲ تربیت^۳ شوند. "مناسب" از این لحاظ که اگر مجموعه داده‌ی^۴ ما دارای موضوعی خاص مثلا بررسی نقدهای یک کالا مانند موبایل باشد داده‌های آموزشی نیز به تبعه باید از همین نوع باشد. استفاده از داده‌های آموزشی نامناسب باعث افت شدیدی در دقت کلاس بندی شده و این اهمیت داده‌های آموزشی مناسب را نشان می‌دهد. به عنوان یکی از شاخه‌های وسیع و پر کاربرد هوش مصنوعی، یادگیری ماشینی به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آن‌ها رایانه‌ها و سامانه‌ها توانایی تعلم و یادگیری پیدا می‌کنند. یکی از تقسیم‌بندی‌های متداول در یادگیری ماشینی بر اساس نوع داده‌های در اختیار عامل هوشمند عبارتند از:

۱. یادگیری با نظارت (با مربی)^۵: یکی از روش‌های یادگیری است که مجموعه‌ای از جفت‌های ورودی- خروجی ارائه شده، و سیستم تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد.
۲. یادگیری بی نظارت (بدون مربی)^۶: یکی از روش‌های یادگیری است که یک مجموعه از مثال‌های یادگیری وجود دارد که در آن فقط مقدار ورودی‌ها مشخص است و اطلاعاتی در مورد خروجی صحیح در دست نیست.
۳. یادگیری نیمه نظارتی^۷: ترکیبی از دو روش یادگیری قبلی که عامل هوشمند هم از داده‌های بدون برچسب و هم از داده‌های با برچسب استفاده می‌کند.

۳-۴- روش‌های ترکیبی

این روش ترکیبی از روش‌های فوق می‌باشد. در شکل ۲ تمامی روش‌های مورد استفاده به صورت خلاصه آورده شده است.

