تعریف متن کاوی

به فرآیند تحلیل متن برای استخراج یا کشف اطلاعات و واقعیت‌های معتبر، جدید و از پیش ناشناخته، پنهان، مفید و قابل درک از داده‌های ساخت نیافته و نیمه ساخت یافته به صورت خودکار (توسط رایانه) گفته می شود. [1]

متن کاوی در مقابل داده کاوی

کاوش متن تا حدی مشابه داده کاوی است و تفاوت در این است که در داده کاوی متدها روی داده‌های ساخت یافته ی پایگاه داده‌ای هستند اما در متن کاوی می‌توان روی داده‌های ساخت نیافته و یا نیمه ساخت یافته مثل ایمیل، اسناد متنی، فایل‌های html و … اعمال شود. [2]

داده ی ساخت یافته، ساخت نیافته و نیمه ساخت یافته

داده ی ساخت یافته به داده‌هایی گویند که به صورت سطری و ستونی مرتب شده‌اند مانند داده‌های پایگاه داده ای ، data wherehouse و جداول

داده ی نیمه ساخت یافته به صورت سطری و ستونی و یا به شکل داده‌های پایگاه داده ی رابطه‌ای نیستند، اما دارای برچسب‌ها و نشان هایی هستند که عناصر معنایی را جدا می‌کند و ساختار سلسله مراتبی پیدا می کنند. مانند فایل‌های xml ، html ، و json

داده‌های ساخت نیافته، ساختار از پیش تعریف شده‌ای وجود ندارد. مانند بدنه ی ایمیل،‌اسناد متنی و یا پیام‌های موجود در شبکه‌های اجتماعی [3]

پردازش زبان‌ طبیعی

از آنجا که افراد برای تعامل و ارتباط با یکدیگر متن رد و بدل می‌کنند و هیچ کدام داده‌های ساخت یافته به هم نمی فرستند، بنابراین برای تحلیل آن‌ها باید از تکنیک های متن کاوی استفاده کرد. یکی از کاربردهای متن کاوی، پردازش زبان‌های طبیعی (NLP) است و حوزه ای است که بر تعامل بین زبان انسان و کامپیوتر مطالعه می کند. پردازش زبان طبیعی یک تکنیک محاسباتی برای کامپیوتر است تا از زبان انسان را به طور معناداری تحلیل کند، معنا را استخراج کند و آن را درک کند. [4]

یکی از کاربردهای پردازش زبان طبیعی و متن کاوی، عقیده کاوی است. در ادامه درباره ی مفاهیم این حوزه، کاربردها، چالش ها و تکنیک های موجود توضیح خواهیم داد.

تعریف عقیده کاوی

عقیده کاوی به استخراج ایده‌ها و تحلیل معنایی آن‌ها در یک متن ساخت نیافته که به زبان طبیعی بیان شده اشاره دارد. در‌واقع این فرآیند به جای رویارویی با متن، تمرکز بر محتوا و احساسات نهفته در آن‌ها دارد و با کشف آن‌ها به نتایج مورد نظر می رسد. هدف اصلی عقیده کاوی کشف رویکردها، لحن، احساسات و درجه ی آگاهی موجود در متن های مورد نظر است. [1]

عقیده یا احساس به صورت یک تاپل پنج تایی تعریف می‌شود که شامل آیتم های زیر است:

منظور از موجودیت i ام است ، خصیصه ی j ام ، فرد k ام است که این نظر را دارد، زمان انتشار این عقیده و نظر یا احساس نسبت به j امین خصیصه ی موجودیت i ام () که فرد در زمان داشته است.

هدف عقیده کاوی یافتن این تاپل برای یک متن است. گاهی مشخص کردن سه آیتم اول یعنی موجودیت و خصیصه ی آن و نظر فرد نبست به این موجودیت کفایت می کند. برای مثال جمله ی زیر را در نظر بگیرید:

“The screen of this mobile phone is good! ”

در این جمله موجودیتmobile phone و خصیصه ی آنscreen است و نظر یا احساس بیان شده مثبت است. [5]

چرایی عقیده کاوی

رشد استفاده از اینترنت و فعالیت‌های آنلاین سبب شده تا اطلاعات زیادی تولید شود و حجم انبوهی از این اطلاعات مربوط به عقاید افراد است که تحلیل آن‌ها دشوار است و نیاز به تکنیک هایی برای خلاصه کردن عقاید وجود دارد. [6]

چالش ها

محتویات موجود در وب می‌تواند به زبان‌های مختلفی نوشته شود. از آنجایی که ایده‌های عمومی بیان شده توسط مردم، در تمام نقاط مختلف جهان صورت می‌گیرد به منابع ایده کاوی و تحلیل معنایی و تکنولوژی احتیاج داریم تا برای زبان‌های مختلف توسعه یابد.

