به نام خدا

مقدمه

تعریف متن کاوی

به فرآیند استخراج داده و دانش جذاب و با اهمیت از متون، متن کاوی (کاوش داده‌های متنی یا کشف دانش در متن) گویند.

متن کاوی در مقابل داده کاوی

کاوش متن تا حدی مشابه داده کاوی است و تفاوت در این است که در داده کاوی (سنتی) 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.

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.

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 بدون ساختار 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.

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 k امین فردی است که این نظر را دارد، t_l زمان انتشار این عقیده و s_{hkji} 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 موبایل (mobile phone) و خصیصه ی آن صفحه ی نمایش (screen) است و نظر یا احساس بیان شده مثبت است.

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.

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.

چالش ها

چالش های عقیده کاوی به طور کلی به سه دسته ی زبانی، محتوایی و پیچیدگی تقسیم می‌شود.

چالش های زبانی که خود به دو دسته ی چند زبانی و استاندارد سازی تقسیم می شود. به این معنا که محتویات موجود در وب می‌تواند به زبان‌های مختلفی نوشته شود. همچنین با استاندارد سازی زبان‌های مربوط به ایده کاوی می‌توان اطلاعات موجود در متن های وبی را همگون‌سازی کرد.

چالش های محتوایی خود به چهار دسته تقسیم می شوند:

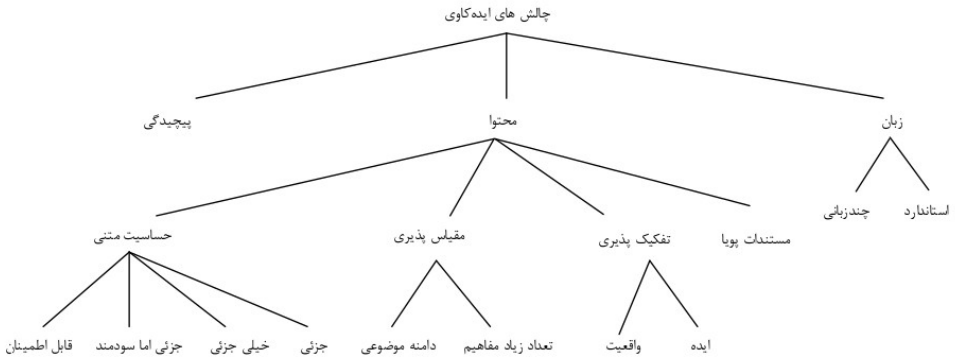
استخراج دانش از وب به دلیل پویا بودن و بروز شدن مطالب با بیشترین چالش همراه است و اطلاعات جدید در هر لحظه به اطلاعات قبلی اضافه می‌شوند و داده‌ها به سرعت تغییر می کنند.

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. عامل دیگری که موجب این مسأله می‌شود وجود مفاهیم زیاد در محتویات متن هاست. کلماتی مانند: عظیم، شگفت انگیز، فقیر،بد و... وجود دارند که به موقعیت ایده اشاره می‌کنند و نحوه ی استفاده از این کلمات در موقعیت های مختلف متفاوت است.

چالش دیگر حساسیت‌های متنی است که شامل حساسیت جزئی (ممکن است مثبت یا منفی باشد)، حساسیت خیلی جزئی (منفی)، جزئی اما سودمند(کلا مثبت است) و قابل اطمینان اما پرهزینه (منفی) است.

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. شکل زیر دسته بندی چالش ها را نشان می دهد.



از دیگر چالش های عقیده کاوی می‌توان به عدم وجود مجموعه داده ی عمومی، تأثیر خطای فرآیند عقیده کاوی(به خصوص در بخش پیش پردازش) و تشخیص عقیده ی هرزنامه 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.

