فصلنامه پژوهش در علوم، مهندسی و فناوری
دوره ۷، شماره ۱، بهار ۱۴۰۰
صفحات ۱۳۶-۱۴۲
ISSN: 2476-5082
www.uctjournals.com

تشخیص قطبیت و تحلیل احساسات در زبان های فارسی و انگلیسی با استفاده از روش ال اس تی ام

سمیه رجبی
دانشجوی ارشد مهندسی نرم افزار از دانشگاه آزاد بندرعباس

چکیده

اطلاعات متنی در دنیا به دو نوع اصلی حقایق[1] و احساسات[2] تقسیم‌بندی می‌شوند [۱، ۲، ۳]. حقایق عبارات واقعی راجع به موجودیت‌ها، رخدادها و خصوصیات آن‌ها هستند، درحالی‌که احساسات عبارات ذهنی هستند که نظرات احساسی افراد، عقاید و اندیشه‌های آن‌ها را راجع به یک موجودیت، رخداد یا خصوصیتی از آن‌ها ارائه می‌کنند. این پژوهش، ازنظر فنی، پر چالش و درعین‌حال در عمل بسیار پرکاربرد می‌باشد. یافتن اطلاعات احساسی در وب، دسته‌بندی کردن و نظم دادن به آن‌ها کار آسانی نیست، چراکه ممکن است منابع و صفحات زیادی از نظرات با حجم بالا در اینترنت موجود باشد و رسیدن به نظرات دلخواه در این صفحات طولانی، یافتن جملات موردنظر و خلاصه‌سازی و جمع‌بندی آن‌ها کار هر خواننده‌ای نباشد [۵،۶]. بنابراین ابزار خودکاری برای کشف و خلاصه کردن نظرات لازم است. تجزیه‌وتحلیل احساسات یا کاوش نظرات[3] مفهومی در علم داده‌کاوی است که با توجه به این نیاز پدید آمده است. در این مقاله ما به صورت کلی جهت پیش بینی آینده راهکاری را با استفاده از LSTM که یک کلاس از شبکه های عصبی است، پیشنهاد می دهیم و سپس نتایج به دست آمده را با روش های دیگر را مورد ارزیابی قرار می دهیم. به صورت جزئی قصد داریم تا تحلیل احساسات در سطح جمله را در مورد نظرات مشتریان هتل انجام دهیم .

کلیدواژه:

[1] Facts
[2] Opinions
[3] Opinion Mining

مقدمه

تا به امروز پژوهش‌های زیادی بر روی اطلاعات از نوع حقایق صورت گرفته است، به‌عنوان نمونه می‌توان بازیابی[۴] و خلاصه‌سازی متون[۵] طبقه‌بندی[۶] و خوشه‌بندی متون[۷] و کاربردهای زیاد دیگری را در علوم متن‌کاوی[۸] و پردازش زبان طبیعی[۹] نام برد. آنالیز احساسات[۱۰] یا کاویدن نظرات[۱۱] به فرآیند محاسباتی و الگو ریتمیک مطالعه‌ی نظرات، احساسات و برخوردها در رابطه با موجودیت‌ها، افراد، اتفاقات، مشکلات، موضوعات و تمامی زیرمجموعه‌های آن‌ها می‌باشد[۴].