استخراج دانش از وب به دلیل پویا بودن و بروز شدن مطالب با بیشترین چالش همراه است و اطلاعات جدید در هر لحظه به اطلاعات قبلی اضافه می‌شوند و داده‌ها به سرعت تغییر می کنند.

چالش دیگر تفکیک واقعیت از ایده‌ها به صورت خودکار است. واقعیت‌ها و ایده‌ها دو موضوع اصلی مضمون های اصلی در وب هستند.

استفاده از زبان‌های محاوره ای، اختصارها و تصویرسازی ها به روش‌های مختلف توسط افراد صورت می گیرد. در‌واقع طرز نوشتن و سطح علمی افراد با هم متفاوت است. همچنین عدم تشخیص جملات و کلمات طنزآمیز و کنایه دار در متن که تشخیص را دچار اشتباه می‌کند و باعث برداشت نادرست می شود.

چاش اصلی وجود فضای تحقیقاتی زیاد است به علت اینکه متن های وابسته به صورت پراکنده هستند از این رو مدل های طبقه بندی شده به پیچیدگی های زیادی نیاز دارند که با نتایج دقیق مطابقت داشته باشند. [7]

چالش دیگری که در عقیده کاوی وجود دارد نبود مجموعه داده‌های عمومی است. [8] در‌ واقع گام اول برای فرایند عقیده کاوی جمع آوری داده ها است و شامل به دست آوردن مجموعه داده هایی است که می خواهیم برای یافتن عقیده ها کاوش کنیم. دو روش برای انجام این کار وجود دارد.

1- از طریق Application Programming Interface ، مانند تویتر که برای دریافت داده از این نرم‌افزار API هایی طراحی کرده است.

2- استفاده از خزنده ی وب[[1]](#footnote-1) ها برای به دست آوردن داده ها از سایت های مورد نظر

هردو این روش ها فواید و مشکلاتی دارند و بین استفاده از این روش‌ها باید تعادل برقرار کرد.

پیاده‌سازی روش مبتنی بر API آسان است و احتمال اینکه ساختار داده ی جمع آوری شده تغییر کند کم است.اما این محدودیت هایی را بر اساس فراهم کننده ی API ایجاد می کند . مثلا سرعت خواندن داده ها توسط کلاینت در تویتر محدود است. این روش البته نیاز به وجود این API ها در آن وب سایت دارد و تمامی وب سایت ها آن را فراهم نمی کنند و حتی در صورت وجود آن ممکن است همه ی کاربرد های مورد نیاز موجود نباشد.

پیاده سازی روش های خزنده وب سخت تر است چون داده های به دست آمده نویزی بوده و ساختار آن ممکن است تغییر کند اما این روش ها مزیت نامحدود بودن را دارند. [5]

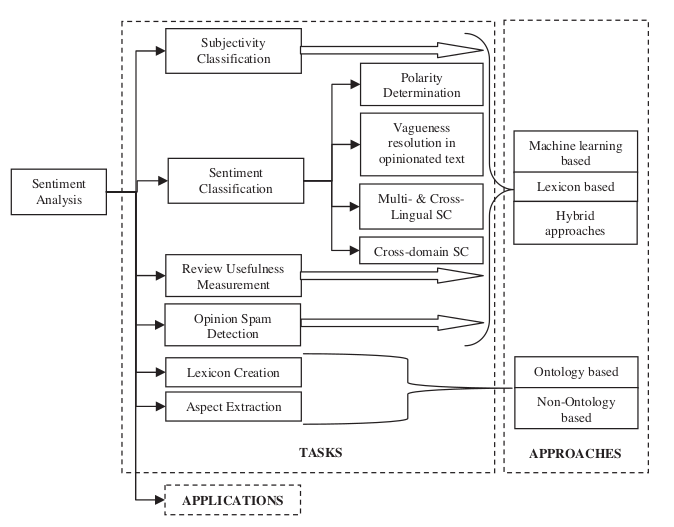
از دیگر چالش های عقیده کاوی می‌توان به تأثیر خطای فرآیند عقیده کاوی(به خصوص در بخش پیش پردازش) اشاره کرد.. [5]