¹ Classifier

² Learning set

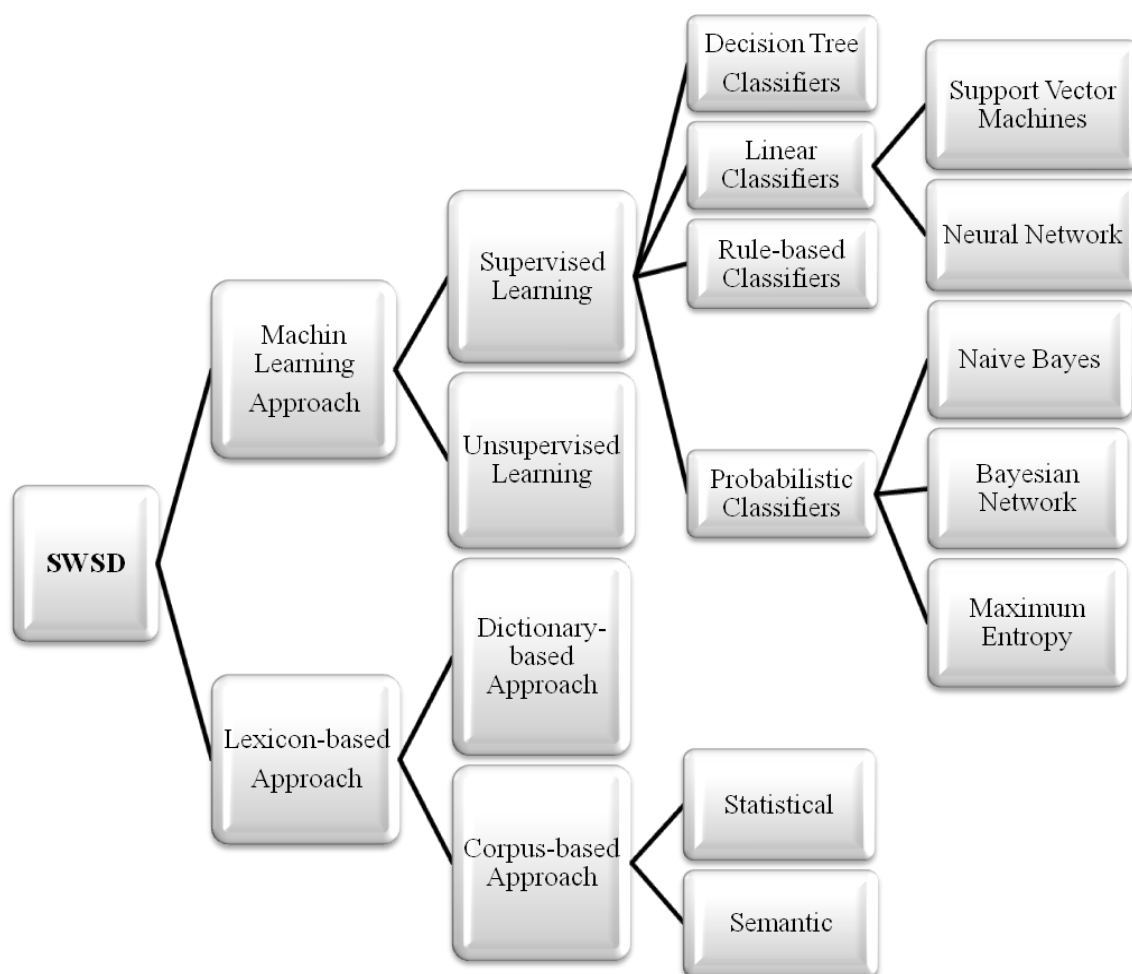
³ Train

⁴ dataset

⁵ Supervised Learning

⁶ Unsupervised Learning

⁷ Semi-supervised Learning



شکل ۲- روش های مورد استفاده در SWSD



۳-۵- روش کار

دیجی کالا نام یکی از وبگاه‌های فعال در زمینه تجارت الکترونیک در ایران است. این وبگاه بزرگ‌ترین فروشگاه آنلاین در کشور است و به طور میانگین روزانه بیش از ۸۵۰ هزار بازدید کننده دارد، دیجی کالا مرجع تخصصی نقد و بررسی و فروش اینترنتی کالا در ایران است. گروه‌های مختلف کالا مانند کالای دیجیتال، لوازم خانگی، لوازم شخصی، فرهنگ و هنر و ورزش و سرگرمی با تنوعی بی نظیر در دیجی کالا عرضه می‌شوند. دیجی کالا با در اختیار داشتن سهم عمده‌ای از بازار عرضه اینترنتی کالا در ایران با تنوعی بسیار بالا از برندهای مختلف، در همراهی مشتریان در فرآیند خرید و حفظ جایگاه بی‌طرفانه در میان آن‌ها، نهایت تلاش خود را می‌کند. دیجی کالا از اولین دریافت کنندگان نماد اعتماد الکترونیک از مرکز توسعه تجارت الکترونیک وزارت صنعت، معدن و تجارت و به عنوان برگزیده‌ی بسیاری از مراجع و نهادهای ذی‌ربط، دارای کلیه مجوزهای رسمی مورد نیاز برای فعالیت در این حوزه است. همچنین، در حال حاضر دیجی کالا با توجه به استقبال چشمگیر کاربران و مشتریان خود، پرمخاطب‌ترین و پربازدیدترین وب سایت تجاری کشور است. (به استناد آمارهای منتشر شده در مراجع آماری مانند Alexa).

از آن جا که یک نمونه باید طوری انتخاب شود که دارای تنوع و به اندازه کافی بزرگ باشد تا نتایج حاصل قابل اعتماد باشد، و به دلیل عدم وجود یک مجموعه داده‌ی مناسب به زبان فارسی جهت تجزیه و تحلیل احساسات موجود در زبان مادریمان در این پژوهش به دلیل دقت بیشتر، صرفه جویی در وقت و کاهش هزینه و به دلیل حجم بالای خدمات دیجی کالا و وجود تعداد نظرات نامحدود در مورد هر یک از کالاهای ارائه شده در سایت دیجی کالا، داده‌های مورد استفاده در این تحقیق در مورد چهارده برند مختلف موبایل می باشد که به صورت دستی جمع آوری شده و نظرات^۱ کاربران در مورد ویژگی‌های مختلف آن‌ها نیز گرد آوری شده است و در نهایت با ویرایش نظرات جمع آوری شده و حذف کلمات زائد^۲ و استفاده از کلمات کلیدی موجود در نظرات، با استفاده از الگوریتم‌های یادگیری نظارت شده به بررسی و تشخیص احساس کاربران^۳ پرداخته ایم.