محققان عمدتاً تجزیه و تحلیل احساسات را در سه سطح بندی مورد مطالعه قرار داده اند: سطح سند[۱۲]، سطح جمله[۱۳] و سطح وجهی یا بعد[۱۴] طبقه بندی احساسات سطح اسناد ، یک دسته نظر را طبقه بندی می کند (به عنوان مثال ، بررسی محصول) به عنوان بیان یک نظر کلی مثبت یا منفی. این کل سند را به عنوان واحد اطلاعاتی اساسی در نظر می گیرد. فرض می کنیم که این اسناد شناخته شده است که باید نظر داده شود و حاوی نظرات در مورد یک موجود واحد باشد (مثلاً خاص تلفن). طبقه بندی احساسات سطح جمله ها جملات فردی را در یک سند طبقه بندی می کند. با این حال ، هر جمله نمی توان به عنوان یک نظر تلقی کرد ، اغلب اوقات یک جمله را به صورت عقیده و یا بدون نظر طبقه بندی می کند ،که به آن طبقه بندی ذهنیت[۱۵] گفته می شود. سپس جملات بیان شده ای که حاوی نظر هستند را به عنوان نظرات مثبت یا منفی طبقه بندی می شوند. طبقه بندی احساسات در سطح جمله می تواند به عنوان طبقه بندی سه طبقه تدوین شود ، یعنی طبقه بندی یک جمله به عنوان خنثی ، مثبت یا منفی. در مقایسه طبقه بندی سطح سند و جمله در تحلیل احساسات، سطح تحلیل احساسات ابعاد دقیق تر هستند. وظیفه آن استخراج و جمع بندی نظرات افراد بیان شده در مورد موجودیت ها و جنبه ها و ویژگی های موجودیت ها است که به آنها اهداف نیز گفته می شود. به عنوان مثال ، در یک بررسی محصول ، هدف آن است که نظرات مثبت و منفی را در مورد جنبه های مختلف از محصول خلاصه کنیم ، اگرچه احساسات عمومی در مورد محصول می تواند مثبت یا منفی باشد [۷].

با توجه به تعداد بسیار بالای سایت ها و محصولات و خدمات و در نتیجه حجم بسیار بالای نظرات کاربران سایت‌ها وشبکه های مجازی ، امکان تحلیل نظرات و ثبت بازخورد ها توسط نیروی انسانی غیر ممکن است . از این رو برای تحلیل احساسات نیاز به روشهای اتوماتیک احساس می شود. روش های اتوماتیک تحلیل احساسات شامل سه روش یادگیری ماشین، مبتنی بر واژگان و روش ترکیبی می باشد[۸].

تحلیل احساسات در روش مبتنی بریادگیری ماشین ، شامل سه رویکرد یادگیری بدون نظارت[۱۶]، یادگیری نظارت شده[۱۷] و یادگیری نیمه نظارت شده[۱۸] می شود .در جدول زیر تکنیک های اتوماتیک تحلیل احساسات و رویکرد های آن ها بصورت خلاصه نمایش داده شده است. [۹]. از ویژگی هایی مانند اصطلاحات و فرکانس مربوط به آنها، بخشی از گفتار، کلمات و عبارات نظر، نفی و وابستگی نحوی در تکنیک های طبقه بندی احساسات استفاده می شود [۱۰].

به طور کلی روش های تحلیل احساسات، به شکل زیر دسته بندی می شوند [۸] :

---

[4] Text retrieval
[5] Text summarization
[6] Text classification
[7] Text clustering
[8] Text Mining
[9] Natural Language Processing
[10] Sentiment Analysis
[11] Opinion Mining
[12] Document Level
[13] Sentence Level
[14] Aspect Level
[15] Subjective Classification
[16] Unsupervised Learning
[17] Supervised Learning
[18] Semi-Supervised Learning

۱- رویکرد یادگیری ماشین
۲-رویکرد واژگانی[۱۹]
۳- رویکرد نوین

همانطور که قبلا در توضیحات ذکر شد، رویکرد یادگیری ماشین، شامل سه روش یادگیری نظارت شده ، یادگیری بدون ناظر و یادگیری نیمه نظارت شده است. از روش های یادگیری بدون نظارت می توان به الگوریتم‌های نزدیک ترین همسایه k[۲۰] و مدلسازی موضوع[۲۱] اشاره کرد. الگوریتم های مرسوم در یادگیری نظارت شده را نیز میتوان در سه دسته تقسیم کرد. شامل:

a-۱) طبقه بندهای غیرخطی[۲۲] مانند الگوریتم درخت تصمیم[۲۳]،ماشین بردار پشتیبان غیر خطی[۲۴]،شبکه عصبی غیر خطی[۲۵] و ...