کاربردها

عقیده کاوی نقش مهمی در فرآیند تصمیم گیری های فردی و سازمانی دارد زیرا افراد از خصیصه ها و باور دیگران تأثیر می گیرند. برای مثال امروزه در سایت‌های تجارت الکترونیک، مشتری‌ها به نظراتی که مشتریان دیگر درباره ی کالا داده‌اند اعتماد می‌کنند و تولیدکنندگان نیز عقاید مشتریان را تحلیل می‌کنند تا کیفیت و استاندارد تولیدات و خدمات خود را بهتر کنند. [6]

از دیگر کاربردهای عقیده کاوی می‌توان به عقیده کاوی در شبکه‌های اجتماعی، قیمت گذاری کالا، پیش‌بینی بازار، پیش‌بینی انتخابات، تحلیل ارتباط قومیت ها، تشخیص ریسک در سیستم‌های بانکی و … اشاره کرد. [5]

وظایف و رویکردها ی عقیده کاوی

به طور کلی وظایف عقیده کاوی را می‌توان در شش دسته قرار داد: subjectivity classification ، طبقه بندی احساس، سنجش نظرهای مفید ،opinion spam detection ، تهیه ی لغت نامه، استخراج جنبه های مختلف . طبق شکل زیر برای وظایف عقیده کاوی چندین رویکرد داریم که در ادامه توضیح خواهیم داد.

Subjectivity classification : به معنای تشخیص حالات شخصی یعنی احساسات، عقاید، ارزیابی ها، باورها و گمان هاست.

طبقه بندی احساس: به معنای طبقه بندی احساسات موجود در متن در دو یا چند کلاس است. کلاس‌ها می‌توانند به صورت باینری(مثبت یا منفی)، سه تایی(مثبت، منفی یا خنثی) یا چند تایی(خوش حال، غمگین، عصبانی و ...) باشند و یا به صورت thumb up و thump down در نظر گرفته شوند. طبقه بندی احساس خود به چند زیر وظیفه تقسیم می شود:

تشخیص گرایش: تشخیص اینکه احساس بیان شده در جمله درباره یک موضوع مثبت، منفی و یا خنثی است.

Vagueness resolution in opinionated text : به معنای تشخیص کنایه و طعنه در نوشته‌ها است.

طبقه بندی احساس در نوشته‌های چند زبانه

طبقه بندی احساس در متونی که درباره ی چند حوزه هستند.

سنجش نظرهای مفید : برخی از مدیران بازاریابی، برای اینکه کالا و خدماتشان رقابت بیشتری داشته باشند، به افرادی اجرت می‌دهند تا به عنوان بازدید کننده ی جعلی، عقیده‌ای ساختگی بنویسد تا به این ترتیب بتوانند خدمات خود را بفروشند. بنابراین سنجش بازدیدهای مفید و تشخیص عقیده ی هرزنامه از حوزه های تحقیقاتی به شمار می‌رود که مورد توجه قرار می گیرد.

Opinion spam detection : به معنای تشخیص عقاید جعلی که توسط بازدیدکنندگان اجیر شده ابراز می‌شود می باشد.

همان‌طور که در شکل آمده، برای وظایفی که تا اینجا بحث شد سه رویکرد کلی وجود دارد: مبتنی بر یادگیری ماشین، مبتنی بر لغت‌نامه و رویکرد ترکیبی.

تهیه ی لغت‌نامه: برای ساخت لغت‌نامه ای از احساسات به کار می‌رود و از لیستی از کلمات شروع شده و با کمک کلمات مترادف گسترش می یابد. این فرآیند تا وقتی ادامه دارد که این لیست دیگر نتواند گسترش یابد.

استخراج جنبه های مختلف : افراد درباره ی جنبه‌های مختلفی از یک کالا (یا هر مفهوم دیگر) نظر می دهند. امتیاز احساسات نسبت به جنبه‌های مختلف کالا می‌تواند تأثیر زیادی روی نظر نهایی درباره ی آن داشته باشد. بنابراین باید مهم‌ترین جنبه ی آن استخراج شود.

رویکردهای کلی دو روش آخر می‌تواند مبتنی بر آنتولوژی باشد و یا مبتنی بر آن نباشد. [6]

آنتولوژی لغتی است که از فلسفه آمده و به معنای هستی شناسی است. منظور از آنتولوژی در علم رایانه به دست آوردن مفاهیم موجود در یک حوزه و ارتباط بین آن هاست. نمونه‌ای از آنتولوژی نمودار ارتباط موجودیت در یک پایگاه داده است که در یک حوزه ی خاص، موجودیت ها و ارتباط آن‌ها را نشان می دهد.