در این مدل مجموعه داده‌ی ما دارای ۱۰۰ نمونه^۴ و ۲۵ ویژگی^۵ می‌باشد. که نهایتاً با استفاده از دو کلاس A و B که دارای مقادیر اسمی^۶ هستند توانسته ایم با استفاده از الگوریتم‌های یادگیری ماشینی به روش نظارت شده و با استفاده از معیارهایی همانند ماتریس در هم ریختگی^۷، دقت یا نرخ دسته بندی^۸، خطای دسته بندی^۹، معیار AUC^{۱۰}، سطح زیر نمودار نمودار ROC^{۱۱}، معیارهای Kappa، Recall و Precision به تجزیه و تحلیل احساسات نظرات کاربران پردازیم. که در نهایت می توان نتایج تمامی معیارها را به صورت زیر در جدول شماره ۲،۱ و ۳ خلاصه نمود.

¹ Comments

² Stopword

³ User

⁴ Instance

⁵ attribute

⁶ Nominal

⁷ Confusion Matrix

⁸ Classification Accuracy - Rate

⁹ Error Rate

¹⁰ Area Under Curve

¹¹ Receiver Operating Characteristic



جدول شماره ۱ نتایج تجزیه و تحلیل داده ها

الگوریتم ها	accuracy	Error Rate	ROC	Kappa
J48	٪.۸۸	٪.۱۲	۰.۶۱۴	۰.۳۹
Random Forest	٪.۸۷	٪.۱۳	۰.۶۵۳	۰
Random Tree	٪.۸۵	٪.۱۵	۰.۵۴۲	۰.۱۴
Lazy IBK	٪.۸۲	٪.۱۸	۰.۵۳۷	۰.۰۸
Naive Bayes	٪.۷۸	٪.۲۲	۰.۷۰۲	۰.۱۴
Lazy k-Star	٪.۷۸	٪.۲۲	۰.۶۱۱	۰.۰۳
Bayes Net	٪.۷۸	٪.۲۲	۰.۵۳۳	۰.۰۹

جدول شماره ۲ مقادیر Precision و Recall کلاس A

الگوریتم ها	Recall	Precision
Random Forest	۱	۰.۸۷
J48	۰.۹۵۴	۰.۹۱۲
Random Tree	۰.۹۵۴	۰.۸۸۳
Lazy IBK	۰.۹۲	۰.۸۷۹
Lazy k-Star	۰.۸۷۴	۰.۸۷۴
Bayes Net	۰.۸۶۲	۰.۸۸۲
Naive Bayes	۰.۸۵۱	۰.۸۹۲

جدول شماره ۳ مقادیر Precision و Recall کلاس B

الگوریتم ها	Recall	Precision
J48	۰.۳۸۵	۰.۵۵۶
Naive Bayes	۰.۳۰۸	۰.۲۳۵
Bayes Net	۰.۲۳۱	۰.۲
Random Tree	۰.۱۵۴	۰.۳۳۳
Lazy IBK	۰.۱۵۴	۰.۲۲۲
Lazy k-Star	۰.۱۵۴	۰.۱۵۴
Random Forest	۰	۰