b- ۱) طبقه بند های خطی[۲۶] شامل الگوریتم هایی مانند ماشین بردار پشتیبان خطی و شبکه عصبی خطی

c- ۱) طبقه بندهای مبتنی بر قانون[۲۷]

d- ۱) طبقه بندهای احتمالاتی[۲۸]

رویکرد واژگانی،خود شامل سه رویکرد مبتنی بر فرهنگ لغت[۲۹]،مبتنی بر کلکسیون کلمات[۳۰] ومبتنی بر روشدستی[۳۱] است که توسط انسان انجام می گیرد .

در رویکرد نوین، روش های ترکیبی وجود دارد که بر اساس ترکیب برخی از روش های بالا ساخته می شوند.

اما مهم ترین روش در رویکردهای نوین، روش یادگیری عمیق است .

با توجه به موفقیت یادگیری عمیق[۳۲] در بسیاری از حوزه های کاربردی، درسال های اخیر ازیادگیری عمیق در حوزه تحلیل احساس نیز استفاده شده است. یادگیری عمیق به عنوان یک تکنیک قدرتمند یادگیری ماشینظهور کرده است که چندین لایه بازنمایی یا ویژگی داده ها را فرا می گیرد و بهترین پیش بینی را تولید می کند [۷].در مقایسه با روش های پایه یادگیری ماشین مانند ماشین بردار پشتیبان و شبکه های عصبی نرمال، روش یادگیری عمیق به دلیل استفاده از لایه های مخفی بیشتر در مقایسه با شبکه عصبی نرمال که از یک یا دو لایه استفاده میکند، عملکرد بهتری را داشته است[۱۱].

از الگوریتم‌هایی که در یادگیری عمیق استفاده می شود، می‌توان به الگوریتم های شبکه عصبی کانولوشنی[۳۳]،الگوریتم حافظه بلند-کوتاه مدت[۳۴]،الگوریتم شبکه عصبی مکرر[۳۵]، الگوریتم شبکه عصبی سلسله مراتبی[۳۶] و الگوریتم حافظه بلند-کوتاه مدت اشاره کرد. [۸].

---

19 Lexicon-Based Approach
20 K-Nearest Neighbor
21 Topic Modeling
22 Non-Linear Classifiers
23 Decision tree
24 Non-Linear SVM
25 Non-Linear Neural Network
26 Linear Classifier
27 Rule-Based
28 Probabilistic
29 Dictionary-Based
30 Corpus-Based
31 Manual approach
32 Deep Learning
33 Convolutional Neural Network(CNN)
34 Long Short-Term Memory (LSTM)
35 Recurrent Neural Network(RNN)
36 Hierarchical Neural Network(HNN)
37 Bidirectional Long Short-Term Memory (Bi-STM)

در سال های اخیر ، مدل های یادگیری عمیق در زمینه پردازش زبان طبیعی به طور گسترده ای مورد استفاده قرار گرفته و پتانسیل های بسیار خوبی را نشان می دهد. [۷].

در این تحقیق قصد داریم با ترکیب دو روش Bi-LSTM و Convolution Neural Network ، تشخیصقطبیت متن را در مورد یک دیتاست به زبان فارسی که شامل نظرات مشتریان هتل درباره سطح کیفی خدمات می باشد، با روش تحلیل احساسات در سطح جمله انجام داده و میزان دقت نتایج پیش بینی بدست آمده را محاسبه کرده و با مقایسه نتایج حاصل با الگوریتم های پایه یادگیری ماشین میزان بهبود نتایج بدست آمدهمقایسه نماییم.همچنین عملیات فوق بر روی یک دیتاست مشابه ، اما به زبان انگلیسی هم انجام خواهد شد تا میزان موفقیت روش استفاده شده را بتوان در زبان فارسی و انگلیسی مقایسه کرد .