راهبرد های مختلف:

دو روش اساسی برای انجام فرایند اصلی عقیده کاوی وجود دارد .

روش lexicon-based :

نام دیگر آن semantic-based است و این روش، تلاش می کند تا بر اساس قوانین وابسته‌ به‌ زبان‌ شناسي‌ گرایش متن را استخراج کند. [5]

این روش بر اساس منابع دانش خارجی عمل می‌کند و برخلاف روش­های supervised شامل مرحله ی آموزش نیست.

گام های این روش به طور کلی به صورت زیر است:

1. استخراج کلمات یا عباراتی که گرایش احساسی متن را بیان می کنند.
2. تعیین گرایش این عبارات استخراج شده.
3. تعیین گرایش متن با اجتماع گرایش تک تک کلمات. [9]

روش supervised learning-based یا machine-learning-based که در آن اطلاعات زیربنایی را از داده های آموزشی یاد می گیرد که به آن امکان کلاس بندی داده های بدون برچسب جدید را می دهد.

این روش معمولا شامل مهندسی feature ها برای نمایش object ی که باید کلاس آن تعیین بشود و سپس استفاده از این نمایش به عنوان ورودی الگوریتم می باشد. برخی از feature هایی که در عقیده کاوی به کار می رود عبارت است از:

تعداد تکرار کلمات، نقش کلمات، کلمات یا عبارات احساسی و...

همچنین تعداد فزاینده ای از مطالعات بر روی ترکیب موفق این دو روش وجود داشته است. [5]

برای تشخیص عقیده، معمولا روش های supervised بهتر از روش های unsupervised عمل می کنند اما کمبود منابع با داده های برچسب دار، تحقیقات را به سمت توسعه ی روش های semisupervised یا unsupervised متمایل کرده است. [10]

# سطوح تحلیل

در زمان توسعه­ی Opinion mining در ابتدا تحلیل متون تنها در سطح داکیومنت انجام می شد و هدف آن پیدا کردن گرایش کلی داکیومنت بود. سپس این فرایند در سطح جمله انجام شد و در نهایت در سطح entity و aspect ها. [5]

سطح داکیومنت: عقیده کاوی در این سطح تحلیل تلاش می کند تا یک داکیومنت را به کلاس های مثبت یا منفی دسته بندی کند. کاربرد این سطح معمولا تحلیل نظرات (مانند بررسی فیلم، نظرات محصول، توییت و بلاگ ها) است.

وظیفه ی این مرحله استخراج گرایش احساس کلی متن است. و فرض می کند که تنها یک target(هدف عقیده) در یک داکیومنت بررسی شده. [10]

سطح جمله: این سطح معادل مورد قبلی است چون می توان یک جمله را یک داکیومنت کوتاه در نظر گرفت.

وظیفه سطح جمله مشابه سطح داکیومت ، مسئله ی کلاس بندی است اما این سطح بر روی هرجمله­ی داکیومت تمرکز می کند و مشخص می کند که یک جمله کدام یک از گرایش های مثبت منفی را بیان می کند.

سطح ویژگی: این سطح ریزدانه ترین سطحی است که در آن عقیده کاوی انجام می شود. در اینجا، وظیفه تنها پیدا کردن گرایش عقیده نیست بلکه یافتن target (موجودیت ، جنبه یا هردو) نیز هست بنابراین 5تایی عقیده کاملا پیدا می شود.

عقیده کاوی در سطح جمله و در سطح داکیومنت هنگامی که متن مورد بررسی تنها شامل یک موجودیت و aspect(ویژگی) باشد به خوبی عمل می کنند ولی وقتی بیشتر از یک مورد وجود داشته باشد دچار اشتباه می شوند. عقیده کاوی در سطح ویژگی تلاش می کند تا این مشکل را با یافتن هر aspect یا ویژگی که در متن ذکر شده و ارتباط دادن آن با یک عقیده حل کند [5]

برای مثال در عقیده کاوی در سطح ویژگی، برای به دست آوردن تاپل عقیده در نظرات مربوط به یک محصول، نام موجودیت به طور صریح مشخص است(برابر نام محصول) و چندین aspect این موجودیت در عقیده ها ذکر شده اند. هدف در این مرحله پیدا کردن target عقیده ها(aspect خاص موجودیت) و گرایش احساس متناظر با آن است.