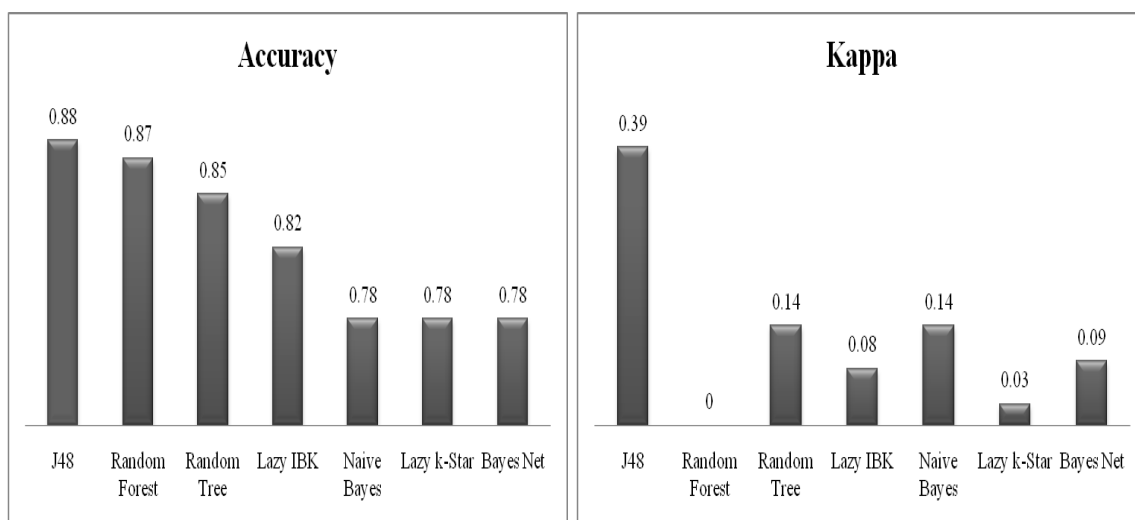


۴- نتیجه گیری

توسعه اینترنت روش‌های خرید ما را به کلی دگرگون کرده است. منافع موجود در خرید اینترنتی هر روز تعداد بیشتری از مردم را به تجربه آن و ایجاد تغییر در الگوهای متداول خرید ترغیب می‌کند. امروز دیگر برای بیشتر نیازهای روزمره راه حل‌های جدیدتری وجود دارد، از جمله برای انتخاب و خرید کالا، دیگر کمتر کسی برای خرید کالا، بدون تحقیق و بررسی دقیق اقدام می‌کند. مشتریان پیشاپیش و قبل از خرید از حق انتخاب گسترده خود و اطلاعات موجود نهایت استفاده را می‌برند. زندگی امروزی بی‌گمان خرید به روش‌های نوتر، هوشمندانه‌تر و البته لذت بخش‌تری را می‌طلبد. از همین روست که هر روز به تعداد استفاده کنندگان سرویس‌های اینترنتی و مشتریان فروشگاه‌های اینترنتی افزوده می‌شود.

امروزه بازار تبلیغ محصولات بسیار گسترده و وسیع است شرکت‌ها برای فروش محصولات بیشتر اقدام به تبلیغ محصولات می‌کنند. یکی از راه‌های تبلیغ نیز جلب رضایت مشتری است. یکی از راه‌های اندازه‌گیری رضایت مشتریان بررسی نظرات مشتریان در فضای مجازی است. شرکت‌ها با بررسی این نظرات نکات مثبت و منفی که در نظرات ارائه شده است را بررسی می‌کنند و از آن‌ها برای بالا بردن کیفیت محصولات خود استفاده می‌کنند. همان‌طور که پیش‌تر نیز اشاره نمودیم با جمع‌آوری نظرات کاربران دیجی کالا بر روی چند برند مختلف موبایل بر آن شده‌ایم که با تجزیه و تحلیل احساسات موجود بتوانیم فروش موفق محصولات را پیش‌بینی نماییم و گامی موثر جهت بهبود تجارت الکترونیک برداریم.