روش کار:

مراحل اصلی پژوهش به شرح ذیل می باشد:

۱-مطالعه مفهومی موضوع با استفاده از منابع اینترنتی از قبیل مقالات مرتبط با موضوع و کارهای انجام شده در این زمینه.

۲-تهیه و تدوین گزارشات متنی در خصوص پیشینه تحقیق و فرضیات مورد نیاز برای مدلسازی.

۳-فراخوانی داده ها و انجام پیش پردازش های لازم با استفاده از زبان برنامه نویسی پایتون.

۳-۱-خواندن مجموعه داده. (اولین گام جهت کار هوش مصنوعی، بارگذاری داده است)

۳-۲-تجزیه و تحلیل داده.از مهمترین قسمت های کار با داده،شناخت داده ورودی است.تسلط بر اینکه داده ورودی ، شامل چه آیتم هایی است، چه ویژگی ها و چه شی ء هایی در داده وجود دارد، تعداد و نوع اشیاء و ویژگی ها و همچنین بررسی برچسب ها، تعداد و نوئع آنها تأثیر بسیاری در نحوه کار با داده و انتخاب روش های مختلف کار با داده دارد.

۳-۳-تقسیم بندی دیتاست.در این تحقیق تقسیم بندی دیتاست بر اساس سه دسته انجام گرفته است:

دسته اول: داده آموزش که ۷۰ درصد از داده اصلی را شامل می شود.

دسته دوم: داده آزمون است که ۲۰ درصد از داده اصلی را تعریف می نماید.

دسته سوم: داده اعتبار سنجی که ۱۰ درصد را شامل می شود.

در واقع در این روش پس از یادگیری از روی داده آموزش و انجام آزمون روی داده آزمون، ارزیابی بر روی داده اعتبار سنجی انجام می گیرد.مهمترین فاکتور در ساخت این سه دسته، رعایت توزیع کلاس ها می باشد به نحوی که وقتی از داده اصلی ۷۰ درصد را به عنوان داده آموزشی استفاده می کنیم، توزیع کلاس ها یا برچسب ها باید به همان میزانی باشد که در داده اصلی موجود بوده است.

۳-۴-پاکسازی داده.

پاکسازی داده یا تمیز کردن داده، فرآیند پیدا کردن، اصلاح کردن(و یا حتی حذف کردن) داده های بی ارزش و استباه از دادگان یا پایگاه داده است.فرآیند تمیز کردن داده ها ممکن است که از طریق ابزارهای داده کاوی انجام شود.پس از پاکسازی، مجموعه داده باید با سایر مجموعه های مشابه در سیستم سازگار باشد چراکه ناسازگاری داده ها شناسایی و حذف(اصلاح) شده ممکن است بر اثر اشتباه انسانی هنگام ورود اطلاعات،انحراف در هنگام انتقال و ذخیره سازی اطلاعات یه به دلیل واژه نامه های داده مختلف باشد.در این پژوهش برای پاکسازی داده، نیاز به انجام عملیات نشانه گذاری،حذف علایم،کوچک کردن تمام حروف،تصحیح املای کلمات، حذف کلمات استُپ وُرد و یافتن ریشه لغات و ... می باشد که با توجه به فارسی بودن دیتاست ، روال انجام کار متفاوت با روش های مرسوم برای زبان انگلیسی خواهد بود.

۴- طراحی و پیاده سازی مدل های یادگیری ماشین با استفاده از زبان برنامه نویسی پایتون.

۴-۱- ابتدا باید عملیات تعیین قطبیت و تحلیل احساسات بر روی داده های پاکسازی شده را بر اساس الگوریتم های پایه SVM و NAÏVE BASE انجام داده تا بتوانیم معیاری جهت مقایسه نتایج به دست آمده در این تحقیق با نتایج به دست آمده در روش های پایه یادگیری ماشین داشته باشیم.