پیچیدگی مسائل از سطح داکیومنت تا سطح ویژگی افزایش میابد. نبود داده های برچسب دار در سطح های دقیق و جزئی تر آن را پیچیده­تر نیز می کند.

تحلیل هایی که ریز دانه­ترند می توانند تجمیع شده تا سطوح بالاتر را بسازند. برای مثال با عقیده کاوی aspect-based می­توان مقدار متوسط احساسات را در یک جمله برای به دست آورد. [5]

# فرایند عقیده کاوی

# preprocess

معمولا داده ی خام جمع آوری شده از منابع مختلف پیش از انجام تحلیل کامل نیاز به پیش پردازش دارد. برخی گام های پیش پردازش عبارت است از:

* Tokenization(نشانه گذاری)

برای تبدیل کردن جمله­ها به کلمات و عبارات و نماد ها یا دیگر توکن های معنادار استفاده می شود و در زبان انگلیسی با جداکردن لغات به وسیله ی space به آسانی انجام می شود. [10]

* حذف stopwordها

فرایند حذف کلماتی که برای ساختاردهی زبان استفاده می شوند اما در محتوای آن نقشی ندارند. مانند a, are, the, was و will

* Stemming(ریشه یابی)

پروسه ی تبدیل کردن کلمه به حالت ریشه ای آن با حذف پسوند و پیشوندهای کلمه و تبدیل به شکل اصلی یا ریشه ی آن

برای مثال person, person’s, personify ,personification تبدیل بهperson می شوند.

* parts of speech (POS) tagging (برچسب گذاری نقش کلمه)

این مرحله برای تشخیص نقش کلمات در متن انجام می شود و در عقیده کاوی بسیار کمک کننده است چون کلمات بیان کننده ی عقیده ها معمولا صفت بوده و و target عقیده ها(موجودیت ها و aspectها) معمولا اسم یا ترکیبی از اسامی هستند و این برچسب­ها می توانند برای استفاده به عنوان features برای فرایند machine learning استفاده شوند [5]

نقش کلمات مانند:adjective, noun, verb, adverb and preposition

همه ی فرایندهای گفته شده همیشه ضروری نیستند و باید بر اساس کاربرد عقیده کاوی برای مسئله انتخاب شوند

روش های unsupervised ممکن است به نقش کلمه ها در جمله برای ساختن قوانین وابستگی داشته باشند که بعدا در فرایند اصلی عقیده کاوی استفاده می شود در نتیجه stopword ها حذف نمی شوند [11]

الگوریتم ها

روش لغت نامه محور

این الگوریتم یک روش لغت نامه محور است و در سطح جمله،گرایش عقیده را تشخیص می دهد. این روش با داشتن موضوع( به عنوان ورودی) و مجموعه ای از متن ها راجع به موضوع، احساسات اظهار شده در مورد موضوع در هر متن را به دست می آورد و افرادی که هر احساس را بیان کرده اند را تشخیص می دهد. احساسات می تواند مثبت یا منفی باشند. در این روش از از کلمات به سمت جمله حرکت می کنیم. هر کلمه را به عنوان واحد نشان دهنده­ی احساسات در نظر گرفته و در ابتدا کلاس هر کلمه را بر اساس احساسات آن دسته بندی می کنیم.

هر جمله ممکن است حتی بیانگر نظرات چند نفر باشد به همین دلیل هنگام ترکیب احساسات در سطح کلمات، ابتدا برای نظرهنده های مختلف بخش مرتبط با او را در جمله مشخص می کنیم و سپس از مدل های مختلف برای ترکیب احساسات کلمات استفاده می کنیم.

با داشتن موضوع و مجموعه ای متن ها سیستم در چهارگام این عملیات را انجام می دهد. اول جمله هایی که شامل عبارت موضوع و کاندیدا های افرادی که نظر را داده اند را انتخاب می کند. سپس مرز ناحیه ی مبتنی بر نظر دهنده را مشخص می کند. پس از آن طبقه بند احساسات کلمات گرایش تک تک کلمات دارای احساسات را محاسبه می کند و در آخر سیستم آن ها را برای تولید احساسات نظر دهنده برای کل جمله را ترکیب می کند.