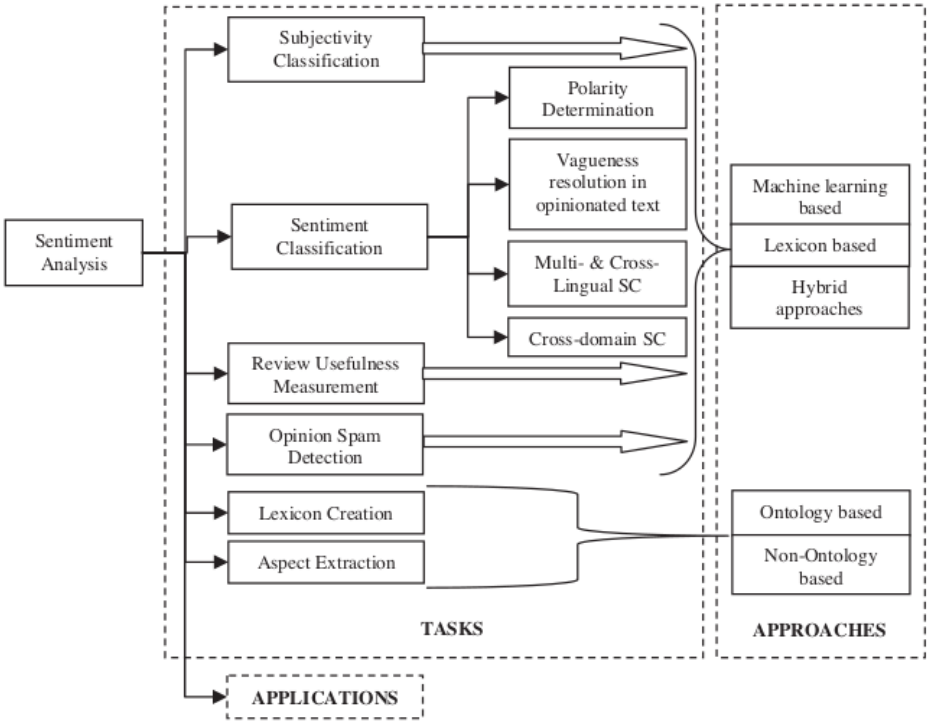
کاربردها

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.

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: طبقه بندی ذهنیت ، طبقه بندی احساس، سنجش بازدیدهای مفید ، تشخیص عقیده ی هرزنامه ، تهیه ی لغت نامه، استخراج نمودهای مختلف . طبق شکل زیر برای وظایف عقیده کاوی چندین رویکرد داریم که در ادامه توضیح خواهیم داد.



طبقه بندی ذهنیت : 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.

تفکیک پذیری ابهام در متون متعصب : 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 بازدیدهای 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.

تشخیص عقیده ی هرزنامه : 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 نمودهای 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.

1. Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A survey of Arabic text mining. *Studies in Computational Intelligence*, *740*, 417–431. doi:10.1007/978-3-319-67056-0_20

2. Chandra, N., Khatri, S. K., & Som, S. (2019). *Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository*. Springer Singapore. doi:10.1007/978-981-10-7323-6

4. Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, *36*, 10–25. doi:10.1016/j.inffus.2016.10.004

5. Balazs, J. A., & Velásquez, J. D. (2016). Opinion Mining and Information Fusion: A survey. *Information Fusion*, *27*, 95–110. doi:10.1016/j.inffus.2015.06.002

6. Tubishat, M., Idris, N., & Abushariah, M. A. M. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing and Management*, *54*(4), 545–563. doi:10.1016/j.ipm.2018.03.008

7. Chiranjeevi, P., Santosh, D. T., & Vishnuvardhan, B. (2019). *Survey on Sentiment Analysis Methods for Reputation Evaluation* (Vol. 768). Springer Singapore. doi:10.1007/978-981-13-0617-4

8. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. doi:10.1016/j.knosys.2015.06.015

کیوانپور, م., حسن زاده, ف., & مرادی, م. (n.d). مباحث پیشرفته در داده کاوی (چاپ دوم.). نشر دانشگاهی کیان.

رحمانی, سمیه و محمدرضا کیوان پور، ۱۳۸۹، دسته بندی و ارزیابی روشهای ایده کاوی، *سومین همایش ملی مهندسی کامپیوتر و فناوری اطلاعات*، همدان، https://www.civilica.com/Paper-CEIC03-CEIC03_168.html