۴-۲- در این مرحله، انجام کار اصلی صورت می‌گیرد. یعنی انجام فرآیند تعیین قطبیت و تحلیل احساسات با استفاده از روش LSTM متمرکز است.هدف اصلی این تحقیق،بررسی نتایج به دست آمده از این قسمت و تحلیل و مقایسه نتایج با الگوریتم های پایه و همچنین نتایج سایر الگوریتم های یادگیری عمیق با دیتاست مشابه میباشد.

۵-اجرای فاز آزمایش خروجی مدل ها با استفاده از داده آموزش.

روش تجزیه و تحلیل اطلاعات و روشهای آماری

داده استفاده شده در این تحقیق شامل دو دیتاست به زبان های فارسی و انگلیسی می باشد:

دیتاست فارسی شامل ۳۰۲۳۲ نظرمی باشد . به عنوان ویژگی، داده هایی مربوط به نظرات مشتریان هتل قرارگرفته است. احساسات براساس شماره ۱ تا ۵ مشخص شده اند. شماره ۱به معنای منفی بودن و شماره ۵به معنای مثبت بودن‌است.میتوان ۲را به معنای‌نسبتاً منفی، ۳را خنثی و ۴را نسبتاً مثبت در نظر گرفت.حجم فایل k ۹۵۳۵ میباشد.

توزیع قطبیت نظرات در دیتاست فارسی نظرات کاربران هتل						
	مثبت	نسبتاً مثبت	خنثی	نسبتاً منفی	منفی	مجموع
تعداد نظرات	۸۵۴۸	۱۳۶۵۳	۵۴۵۲	۱۸۴۲	۷۳۷	۳۰۲۳۲
امتیاز	۵	۴	۳	۲	۱	-

شکل ۱-توزیع قطبیت نظرات در دیتاست فارسی

که با توجه به ماهیت نظرات، شماره های ۱ و ۲ را به عنوان نظرات منفی و شماره های ۴ و ۵ را به عنوانت نظرات مثبت و شماره ۳ را به عنوان نظر خنثی در نظر گرفته ایم که در این صورت تعداد نظارت مثبت ۲۲۲۰۱ ، خنثی ۵۴۵۲ و منفی ۲۵۷۹ می باشد. دیتاست انگلیسی شامل نظرات مشتریان درباره ۱۰ هتل برتر لندن که گرانترین و ارزانترین هتل ها را در برگرفته می باشد. احساسات شماره ۱ تا ۵ مشخص شده اند.شماره ۱ به معنای منفی بودن و ۵ به معنای مثبت بودن است.میتوان ۲ را نسبتاً منفی ، ۳ را خنثی و ۴ را نسبتاً مثبت در نظر گرفت.این مجموعه داده شامل بیش از ۲۷۳۳۳ نظر می باشد. حجم فایل k ۲۲۳۳۸ است.

توزیع قطبیت نظرات در دیتاست انگلیسی نظرات کاربران هتل						
	مثبت	نسبتاً مثبت	خنثی	نسبتاً منفی	منفی	مجموع
تعداد نظرات	۱۸۳۲۷	۶۰۲۰	۱۷۶۵	۶۹۱	۶۱۷	۲۷۳۳۰
امتیاز	۵	۴	۳	۲	۱	-

شکل ۲- توزیع قطبیت نظرات در دیتاست انگلیسی

در مرحله ارزشیابی مدل های ایجاد شده مقایسه و ارزیابی می شوند. معیارهای ارزیبای دقت پیش بیتی بر اساس معیار F1-Score می باشد. در واقع F1-Measure یک نوع میانگین بین پارامتر Precision (دقت) و پارامتر Recall (یادآوری) است. در واقع:

معیار صحت (Precision): نسبت مقداری موارد صحیح طبقه بندی شده از یک کلاس مشخص، به کل تعداد مواردی که چه به صورت صحیح و چه به صورت غلط، در آن کلاس طبقه بندی شده است که به صورت زیر محاسبه میشود:

$$\text{صحت (Precision)} = \frac{\text{تعداد های نمونه تشخیصی درست مثبت}}{\text{تعداد کل های نمونه تشخیصی مثبت}} = \frac{TP}{TP+FP}$$