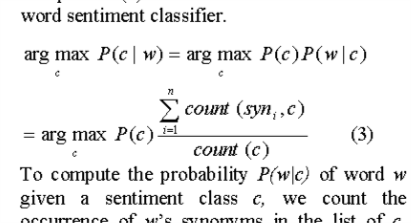
word sentiment classification

در روش اصلی برای به دست آوردن احساسات کلمات ، مجموعه ی کوچکی از کلمات را به صورت دستی جمع آوری می کنیم و بر اساس گرایش به دو دسته تقسیم می کنیم.(مثبت ومنفی) سپس این مجموعه را با اضافه کردن کلمات مترادف و متضاد آن ها از WordNet گسترش می دهیم با فرض اینکه اکثر مترادف های یک کلمه مثبت هم مثبت بوده و متضاد های آن منفی هستند . (در صورتی که کلمه ای بیشتر از یکبار تکرار شد، تعداد آن نیز ذخیره می شود)

WordNet یک مرجع آنلاین لغات است که در آن اسم ها و فعل ها و صفت های انگلیسی در گروه های مترادف قرار دارند. [5]

برخی کلمات مانند great, strong , take چندین با در دسته های مثبت و منفی دید شدند که نشان می دهد نیاز به معیار قدرت احساسات گرایش هر کلمه داریم. این روش به ما امکان این را می دهد که کلماتی که از لحاظ احساسی مبهم هستند را دور انداخته و تنها کلماتی که قدرت آن ها از مقدار آستانه بیشتر است را نگه داریم.

با این روش می توانیم قدرت احساسی کلماتی که تا کنون مشاهده نشده اند را نیز تعیین کنیم و به این صورت عمل می کنیم که با استفاده از WordNet مجموعه کلمات مترادف آن کلمه جدید را تعیین می کنیم و به صورت زیر عمل می کنیم:



که در آن C دسته بندی احساسات است(مثبت و منفی) و w کلمه ی جدید است. synn کلمات مترادف w استخراج شده از WordNet اند.

در این روش برای محاسبه ی قدرت کلمه ی w با داشتن کلاس c (مقدارP(w|c)) ، تعداد رخ داد مترادف های w را در لیست c می شماریم ( ایده اصلی این است که هرچه بیشتر مترادف ها در این کلاس باشند، با احتمال بیشتری این کلمه عضو این کلاس است). برای هرکلمه قدرت مثبت و منفی را محاسبه می کنیم و قدر مطلق آن ها را مقایسه می کنیم.

مثال:

abysmal : NEGATIVE

[+ : 0.3811][- : 0.6188]

adequate : POSITIVE

[+ : 0.9999][- : 0.0484e-11]

afraid : NEGATIVE

[+ : 0.0212e-04][- : 0.9999]

ailing : NEGATIVE

[+ : 0.0467e-8][- : 0.9999]

amusing : POSITIVE

[+ : 0.9999][- : 0.0593e-07]

answerable : POSITIVE

[+ : 0.8655][- : 0.1344]

apprehensible: POSITIVE

[+ : 0.9999][- : 0.0227e-07]

averse : NEGATIVE

[+ : 0.0454e-05][- : 0.9999]

blame : NEGATIVE

[+ : 0.2530][- : 0.7469]

Sentence Sentiment Classifier

ما در جستجوی احساساتی هستیم که نظر دهنده ابراز کرده است. تحلیل ها نشان داده اند که این احساسات به طور قابل اطمینان تری در جمله نزدیک به نظردهنده به دست می آیند. به همین دلیل در این مراحل ، گام هایی برای یافتن موضوع با جستجوی خود کلمه در متن و هر نظردهنده ی ممکن ،اضافه شده است. سپس نزدیک هر نظردهنده، منطقه ای که احساسات آن باید در نظر گرفته شود را مشخص کنیم.

تشخیص نظردهنده

در اینجا از ابزاری برای مشخص کردن Named entity ها برای تشخیص نظردهندگان یک عقیده استفاده شده است و فقط فرد یا سازمان را به عنوان نظردهنده های ممکن در نظر گرفته است. در جمله هایی با بیشتر از یک نظر دهنده، موردی که به عبارت موضوع نزدیکتر بود را انتخاب می کند.

Named entity یک مفهوم در دنیای واقعی است مانند افزاد مکان ها سازمان ها محصولات و..که می تواند با یک نام خاص نشان داده شود وNamed-entity recognition در یک متن بدون ساختارمحل Named entity ها را یافته و آن ها را در کلاس های از قبل مشخص شده، کلاس بندی می کند .

Sentiment Region

به صورت های مختلفی می توان ناحیه ای که احساسات مورد نظر در آن بیان شده را مشخص کرد. برای مثال پنجره های زیر را می توان درنظر گرفت.

