به نام خدا

مقدمه

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.(کیوانپور, حسن زاده, و مرادی ۱۳۹۷)

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 بدون ساختار یا نیمه ساخت یافته مثل ایمیل، اسناد متنی، فایل‌های html و ... اعمال شود.(Salloum و دیگران ۲۰۱۸)

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 csv ، html ، xml 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(Patibandla و Veeranjaneyulu 2018)

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.(Chandra, Khatri, و Som 2019)

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.(کیوانپور و دیگران ۱۳۹۷)

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

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

منظور از e_i موجودیت i ام است ، a_{ji} خصیصه ی j ام e_i ، h_k امین فردی است که این نظر را دارد، t_l زمان انتشار این عقیده و s_{kji} نظر یا احساس نسبت به j امین خصیصه ی موجودیت i ام (e_i) که فرد h_k در زمان t_l 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 موجودیت mobile phone و خصیصه ی آن screen است و نظر یا احساس بیان شده مثبت است.(Sun, Luo, و Chen 2017)

عقیده کاوی و تحلیل احساس

برخی از محققان معتقدند که نام دیگر عقیده کاوی، تحلیل احساسات است و این دو را معادل یکدیگر می دانند. در مقابل برخی دیگر می گویند که عقیده کاوی به دنبال نظرات افراد در مورد موضوعی خاص است و تحلیل احساس، جهت گیری احساسات موجود در عبارات را بررسی می کند. اما به طور کلی می‌توان گفت که عقیده و احساسات معادل یکدیگرند و عقیده کاوی و تحلیل احساسات اشاره به یک حوزه دارند.(Chiranjeevi, Santosh, و Vishnuvardhan 2019)

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.(Ravi و Ravi 2015)

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.(کیوانپور و رحمانی ۱۳۸۹)

چالش دیگری که در عقیده کاوی وجود دارد نبود مجموعه داده‌های عمومی است.(Tubishat, Idris, و Abushariah 2018) 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.

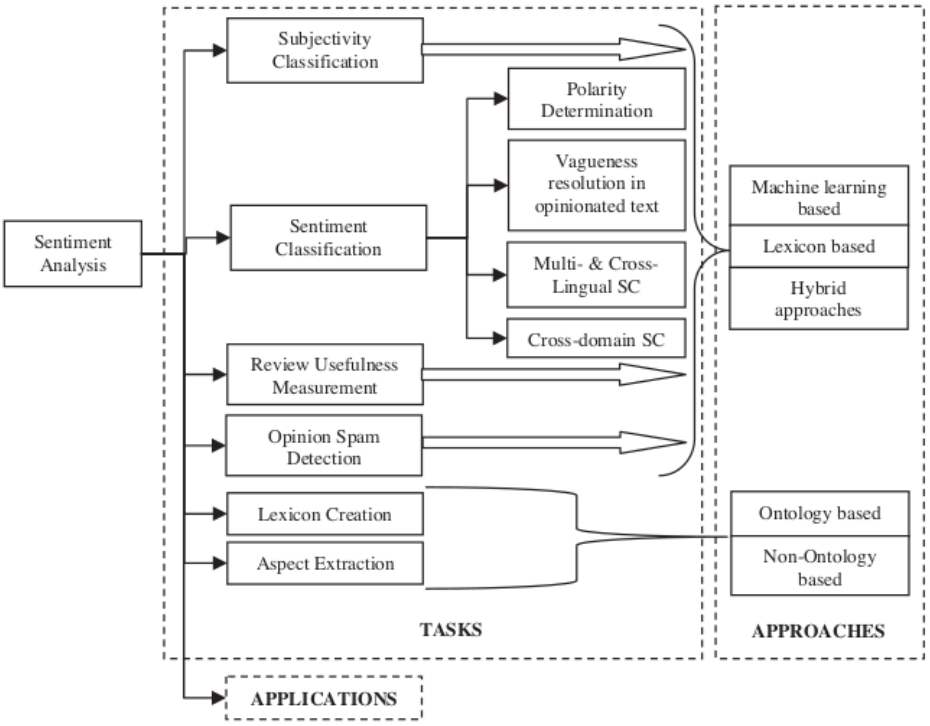
۱- از طریق Application Programming Interface ، مانند تویتر که برای دریافت داده از این نرم‌افزار API هایی طراحی کرده است.
۲- استفاده از خزنده ی وب[1] ها برای به دست آوردن داده ها از سایت های مورد نظر
هردو این روش ها فواید و مشکلاتی دارند و بین استفاده از این روش‌ها باید تعادل برقرار کرد.
پیاده‌سازی روش 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 cخزنده 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.(Balazs و Velásquez 2016)

از دیگر چالش های عقیده کاوی می‌توان به تأثیر خطای فرآیند عقیده کاوی(به خصوص در بخش پیش پردازش) اشاره کرد. همچنین گرچه روش‌های یادگیری عمیق به برخی وظایف عقیده کاوی اعمال شده‌اند اما هنوز مسائل حل نشده ی زیادی در این حوزه وجود دارد که می‌توان با مطالعه بیشتر، کاربردهای گسترده تری از یادگیری عمیق را در عقیده کاوی پیدا کرد.(Sun و دیگران ۲۰۱۷)

کاربردها

1web crawler

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.
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.
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 احوال خصوصی 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.

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.(Ravi و Ravi 2015)

Balazs, Jorge A. و Juan D. Velásquez. 2016. “Opinion Mining and Information Fusion: A survey.” *Information Fusion* 27:95–110.

Chandra, Nidhi, Sunil Kumar Khatri, و Subhranil Som. 2019. *Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository*. Springer Singapore.

Chiranjeevi, P., D. Teja Santosh, و B. Vishnuvardhan. 2019. *Survey on Sentiment Analysis Methods for Reputation Evaluation*. ج 768. Springer Singapore.

Patibandla, R. S. M. Lakshmi و N. Veeranjaneyulu. 2018. “Survey on Clustering Algorithms for Unstructured Data.” در ج 695 صص 421–29. Springer Singapore.

Ravi, Kumar و Vadlamani Ravi. 2015. “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications.” *Knowledge-Based Systems* 89:14–46.

Salloum, Said A., Ahmad Qasim AlHamad, Mostafa Al-Emran, و Khaled Shaalan. 2018. “A survey of Arabic text mining.” *Studies in Computational Intelligence* 740:417–31.

Sun, Shiliang, Chen Luo, و Junyu Chen. 2017. “A review of natural language processing techniques for opinion mining systems.” *Information Fusion* 36:10–25.

Tubishat, Mohammad, Norisma Idris, و Mohammad A. M. Abushariah. 2018. “Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges.” *Information Processing and Management* 54(4):545–63.

کیوانپور, محمدرضا, فرانک حسن زاده, و محمد مرادی. 1397. مباحث پیشرفته در داده کاوی. چاپ دوم. نشر دانشگاهی کیان.

کیوانپور, محمدرضا و سمیه رحمانی. 1389. "دسته بندی و ارزیابی روشهای ایده کاوی." در سومین همایش ملی مهندسی کامپیوتر و فناوری اطلاعات، همدان.