معیار بازخوانی (Recall) : نسبت موارد صحیح طبقه بندی شده از یک کلاس به تعداد موارد حاضر در کلاس مذکور که به صورت زیر محاسبه می شود:

$$\text{بازخوانی (Recall)} = \frac{\text{تعداد های نمونه تشخیصی درست مثبت}}{\text{کل های نمونه واقعاً مثبت}} = \frac{TP}{TP+FN}$$

معیارهای مورداستفاده در این دیدگاه به شرح زیر میباشند:

TP_i: تشخیص مثبت صحیح از کلاس i اُم . – درست مثبت.

TN_i: تشخیص منفی صحیح از کلاس i اُم – درست منفی.

FP_i: تشخیص مثبت اشتباه از کلاس i اُم – نادرست مثبت.

FN_i: تشخیص منفی اشتباه از کلاس i اُم – نادرست منفی.

معیار F1-Measure : استفاده از معیاری ترکیبی از دو معیار precision و recall برای سنجش کیفیت دسته بندی و تمرکز بر آن به جای بررسی همزمان این دو، مناسب تر خواهد بود.با توجه به محاسبات انجام گرفته برای معیار های precision و recall می توان مقدار کمیت وزن دار F-Measure را محاسبه نمود. F-Measure پارامتر مناسبی برای ارزیابی کیفیت کلاس بندی می باشد و همچنین توصیف کننده میانگین وزن دار مابین دو کمیت precision و recall می باشد. برای یک الگوریتم کلاس بندی کننده در شرایط ایده آل، مقدار این کمیت برابر با ۱ می باشد و در برترین وضعیت برابر با صفر است.این پارامتر با توجه به رابطه زیر محاسبه می شود:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

نتیجه‌گیری

در این مقاله، یک مدل شبکه عصبی جدید برای دسته‌بندی رابطه پیشنهاد شده‌است. این مدل بر ابزار nlp یا منابع لغوی تکیه ندارد و از متن خام به عنوان ورودی استفاده می‌کند . کارایی این مدل با ارزیابی مدلهای پیشین مقایسه شده و نتایج بهتری به نسبت روش های قبلی به دست آمده است.

Refrence:

1. Ali Mardani, S. & Aghayi, A. (2015). Opinion Mining in Persian Language. Journal of Information technology management, 2(7), 345-362. (in Persian)
2. Balahur, A . & Turchi, M. (2014). Comparative experiments for multilingual sentiment analysis using machine translation. Computer Speech and Language, 28(1), 56–75
3. Banea, C. & Mihalcea, R. & Wiebe, J. (2014). Sense-level subjectivity in a multilingual setting. Computer Speech and Language, 28(1), 7–19.
4. Barawi Hardyman, M. & Seng , Y. (2013). Evaluation of resource creations accuracy by using sentiment Analysis. Procedia - Social and Behavioral Sciences,97(11), 522 – 527.
5. Yu, L. & Wu, Ch. & Jang, F. (2009). Psychiatric document retrieval using a discourse-aware model. Artificial Intelligence, 173(7), 817–829.
6. Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2013. p. 1631–1642.
7. Lei Zhang, S. W. (2018). Deep learning for sentiment analysis: A survey. Wiley Online Library, 1-25.
8. Yong Shi, L. Z. (2019). Survey on Classic and Latest Textual Sentiment Analysis. International Journal of Information Technology & Decision Making, 1243–1287.

9. Verma B, T. R. (2018). Sentiment analysis using lexicon and machine learning-based approaches: A survey. In Proceedings of international conference on recent advancement on computer and communication, 441-447.
10. Kowsari K, J. M. (2019). Text classification algorithms: A survey. Information.
11. Ain, Q. T. (2017). Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 424.