پنجره 1: کل جمله

پنجره 2: لغات بین نظردهنده وعبارت موضوع

پنجره 3: همان پنجره2 ± دوکلمه

پنجره 4: همان پنجره 2 تا پایان جمله

مدل های کلاس بندی:

برای انتساب یک کلاس احساسی به جمله ی داده شده می توان از راه های زیر استفاده کرد:

مدل 1:

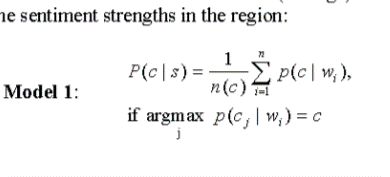
Π (signs in region)

ایده ی این مدل براساس این است که منفی در منفی برابر مثبت است. در این مدل کلماتی مانند not و neverمی تواند گرایش احساسی را وارونه کند. این روش به این دو جمله یک احساس را نسبت می دهد:

"The California Supreme Court ***agreed*** that the state’s new term-limit law was ***constitutional***”

“the California Supreme Court ***disagreed*** that the state’s new term-limit law was ***unconstitutional***”

مدل دوم میانگین قدرت احساسات در جمله است



در این مدل n( c) تعداد کلمات در ناحیه است که دسته بندی احساسی آن برابر c است.در صورتی که جمله ای دارای کلمات مثبت قوی­تری باشد ، جمله مثبت خواهد بود. [12]

مثال:

For that reason and others, the Constitutional

Convention unanimously rejected term limits

and the ***First Congress*** soundly defeated two

subsequent ***term-limit*** proposals.

TOPIC : term limit

HOLDER : First Congress

OPINION REGION: soundly, defeated ,two,subsequent, term-limit,proposals.

SENTIMENT\_POLARITY: negative

Naive Bayes Classifier

Naive Bayes یکی از ساده ترین و معمول ترین classifierها برای عقیده کاوی است و روشی supervised بوده که در سطح داکیومنت گرایش را تشخیص می دهد. این مدل احتمال پسین هر کلاس(مثبت یا منفی بودن عقیده) را بر اساس پراکندگی کلمات در داکیومنت به دست می آورد و براساس روش استخراج ویژگی bag of words کار می کند.

مدل bag of words روشی برای نمایش است که در پردازش زبان طبیعی استفاده می شود که در آن متن به صورت مجموعه ای از کلمات بدون ترتیب نمایش داده می شود و قوانین گرامری و ترتیب کلمات را نادیده می گیرد.

کلاس داکیومت برابر برچسب مربوط به بیشترین احتمال زیر است.

P(label|document)=

با فرض مستقل بودن رخ دادن کلمات(ویژگی ها) از یکدیگر، عبارت بالا به صورت زیر تبدیل می شود:

P(label|document)=

P(document) تاثیری در انتخاب label ندارد پس می توان آن را حذف کرد. [13]

مرحله ی آموزش محاسبه ی P(label) :

P(label)=

P(fi|label)=

در محاسبه ی P(fi|label) برای کلماتی که تا کنون مشاهده نشده اند مقدار 0 به دست می آید که باعث می شود محاسبات با مشکل مواجه شود پس برای رفع این مشکل از فرمول زیر برای محاسبه ی این احتمال استفاده می کنیم. با این راه (با نام add-one smoothing ) به کلماتی که تا کنون مشاهده نشده اند احتمالی اختصاص داده می شود و جمع احتمال همه ی کلمات در کلمات یک برچسب همچنان یک باقی خواهد ماند. [14]

روش محاسبه ی P(fi|label) با استفاده از به شکل زیر خواهد بود:

P(fi|label)=

مثال

برای آموزش سیستم داکیومنت هایی مانند نظرات راجع به محصولات (در این مثال بازی کامپیوتری)که دارای امتیاز نیز هستند انتخاب می شود و امتیاز های بالا(4و5) به برچسب مثبت و امتیاز های پایین(1و2) به برچسب منفی تبدیل شد نظرها با امتیاز 3 به عنوان نمونه های خنثی حذف شد.

برای آماده سازی document ها علامت NOT\_ به کلماتی که بین عبارت منفی کننده مانند not, isn’t , didn’t و اولین علامت نگارشی پس از عبارت منفی کننده، اضافه شد سپس tokenization انجام شد.

مثال اعمال منفی کننده در جمله:

This place isn’t OK.

Features={ This ,place, NOT\_OK)

برای حفظ سادگی این راه و بررسی تاثیر مستقیم Naïve Bayes ، از موارد پیش پردازش دیگر مانند بردن به حالت ریشه ای و حذف برخی کلمات خودداری شده است.