با توجه به جدول رسم شده در بخش قبل الگوریتم J48 دارای بیش‌ترین مقدار Accuracy و کم‌ترین مقدار Error Rate می‌باشد، همچنین در معیارهای Kappa، Recall و Precision این الگوریتم نیز دارای ماکزیمم مقدار می‌باشد و فقط در معیار ROC پس از الگوریتم‌های Naive Bayes و Random Forest دارای بیش‌ترین مقدار است اما با توجه به اینکه اختلاف مقادیر به دست آمده بسیار ناچیز می‌باشد و الگوریتم‌های ذکر شده در سایر معیارها دارای نتایج چشمگیری نبوده‌اند پس می‌توان نتیجه گرفت که با توجه به سایر الگوریتم‌های پیشنهادی بهترین الگوریتم با توجه به این نوع داده آموزشی جهت تجزیه و تحلیل احساسات موجود در نظرات و پیش‌بینی فروش موفق محصولات الگوریتم J48 می‌باشد. به عنوان نمونه نمودارهای Accuracy و Kappa با توجه به جدول شماره ۱ در زیر آورده شده است.



نمودار ۱- نمایش نتایج به دست آمده در فاز Accuracy و Kappa



امید است که در آینده بتوان با ایجاد یک واژه نامه همانند ورد نت به زبان فارسی و پیاده سازی روش مبتنی بر واژگان و استفاده از روش مبنی بر کورپوس نیز بتوان به تجزیه و تحلیل احساسات پرداخت و در نهایت با ترکیب دو روش نظارت شده و مبنی بر واژگان بتوان گامی بلند در راستای زبان فارسی برداشت.

مراجع

1. Agirre, D. L. Eneko, S. Oier Lopez and Aitor, "Random walks for knowledge-based word sense disambiguation.," Computational Linguistics , vol. 40, no. 1, pp. 57-84, 2014.
2. Akkaya and Cem, "SUBJECTIVITY WORD SENSE DISAMBIGUATION: A TOOL FOR SENSE-AWARE SUBJECTIVITY ANALYSIS," Diss. University of Pittsburgh, 2013.
3. Akkaya Cem, Wiebe Janyce and Mihalcea Rada "Subjectivity Word Sense Disambiguation" Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 190–199, Singapore, 6-7 August 2009
4. Christopher D. Manning, and Hinrich Schutze "Foundations of Statistical Natural Language Processing," MIT Press, Cambridge, Massachusetts London, England 1999.
5. Golkar, J. Ali, D. Shahram, F. Seyed Mohammad Sadegh and S. Mostafa, "Improve Word Sense Disambiguation by Proposing a Pruning Method for Optimizing Conceptual Density's Contexts," in International Conference on Artificial Intelligence and Signal Processing, Shiraz, Iran, 2015.
6. G. R. Eneko Agire, "Word Sense Disambiguation using Conceptual Density," in 6 th conference on Computational linguistics, 1996.
7. J. Park, V. Barash, C. Fink, and M. Cha. Emoticon style: Interpreting differences in emoticons across cultures. In International AAAI Conference on Weblogs and Social Media (ICWSM), 2013
8. Lesk, M. (1986). Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC Conference, Association for Computing Machinery, pp. 24–26 Toronto, Canada.
9. MaiteTaboada Simon Fraser University Julian Brooke University of Toronto Milan Tofiloski Simon Fraser University Kimberly Voll University of British Columbia Manfred Stede University of Potsdam Lexicon-Based Methods for Sentiment Analysis, 2011.
10. Montoyo, S. Andres, R. Armando, German, Palumar and Manuel, "Combining Knowledge-and Corpus-based Word-Sense-Disambiguation Methods," Journal of Artificial Intelligence , vol. 23, pp. 299-330, 2005.
11. Navigli and Roberto, "Word Sense Disambiguation: A survey," ACM Computing Surveys, vol.41, no. 10, 2009.
12. Navigli, P. Roberto and MohammadTaher, "A Large-Scale Pseudoword-Based Evaluation



13. Framework for State-of-the-Art Word Sense Disambiguation,” Computational Linguistics, vol. 40,no. 4, pp. 837-881, 2014.
14. N. Ide and J. Veronic, “Introduction to the special issue on word sense disambiguation: the state of the art,” Computational Linguistics Special issue on word sense disambiguation, vol. 24, no. 1, pp. 2-40, 1998.
15. Shallu, G. Shallu and Vishal, “A Survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages,” Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 4, pp. 354-360 , 2013.