برای مرحله ی آموزش، جملات زیر را بررسی می کنیم:

1. This is a fun game 4
2. This game is wonderful 5
3. started crashing for no reason. 2
4. I have to restart everytime before new game. 1

برای هر جمله، در ابتدا امتیاز هر کدام تبدیل به برچسب مثبت و منفی شد وبه جملات 1و2 برچسب مثبت وبه جملات 3و4 برچسب منفی داده شد.

بعد از انجام tokenization و اعمال منفی کننده ها( مانند NO\_reason ) به کلمات جملات مثبت برچسب مثبت زده و به کلمات جملات منفی برچسب منفی می زنیم. حالا با استفاده از اطلاعات به دست آمده جدول زیر را ایجاد می کنیم که در هر سطر تعداد دفعات تکرار برچسب آن ستون برای آن کلمه ذخیره شده است.

Table 1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | This | Is | A | Fun | game | wonderful | Started | crashing | for | NO\_reason | I | Have | to | restart | everyTime | Before | new |  |
| positive | 2 | 2 | 1 | 1 | 2 | 1 |  |  |  |  |  |  |  |  |  |  |  | 9 |
| negative |  |  |  |  | 1 |  | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 12 |

برای مشاهده ی روند کلی این فرایند گرایش جمله ی زیر را بررسی می کنیم

Document= Everytime I have started the game it is crashing.

بر اساس تعداد بار تکرار کلمات در جدول شماره 1 ، برای جمله تست ، جدول زیر را تشکیل می دهیم:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Everytime | I | have | started | the | game | it | is | crashing |
| positive | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 |
| negative | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |

مقدار عبارت P(label|document) را برای برچسب های مثبت و منفی محاسبه می کنیم.

P(pos)= =

P(Everytime |document)= =

P(game|document)= =

P(positive|document)= = =.0.004

P(neg)= =

P(the|neg)= =

P(Everytime | neg)= =

P(neg|document)= ==0.038

به دلیل اینکه مقدار عبارت برای برچسب منفی بیشتر است، برچسب جمله ی تست بر اساس کلاسیفایر naïve bayes منفی است.

References

|  |  |
| --- | --- |
| [1] | م. کیوانپور, ف. حسن زاده and م. مرادی, مباحث پیشرفته در داده کاوی, نشر دانشگاهی کیان, 1397. |
| [2] | S. A. Salloum, A. Q. AlHamad, M. Al-Emran and K. Shaalan, "A Survey of Arabic Text Mining," in *Intelligent Natural Language Processing: Trends and Applications*, Springer, 2018, pp. 417-431. |
| [3] | R. Patibandla and N. Veeranjaneyulu , "Survey on Clustering Algorithms for Unstructured Data," in *Intelligent Engineering Informatics*, Singapore, Springer, 2018. |
| [4] | N. Chandra, S. K. Khatri and S. Som, "Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository," in *System Performance and Management Analytics*, Springer, 2018, pp. 65-76. |
| [5] | "J.A. Balazs , J.D. Velásquez , Opinion mining and information fusion: a survey, Inf. Fusion 27 (2016) 95–110 .". |
| [6] | K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems,* vol. 89, pp. 14-46, 2015. |
| [7] | م. کیوانپور and س. رحمانی, "دسته بندی و ارزیابی روش های ایده کاوی," in *سومین همایش ملی مهندسی کامپیوتر و فناوری اطلاعات*, همدان, 1389. |
| [8] | M. Tubishat, N. Idris and M. A. Abushariah, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges," *Information Processing & Management,* vol. 54, no. 4, pp. 545-563, 2018. |
| [9] | "Ontology-Supported Polarity Mining". |
| [10] | S. Sun, C. Luo and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion 36 (2017) 10–25.* |
| [11] | " a review of nlp tecch for om systems". |
| [12] | "S.-M. Kim , E. Hovy , Determining the sentiment of opinions, in: Proceed- ings of the 20th". |
| [13] | "W. Medhat et al., Sentiment analysis algorithms and applications: a survey,". |
| [14] | "B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, 2002, p". |
| [15] | "Miller, G.A., R. Beckwith, C. Fellbaum, D.Gross, and K. Miller. 1993. Introduction to WordNet: An On-Line Lexical Database. http://www.cosgi.princeton.edu/~wn.". |

1. web crawler [↑](#footnote-ref-1)